Springer Undergraduate Mathematics Series

SUMS

Vilmos Komornik

Topology, Calculus and Approximation

Springer

Springer Undergraduate Mathematics Series

More information about this series at http://www.springer.com/series/3423

Vilmos Komornik

Topology, Calculus and Approximation

Springer

Vilmos Komornik
Department of Mathematics
University of Strasbourg
Strasbourg, France

Translation from the French language edition:
Précis d'analyse réelle - Topologie - Calcul différentiel - Méthodes d'approximations, vol-1
by Vilmos Komornik

ISSN 1615-2085 ISSN 2197-4144 (electronic)
Springer Undergraduate Mathematics Series
ISBN 978-1-4471-7315-1 ISBN 978-1-4471-7316-8 (eBook)
DOI 10.1007/978-1-4471-7316-8

Library of Congress Control Number: 2017935974

Mathematics Subject Classification (2010): 41-01, 41A05, 41A10, 41A15, 41A50

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer-Verlag London Ltd.
The registered company address is: 236 Gray's Inn Road, London WC1X 8HB, United Kingdom

For God's sake, I beseech you, give it up. Fear it no less than sensual passions because it too may take all your time and deprive you of your health, peace of mind and happiness in life…

Letter of Farkas Bolyai to his son, April 4, 1820.

I created a new, different world out of nothing…

Letter of János Bolyai to his father, November 2, 1823.

Preface

This textbook contains the lecture notes of three one-semester courses given by the author to third year students at the University of Strasbourg. We assume that the reader is familiar with the calculus of one real variable. The first part on *Topology* is used everywhere in the sequel. The following two parts on *Differential calculus* and *Approximation methods* are logically independent.

We have made much effort to select the material covered by the lectures, to formulate aesthetical and general statements, to seek short and elegant proofs, and to illustrate the results with simple but pertinent examples. (See also the remarks on p. 369.) Our work is strongly influenced by the beautiful lectures of Professors Ákos Császár and László Czách during the 1970s at the Eötvös Loránd University in Budapest, and more generally by the mathematical tradition created by Leopold Fejér, Frédéric Riesz, Paul Turán, Paul Erdős and others.

On p. 337 we cite many papers of historical importance, indicating the origin of most of the notions and theorems treated here. They often contain different versions of the theorems we treat, illustrating the genesis of mathematical interest.

We suggest that, on the first reading, the reader should skip the material marked by $*$. At the end of each chapter we give some exercises. However, the most important exercises are incorporated into the text as examples and remarks, and the reader is expected to fill in the missing details.

We list on p. ix some books of general mathematical interest.

We thank Á. Besenyei, C. Baud, L. Czách, C. Disdier, D. Dumont, J. Gerner, P. Loreti, C.-M. Marle, P. Martinez, M. Mehrenberger, P.P. Pálfy, M. Pedicini, P. Pilibossian, J. Saint Jean Paulin, Z. Sebestyén, A. Simonovits, L. Simon, Mrs. B. Szénássy, G. Szigeti, J. Vancostenoble, Zs. Votisky and the editors of Springer for their precious help.

This book is dedicated to the memory of Paul Erdős.

Strasbourg, France
March 26, 2017

Vilmos Komornik

Some Books of General Interest

1. M. Aigner, G.M. Ziegler, *Proofs from the Book*, 4th ed., Springer, 2010.
2. P.S. Alexandroff, *Elementary Concepts of Topology*, Dover, New York, 1961.
3. E.T. Bell, *Men of Mathematics*, Simon and Schuster, New York, 1965.
4. V.G. Boltyanskii, V.A. Efremovich, *Intuitive Combinatorial Topology*, Springer, 2001.
5. J. Bolyai, *Appendix. The Theory of Space. With Introduction, Comments, and Addenda edited by F. Kárteszi*, Akadémiai Kiadó, Budapest, 1987.
6. D. Bressoud, *A Radical Approach to Real Analysis*, The Mathematical Association of America, Washington, 1994.
7. R. Courant, H. Robbins, *What is Mathematics? An Elementary Approach to Ideas and Methods*, Oxford Univ. Press, 1941.
8. H.S.M. Coxeter, S.L. Greitzer, *Geometry Revisited*, The Mathematical Association of America, Washington, 1967.
9. W. Dunham, *Journey Through Genius. The Great Theorems of Mathematics*, John Wiley & Sons, New York, 1990.
10. P. Erdős, J. Surányi, *Topics in the Theory of Numbers*, Springer, 2003.
11. E. Hairer, G. Wanner, *Analysis by Its History*, Springer, New York, 1996.
12. G.H. Hardy, *A Mathematician's Apology*, Cambridge Univ. Press, 1940.
13. P. Hoffman, *The Man Who Loved Only Numbers*, Hyperion, New York, 1998.
14. M. Kac, S.M. Ulam, *Mathematics and Logic*, Dover, New York, 1992.
15. A.Y. Khintchin, *Three Pearls of Number Theory*, Graylock, Rochester, 1952.
16. T.W. Körner, *Fourier Analysis*, Cambridge Univ. Press, 1988.
17. J. Kürschák, G. Hajós, G. Neukomm, J. Surányi, *Hungarian Problem Book I-IV*, Math. Assoc. of America, 1963–2011.
18. M. Laczkovich, *Conjecture and proof*, Cambridge University Press, 2001.
19. J. Muir, *Of Men and Numbers. The Story of Great Mathematicians*, Dover, New York, 1996.
20. J. Newman (ed.), *The World of Mathematics I-IV*, Dover, New York, 2000.
21. B. Schechter, *My Brain is Open: The Mathematical Journeys of Paul Erdős*, Simon & Schuster, 1998.

22. H. Steinhaus, *Mathematical Snapshots*, Dover, New York, 1999.
23. I. Stewart, *The Problems of Mathematics*, Oxford University Press, New York, 1992.
24. D.J. Struik, *A Concise History of Mathematics*, Dover, New York, 1987.

Contents

Part III Approximation Methods

Part I
Topology

Topology is the science of continuity. Following the polyhedron formula of Descartes [127][1] in 1639 and Euler [153] in 1750, Euler's paper [149] in 1736 on the bridges of Königsberg, and the works of Bolzano [53] (in 1817) and Cauchy [87] (in 1821), J. Bolyai [50] (in 1831) and Lobachevsky [337] (in 1829) stepped out of the two thousand year-old framework of Euclidean geometry. Their work was extended by Riemann [409, 410] (in 1854 and 1857).

Weierstrass [507–509] (1841–1874) and his students completed and clarified the earlier results. Most of today's basic notions (cluster points, interior and boundary points, density, closed sets, neighborhoods, ...) are due to Cantor [75–79] (1872–1884).

Grassmann [202] introduced the spaces $\mathbb{R}^n$ in 1862; after Jordan [263] and Peano [380, 382] (1882–1890), they were thoroughly studied by Poincaré [392] in 1895.

In order to overcome various difficulties occurring in the calculus of variations and in the study of Fourier series and partial differential equations, Ascoli [20] (1883), Arzelà [18, 19] (1889–1895), Volterra [500, 501] (1896–1897), Fredholm [179, 180] (1900–1903), Hilbert [238, 239] (1904–1906) and others started to study infinite-dimensional spaces.

Fréchet [174] introduced in 1906 *metric spaces* and the notions of completeness, compactness and separability.

Riesz [412] introduced *topological spaces* in 1906. The general theory of topological and metric spaces was developed and greatly enriched by Hausdorff in his influential monograph [225] in 1914. It already contains almost all the results of the first two chapters of this book.

Riesz [416] also introduced *normed spaces* in 1917, which are very useful in Differential Calculus and Functional Analysis.[2]

[1] The references refer to the bibliography at the end of this volume, on page 349.

[2] Analysis in infinite-dimensional spaces.

Readers interested in historical aspects of the subject may find much more information in the following works: [21, 31, 60, 62, 72, 81, 131, 137, 139, 217, 225, 277, 341, 429, 485].

The following books contain many interesting exercises and additional results: [22, 102, 106, 118, 131, 273, 278, 297, 431].

Most of the definitions and notations of this book are traditional; a few exceptions will be stated explicitly. The notation $(x_n) \subset A$ means that (x_n) is a sequence of elements of the set A.

The *domain* and the *range* of a function f will be denoted by $D(f)$ and $R(f)$, respectively. When it is not necessary to indicate the domain of f explicitly, we sometimes write $f : X \hookrightarrow Y$ instead of $f : D \to Y$, $D \subset X$.

Chapter 1
Metric Spaces

Metric spaces are very convenient in the study of continuity and uniform continuity. In this chapter we generalize a number of results on real sequences and functions of a real variable to sequences and functions defined on arbitrary metric spaces.

The reader may notice, however, the absence of *monotone* sequences: metric spaces are not suited to this notion.

Bounded sets may be defined easily in metric spaces, but they play a less important role than in $\mathbb{R}$. The related notion of *totally (or completely) bounded* sets proves to be more useful.

1.1 Definitions and Examples

Definitions

- By a *distance* or *metric* on a non-empty set X we mean a function $d : X \times X \to \mathbb{R}$ satisfying the following four properties for all $x, y, z \in X$:

  - $d(x, y) \ge 0,$
  - $d(x, y) = 0 \Longleftrightarrow x = y,$
  - $d(x, y) = d(y, x),$
  - $d(x, y) \le d(x, z) + d(z, y).$

  The last property is called the *triangle inequality* (see Fig. 1.1).
- If d is a metric on X, then (X, d) is called a *metric space*. When there is no risk of confusion, we write X instead of (X, d). The elements of X are also called *points*.

V. Komornik, *Topology, Calculus and Approximation*, Springer Undergraduate Mathematics Series, DOI 10.1007/978-1-4471-7316-8_1

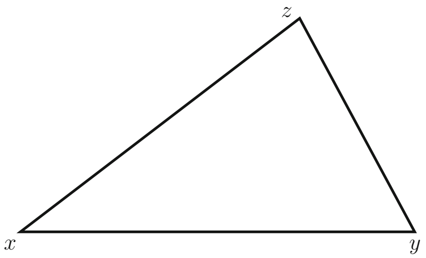


Fig. 1.1 The triangle inequality

Examples

- $X = \mathbb{R}$, $d(x, y) = |x - y|$. This is the *usual metric* of $\mathbb{R}$. Unless stated otherwise, we always consider this metric on $\mathbb{R}$.
- Given a non-empty set X, the formula
$$d(x, y) := \begin{cases} 0 & \text{if } x = y, \\ 1 & \text{if } x \neq y \end{cases}$$
defines the *discrete* metric, and (X, d) is called a *discrete metric space*.
- Given a non-empty set K, the set of bounded functions $f : K \to \mathbb{R}$ is denoted by $\mathcal{B}(K)$.

  The formula
$$d_\infty(f, g) := \sup_{t \in K} |f(t) - g(t)|$$
defines a metric on $\mathcal{B}(K)$. Henceforth $\mathcal{B}(K)$ will be endowed with this metric.
- We generalize the previous example. Given a metric space (X, d), the *diameter* of a set $A \subset X$ is defined by the formula[1]
$$\operatorname{diam} A := \sup \{d(x, y) \ : \ x, y \in A\}.$$
A is *bounded* if $\operatorname{diam} A < \infty$, and a function $f : K \to X$ is *bounded* if its range is a bounded set in X.

  The formula
$$d_\infty(f, g) := \sup \{d(f(t), g(t)) \ : \ t \in K\}$$
defines a metric on $\mathcal{B}(K, X)$. Henceforth $\mathcal{B}(K, X)$ will be endowed with this metric.

[1] We have $\operatorname{diam} \varnothing = -\infty$, and $\operatorname{diam} A \in [0, \infty]$ otherwise.

Definition Let (X, d) be a metric space, $y \in X$ and $r > 0$. The set

$$B_r(y) := \{x \in X \;:\; d(x, y) < r\}$$

is called the *(open) ball of radius r, centered at y*.

Remarks

- If $r < s$, then $B_r(y) \subset B_s(y)$. But we may have equality! For example, in a discrete metric space X we have $B_r(y) = \{y\}$ for all $r \le 1$ and $B_r(y) = X$ for all $r > 1$.
- The preceding example also shows that the radius and the center of a ball are not always uniquely defined.
- The diameter of a ball of radius r is at most $2r$ by the triangle inequality.
- A set is bounded if and only if it is contained in some ball.

Definition In a metric space a set U is *open* if for each $y \in U$ there exists an $r > 0$ such that $B_r(y) \subset U$.

Let us collect the basic properties of open sets:

Proposition 1.1 *Let (X, d) be a metric space.*

(a) *The sets $\varnothing$ and X are open.*
(b) *The intersection of finitely many open sets is open.*
(c) *Any union of open sets is open.*
(d) *The open balls are open sets.*
(e) *Any two points $x, y \in X$ may be separated by open sets U and V:*

$$x \in U, \quad y \in V \quad \textit{and} \quad U \cap V = \varnothing.$$

Proof

(a) X is open because $x \in B_1(x) \subset X$ for each $x \in X$. The empty set is open because no condition has to be verified.
(b) If $y \in U_1 \cap \cdots \cap U_n$, where $U_1, \dots, U_n$ are open sets, then for each i there exists an $r_i > 0$ such that $B_{r_i}(y) \subset U_i$. Then $r := \min\{r_1, \dots, r_n\} > 0$, and $B_r(y) \subset U_1 \cap \cdots \cap U_n$.
(c) If y belongs to the union U of a family $\{U_i\}$ of open sets, then $y \in U_i$ for a suitable index i. Since U_i open, there exists an $r > 0$ such that $B_r(y) \subset U_i$, and then $B_r(y) \subset U$.
(d) Given $y \in B_r(x)$, we seek $s > 0$ such that $B_s(y) \subset B_r(x)$. We may choose $s := r - d(x, y)$ (see Fig. 1.2). Indeed, then $s > 0$ by the definition of $B_r(x)$. Furthermore, if $z \in B_s(y)$, then

$$d(x, z) \le d(x, y) + d(y, z) < d(x, y) + s = r,$$

so that $z \in B_r(x)$.

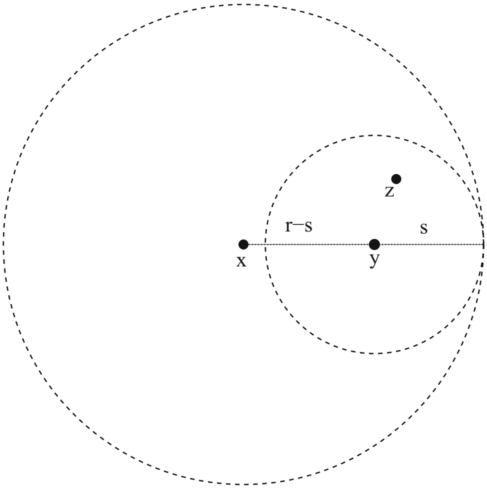


Fig. 1.2 The openness of balls

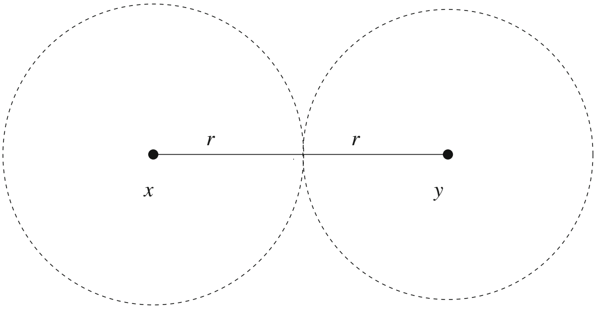


Fig. 1.3 Separation of points

(e) We may choose, for example, $U = B_r(x)$ and $V = B_r(y)$ with $r := d(x,y)/2 > 0$. Indeed, if $z \in U$, then $d(y,z) \geq d(x,y) - d(x,z) > r$ by the triangle inequality (see Fig. 1.3), and hence $z \notin V$. □

Example The open intervals of $\mathbb{R}$ are indeed open sets. First, the (non-empty) *bounded* open intervals are open balls: $(a,b) = B_r(y)$ with

$$y = \frac{a+b}{2} \quad \text{and} \quad r = \frac{b-a}{2}.$$

Next, every unbounded open interval is the union of (countably many) bounded open intervals: for any $a \in \mathbb{R}$ we have

$$(a, \infty) = \bigcup_{n=1}^{\infty} (a, a+n), \quad (-\infty, a) = \bigcup_{n=1}^{\infty} (a-n, a) \quad \text{and} \quad \mathbb{R} = \bigcup_{n=1}^{\infty} (-n, n).$$

We introduce two common constructions of new metric spaces.

Definitions

- By a *subspace* of a metric space (X, d) we mean a non-empty subset Y of X endowed with the restriction of the metric d to $Y \times Y$.
- By the *product* of finitely many metric spaces $(X_1, d_1), \ldots, (X_m, d_m)$ we mean the product set $X = X_1 \times \cdots \times X_m$ endowed with the metric

$$d(x, y) := d_1(x_1, y_1) + \cdots + d_m(x_m, y_m), \quad x, y \in X,$$

where we use the notations

$$x = (x_1, \ldots, x_m) \quad \text{and} \quad y = (y_1, \ldots, y_m).$$

One may readily verify that both definitions do indeed yield metric spaces.

Applying these constructions to $\mathbb{R}$, $\mathbb{R}^m$ becomes a metric space for the metric

$$d(x, y) := |x_1 - y_1| + \cdots + |x_m - y_m|,$$

and all non-empty subsets of $\mathbb{R}^m$ also become metric spaces.

Finally, we generalize the closed intervals:

Definition A set in a metric space is *closed* if its complement is open.

Examples

- A *closed interval* of $\mathbb{R}$ is indeed a closed set: the complement of an unbounded closed interval is either empty or an open interval:

$$\mathbb{R} \setminus \mathbb{R} = \varnothing, \quad \mathbb{R} \setminus (-\infty, b] = (b, \infty), \quad \mathbb{R} \setminus [a, \infty) = (-\infty, a),$$

while the complement of a bounded closed interval is the union of two open intervals:

$$\mathbb{R} \setminus [a, b] = (-\infty, a) \cup (b, \infty),$$

hence is again an open set.

- The bounded half-closed intervals $[a, b)$ and $(a, b]$ are neither open, nor closed.

1.2 Convergence, Limits and Continuity

The convergence of a sequence in a metric space is defined by using the convergence of number sequences:

Definition A sequence (x_n) in a metric space (X, d) is *convergent* if there exists a point $x \in X$ satisfying $d(x_n, x) \to 0$. The point x is called the *limit* of the sequence. We also say that x_n *converges* or *tends* to x, and we write $x_n \to x$ or $\lim x_n = x$.

Examples

- In $X = \mathbb{R}$ the definition reduces to the usual convergence.
- If $x_n = x$ for all sufficiently large indices n, then $x_n \to x$.[2] In discrete metric spaces there are no other convergent sequences.
- If Y is a subspace of a metric space X, $y \in Y$ and $(x_n) \subset Y$, then

$$x_n \to y \quad \text{in} \quad Y \Longleftrightarrow x_n \to y \quad \text{in} \quad X.$$

- In $\mathcal{B}(K)$ we have $f_n \to f$ if and only if the function sequence (f_n) converges *uniformly* to f on K, i.e., for each $\varepsilon > 0$ there exists an index N such that

$$|f_n(t) - f(t)| < \varepsilon$$

for all $n \geq N$ and $t \in K$.
- More generally, in $\mathcal{B}(K, X)$ we have $f_n \to f$ if and only if for each $\varepsilon > 0$ there exists an index N such that

$$d(f_n(t), f(t)) < \varepsilon$$

for all $n \geq N$ and $t \in K$.
- For the product (X, d) of the metric spaces $(X_1, d_1), \dots, (X_m, d_m)$ we have

$$x_n \to a \quad \text{in} \quad X \Longleftrightarrow x_{nj} \to a_j \quad \text{in} \quad X_j \quad \text{for each} \quad j = 1, \dots, m,$$

where we use the notation

$$x_n = (x_{n1}, \dots, x_{nm}) \quad \text{and} \quad a = (a_1, \dots, a_m)$$

(*component-wise convergence*). An important case is $X = \mathbb{R}^m$.

[2] Such sequences are called *eventually constant*.

The usual properties of convergent number sequences (except those related to monotonicity) remain valid in metric spaces:

Proposition 1.2

(a) *The limit of a convergent sequence is unique.*
(b) *If $x_n \to x$ and $y_n \to y$, then $d(x_n, y_n) \to d(x, y)$.*
(c) *If $x_n \to x$, then $x_{n_k} \to x$ for each subsequence (x_{n_k}) of (x_n).*
(d) *(Cantor) If $x_n \not\to x$, then there exists a subsequence (x_{n_k}) such that $x_{n_{k_l}} \not\to x$ for every further subsequence $(x_{n_{k_l}})$ of (x_{n_k}).*

Proof

(a) If $x_n \to x$ and $x_n \to y$, then

$$d(x, y) \le d(x, x_n) + d(x_n, y) \to 0$$

by the triangle inequality. Hence $d(x, y) \le 0$, so that $x = y$.
(b) By the triangle inequality, we have

$$d(x_n, y_n) \le d(x_n, x) + d(x, y) + d(y, y_n)$$

and

$$d(x, y) \le d(x, x_n) + d(x_n, y_n) + d(y_n, y),$$

so that

$$|d(x_n, y_n) - d(x, y)| \le d(x_n, x) + d(y_n, y).$$

We conclude by observing that the right-hand side tends to zero.
(c) If $d(x_n, x) \to 0$, then $d(x_{n_k}, x) \to 0$, because the *number* sequence $(d(x_{n_k}, x))$ is a subsequence of $(d(x_n, x))$.
(d) If $x_n \not\to x$, then there exists an $\varepsilon > 0$ and a subsequence (x_{n_k}) such that $d(x_{n_k}, x) > \varepsilon$ for all k. Then no subsequence $(x_{n_{k_l}})$ of (x_{n_k}) converges to x, because $d(x_{n_{k_l}}, x) > \varepsilon$ for all l. □

There is a sequential characterization of closed sets (see Fig. 1.4). First we prove a lemma:

Lemma 1.3 *Given a set $D \subset X$ and a point $a \in X$ in a metric space X, the following properties are equivalent:*

(a) *every open set containing a meets D;*
(b) *every ball $B_r(a)$ meets D;*
(c) *there exists a sequence (x_n) in D, converging to a.*

Proof

(a) ⇒ (b) because the balls $B_r(a)$ are open.

(b) ⇒ (c) Choose a point $x_n \in B_{1/n}(a) \cap D$ for each n, then $(x_n) \subset D$, and $x_n \to a$ because $0 \le d(x_n, a) < 1/n \to 0$.

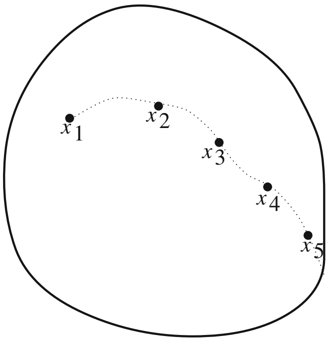


Fig. 1.4 A sequence in a closed set

(c) ⇒ (a) Choose a sequence (x_n) in D, converging to a. For each open set U containing a, choose a ball $B_r(a) \subset U$. If n is sufficiently large, then $x_n \in B_r(a) \cap D \subset U \cap D$. □

Proposition 1.4 *A set F in a metric space X is closed if and only if*

$$x_n \to a \quad in \quad X \quad and \quad (x_n) \subset F \quad \Longrightarrow \quad a \in F.$$

Proof F is not closed if and only if $X \setminus F$ is not open, i.e., there exists an $a \in X \setminus F$ such that all balls $B_r(a)$ meet F. By Lemma 1.3 this is equivalent to the existence of a sequence $(x_n) \subset F$ converging to a. □

Next we investigate the limit of a *function*:

Definition Let (X, d) and (X', d') be metric spaces, $a \in X$ and $a' \in X'$. A function $f : X \to X'$ has the *limit a'* at a if for each $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$x \in X \quad \text{and} \quad 0 < d(x, a) < \delta \quad \Longrightarrow \quad d'(f(x), a') < \varepsilon.$$

In this case we write $\lim_a f = a'$ or $\lim_{x \to a} f(x) = a'$.

Remarks

- Even if f is defined at a, neither the existence, nor the value of the limit depends on $f(a)$.
- We say that a is an *accumulation point* (or a *cluster point*) of X if every ball $B_r(a)$ contains at least one point, other than a. In this case f has at most one limit in a. Indeed, by Lemma 1.3 there exists a sequence $(x_n) \subset X \setminus \{a\}$ converging to a. If $\lim_a f = a'$, then $f(x_n) \to a'$ by the definition of the limit of a sequence. We conclude by recalling that a sequence has at most one limit.
- We say that a is an *isolated point* of X if it is not an accumulation point, i.e., if there exists an $r > 0$ such that $B_r(a) = \{a\}$. In this case we have $\lim_a f = a'$ for *all* $a' \in X'$. In order to avoid this pathological case the limit is usually defined only at accumulation points.

Our broader definition avoids the notion of accumulation points. More importantly, it leads to some simplifications in the sequel.[3]

Now we study *continuity*.

Lemma 1.5 *Let (X, d) and (X', d') be metric spaces, $a \in X$ and $f : X \to X'$. The following three properties are equivalent:*

(a) $\lim_a f = f(a)$.
(b) *For each $\varepsilon > 0$ there exists a $\delta > 0$ such that*

$$x \in X \quad and \quad d(x,a) < \delta \quad \Longrightarrow \quad d'(f(x), f(a)) < \varepsilon.$$

Equivalently, each ball $B_\varepsilon(f(a))$ contains the image $f(B_\delta(a))$ of some ball $B_\delta(a)$.
(c) *If $x_n \to a$ in X, then $f(x_n) \to f(a)$ in X'.*

Proof

$(a) \Rightarrow (b)$. For any fixed $\varepsilon > 0$ we choose $\delta > 0$ according to the definition of $\lim_a f = f(a)$. Thus $x \neq a$ and $d(x,a) < \delta$ imply $d'(f(x), f(a)) < \varepsilon$. The last inequality holds for $x = a$, too.
$(b) \Rightarrow (c)$. For any fixed $\varepsilon > 0$ we have to find N such that

$$d'(f(x_n), f(a)) < \varepsilon$$

for all $n \geq N$. Choose $\delta > 0$ according to (b). Since $x_n \to 1$, there exists an N such that $d(x_n, a) < \delta$ for all $n \geq N$. Then $d'(f(x_n), f(a)) < \varepsilon$ for all $n \geq N$.
$(c) \Rightarrow (a)$. If (a) is not satisfied, then we may fix $\varepsilon > 0$ such that for each $\delta > 0$ there exists an $x \in X$ satisfying

$$0 < d(x,a) < \delta \quad \text{and} \quad d'(f(x), f(a)) \geq \varepsilon.$$

Applying this with $\delta = 1/n$ we obtain a sequence $x_1, x_2, \ldots$ such that

$$0 < d(x_n, a) < 1/n \quad \text{and} \quad d'(f(x_n), f(a)) \geq \varepsilon$$

for all n. Then $d(x_n, a) \to 0$, but $d'(f(x_n), f(a)) \not\to 0$, so that (c) is not satisfied either. □

Definition Let (X, d), (X', d') be metric spaces and $a \in X$. A function $f : X \to X'$ is *continuous* at a if the equivalent conditions of the preceding lemma are satisfied.

Proposition 1.6 *Consider two functions $g : X \to X'$ and $f : X' \to X''$, where X, X', X'' are metric spaces. If g is continuous at $a \in X$ and f is continuous at $g(a) \in X'$, then the composite function $f \circ g : X \to X''$ is continuous at a.*

[3] See, e.g., Lemma 1.5 below.

Proof We use definition (c) of Lemma 1.5. We have to show that if $x_n \to a$ in X, then $(f \circ g)(x_n) \to (f \circ g)(a)$ in X''.

Since g is continuous at a, $g(x_n) \to g(a)$ in X'. Therefore, since f is continuous at $g(a), f(g(x_n)) \to f(g(a))$ in X''. □

There are several useful notions of *global* continuity:

Definitions Let (X, d) and (X', d') be metric spaces. A function $f : X \to X'$ is

- *continuous* if it is continuous at each point $a \in X$;
- *uniformly continuous* if for each $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$x, y \in X \quad \text{and} \quad d(x, y) < \delta \quad \Longrightarrow \quad d'(f(x), f(y)) < \varepsilon;$$

- *Lipschitz continuous* if there exists a constant $L \geq 0$ such that

$$d'(f(x), f(y)) \leq Ld(x, y)$$

  for all $x, y \in X$.

Remark Every Lipschitz continuous function is uniformly continuous (choose $\delta = \varepsilon/L$), and every uniformly continuous function is continuous (choose the same δ for each $a \in X$). The converse implications may fail:

Examples (See Fig. 1.5)

- The function $f : \mathbb{R} \to \mathbb{R}, f(x) = x^2$ is continuous, but not uniformly continuous.
- The function $f : \mathbb{R} \to \mathbb{R}, f(x) = |x|^{1/2}$ is uniformly continuous, but not Lipschitz continuous.
- The function $f : \mathbb{R} \to \mathbb{R}, f(x) = x$ is Lipschitz continuous.
- If (X, d) is a metric space, then the distance function $d : X \times X \to \mathbb{R}$ is Lipschitz continuous. Indeed, applying the triangle inequality as in the proof of Proposition 1.2 (b) (p. 9), we obtain that

$$|d(x_1, x_2) - d(y_1, y_2)| \leq d(x_1, y_1) + d(x_2, y_2)$$

  for all pairs $(x_1, x_2), (y_1, y_2) \in X \times X$. Since the right-hand side is the distance between these pairs in the product metric of $X \times X$ by definition, this estimate shows the Lipschitz continuity of d with $L = 1$.

Proposition 1.7 *Consider two functions $g : X \to X'$ and $f : X' \to X''$ where (X, d), (X', d'), (X'', d'') are metric spaces.*

(a) *If f and g are continuous, then $f \circ g$ is continuous.*
(b) *If f and g are uniformly continuous, then $f \circ g$ is uniformly continuous.*
(c) *If f and g are Lipschitz continuous, then $f \circ g$ is Lipschitz continuous.*

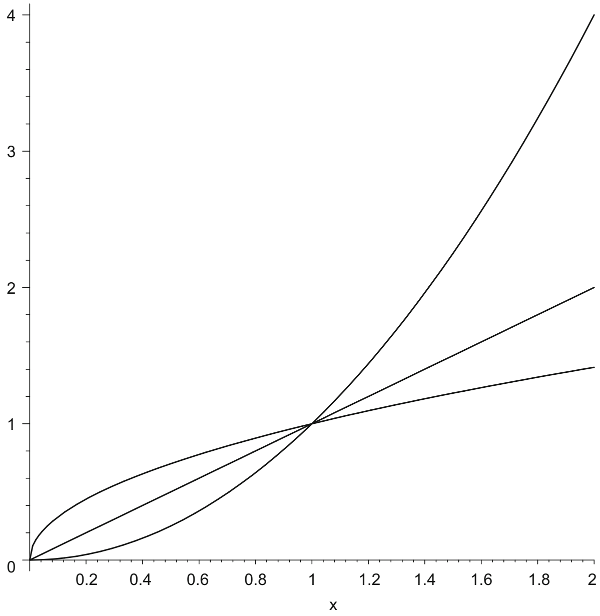


Fig. 1.5 Different versions of continuity

Proof

(a) Apply the preceding proposition for each $a \in X$.

(b) For any fixed $\varepsilon > 0$ we seek $\delta > 0$ such that

$$\text{if} \quad x, y \in X \quad \text{and} \quad d(x,y) < \delta, \quad \text{then} \quad d''(f(g(x)), f(g(y))) < \varepsilon.$$

Since f is uniformly continuous, there exists a $\sigma > 0$ such that

$$\text{if} \quad x', y' \in X' \quad \text{and} \quad d'(x', y') < \sigma, \quad \text{then} \quad d''(f(x'), f(y')) < \varepsilon.$$

Furthermore, since g is uniformly continuous, there exists a $\delta > 0$ such that

$$\text{if} \quad x, y \in X \quad \text{and} \quad d(x,y) < \delta, \quad \text{then} \quad d'(g(x), g(y)) < \sigma.$$

Applying the definition of σ with $x' = g(x)$ and $y' = g(y)$ we get the desired result.

(c) If f and g are Lipschitz continuous with constants L_1 and L_2, then $f \circ g$ is Lipschitz continuous with constant $L_1 L_2$ because

$$d''(f(g(x)), f(g(y))) \le L_1 d'(g(x), g(y)) \le L_1 L_2 d(x,y)$$

for all $x, y \in X$. □

1.3 Completeness: A Fixed Point Theorem

The classical *Cauchy criterion* often allows us to prove the convergence of a numerical sequence without knowing its limit. In this section we study the metric spaces where this useful property still holds.

Definition A sequence (x_n) in a metric space (X, d) is a *Cauchy sequence* if

$$\operatorname{diam}\{x_n \ : \ n \ge N\} \to 0 \quad \text{as} \quad N \to \infty.$$

In other words, for each $\varepsilon > 0$ there exists an N such that $d(x_m, x_n) < \varepsilon$ for all $m, n \ge N$. We express this property by writing

$$d(x_m, x_n) \to 0 \quad \text{as} \quad m, n \to \infty.$$

Example Every convergent sequence is a Cauchy sequence. Indeed, if $x_n \to x$, then letting $m, n \to \infty$ we have

$$0 \le d(x_m, x_n) \le d(x_m, x) + d(x, x_n) \to 0.$$

Definition A metric space is *complete* if every Cauchy sequence is convergent.

Examples

- $\mathbb{R}$ is complete by Cauchy's classical theorem.
- The discrete metric spaces are complete because every Cauchy sequence is eventually constant.
- The spaces $\mathcal{B}(K)$ are complete. More generally:

Proposition 1.8 *If (X, d) is a complete metric space, then the metric spaces $\mathcal{B}(K, X)$ are also complete.*

Proof Let (f_n) be a Cauchy sequence in $\mathcal{B}(K, X)$: for each fixed $\varepsilon > 0$ there exists an N such that

$$d(f_n(t), f_m(t)) < \varepsilon \tag{1.1}$$

for all $t \in K$ and $m, n \ge N$. We have to find a bounded function $f : K \to X$ such that f_n converges uniformly to f on K.

It follows from our assumptions that for each fixed $t \in K$, $(f_n(t))$ is a Cauchy sequence in X. Since X is complete, this sequence converges to some limit $f(t) \in X$. We thus obtain a function $f : K \to X$ such that

$$f_n(t) \to f(t)$$

for each $t \in K$.

Letting $m \to \infty$ in (1.1), and using the continuity of the metric we obtain that

$$d(f_n(t), f(t)) \le \varepsilon \tag{1.2}$$

for all $t \in K$ and $n \ge N$. Hence f is bounded, i.e., $f \in \mathcal{B}(K, X)$. Indeed, since f_N is bounded, there exists a constant M such that

$$d(f_N(s), f_N(t)) \le M$$

for all $s, t \in K$. Applying (1.2) with $n = N$ and using the triangle inequality we obtain the boundedness of f:

$$d(f(s), f(t)) \le M + 2\varepsilon \quad \text{for all} \quad s, t \in K.$$

Finally, (1.2) shows that f_n converges to f in $\mathcal{B}(K, X)$. □

The following proposition often enables us to prove the completeness of metric spaces:

Proposition 1.9

(a) *The product of finitely many complete metric spaces is complete.*
(b) *The* closed *subspaces of complete metric spaces are complete.*

Proof

(a) Let (x_n) be a Cauchy sequence in the product (X, d) of the complete metric spaces

$$(X_1, d_1), \dots, (X_m, d_m).$$

Writing $x_n = (x_{n1}, \dots, x_{nm})$, we deduce from the equality

$$d(x_n, x_k) = d_1(x_{n1}, x_{k1}) + \cdots + d_m(x_{nm}, x_{km})$$

that (x_{nj}) is a Cauchy sequence in (X_j, d_j) for each $j = 1, \dots, m$. These spaces being complete, there exist points $a_j \in X_j$ satisfying $d_j(x_{nj}, a_j) \to 0$. Setting $a := (a_1, \dots, a_m) \in X$, we deduce from the equality

$$d(x_n, a) = d_1(x_{n1}, a_1) + \cdots + d_m(x_{nm}, a_m)$$

that $d(x_n, a) \to 0$, i.e., $x_n \to a$ in (X, d).
(b) If (x_n) is a Cauchy sequence in Y, then it is also a Cauchy sequence in X. Since X is complete, (x_n) converges to some point $x \in X$. Since Y is a closed set, $x \in Y$, and hence $x_n \to x$ in Y. □

Example $\mathbb{R}^m$ (with the product metric) and more generally every non-empty *closed* subset of $\mathbb{R}^m$ (with the subspace metric) is a complete metric space.

The following existence theorem is often used to solve equations.[4]

Definitions Consider a function $f : X \to X$ on a metric space (X, d).

- f is a *contraction* if it is Lipschitz continuous with some constant $L < 1$:

$$d(f(x), f(y)) \le Ld(x, y)$$

  for all $x, y \in X$.
- $x \in X$ is a *fixed point* of f if $f(x) = x$.

Theorem 1.10 (Banach–Cacciopoli Fixed Point Theorem) *In a* complete *metric space X every contraction f has a unique fixed point x.*
Moreover, starting from an arbitrary point $x_0 \in X$, the formula $x_n := f(x_{n-1})$ defines a sequence converging to x, and

$$d(x, x_n) \le L^n (1 - L)^{-1} d(x_1, x_0)$$

for all $n = 0, 1, \ldots$.

Remarks

- The function $f(x) := x + 1$ on $\mathbb{R}$ shows the necessity of the hypothesis $L < 1$.
- The function $f(x) := x/2$ on $\mathbb{R}^* := \mathbb{R} \setminus \{0\}$ shows the necessity of the hypothesis of completeness.
- A more general result will be given in Proposition 12.4, p. 292.
- There is another important fixed point theorem, due to Brouwer; see also the comments on p. 338.

Proof Let f be a contraction on a complete metric space X. Fix an arbitrary point $x_0 \in X$ and define a sequence by the recursive formula $x_n := f(x_{n-1})$. This is a Cauchy sequence because for any $m > n \ge 0$ we have

$$\begin{aligned} d(x_m, x_n) &\le d(x_m, x_{m-1}) + \cdots + d(x_{n+1}, x_n) \\ &\le (L^{m-1} + \cdots + L^n) d(x_1, x_0) \\ &\le L^n (1 - L)^{-1} d(x_1, x_0), \end{aligned}$$

and $L^n \to 0$ as $n \to \infty$. Since X is complete, the sequence converges to some point $x \in X$. Letting $n \to \infty$ in the equation $x_n = f(x_{n-1})$ and using the continuity of f at x we obtain that $x = f(x)$.

If y is an arbitrary fixed point of f, then

$$d(x, y) = d(f(x), f(y)) \le Ld(x, y).$$

[4] See, e.g., the proof of Theorems 6.2, 7.7, and Exercise 11.4 below, pp. 148, 179 and 281. See also Proposition 12.4, p. 292.

Since $L < 1$, we have $d(x, y) \leq 0$ and therefore $x = y$. The fixed point is thus unique.

Finally, the estimate of $d(x, x_n)$ is obtained by letting $m \to \infty$ in the inequality at the beginning of the proof. □

Example The function $f : [1, \infty) \to [1, \infty)$ defined by the formula

$$f(x) := \frac{1}{2}\Big(x + \frac{2}{x}\Big), \quad x \geq 1$$

is a contraction because

$$|f(x) - f(y)| = \frac{|xy - 2|}{2xy}|x - y| \leq \frac{xy}{2xy}|x - y| = \frac{1}{2}|x - y|$$

for all $x, y \geq 1$.[5] Since the interval $[1, \infty)$ is closed in $\mathbb{R}$, it is a complete metric space. Therefore f has a unique fixed point by the theorem, and starting from an arbitrary number $x_0 \geq 1$, the sequence

$$x_{n+1} := \frac{1}{2}\Big(x_n + \frac{2}{x_n}\Big), \quad n = 0, 1, \ldots$$

converges to this unique fixed point (equal to $\sqrt{2}$).[6]

Definition A non-empty set F in a metric space is *complete* if the corresponding metric subspace is complete, i.e., if every Cauchy sequence in F converges to some element of F.

The empty set is also considered to be complete.

Proposition 1.11 *The complete subsets of a metric space are closed.*

Proof Let F be a complete set in a metric space X, and consider a sequence (x_n) in F, converging to some point $x \in X$. We have to show that $x \in F$.

Being convergent, (x_n) is a Cauchy sequence. Since F is complete, it converges to some $y \in F$. By the uniqueness of the limit of a sequence we conclude that $y = x$, and therefore $x \in F$. □

Proposition 1.12 (Cantor's Intersection Theorem) *Let (F_n) be a non-increasing sequence of non-empty closed sets in a complete metric space. If the diameters* diam F_n *tend to zero, then the sets F_n have at least one common point. (See Fig. 1.6.)*

[5] If $xy \leq 2$, then we have $-1 \leq xy - 2 \leq 0$, and therefore $|xy - 2| \leq 1 \leq xy$.

[6] See Exercise 1.1 for a generalization, p. 280.

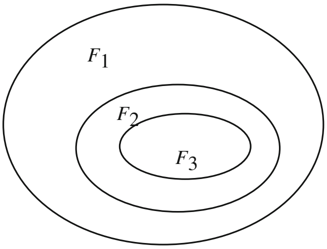


Fig. 1.6 Cantor's intersection theorem

Remarks

- Since diam $(\cap F_n) = 0$, there is a *unique* common point.
- The sets $F_n = [n, \infty)$ in $\mathbb{R}$ show the necessity of the assumption diam $F_n \to 0$.
- The sets $F_n = (0, 1/n]$ in $\mathbb{R}$ show the necessity of the closedness of the sets F_n.
- Cantor used this result to prove the uncountability of $\mathbb{R}$ as follows. We have to show that no sequence (x_n) of real numbers contains all elements of $\mathbb{R}$. For this we construct by induction a non-increasing sequence of closed intervals F_n satisfying[7] $0 < \operatorname{diam} F_n < 1/n$ and $x_n \notin F_n$ for each n. There exists a common point x of these intervals, and $x \neq x_n$ for all n by construction.

Proof Choose for each n a point $x_n \in F_n$. Then (x_n) is a Cauchy sequence. Indeed, if $m > n$, then $x_m \in F_m \subset F_n$ and $x_n \in F_n$, so that

$$d(x_m, x_n) \le \operatorname{diam} F_n,$$

and the right-hand side tends to zero as $n \to \infty$.

Since X is complete, (x_n) converges to some point x. It remains to show that $x \in F_m$ for each m.

This follows by observing that the subsequence $x_m, \ x_{m+1}, \dots$ belongs to the closed set F_m, and converges to x. □

The remainder of this section is related to the notion of density. Given a set D in a metric space X, by Lemma 1.3 the following three conditions are equivalent:

- every non-empty open set meets D;
- every ball meets D;
- for each $a \in X$ there exists a sequence (x_n) in D, converging to a.

Definition A set D in a metric space X is *dense* if the above equivalent conditions are satisfied.

The next result is often applied in Functional Analysis.[8]

[7] Of course, diam F_n is the length of this interval.

[8] See also Exercise 1.14 (iii) (p. 34) for one of its first applications.

***Proposition 1.13 (Baire)** *Let (X, d) be a complete metric space.*

(a) *If (G_k) is a sequence of open dense sets, then their intersection is also dense.*
(b) *If X is the union of countably many closed sets, then at least one of them contains a non-empty open set.*

Remark The completeness assumption is essential: consider in $X = \mathbb{Q}$ (with the usual distance) the (closed) one-point sets and their open complements.

Proof

(a) We define the *closed balls*

$$\overline{B}_r(x) := \{y \in X \ : \ d(x, y) \le r\}, \quad r > 0,\ x \in X.$$

Their complements are open,[9] so that the closed balls are indeed closed sets. Let (G_n) be a sequence of open dense sets and $B_{r_0}(x_0)$ an arbitrary (open) ball in X. We have to show that $\cap G_n$ meets $B_{r_0}(x_0)$.

Since G_1 is dense, there exists a point $x_1 \in G_1 \cap B_{r_0}(x_0)$. Since the latter set is open, there exists $0 < r_1 < 1$ such that

$$\overline{B}_{r_1}(x_1) \subset G_1 \cap B_{r_0}(x_0).$$

Since G_2 is dense, there exists a point $x_2 \in G_2 \cap B_{r_1}(x_1)$. Since the latter set is open, there exists $0 < r_2 < \frac{1}{2}$ such that

$$\overline{B}_{r_2}(x_2) \subset G_2 \cap B_{r_1}(x_1).$$

Continuing by induction we obtain a non-increasing sequence of closed balls such that

$$0 < r_k < \frac{1}{k} \quad \text{and} \quad \overline{B}_{r_k}(x_k) \subset G_k \cap B_{r_{k-1}}(x_{k-1})$$

for all k. We may apply Proposition 1.12 with $F_k := \overline{B}_{r_k}(x_k)$ to find a point $x \in \cap \overline{B}_{r_k}(x_k)$. Then $x \in \cap G_n$ and $x \in B_{r_0}(x_0)$ by construction.

(b) If none of the closed sets F_k contains a non-empty open set, then their complements G_k are dense open sets, and hence they have a common point x by (a). Then x does not belong to any F_k, so that $X \ne \cup F_k$. □

Let (X, d), (X', d') be two metric spaces, $D \subset X$ a non-empty set, and $f : D \to X'$ a continuous function (for the subspace metric of D). It is natural to ask whether f may be extended continuously to the whole space X. The example of the function

$$f(x) = 1/x, \quad X = X' = \mathbb{R}, \quad D = \mathbb{R} \setminus \{0\}$$

[9] Indeed, if $z \in X \setminus \overline{B}_r(x)$, then $s := d(z, x) - r > 0$ and $B_s(z) \subset X \setminus \overline{B}_r(x)$.

shows that some extra hypotheses are needed. Here we study only the simpler case of *uniformly* continuous functions.[10]

Proposition 1.14 *Let (X,d), (X',d') be two metric spaces, $D \subset X$ a dense set, and $f : D \to X'$ a uniformly continuous function.*

If (X',d') is complete, then f may be uniquely extended to a continuous function $F : X \to X'$. Moreover, F is also uniformly continuous.

If f is Lipschitz continuous, then F is also Lipschitz continuous with the same constant.

Finally, if f is an isometry, *i.e., $d'(f(x),f(y)) = d(x,y)$ for all $x,y \in D$, then F is also an isometry: $d'(F(x),F(y)) = d(x,y)$ for all $x,y \in X$.*

Proof If there is a continuous extension F, then for every sequence $(x_n) \subset D$ converging to some point $a \in X$ the relation $f(x_n) \to F(a)$ holds. (Since D is dense, there exists such a sequence for each $a \in X$.) This proves the uniqueness of F.

For the existence we show that the formula $F(a) := \lim f(x_n)$, where $(x_n) \subset D$ is an arbitrary sequence converging to a, defines a uniformly continuous extension $F : X \to X'$ of f.

The limit always exists because X' is complete, and $(f(x_n))$ is a Cauchy sequence. To show the latter we have to find for each fixed $\varepsilon > 0$ an index N such that

$$d'(f(x_n),f(x_k)) < \varepsilon \tag{1.3}$$

for all $n,k \geq N$. Choose a $\delta > 0$ for ε by the uniform continuity of f, and then choose a sufficiently large N such that $d(x_n,x_k) < \delta$ for all $n,k \geq N$. Then (1.3) is satisfied.

The limit does not depend on the special choice of the sequence (x_n). Indeed, if $(y_n) \subset D$ is another sequence converging to a, then the combined sequence $x_1,y_1,x_2,y_2,\dots$ also has this property. By the preceding assertion the sequence $f(x_1),f(y_1),f(x_2),f(y_2),\dots$ converges in X'. We conclude that its *subsequences* $(f(x_n))$ and $(f(y_n))$ converge to the same limit.

If $a \in D$, then choosing the constant sequence $x_n := a$ we get $F(a) = f(a)$, i.e., F is an extension of f.

We claim that F is uniformly continuous. Given $\varepsilon > 0$, choose $\delta > 0$ by the uniform continuity of f. It suffices to show that

$$\text{if} \quad a,b \in X \quad \text{and} \quad d(a,b) < \delta, \quad \text{then} \quad d'(F(a),F(b)) \leq \varepsilon.$$

Choose two sequences (x_n) and (y_n) in X with $x_n \to a$ and $y_n \to b$, and then choose N such that $d(x_n,y_n) < \delta$ for all $n \geq N$. Then

$$d'(f(x_n),f(y_n)) < \varepsilon$$

[10] See Exercise 1.15 (p. 34) and the comments on p. 339 on the continuous extension of continuous functions.

for all $n \geq N$. Letting $n \to \infty$, and using the continuity of the metric d', we obtain the inequality $d'(F(a), F(b)) \leq \varepsilon$.

If f is Lipschitz continuous with some constant L, then consider two arbitrary points $a, b \in X$. Choosing again two sequences (x_n) and (y_n) in X with $x_n \to a$ and $y_n \to b$, letting $n \to \infty$ in the inequality

$$d'(f(x_n), f(y_n)) \leq Ld(x_n, y_n)$$

we obtain

$$d'(F(a), F(b)) \leq Ld(a, b)$$

as desired.

Finally, if f is an isometry, then using the above notations, letting $n \to \infty$ in the identity

$$d'(f(x_n), f(y_n)) = d(x_n, y_n)$$

we obtain that

$$d'(F(a), F(b)) = d(a, b)$$

for all $a, b \in X$. □

The following important property was essentially established at the beginning of the above proof:

Proposition 1.15 *Let X, X' be two metric spaces and F, $G : X \to X'$ two continuous functions. If $F = G$ on a dense set D of X, then $F = G$ on X.*

Proof Given any $a \in X$, by the density of D there exists a sequence $(x_n) \subset D$ converging to a. Letting $n \to \infty$ in the equality $F(x_n) = G(x_n)$ and using the continuity of F and G in a we conclude that $F(a) = G(a)$. □

In view of the usefulness of completeness it is important to know that every metric space may be completed. More precisely, we have the following

Proposition 1.16 (Hausdorff) *Given a non-complete metric space (X, d), there exists a complete metric space (X', d') and an isometry $h : X \to X'$.*

Remark Using the isometry h we may identify (X, d) with the subspace $h(X)$ of (X', d').

Proof Consider the complete metric space $(X', d') = \mathcal{B}(X, \mathbb{R})$ and fix an arbitrary point $a \in X$. For each $x \in X$ we define a function $h(x) = h_x : X \to \mathbb{R}$ by setting

$$h_x(y) := d(x, y) - d(a, y), \quad y \in X.$$

Then $h_x \in \mathcal{B}(X, \mathbb{R})$ because

$$|h_x(y) - h_x(z)| \le |h_x(y)| + |h_x(z)| \le 2d(x,a)$$

for all $y, z \in X$ by the triangle inequality.

Since

$$|h_x(z) - h_y(z)| = |d(x,z) - d(y,z)| \le d(x,y)$$

for all $z \in X$, we have

$$d'(h_x, h_y) \le d(x,y) \tag{1.4}$$

for all $x, y \in X$. The converse inequality also holds because

$$d'(h_x, h_y) \ge |h_x(y) - h_y(y)| = d(x,y).$$

□

Remark If the metric d is bounded, then the proof may be simplified by setting $h_x(y) := d(x,y)$.

If (X', d') is replaced by the set of limits of the convergent sequences in $h(X)$,[11] then we may also assume in the preceding proposition that the range of h is dense in (X', d'). Then (X', d') is essentially unique:

***Proposition 1.17** *Let X be a metric space and X_1, X_2 two complete metric spaces. Assume that there exist two isometries $h_i : X \to X_i$ such that $h_i(X)$ is dense in X_i $(i = 1, 2)$. Then there exists an isometric bijection $f : X_1 \to X_2$ such that $f \circ h_1 = h_2$.*

Proof Applying Proposition 1.14 the isometries

$$h_2 \circ (h_1)^{-1} : h_1(X) \to h_2(X) \quad \text{and} \quad h_1 \circ (h_2)^{-1} : h_2(X) \to h_1(X)$$

may be extended to two isometries $f : X_1 \to X_2$ and $g : X_2 \to X_1$. Since

$$(g \circ f)(x) = x \text{ for all } x \in h_1(X), \text{ and } h_1(X) \text{ is dense in } X_1,$$
$$(f \circ g)(x) = x \text{ for all } x \in h_2(X), \text{ and } h_2(X) \text{ is dense in } X_2,$$

applying Proposition 1.15 we have

$$g \circ f = \mathrm{id}_{X_1} \quad \text{and} \quad f \circ g = \mathrm{id}_{X_2}.$$

This shows that f is a bijection between X_1 and X_2. □

[11] This is called the *closure* of $h(X)$, see the end of Sect. 2.1 below, p. 42.

1.4 Compactness

We start by recalling a basic theorem of classical analysis:

Proposition 1.18 (Bolzano–Weierstrass) *Every bounded real sequence has a convergent subsequence.*

Proof It follows from the usual axioms of real numbers that every bounded and *monotone* sequence is convergent. Therefore the theorem follows from the next proposition. □

Proposition 1.19 (Kürschák) *Every real sequence has a monotone subsequence.*

Proof An element x_n of the sequence (x_n) is called a *peak* if $x_n > x_m$ for all $m > n$.[12] If there are infinitely many peaks, then they form a decreasing subsequence.

Otherwise there are no peaks after some index N. Starting with x_{N+1} we may therefore define by induction a non-decreasing subsequence. □

It follows from Proposition 1.18 that in a bounded closed set of real numbers every sequence has a convergent subsequence. This may fail even in complete metric spaces:

Example Let K be an infinite set in a discrete (and hence complete) metric space, and (x_n) a sequence formed by distinct elements of K. Then K is bounded and closed, but (x_n) has no convergent subsequence.

A more useful generalization of bounded closed number sets is provided by the *compact* sets:

Definitions Let X be a metric space.

- $a \in X$ is a *cluster point* of a sequence $(x_n) \subset X$ if (x_n) has a subsequence converging to a.
- X is *compact* if every sequence in X has at least one cluster point.
- More generally, a set $K \subset X$ is *compact* if every sequence in K has at least one cluster point belonging to K.[13]

Examples

- By the Bolzano–Weierstrass theorem the compact sets of $\mathbb{R}$ are the bounded closed sets.
- In discrete metric spaces the compact sets are the finite sets.

[12] Representing x_n by the vertical halfline $\{(n,t) \in \mathbb{R}^2 \ : \ t \le x_n\}$, x_n is a peak if (n, x_n) is “visible from the right”.

[13] The empty set is compact because there is no sequence to check this property.

Remarks

- The limit of a convergent sequence is also a cluster point. Moreover, by Proposition 1.2 (c) it is its *unique* cluster point.
- In compact metric spaces the converse of the preceding property also holds. For if a is a cluster point, but not a limit of a sequence (x_n), then by Proposition 1.2 (d) there exists a subsequence (x_{n_k}) of (x_n) for which a is no longer a cluster point. By compactness this sequence has a cluster point b that is necessarily different from a, and b is also a cluster point of the original sequence (x_n).
- For Cauchy sequences the two notions coincide. Indeed, let a be a cluster point of a Cauchy sequence (x_n), and consider a subsequence $x_{n_k} \to a$. For any given $\varepsilon > 0$ choose N such that $d(x_n, x_m) < \varepsilon/2$ for all $n, m \geq N$, and then choose a large index $n_k \geq N$ such that $d(x_{n_k}, a) < \varepsilon/2$. Then

  $$d(x_n, a) \leq d(x_n, x_{n_k}) + d(x_{n_k}, a) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

  for all $n \geq N$. This shows that $x_n \to a$.

The following result is analogous to Lemma 1.3, p. 9:

Lemma 1.20 *Given a sequence (x_n) and a point a in a metric space, the following conditions are equivalent:*

(a) *every open set U containing a contains infinitely many elements of (x_n);*
(b) *every ball $B_r(a)$ contains infinitely many elements of (x_n);*
(c) *a is a cluster point of (x_n), i.e., there exists a subsequence $x_{n_k} \to a$.*

Proof

(a) $\Rightarrow$ (b) because the balls $B_r(a)$ are open.
(b) $\Rightarrow$ (c) We may define by induction a strictly increasing sequence (n_k) of positive integers such that $x_{n_k} \in B_{1/k}(a)$ for all k.
(c) $\Rightarrow$ (a) Choose a ball $B_r(a) \subset U$. There exists an index k_0 such that $d(a, x_{n_k}) < r$ for all $k \geq k_0$, and then $x_{n_k} \in U$ for all $k \geq k_0$. □

There is a variant of Proposition 1.9 (p. 15):

Proposition 1.21

(a) *The product of finitely many compact metric spaces is compact.*
(b) *The* closed *subspaces of compact metric spaces are compact.*

Proof

(a) Consider a sequence (x_n) in the product (X, d) of the compact metric spaces $(X_1, d_1), \ldots, (X_m, d_m)$, and write $x_n = (x_{n1}, \ldots, x_{nm})$.

Since (X_1, d_1) is compact, there exists a point $a_1 \in X_1$ and a subsequence (x_n^1) of (x_n) such that $d_1(x_{n1}^1, a_1) \to 0$.

Next, since (X_2, d_2) is compact, there exists a point $a_2 \in X_2$ and a subsequence (x_n^2) of (x_n^1) such that $d_2(x_{n2}^2, a_2) \to 0$.

Since (x_{n1}^2) is a subsequence of (x_{n1}^1), we also have $d_1(x_{n1}^2, a_1) \to 0$.

Continuing, after m steps we obtain a subsequence (x_n^m) of (x_n) and suitable points $a_j \in X_j$ such that

$$d_j(x_{nj}^m, a_j) \to 0, \quad j = 1, \dots, m.$$

Setting $a := (a_1, \dots, a_m) \in X$ we conclude that

$$d(x_n^m, a) = d_1(x_{n1}^m, a_1) + \cdots + d_m(x_{nm}^m, a_m) \to 0,$$

i.e., $x_n^m \to a$ in (X, d).

(b) Let F be a closed subspace of a compact metric space (X, d), and (x_n) a sequence in F. Since X is compact, (x_n) has a subsequence (x_{n_k}) converging to some point $x \in X$. Since F is closed, $x \in F$. □

Proposition 1.22 *The compact sets of metric spaces are closed.*

Proof Consider a convergent sequence $x_n \to x$ in a metric space X, and assume that the elements x_n belong to some compact set K. We have to show that $x \in K$.

Since K is compact, there exists a convergent subsequence, say $x_{n_k} \to y \in K$. Since we also have $x_{n_k} \to x$, by the uniqueness of the limit we conclude that $x = y$. Therefore $x = y \in K$. □

Next we establish the following version of Proposition 1.12 (p. 17):

Proposition 1.23 (Cantor's Intersection Theorem) *Let (F_n) be a non-increasing sequence of non-empty compact sets in a metric space X. Then the sets F_n have at least one common point.*[14]

Proof We select a point x_n in each set F_n. Since $(x_n) \subset F_1$ and F_1 is compact, there exists a point $x \in F_1$ and a subsequence $x_{n_k} \to x \in X$. Since F_m is closed and $(x_{n_k})_{k \geq m} \subset F_m$ for each m, x belongs to each F_m. □

Next we study the continuous functions defined on compact sets.

Theorem 1.24 (Weierstrass) *If $f : X \to \mathbb{R}$ is a continuous function on a compact metric space X, then f is bounded; moreover, it has maximal and minimal values.*

Proof Choose a *minimizing sequence,* i.e., a sequence $(x_n) \subset X$ satisfying

$$f(x_n) \to \inf_{x \in X} f(x) =: m.$$

Since X is compact, there exists a convergent subsequence $x_{n_k} \to a$. We clearly have $f(x_{n_k}) \to m$.

[14] By Proposition 1.21 (b) the hypotheses are satisfied if (F_n) is a non-increasing sequence of non-empty *closed* sets in a *compact* metric space X.

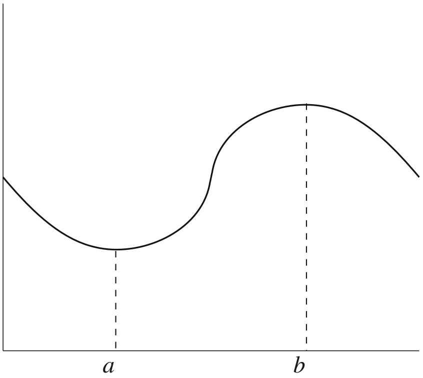


Fig. 1.7 Weierstrass' theorem

Since f is continuous at a, we also have $f(x_{n_k}) \to f(a)$, and hence $f(a) = m$ by the uniqueness of the limit (Fig. 1.7).

The proof for the maximum is analogous. □

We illustrate the usefulness of this theorem by studying the diameter

$$\operatorname{diam} K := \sup \{d(x,y) \ : \ x, y \in K\}$$

of a set and the *distance*

$$\operatorname{dist}(K,L) := \inf \{d(x,y) \ : \ x \in K, y \in L\}$$

of two sets in a metric space (X,d).[15]

For non-empty compact sets these upper and lower bounds are attained:

***Proposition 1.25** *Let K and L be non-empty compact sets in a metric space (X,d).*

(a) *There exist points $a, b \in K$ such that* $\operatorname{diam} K = d(a,b)$.
(b) *There exist points $a \in K$ and $b \in L$ such that* $\operatorname{dist}(K,L) = d(a,b)$.

Proof

(a) The restriction of the metric d to the product set $K \times K$ is continuous. Since $K \times K$ is compact, this function has a maximal value.
(b) The restriction of the metric d to the product set $K \times L$ is continuous. Since $K \times L$ is compact, this function has a minimal value. □

[15] If one of the sets has only one point, e.g., if $K = \{a\}$, then we write $\operatorname{dist}(a,L)$ instead of $\operatorname{dist}(\{a\}, L)$.

The following important generalization of Theorem 1.24 states that *the continuous image of a compact set is compact:*

Theorem 1.26 (Hausdorff) *Let X, Y be metric spaces and $f : X \to Y$ a continuous function. If K is a compact set in X, then $f(K)$ is a compact set in Y.*

Proof Given an arbitrary sequence (x_n) in K, we have to find a subsequence (x_{n_k}) and a point $a \in K$ such that $f(x_{n_k}) \to f(a)$.

Since K is compact, there exists a subsequence (x_{n_k}) and a point $a \in K$ such that $x_{n_k} \to a$. Since f is continuous at a, this implies that $f(x_{n_k}) \to f(a)$. □

Another important property is that in compact spaces the notions of continuity and uniform continuity coincide:

Theorem 1.27 (Heine) *Let X, X' be metric spaces and $f : X \to X'$ a continuous function. If X is compact, then f is uniformly continuous.*

Proof We denote by d and d' the metrics of X and X'. Assume on the contrary that f is not uniformly continuous. Then there exist $\varepsilon > 0$ and two sequences (x_n), (y_n) in X such that $d(x_n, y_n) \to 0$, but $d'(f(x_n), f(y_n)) \ge \varepsilon$ for all n. Since X is compact, (x_n) has a convergent subsequence $x_{n_k} \to a$. Since $d(x_{n_k}, y_{n_k}) \to 0$, using the triangle inequality we also have $y_{n_k} \to a$:

$$d(a, y_{n_k}) \le d(a, x_{n_k}) + d(x_{n_k}, y_{n_k}) \to 0.$$

Since f is continuous at a, we have $f(x_{n_k}) \to f(a)$ and $f(y_{n_k}) \to f(a)$, so that $d'(f(x_{n_k}), f(y_{n_k})) \to 0$ by the triangle inequality in X'. This contradicts the choice of the sequences (x_n) and (y_n). □

Next we give two important characterizations of compactness.

Definition A set K in a metric space is *totally bounded* (or *completely bounded* or *precompact*)[16] if for any given $r > 0$ it has a finite cover by balls of radius r.

Since a ball of radius r has a diameter $\le 2r$ and since every set of diameter $< r$ is contained in a ball of radius r, an equivalent definition is obtained by replacing the words "balls of radius r" by "sets of diameter $< r$".

Examples

- The implications

$$\text{finite} \Longrightarrow \text{totally bounded} \Longrightarrow \text{bounded}$$

[16] See the footnote to Corollary 1.29 below on the terminology "precompact".

always hold. The first one is obvious. To prove the second we fix $\varepsilon > 0$ arbitrarily and we cover K by the balls $B_r(x_1), \dots, B_r(x_n)$. Then

$$\operatorname{diam} K \le 2r + \max \left\{ d(x_i, x_j) \ : \ i,j = 1, \dots, n \right\} < \infty.$$

- In discrete metric spaces we have the equivalence

$$\text{finite} \iff \text{totally bounded.}$$

- In $\mathbb{R}^N$ we have the equivalence

$$\text{totally bounded} \iff \text{bounded;}$$

see Theorem 3.10 (c) below (p. 79) for a more general result.

Theorem 1.28 *Given a set K in a metric space (X, d), the following properties are equivalent:*

(a) *every open cover of K has a finite subcover;*
(b) *K is compact, i.e., every sequence in K has a subsequence converging to some element of K;*
(c) *K is totally bounded and complete.*

Property (a) means that if $K \subset \cup U_\alpha$ for an arbitrary family $\{U_\alpha\}$ of open sets, then

$$K \subset U_{\alpha_1} \cup \dots \cup U_{\alpha_n}$$

for a suitable *finite* subfamily.

Proof

(a) $\Rightarrow$ (b). If K is not compact, then there exists a sequence $(x_n) \subset K$ having no cluster point in K. Using Lemma 1.20, p. 24, we may fix for each $x \in K$ an open ball U_x centered at x that contains at most finitely many elements of the sequence (x_n). These balls form an open cover of K without any finite subcover, because any finite union of sets U_x contains at most finitely many elements of the sequence $(x_n) \subset K$. Hence K does not have property (a) either.
(b) $\Rightarrow$ (c). First we show that K is complete. If (x_n) is a Cauchy sequence in K, then it has a cluster point in K by (b). Since (x_n) is a Cauchy sequence, this cluster point is also its limit by the last remark on p. 24.
Now assume that K is not totally bounded, and fix $r > 0$ such that K cannot be covered by finitely many balls of radius r. Fix an arbitrary point $x_1 \in K$, and construct by recursion a sequence (x_n) satisfying

$$x_n \in K \setminus \bigcup_{i=1}^{n-1} B_r(x_i)$$

for all $n = 2, 3, \ldots$. Then $d(x_n, x_m) \geq r$ whenever $m \neq n$, so that the sequence $(x_n) \subset K$ cannot have cluster points, contradicting the compactness of K.
(c) $\Rightarrow$ (a). Assume on the contrary that K has an open cover $\{U_\alpha\}$ without any finite subcover. We say that a closed subset of K is *bad* if it cannot be covered by finitely many sets U_α. For example, K itself is a bad set.
Each bad set F has arbitrarily small bad subsets. Indeed, since K is totally bounded, for each $r > 0$ it has a finite cover by closed balls B_j of diameter $< r$. At least one of the sets $B_j \cap F$ is bad, for otherwise each of them and hence also F would have a finite cover by sets U_α, contradicting our assumption.
Using this property, starting with $F_1 := K$ we may construct a non-increasing sequence (F_n) of bad sets, satisfying $\operatorname{diam} F_n \to 0$. By Proposition 1.12 (p. 17) they have a common point x. Since $x \in F_1 = K$, it belongs to some set U_β. Since U_β is open and $\operatorname{diam} F_n \to 0$, we have $F_n \subset U_\beta$ if n is sufficiently large, contradicting the badness of F_n. □

Now we generalize the Bolzano–Weierstrass theorem:

Corollary 1.29 *In a* complete *metric space a set is compact if and only if it is totally bounded and closed.*[17]

Proof If K is totally bounded and closed, then it is complete by Proposition 1.9 (b) (p. 15), and hence compact by the preceding theorem.

If K is compact, then it is closed by Proposition 1.22 (p. 25) and totally bounded by the preceding theorem. □

We end this section by proving that compact metric spaces are not "too big".

Definition A metric space is *separable* if it contains a countable dense set.

Example $\mathbb{R}$ is separable because the rational numbers are dense in $\mathbb{R}$.

Proposition 1.30

(a) *Every compact metric space is separable.*
(b) *All subspaces of a separable metric space are separable.*
(c) *Let X, Y be metric spaces and $f : X \to Y$ a continuous function. If A is a dense set in X, then $f(A)$ is a dense set in the subspace $f(X)$ of Y.*
(d) *The continuous image of a separable metric space is separable.*

Proof

(a) If X is a compact metric space, then for each positive integer n it has a finite cover by balls of radius $1/n$:

$$X = B_{1/n}(a_{n1}) \cup \cdots \cup B_{1/n}(a_{nk_n}), \quad n = 1, 2, \ldots.$$

[17] Since the closure of a totally bounded set is still totally bounded (see the footnote on p. 22 on the notion of closure), hence a set in a complete metric space is precompact if and only if its closure is compact.

Then the countable set

$$A := \{a_{nj} \ : \ n = 1, 2, \ldots, \quad j = 1, \ldots, k_n\}$$

is dense in X because each ball $B_r(a) \subset X$ meets A.

Indeed, fix an integer $n > 1/r$ and then choose an index j such that $a \in B_{1/n}(a_{nj})$. Then $a_{nj} \in B_r(a)$.

(b) Let A be a countable dense set in a separable metric space X, and consider a subspace Y. If the ball $B_{1/n}(a)$ meets Y for some $a \in A$ and $n \geq 1$, then choose a point $y_{an} \in Y \cap B_{1/n}(a)$. We claim that the countable set $\{y_{an}\} \subset Y$ meets each ball $B_r(y) \subset Y$.

Fix an integer $n > 2/r$, and then choose a point $a \in A \cap B_{1/n}(y)$. Then $y \in Y \cap B_{1/n}(a)$, so that we have chosen earlier a point $y_{an} \in Y \cap B_{1/n}(a)$. Since

$$d(y, y_{an}) \leq d(y, a) + d(a, y_{an}) < 2/n < r,$$

we conclude that $y_{an} \in B_r(y)$.

(c) It suffices to show that if A is dense in X, then $f(A)$ is dense in $f(X)$. Given $f(x) \in f(X)$ arbitrarily we choose a sequence $(a_n) \subset A$ converging to x. Then $(f(a_n)) \subset f(A)$, and $f(a_n) \to f(x)$ by the continuity of f.

(d) Apply (c) with a countable dense set A in X. □

1.5 Exercises

Exercise 1.1 If a function $d : X \times X \to \mathbb{R}$ satisfies the conditions

$$d(x, y) = 0 \Longleftrightarrow x = y$$

and

$$d(x, y) \leq d(z, x) + d(z, y)$$

for all $x, y, z \in X$, then d is a metric on X.

Exercise 1.2 An *ultrametric* is a metric satisfying the following strengthening of the triangle inequality:

$$d(x, y) \leq \max\{d(x, z), d(z, y)\}.$$

If d is an ultrametric on X, then (X, d) is called an *ultrametric space*. Prove that the following metrics are ultrametrics:

(i) The discrete metrics.

(ii) We define the distance between two sequences of real numbers to be equal to 2^{-n} if they first differ at their nth elements.

(iii) Fix a prime number p, and define the distance between two rational numbers $x \neq y$ to be p^{-n} if $x - y = ap^n/b$ with integers a, b, neither of which is a multiple of p.[18]

Henceforth, let (X, d) be an ultrametric space. Prove the following:

(iv) All triangles are either equilateral, or isosceles with the unequal side being the shortest.
(v) Every point in a given ball is a center of that ball, with the same radius.
(vi) If two balls intersect, then one of them contains the other.
(vii) All open balls are closed and all closed balls are open.

Exercise 1.3 Prove the following results concerning the fixed point Theorem 1.10, p. 16.

(i) We cannot take $L = 1$ in the theorem, even if X is compact.
(ii) We cannot weaken the contraction assumption to

$$d(f(x), f(y)) < d(x, y) \quad \text{whenever} \quad x \neq y.$$

(iii) However, the weaker assumption of (ii) implies the existence of a unique fixed point in *compact* metric spaces.
(iv) The function $f(x) := x^2$ on the compact interval $[0, 1]$ has a unique fixed point, although the above assumption is not satisfied.
(v) Every monotone and continuous function $f : I \to I$, where I is a non-empty compact interval, has at least one fixed point.

Exercise 1.4 Let X be a metric space.

(i) The accumulation points of a set in X form a closed set.
(ii) The cluster points of a sequence in X form a separable closed set.
(iii) Conversely, every non-empty separable closed subset of X is the set of cluster points of a suitable sequence in X.
(iv) X is compact if and only if every infinite subset of X has an accumulation point in X.
(v) If a sequence (x_n) of real numbers satisfies $x_{n+1} - x_n \to 0$, then its cluster points form a closed interval.

Exercise 1.5 Let $f : [0, \infty) \to [0, \infty)$ be a non-decreasing, continuous function satisfying $f(a) = 0 \Longleftrightarrow a = 0$, and $f(a + b) \leq f(a) + f(b)$ for all a, b.[19] Prove that if d is a metric on X, then $f \circ d$ is also a metric on X.

Furthermore, the convergent or Cauchy sequences, and the closed, open, complete and compact sets are the same for the two metrics.

Exercise 1.6 (Lebesgue) Let (U_i) be an open cover of a compact set K in a metric space (X, d). Prove that there exists a $\delta > 0$ such that if $x, y \in K$ and $d(x, y) < \delta$, then there exists an index i satisfying $x, y \in U_i$.

[18] The completion of $\mathbb{Q}$ for this metric is the field of *p-adic numbers*.

[19] A function satisfying the last condition is called *subadditive*.

Exercise 1.7 Prove that we may define the product of countably many metric spaces by the formula

$$d(x,y) := \sum_{i=1}^{\infty} \frac{d_i(x_i,y_i)}{1+2^i d_i(x_i,y_i)}.$$

Prove that the metric convergence in the product space is equivalent to the component-wise convergence.

Exercise 1.8 (Cantor's Ternary Set) Starting with $C_1 := [0,1]$, we define a decreasing family (C_n) of compact sets as follows. First we remove from C_1 its open middle third $(\frac{1}{3}, \frac{2}{3})$ to get $C_2 = [0, \frac{1}{3}] \cup^* [\frac{2}{3}, 1]$: the union of two disjoint closed intervals.

Next we remove the open middle third $(\frac{1}{9}, \frac{2}{9})$ and $(\frac{7}{9}, \frac{8}{9})$ from the latter two closed intervals to get C_3: the union of four disjoint closed intervals.

Continuing by induction, in the nth step we remove 2^{n-1} open middle thirds to get C_{n+1}: the union of 2^n disjoint closed intervals.

The intersection $C := \cap C_n$ is called *Cantor's ternary set*. Prove the following:

(i) C consists of those points $t \in [0,1]$ that can be written in the form

$$t = 2\Big(\frac{t_1}{3} + \frac{t_2}{3^2} + \cdots + \frac{t_n}{3^n} + \cdots\Big) \tag{1.5}$$

for suitable integers $t_n \in \{0,1\}$.

(ii) C is closed, and has no isolated points.

(iii) C has the power of the continuum.

Exercise 1.9 A set $A \subset \mathbb{R}$ is called *perfect* if it is closed and has no isolated points. For example, every non-degenerate[20] closed interval is perfect, and Cantor's ternary set in the preceding exercise is perfect. Prove the following:

(i) A set is perfect if and only if it is equal to the set of its accumulation points.

(ii) Every open set $U \subset \mathbb{R}$ is the union of countably many disjoint open intervals.

(iii) Every closed set $F \subset \mathbb{R}$ has the form

$$F = \mathbb{R} \setminus \cup^*_{i\in I}(a_i,b_i) \tag{1.6}$$

with countably many disjoint open intervals (a_i,b_i).[21]

(iv) A closed set $F \subset \mathbb{R}$ is perfect if and only if the *closed* intervals $[a_i,b_i]$ in (1.6) are also pairwise disjoint.

(v) (Cantor–Bendixson theorem) Every closed set $F \subset \mathbb{R}$ is the union of a perfect set and a countable (maybe finite or empty) set.

[20] An interval is *non-degenerate* if it has at least two (and hence infinitely many) points.

[21] The superscript $*$ indicates that the intervals (a_i,b_i) are pairwise disjoint. The union may be finite or even empty.

Exercise 1.10 A set $A \subset \mathbb{R}^n$ is called *perfect* if it is closed, and has no isolated points. For example, every closed ball is perfect. We say that $x \in \mathbb{R}^n$ is a *condensation point* of $A \subset \mathbb{R}^n$ if each neighborhood of x contains uncountably many points of A. Prove the following:

(i) A set is perfect if and only if it is equal to the set of its accumulation points.
(ii) (Lindelöf) Every uncountable set $A \subset \mathbb{R}^n$ has at least one condensation point.
(iii) The condensation points of a set $A \subset \mathbb{R}^n$ form a perfect set P, and $P \setminus A$ is countable (maybe finite or empty).
(iv) (Cantor–Bendixson theorem) Every closed set $F \subset \mathbb{R}^n$ is the union of a perfect set and a countable set.

Exercise 1.11 (Alexandroff's Theorem) Let (X, d) be a complete metric space. Prove the following:

(i) If G is a non-empty proper open subset of X, then the formula

$$D(x,y) := d(x,y) + \left| \frac{1}{\operatorname{dist}(x, X \setminus G)} - \frac{1}{\operatorname{dist}(y, X \setminus G)} \right|$$

defines a complete metric on G.

Furthermore, (G, D) and the metric subspace (G, d) have the same convergent sequences, and the same open, closed and compact sets. We say that the metrics d and D are *equivalent* on G.

(ii) More generally, if $Y = \cap_{i=1}^{\infty} G_i$ where $G_1, G_2, \dots$ is a sequence of open sets in X (such sets are called G_δ *sets*), then the formula

$$D(x,y) := d(x,y) + \sum_{i=1}^{\infty} 2^{-i} \min \left\{ 1, \left| \frac{1}{\operatorname{dist}(x, X \setminus G_i)} - \frac{1}{\operatorname{dist}(y, X \setminus G_i)} \right| \right\}$$

defines a complete metric on Y that is equivalent to d on Y. In other words, every non-empty G_δ subset of a complete metric space is completely metrizable.

(iii) Every closed set in a metric space is also a G_δ set.
(iv) Only the non-empty G_δ subsets of complete metric spaces are completely metrizable.

Exercise 1.12 (Hausdorff Distance) Consider the family $\mathcal{A}$ of *bounded closed* sets in a metric space. For $A \in \mathcal{A}$ and $r > 0$ we set $V_r(A) = \{x : \operatorname{dist}(x, A) < r\}$. Prove that the formula

$$d(A,B) := \inf\{r : A \subset V_r(B) \quad \text{and} \quad B \subset V_r(A)\}$$

defines a metric on $\mathcal{A}$.

Exercise 1.13 The null set $f^{-1}(0)$ of a continuous function $f : \mathbb{R} \to \mathbb{R}$ is closed. Conversely, every closed set of real numbers is the null-set of a suitable continuous function $f : \mathbb{R} \to \mathbb{R}$.

Exercise 1.14 We define the *oscillation function* $\omega_f : \mathbb{R} \to \mathbb{R}$ of a function $f : \mathbb{R} \to \mathbb{R}$ by the formula

$$\omega_f(x) := \lim_{r \to 0} \left(\sup_{B_r(x)} f - \inf_{B_r(x)} f \right).$$

Prove the following:

(i) ω_f is well-defined, nonnegative, *upper semi-continuous*,[22] and

$$f \quad \text{is continuous in} \quad x \Longleftrightarrow \omega_f(x) = 0.$$

(ii) f is continuous in x if and only if

$$x \in \bigcap_{n=1}^{\infty} \left\{ y \in \mathbb{R} \; : \; \omega_f(y) < \frac{1}{n} \right\}.$$

Hence the set of continuity of f is a G_δ set.

(iii) There is no function $f : \mathbb{R} \to \mathbb{R}$ that is continuous exactly at the rational points.

(iv) There exists a function $f : \mathbb{R} \to \mathbb{R}$ that is continuous exactly at the irrational points.

(v) Every G_δ set of real numbers is the set of continuity of a suitable function $f : \mathbb{R} \to \mathbb{R}$.

Exercise 1.15 Let X be a metric space.

(i) (Urysohn) If A and B are non-empty disjoint closed sets in X, then the formula

$$f(x) := \frac{\operatorname{dist}(x, A) - \operatorname{dist}(x, B)}{\operatorname{dist}(x, A) + \operatorname{dist}(x, B)}$$

defines a continuous function $f : X \to \mathbb{R}$ satisfying $f = -1$ on A and $f = 1$ on B.

(ii) Given a closed set $F \subset X$ and a continuous function $g : F \to [-M, M]$, there exists a continuous function

$$f_1 : X \to [-\frac{M}{3}, \frac{M}{3}] \quad \text{satisfying} \quad |g - f_1| \le \frac{2M}{3} \quad \text{on} \quad F.$$

[22] The upper semi-continuity means that if $\omega_f(x) < A$, then $\omega_f(y) < A$ for all y in a neighborhood of x. See also Exercise 2.11, p. 64.

(iii) Given a closed set $F \subset X$ and a continuous function $g : F \to [-M, M]$, there exists a continuous function $f : X \to [-M, M]$ satisfying $f = g$ on F.

(iv) (Tietze) Given a closed set $F \subset X$, every continuous function $g : F \to \mathbb{R}$ may be extended continuously to X.

Chapter 2
Topological Spaces

Metric spaces are convenient for many applications in geometry and physics, but sometimes we need a broader framework: that of topological spaces. In this chapter we give a short introduction to this theory.

The results of the preceding chapter on continuous functions remain valid. On the other hand, the reader may observe the absence of *sequences*: they retain their usefulness only in special topological spaces.[1]

Topological spaces are not suitable for the study of *uniform* continuity either.[2]

2.1 Definitions and Examples

Definitions By a *topology* on a non-empty set X we mean a family $\mathcal{T}$ of subsets of X satisfying the following conditions:

(a) $\varnothing \in \mathcal{T}$ and $X \in \mathcal{T}$;
(b) the intersection of *finitely many* sets of $\mathcal{T}$ still belongs to $\mathcal{T}$;
(c) the union of an *arbitrary* subfamily of $\mathcal{T}$ still belongs to $\mathcal{T}$.

If $\mathcal{T}$ is a topology on X, then the pair $(X, \mathcal{T})$ is called a *topological space*, and the elements of $\mathcal{T}$ are called the *open sets* of $(X, \mathcal{T})$. When the topology is evident from the context, we speak simply of the open sets of X. The elements of X are also called *points*.

[1] In the last, optional section of this chapter we introduce a generalization of sequences that is well suited for all topological spaces.

[2] The convenient framework for such studies is provided by the *uniform spaces*, see, e.g., Császár [118] or Kelley [273].

V. Komornik, *Topology, Calculus and Approximation*, Springer
Undergraduate Mathematics Series, DOI 10.1007/978-1-4471-7316-8_2

If $\mathcal{T}$, $\mathcal{S}$ are two topologies on the same set X and $\mathcal{T} \subset \mathcal{S}$, then we say that $\mathcal{T}$ is *coarser* or *weaker* than $\mathcal{S}$, or that $\mathcal{S}$ is *finer* or *stronger* than $\mathcal{T}$.[3]

Examples Let X be a non-empty set.

- The family of *all* subsets of X is called the *discrete topology* on X: every subset of X is open. This is the finest topology on X.
- The family $\mathcal{T} = \{\varnothing, X\}$ is called the *antidiscrete topology* on X: there are only two open sets. This is the coarsest topology on X.

We introduce an important special class of topological spaces:

Definition A topological space X is *separated* or is a *Hausdorff space* if any two points $x, y \in X$ may be separated by disjoint open sets U and V:

$$x \in U, \quad y \in V \quad \text{and} \quad U \cap V = \varnothing.$$

Examples

- Every discrete topological space is a Hausdorff space.
- The antidiscrete topology on a set of at least two points is not separated.

Proposition 1.1 (p. 5) shows that the open sets of a metric space form a separated topology. Henceforth we associate this topology with the metric. Then we have the

Proposition 2.1 *Every metric space is a Hausdorff space.*

***Remarks**

- The topology associated with a discrete metric is the discrete topology, so that our terminology is consistent.
- Two metrics d_1 and d_2 on the same set X are called *equivalent* if they are associated with the same topology. This is the case, for example, if there exist two positive constants c_1, c_2 such that

  $$c_1 d_1(x,y) \le d_2(x,y) \le c_2 d_1(x,y) \tag{2.1}$$

  for all $x, y \in X$.
- The condition (2.1) is sufficient, but not necessary for the equivalence. For example, on the set of positive integers the two metrics

  $$d_1(x,y) = |x-y| \quad \text{and} \quad d_2(x,y) = |x^{-1} - y^{-1}|$$

  define the same (discrete) topology, although (2.1) is not satisfied.
- Although the above two metrics define the same topology, d_1 is complete, while d_2 is not. This shows that completeness is not a topological property.

[3] We do not exclude equality.

- Not all Hausdorff spaces are metrizable[4]: see, e.g., the second example on p. 45 and the examples on p. 58. Other, natural examples appear in the study of weak topologies in *Functional Analysis*.

Next we define products and subspaces of topological spaces.

Proposition 2.2 *If $\mathcal{T}$ is a topology on X and $Y \subset X$ is a non-empty set, then*

$$\mathcal{T}_Y := \{U \cap Y \ : \ U \in \mathcal{T}\}$$

is a topology on Y.

Proof We check the three properties of topologies.

(a) $\varnothing = \varnothing \cap Y \in \mathcal{T}_Y$ and $Y = X \cap Y \in \mathcal{T}_Y$.
(b) If $U_i \in \mathcal{T}$ for $i = 1, \dots, n$, then

$$\cap_{i=1}^n (U_i \cap Y) = \left(\cap_{i=1}^n U_i\right) \cap Y \in \mathcal{T}_Y.$$

(c) If $\{U_i\}_{i \in I}$ is an arbitrary subfamily of $\mathcal{T}$, then

$$\cup_{i \in I} (U_i \cap Y) = (\cup_{i \in I} U_i) \cap Y \in \mathcal{T}_Y.$$

□

Definition By a *(topological) subspace* of a topological space X we mean a non-empty subset $Y \subset X$ endowed with the topology $\mathcal{T}_Y$.[5]

If $\mathcal{T}_i$ is a topology on X_i for $i = 1, \dots, m$, then we denote by $\mathcal{B}$ the family of *base sets*

$$B = U_1 \times \cdots \times U_m$$

in $X := X_1 \times \cdots \times X_m$, where U_i runs over $\mathcal{T}_i$ for each i.

Furthermore, we denote by $\mathcal{T}$ the family of arbitrary (finite or infinite) unions of base sets:

$$\mathcal{T} = \{\cup_{\alpha \in A} B^\alpha \ : \ B^\alpha \in \mathcal{B} \quad \text{for all} \quad \alpha \in A\}.$$

Proposition 2.3 *$\mathcal{T}$ is a topology on X.*

Proof We check again the three properties of topologies.

(a) $\varnothing = \varnothing \times \cdots \times \varnothing \in \mathcal{T}$ and $X = X_1 \times \cdots \times X_m \in \mathcal{T}$.

[4] See, e.g., Császár [118] or Kelley [273] for the characterization of metrizable topological spaces.

[5] This is consistent with our terminology on metric spaces: the topology of a metric subspace coincides with the subspace topology.

(b) If $U_1^i \times \cdots \times U_m^i \in \mathcal{B}$ for $i = 1, \dots, n$, then

$$\bigcap_{i=1}^{n} (U_1^i \times \cdots \times U_m^i) = \left(\bigcap_{i=1}^{n} U_1^i \right) \times \cdots \times \left(\bigcap_{i=1}^{n} U_m^i \right) \in \mathcal{B}.$$

Next, if $U^1, \dots, U^n \in \mathcal{T}$, then each U^i has the form

$$U^i = \bigcup_{\alpha_i \in A_i} B^{\alpha_i}$$

with suitable sets $B^{\alpha_i} \in \mathcal{B}$. Hence

$$\bigcap_{i=1}^{n} U^i = \bigcup_{\alpha_1 \in A_1} \cdots \bigcup_{\alpha_n \in A_n} (B^{\alpha_1} \cap \cdots \cap B^{\alpha_n}) \in \mathcal{T}.$$

(c) Each U^i again has the form

$$U^i = \bigcup_{\alpha_i \in A_i} B^{\alpha_i}$$

with suitable sets $B^{\alpha_i} \in \mathcal{B}$. Hence

$$\bigcup_{i \in I} U^i = \bigcup_{i \in I} \bigcup_{\alpha_i \in A_i} B^{\alpha_i} \in \mathcal{T}.$$

□

Definition The pair $(X, \mathcal{T})$ is called the *(topological) product* of $(X_1, \mathcal{T}_1), \dots, (X_m, \mathcal{T}_m)$.[6]

Different topological spaces may have the same structure:

Definition X and Y are *homeomorphic* if there exists a bijection f between X and Y such that

$$U \text{ is open in } X \Longleftrightarrow f(U) \text{ is open in } Y.$$

The map f is then called a *homeomorphism*.

It is clear that homeomorphism is an equivalence relation between topological spaces. Homeomorphic topological spaces have the same topological properties. For example, if X is compact, separable or separated, then every homeomorphic topological space is also compact, separable or separated.

[6] The topology of a product of metric spaces coincides with the product of the corresponding topologies. We will also define the product of infinitely many spaces in Sect. 2.4, p. 53.

The following result asserts that the topological product is essentially commutative, associative, and the projections are homeomorphisms.

Proposition 2.4 *Let X_1, X_2, X_3 be topological spaces.*

(a) *$X_1 \times X_2$ is homeomorphic to $X_2 \times X_1$.*
(b) *$(X_1 \times X_2) \times X_3$ is homeomorphic to $X_1 \times X_2 \times X_3$.*
(c) *$X_1 \times \{a_2\}$ is homeomorphic to X_1 for each fixed $a_2 \in X_2$.*

Proof It follows from the definition of the product topology that the maps $(x_1, x_2) \mapsto (x_2, x_1)$, $((x_1, x_2), x_3) := (x_1, x_2, x_3)$ and $(x_1, a_2) \mapsto x_1$ are suitable homeomorphisms. □

***Proposition 2.5**

(a) *All subspaces of Hausdorff spaces are Hausdorff spaces.*
(b) *The products of Hausdorff spaces are Hausdorff spaces.*

Proof

(a) Let Y be a subspace of X, and $a, b \in Y$ two distinct points. Since X is separated, there exist two disjoint open sets U, V in X satisfying $a \in U$ and $b \in V$. Then $U \cap Y$ and $V \cap Y$ are disjoint open sets in Y, and $a \in U \cap Y$, $b \in V \cap Y$.
(b) Let $(X, \mathcal{T})$ be the product of the Hausdorff spaces

$$(X_1, \mathcal{T}_1), \dots, (X_m, \mathcal{T}_m),$$

and $a = (a_1, \dots, a_m), b = (b_1, \dots, b_m)$ be two distinct points in X. There exists an index j such that $a_j \neq b_j$, and then there exist two disjoint open sets U_j, V_j in X_j such that $a_j \in U_j$ and $b_j \in V_j$. Then

$$U := \left\{(x_1, \dots, x_m) \in X \ : \ x_j \in U_j\right\}$$

and

$$V := \left\{(x_1, \dots, x_m) \in X \ : \ x_j \in V_j\right\}$$

are disjoint open sets in X with $a \in U$ and $b \in V$. □

We define the closed sets in the same way as in metric spaces:

Definition A set in a topological space is *closed* if its complement is open.

The definition yields at once the following

Proposition 2.6 *Let X be a topological space.*

(a) *$\varnothing$ and X are closed sets.*
(b) *The union of finitely many closed sets is closed.*
(c) *The intersection of any (finite or infinite) family of closed sets is closed.*

The closed sets of (topological) subspaces are easily characterized:

Proposition 2.7 *Let Y be a subspace of a topological space X. The closed sets of Y are the sets $F \cap Y$ where F runs over the closed sets of X.*

Proof The closed sets of Y are the complements of the open sets of Y, i.e. the sets $Y \setminus (Y \cap U)$, where U runs over the open sets of X. Since

$$Y \setminus (Y \cap U) = Y \cap (X \setminus U),$$

we conclude by observing that $X \setminus U$ runs over the closed sets of X. □

Definitions Let A be a set in a topological space X.

- The union of all open subsets $A \subset X$ is the largest open subset of A. It is called the *interior* of A, and is denoted by $\operatorname{int} A$. Its elements are called the *interior points* of A.
- The intersection of all closed sets $F \supset A$ is the smallest closed set containing A as a subset. It is called the *closure* of A, and is denoted by $\overline{A}$.
- We have

  $$\operatorname{int} A \subset A \subset \overline{A}$$

  by definition. The set $\partial A := \overline{A} \setminus \operatorname{int} A$ is called the *boundary* of A. Its elements are called the *boundary points* of A.
- The union of all open sets disjoint from A is the largest open set disjoint from A. It is called the *exterior* of A, and is denoted by $\operatorname{ext} A$. Its elements are called the *exterior points* of A.

Remarks

- It follows from the definitions that the sets $\operatorname{ext} A$, $\operatorname{int} A$, ∂A form a (disjoint) partition of X.
- Since $\overline{A} = X \setminus \operatorname{ext} A$,

  $$a \in \overline{A} \Longleftrightarrow \quad \text{each ball } B_r(a) \text{ meets } A.$$

  By Lemma 1.3 there is a sequential characterization of $\overline{A}$ in *metric spaces*:

  $$a \in \overline{A} \Longleftrightarrow A \quad \text{contains a sequence converging to} \quad a.$$

- Since $\partial A = \overline{A} \cap (X \setminus \operatorname{int} A)$ is the intersection of two closed sets, the boundary of a set is always closed.

In the rest of this chapter the letters X, Y, Z will always denote topological spaces.

2.2 Neighborhoods: Continuous Functions

In order to define continuity we introduce the neighborhoods of a point.

Definition Let $a \in X$ and $V \subset X$. We say that V is a *neighborhood* of a if there is an open set U in X satisfying $a \in U \subset V$.

Lemma 2.8 *Let (X, d) and (Y, d') be two metric spaces, $f : X \to Y$ and $a \in X$. The following two properties are equivalent:*

(a) *for every $\varepsilon > 0$ there exists a $\delta > 0$ such that $f(B_\delta(a)) \subset B_\varepsilon(f(a))$, i.e.,*

$$x \in X \quad \textit{and} \quad d(x, a) < \delta \quad \Longrightarrow \quad d'(f(x), f(a)) < \varepsilon;$$

(b) *for each neighborhood V of $f(a)$, $f^{-1}(V)$ is a neighborhood of a.*

Proof

(a) ⇒ (b). Since V is a neighborhood of $f(a)$, there exists an open set U in Y such that $f(a) \in U \subset V$. By the definition of open sets in a metric space there exists an $\varepsilon > 0$ such that $B_\varepsilon(f(a)) \subset U$. Choose δ according to property (a), then $f(B_\delta(a)) \subset B_\varepsilon(f(a))$, so that

$$a \in B_\delta(a) \subset f^{-1}(B_\varepsilon(f(a))) \subset f^{-1}(U) \subset f^{-1}(V).$$

Since the ball $B_\delta(a)$ is open, $f^{-1}(V)$ is a neighborhood of a in X by definition.
(b) ⇒ (a). For any fixed $\varepsilon > 0$ the open ball $B_\varepsilon(f(a))$ is a neighborhood of $f(a)$ in Y by definition. Then $f^{-1}(B_\varepsilon(f(a)))$ is a neighborhood of a in X by property (b). There exists therefore an open set U satisfying $a \in U \subset f^{-1}(B_\varepsilon(f(a)))$. Using the definition of open sets in a metric space, there exists a $\delta > 0$ such that $B_\delta(a) \subset U$. Then $B_\delta(a) \subset f^{-1}(B_\varepsilon(f(a)))$, whence $f(B_\delta(a)) \subset B_\varepsilon(f(a))$. □

Comparing Lemmas 1.5 and 2.8 we see that the following definitions are consistent with those given earlier for metric spaces:

Definitions A function $f : X \to Y$ is *continuous* at $a \in X$ if for each neighborhood V of $f(a)$ (in Y), its inverse image $f^{-1}(V)$ is a neighborhood of a (in X).
The function f is *continuous* if it is continuous at every $a \in X$.

The earlier results on the continuity of composite functions remain valid:

Proposition 2.9 *Consider two functions $g : X \to Y$ and $f : Y \to Z$.*

(a) *If g is continuous at a and f is continuous at $g(a)$, then $f \circ g$ is continuous at a.*
(b) *If f and g are continuous, then $f \circ g$ is continuous.*

Proof

(a) If V is a neighborhood of $(f \circ g)(a) = f(g(a))$ in Z, then $f^{-1}(V)$ is a neighborhood of $g(a)$ in Y because f is continuous at $g(a)$, and then $g^{-1}(f^{-1}(V))$ is a neighborhood of a in X because g is continuous at a. This proves our claim because

$$(f \circ g)^{-1}(V) = g^{-1}(f^{-1}(V)).$$

(b) Apply (a) to each $a \in X$. □

There is an elegant characterization of continuity in terms of open or closed sets:

Proposition 2.10 (Hausdorff) *Given a function $f : X \to Y$, the following properties are equivalent:*

(a) *f is continuous;*
(b) *the inverse image $f^{-1}(U) \subset X$ of each open set $U \subset Y$ is open;*
(c) *the inverse image $f^{-1}(F) \subset X$ of each closed set $F \subset Y$ is closed.*

As a consequence, a bijection *f between X and Y is a homeomorphism if and only if both f and f^{-1} are continuous.*

Proof

(a) ⇒ (b). Let U be an open set in Y. We have to show that $f^{-1}(U)$ is open in X, i.e., it is a neighborhood of each $a \in f^{-1}(U)$. Since U is open and $f(a) \in f\bigl(f^{-1}(U)\bigr) \subset U$, U is a neighborhood of $f(a)$. Using our assumption we conclude that $f^{-1}(U)$ is a neighborhood of a.
(b) ⇒ (a). If V is a neighborhood of some point $f(a) \in Y$, then there exists an open set U in Y such that $f(a) \in U \subset V$. It follows that $a \in f^{-1}(U) \subset f^{-1}(V)$. Since $f^{-1}(U)$ is open in X by (b), $f^{-1}(V)$ is a neighborhood of a in X.
(b) ⇒ (c). If F is closed in Y, then $Y \setminus F$ is open in Y, and therefore $f^{-1}(Y \setminus F)$ is open in X by assumption. Using the equality

$$f^{-1}(F) = X \setminus f^{-1}(Y \setminus F)$$

we conclude that $f^{-1}(F)$ is closed.
(c) ⇒ (b). If U is open in Y, then $Y \setminus U$ is closed in Y, and therefore $f^{-1}(Y \setminus U)$ is closed in X by assumption. But then its complement

$$f^{-1}(U) = X \setminus f^{-1}(Y \setminus U)$$

is open.
The last assertion follows from the equivalence (a) $\Longleftrightarrow$ (b). □

Next we generalize two more notions from the theory of metric spaces.

Definitions Let X be a topological space.

- A set $D \subset X$ is *dense* if it meets every non-empty open set, i.e., if $\overline{D} = X$.[7]
- X is *separable* if it contains a countable dense set.

Two properties in Proposition 1.30 (p. 29) remain valid:

Corollary 2.11 *Let X, Y be topological spaces and $f : X \to Y$ a continuous map* onto *Y.*

(a) *If D is dense in X, then $f(D)$ is dense in Y.*
(b) *If X is separable, then Y is also separable.*

Proof

(a) Given any non-empty open set V in Y, its preimage $f^{-1}(V)$ is a *non-empty* open set in X. By our assumption $f^{-1}(V) \cap D$ contains at least one point x, and then $f(x) \in V \cap f(D)$.
(b) If D is a countable dense set in X, then $f(D)$ is a countable dense set in Y by (a). □

The two other properties in Proposition 1.30 may fail in topological spaces:

Examples

- The empty set and the sets containing 0 form a separable topology on $\mathbb{R}$ because the one-point set $\{0\}$ is dense. But its subspace $\mathbb{R} \setminus \{0\}$ is not separable because it is an uncountable set endowed with the discrete topology, so that no countable set is dense.[8]
- A *compact* Hausdorff topology is defined on $\mathbb{R}$ by the complements of the finite sets and the subsets of $\mathbb{R} \setminus \{0\}$. Then the closed sets are the finite sets and the sets containing 0. Hence $\overline{D} \subset D \cup \{0\}$ for all sets, and therefore the closure of any countable set is also countable. Since $\mathbb{R}$ is uncountable, no countable set is dense.

Corollary 2.12 *Let $f : X \to \mathbb{R}$ be a continuous function and $c \in \mathbb{R}$.*

(a) *If U is an open set in X, then the sets*

$$\{x \in U : f(x) < c\}, \quad \{x \in U : f(x) > c\}, \quad \{x \in U : f(x) \neq c\}$$

are open in X.

(b) *If F is a closed set in X, then the sets*

$$\{x \in F : f(x) \le c\}, \quad \{x \in F : f(x) \ge c\}, \quad \{x \in F : f(x) = c\}$$

are closed in X.

[7] This definition is consistent with the definition of density in metric spaces; see pp. 18 and 42.

[8] All sets are closed in the discrete topology.

Proof Introducing the open interval $I := (-\infty, c)$, the set

$$\{x \in U \ : f(x) < c\}$$

is the intersection of the open sets $f^{-1}(I)$ and U, hence it is also open.

The other proofs are similar. □

Before giving an important example we generalize a classical theorem stating that the *uniform limit of a sequence of continuous functions is also continuous*:

Proposition 2.13 *Let K be a topological space, (X, d) a metric space, and (f_n) a sequence of functions $f_n : K \to X$. Assume that (f_n) converges uniformly to some function $f : K \to X$.*

If each f_n is continuous at some point $a \in K$, then f is also continuous at a.

Proof For any given $\varepsilon > 0$ we choose n such that

$$d(f(t), f_n(t)) < \varepsilon/3 \quad \text{for all} \quad t \in K,$$

and then we choose a small neighborhood U of a such that

$$d(f_n(t), f_n(a)) < \varepsilon/3 \quad \text{for all} \quad t \in U.$$

Then

$$\begin{aligned} d(f(t), f(a)) &\le d(f(t), f_n(t)) + d(f_n(t), f_n(a)) + d(f_n(a), f(a)) \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon \end{aligned}$$

for every point $t \in U$. □

Example Let K be a topological space and X a metric space. By the preceding proposition the *continuous* and *bounded* functions $f : K \to X$ form a *closed* subspace $C_b(K, X)$ of the metric space $\mathcal{B}(K, X)$. Consequently, if X is complete, then $C_b(K, X)$ is also complete.

When $X = \mathbb{R}$ we often write $C_b(K)$ instead of $C_b(K, X)$ for brevity.

2.3 Connectedness

The intervals C of the real line are connected in the sense that whenever $x, z \in C$ and $x < y < z$, we also have $y \in C$. There is also a topological characterization:

Proposition 2.14 *A non-empty set $A \subset \mathbb{R}$ is an interval if and only if in its subspace topology only A and $\varnothing$ are both open and closed.*

Proof If A is not an interval, then there exist points $x < y < z$ such that $x, z \in A$, but $y \notin A$. Then $B := \{a \in A \ : \ a < y\}$ is a non-empty proper subset of A. Since the relation $y \notin A$ implies that

$$B = A \ \cap \ (-\infty, y) = A \ \cap \ (-\infty, y],$$

B is both open and closed by the definition of the subspace topology and by Proposition 2.7 (p. 42).

Now let B be a non-empty proper subset of an interval A. Fix two points $x \in B$ and $z \in A \setminus B$. Assume that $x < z$ (the other case is similar), and consider the point $y = \sup \{a \in B \ : \ a < z\}$. Then $x \le y \le z$, so that $y \in A$. If $y \notin B$, then B is not closed. If $y \in B$, then $y < z$ and $(y, z] \cap B = \varnothing$, so that B is not open. □

The preceding proposition suggests the following

Definition A topological space X (and its topology) is *connected* if only X and $\varnothing$ are both open and closed.

Equivalently, X is connected if every non-empty proper subset of X has a boundary point.

A subset of X is *connected* if it is empty, or if its subspace topology is connected.

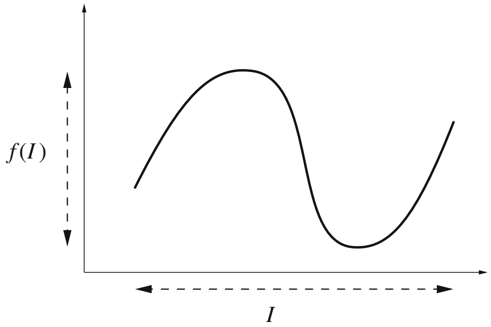
Examples

- Every antidiscrete topological space is connected.
- The discrete topological spaces having at least two points are not connected.

A well-known theorem of Bolzano states that continuous images of intervals are also intervals: see Fig. 2.1. More generally, *continuous images of connected sets are also connected*:

Theorem 2.15 *Let $f : X \to Y$ be a continuous function. If X is connected, then $f(X)$ is also connected.*

Fig. 2.1 Bolzano's theorem



Proof We have to show that if B is a non-empty, open and closed subset of the subspace $f(X)$, then $B = f(X)$. Since f is continuous, $f^{-1}(B)$ is a non-empty, open and closed subset of X. Since X is connected, this implies that $f^{-1}(B) = X$. We conclude by using the relations

$$f(X) = f\left(f^{-1}(B)\right) \subset B \subset f(X).$$

□

Let us prove some basic properties of connected sets:

Proposition 2.16

(a) *Let $\{A_\alpha\}_{\alpha\in I}$ be a (finite or infinite) family of connected sets in a topological space. If $\cap A_\alpha \neq \varnothing$, then $\cup A_\alpha$ is connected.*
(b) *The product of finitely many connected topological spaces is connected.*
(c) *The closure of a connected set is connected.*

Proof

(a) Let C be a non-empty, open and closed set in $\cup A_\alpha$. We have to show that $A_\alpha \subset C$ for each α.

If C meets one of the sets A_α, then $A_\alpha \subset C$. Indeed, $C \cap A_\alpha$ is a non-empty, open and closed set in the subspace topology of A_α. Since A_α is connected, $C \cap A_\alpha = A_\alpha$, i.e., $A_\alpha \subset C$.

Since C is non-empty, it meets at least one of the sets A_β. Then it contains this set, and hence it contains $\cap A_\alpha \neq \varnothing$ as well. Therefore C meets each A_α, and thus $A_\alpha \subset C$ for each α.

(b) Using induction on the number of factors, by Proposition 2.4 (b) (p. 41) it suffices to prove that the product Z of two connected sets X and Y is connected.

Fix $a \in X$ arbitrarily, and consider the sets

$$Z_b = (\{a\} \times Y) \cup (X \times \{b\}),\ b \in Y;$$

see Fig. 2.2.

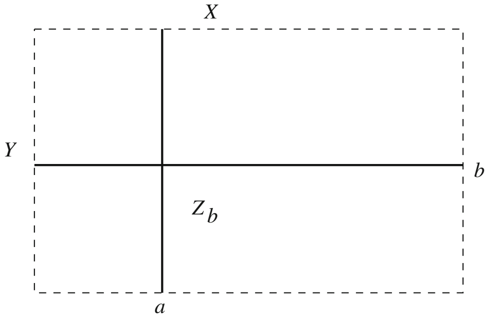


Fig. 2.2 A product of connected spaces

By Proposition 2.4 (c) the subspaces $\{a\} \times Y$ and $X \times \{b\}$ of Z are homeomorphic to Y and X, respectively, and hence connected. Since they have a common point (a, b), Z_b is connected for all b by (a).

The intersection of the sets Z_b contains $\{a\} \times Y$, hence it is non-empty. Applying (a) again we conclude that $Z = \cup_{b \in Y} Z_b$ is connected.

(c) Let C be a non-empty, open and closed set in $\overline{A}$. We have to show that $C = \overline{A}$. Since $\overline{A}$ is the smallest closed set containing A as a subset, it suffices to show that $A \subset C$.

C is a non-empty open set in $\overline{A}$, hence it meets A by the definition of the closure. Therefore $C \cap A$ is non-empty, open and closed in A. Since A is connected, we conclude that $C \cap A = A$, i.e., $A \subset C$. □

Remarks

- We say that two points of a topological space X *may be connected* if they belong to some connected set. By the preceding proposition this is an equivalence relation, and the corresponding equivalence classes are connected and closed. They are the maximal connected sets of X, called its *connected components*.
- Each open and closed subset A of X is a union of connected components. Indeed, if A meets a component C, then $A \cap C$ is a non-empty open and closed subset of the subspace C, and hence equal to C by its connectedness. Equivalently, $C \subset A$.
- A connected component is not necessarily open. To see this we consider the subspace of $\mathbb{R}$ formed by 0 and the numbers $1/n$, $n = 1, 2, \ldots$. The one-point sets $\{1/n\}$ are both open and closed, hence connected components by our preceding remark. Then the remaining set $\{0\}$ is also a connected component, and it is not open.

We end this section by giving a useful and intuitive *sufficient* condition for connectedness.

Definition

- Two points a and b in a topological space X *may be connected by a path* if there exists a continuous function $f : [0, 1] \to X$ satisfying $f(0) = a$ and $f(1) = b$. The function f is also called a *path*.
- X is *pathwise connected* if any two points of X may be connected by a path.

Proposition 2.17 *Every pathwise connected space is connected.*

Proof Fix $a \in X$ arbitrarily, and choose for each $x \in X$ a path f_x connecting a and x. Their ranges have a common point a, hence their union, i.e., the space X, is connected by Theorem 2.15 and Proposition 2.16 (a). □

2.4 * Compactness

Compactness may be defined in every topological space. Moreover, all results of Sect. 1.4 remain valid except those using the notions of metric, completeness, boundedness or complete boundedness. However, new proofs are needed that avoid the use of sequences.

Definition A set K in a topological space X is *compact* if every open cover of K has a finite subcover.

If X itself is compact, then X is called a *compact topological space*.

Remark Theorem 1.28 (p. 28) shows that the above definition is consistent with the earlier one for metric spaces. The other two properties of Theorem 1.28 cannot be used here. Indeed, as we observed on page 38, completeness is not a topological property. Furthermore, we should modify the definition of cluster points to avoid the countable nature of sequences.[9]

Proposition 2.18

(a) *In a compact topological space all closed subsets are compact.*
(b) *The product of finitely many compact topological spaces is compact.*

Proof

(a) Let F be a closed subset of a compact topological space X, and $\{U_i\}$ an open cover of F. Then adding $X \setminus F$ we obtain an open cover of X. By assumption the latter has a finite subcover:

$$X = U_{i_1} \cup \cdots \cup U_{i_n}$$

or

$$X = (X \setminus F) \cup U_{i_1} \cup \cdots \cup U_{i_n}.$$

In both cases we have

$$F \subset U_{i_1} \cup \cdots \cup U_{i_n},$$

so that $\{U_i\}$ has a finite subcover of F.

(b) Using induction on the number of factors, by Proposition 2.4 (b) (p. 41) it suffices to consider the product $Z = X \times Y$ of two compact topological spaces. Let $\{W_\alpha\}$ be an open cover of Z.

[9] See Sect. 2.5 for such a modification.

By the definition of the product topology each point $z = (x, y) \in Z$ has a neighborhood $W'_z = U_z \times V_z$ such that $W'_z \subset W_\alpha$ for some $\alpha = \alpha(z)$. Then $\{W'_z \ : \ z \in Z\}$ is an open cover of Z. It suffices to find a finite subcover

$$Z = W'_{z_1} \cup \cdots \cup W'_{z_n}$$

of the latter because then we will also have

$$Z = W_{\alpha(z_1)} \cup \cdots \cup W_{\alpha(z_n)}.$$

Fix $a \in X$ arbitrarily and consider the sets $W'_z = U_z \times V_z$ meeting $\{a\} \times Y$. The corresponding open sets V_z cover the compact set Y, hence there exists a finite subcover: $Y = \cup_{i=1}^{k(a)} V_{a,z_i}$. Setting $U'_a := \cap_{i=1}^{k(a)} U_{a,z_i}$, we also have $U'_a \times Y \subset \cup_{i=1}^{k(a)} W'_{a,z_i}$.

The sets U'_a form an open cover of the compact set X, hence there exists a finite subcover $X = \cup_{j=1}^{m} U_{a_j}$. Then

$$Z = \cup_{j=1}^{m} \cup_{i=1}^{k(a_j)} W'_{a_j,z_i}$$

as required. □

Proposition 2.19 *In* Hausdorff *spaces every* compact *set is closed.*

Example The Hausdorff property cannot be omitted: for example, in an antidiscrete topological space X all sets are compact, but only $\varnothing$ and X are closed.

Proof Let K be a compact set in a Hausdorff space X. For any fixed point $a \in X \setminus K$ we have to find an open set U satisfying $a \in U \subset X \setminus K$: this will prove that $X \setminus K$ is open, i.e., K is closed.

By the separation hypothesis, for each point $b \in K$ there exist open sets U_b and V_b satisfying $a \in U_b$, $b \in V_b$ and $U_b \cap V_b = \varnothing$. Since K is compact, the open cover $K \subset \cup_{b \in K} V_b$ has a finite subcover, say

$$K \subset V_{b_1} \cup \cdots \cup V_{b_n}.$$

Then $U := U_{b_1} \cap \cdots \cap U_{b_n}$ is an open neighborhood of a, and $U \cap K = \varnothing$ because U does not meet any of the sets V_{b_i}. □

Proposition 2.20 (Cantor's Intersection Theorem) *Let (F_n) be a non-increasing sequence of non-empty compact sets in a topological space X. Then the sets F_n have at least one common point.*[10]

[10] By Proposition 2.18 (a) the hypotheses are satisfied if (F_n) is a non-increasing sequence of non-empty *closed* sets in a *compact* topological space.

Proof Assume on the contrary that $\cap F_n = \varnothing$. Then their complements form an open cover of X, and hence also of F_1. Since F_1 is compact, it has a finite subcover. Furthermore, since the complements form a non-decreasing sequence, we have $F_1 \subset X \setminus F_m$ for a sufficiently large index m. Since $F_m \subset F_1$, this implies that F_m is empty, contradicting our assumption. □

Theorem 2.21 (Hausdorff) *Let $f : X \to Y$ be a continuous function. If K is a compact set in X, then $f(K)$ is compact in Y.*

Proof If $\{U_i\}$ is an open cover of $f(K)$ in Y, then $\{f^{-1}(U_i)\}$ is an open cover of K in X by the continuity of f. Since K is compact, there exists a finite subcover:

$$K \subset f^{-1}(U_{i_1}) \cup \cdots \cup f^{-1}(U_{i_n}).$$

Hence

$$f(K) \subset U_{i_1} \cup \cdots \cup U_{i_n},$$

i.e., $\{U_i\}$ contains a finite subcover of $f(K)$. □

Theorem 2.22 (Weierstrass) *If $f : K \to \mathbb{R}$ is a continuous function on a compact topological space K, then f is bounded; moreover, it has maximal and minimal values.*

The result may be deduced from the preceding theorem: the set $f(K) \subset \mathbb{R}$ is compact, hence bounded and closed. Since K and hence $f(K)$ is non-empty, it has a maximal and a minimal element.

Let us also give a direct proof:

Proof Assume on the contrary that $\alpha := \inf_K f$ is not attained (the proof for $\sup_K f$ is similar).

Then the open sets

$$U_n := \left\{ x \in K \ : \ f(x) > \alpha + \frac{1}{n} \right\}, \quad n = 1, 2, \ldots$$

cover K. By the compactness of K there exists a finite subcover. Since (U_n) is a non-decreasing set sequence, we have $K = U_N$ for some N. This implies that $f(x) > \alpha + N^{-1}$ for all $x \in K$, contradicting the definition of α. □

Remark Consider the spaces $C_b(K, X)$ (p. 46) where K is compact. Then every continuous function $f : K \to X$ is bounded, so that we often write $C(K, X)$ and $C(K)$ instead of $C_b(K, X)$ and $C_b(K)$. Furthermore, we may write

$$d_\infty(f, g) = \max_{t \in K} d_\infty(f(t), g(t))$$

instead of sup for all $f, g \in C(K, X)$.

The following sufficient condition is often used to check that a given map is a homeomorphism:

Proposition 2.23 (Hausdorff) *Let $f : X \to Y$ be a continuous bijection of a compact topological space onto a Hausdorff space. Then f is a homeomorphism.*

Proof We have to prove that the inverse map $f^{-1} : Y \to X$ is continuous. Equivalently, we prove that if A is a closed set in X, then $f(A)$ is closed in Y.

If A is closed in X, then it is also compact because X is compact. Then the continuous image $f(A)$ of A is also compact. Finally, since Y is a Hausdorff space, $f(A)$ is closed in Y. □

Example By a *simple plane curve* we mean the range of a continuous one-to-one function $f : [0, 1] \to \mathbb{R}^2$. Such curves are homeomorphic to the interval $[0, 1]$. Indeed, $[0, 1]$ is compact, and a simple plane curve, as a subspace of the Hausdorff space $\mathbb{R}^2$ is also a Hausdorff space.

Remark In view of the importance of compactness we remark that every topological space may be considered as a dense subspace of a compact topological space. There are usually many ways to do this.[11]

We may also define infinite products of topological spaces:

Definition Let $(X_i)_{i\in I}$ be an arbitrary family of topological spaces. Let us denote by X the set of points of the form $x = (x_i)_{i\in I}$ where $x_i \in X_i$ for each $i \in I$.[12] We endow X with the following topology: a set $U \subset X$ belongs to $\mathcal{T}$ if for each $a \in U$ there exists a *finite* index set $J \subset I$ and for each $j \in J$ an open set U_j in X_j such that

$$a \in \bigl\{x \in X \ : \ x_j \in U_j \quad \text{for all} \quad j \in J\bigr\} \subset U.$$

One can readily check that $\mathcal{T}$ is indeed a topology on X, and that for finite index sets I this definition is equivalent to the earlier one.

This notion is justified by the following important theorem:

Theorem 2.24 (Tychonoff) *An arbitrary product of compact topological spaces is compact.*

For the proof we need two lemmas on a generalization of the method of induction to the uncountable case and on the verification of compactness by considering only special open covers.

[11] See Exercise 2.8 (p. 63) for an example, and Császár [118] or Kelley [273] for a general treatment.

[12] More precisely, but less intuitively, we could consider the functions $x : I \to X$ defined by the formula $x(i) := x_i$.

Definitions Let $\mathcal{A}$ be a family of sets and $\mathcal{B} \subset \mathcal{A}$ a subfamily. We say that

- $A \in \mathcal{A}$ is a *maximal* element if it is not a proper subset of another element of $\mathcal{A}$;
- $\mathcal{B}$ is *monotone* if for any two sets $B_1, B_2 \in \mathcal{B}$ we have either $B_1 \subset B_2$ or $B_2 \subset B_1$;
- $A \in \mathcal{A}$ is an *upper bound* of $\mathcal{B}$ if $B \subset A$ for all $B \in \mathcal{B}$.
- We similarly define the *lower bounds* and the *minimal* elements.

The following lemma is equivalent to the *axiom of choice* in set theory:

Lemma 2.25 (Zorn) *Let $\mathcal{A}$ be a family of sets. If every monotone subfamily has an upper bound, then $\mathcal{A}$ has at least one maximal element.*

Similarly, if every monotone subfamily has a lower bound, then $\mathcal{A}$ has at least one minimal element.

Remark Applying the hypotheses of the lemma to the empty subfamily we see that $\mathcal{A}$ has at least one element.

Proof See, e.g., Hajnal and Hamburger [218] or Kelley [273]. □

Definition A family $\mathcal{S}$ of open sets in a topological space X is a *subbase* if for each open set V and for each $a \in V$ there exist finitely many sets $S_1, \dots, S_n \in \mathcal{S}$ such that

$$a \in S_1 \cap \cdots \cap S_n \subset V.$$

Example The *unbounded* open intervals form a subbase in $\mathbb{R}$.

***Proposition 2.26 (Alexander)** *Let $\mathcal{S}$ be a subbase in a topological space X. If every cover $\mathcal{U} \subset \mathcal{S}$ of X has a finite subcover, then X is compact.*

Proof Assume on the contrary that every cover $\mathcal{U} \subset \mathcal{S}$ of X has a finite subcover, but X is non-compact. Then it has an open cover without any finite subcover.

The family of such open covers satisfies the hypothesis of Zorn's lemma. Indeed, the union of any monotone subfamily is still an open cover without any finite subcover. For otherwise the finite subcover would already belong to some open cover of the subfamily. Applying Zorn's lemma we obtain the existence of a *maximal* open cover $\mathcal{V}$ without any finite subcover.

In particular, $\mathcal{V} \cap \mathcal{S}$ has no finite subcover, and therefore it cannot cover X by our assumption on the subbase. We may thus fix a point $a \in X$ not covered by these sets, and then an open set $V \in \mathcal{V}$ containing a. Finally, since $\mathcal{S}$ is a subbase, we may choose finitely many sets $S_1, \dots, S_n \in \mathcal{S}$ such that

$$a \in S_1 \cap \cdots \cap S_n \subset V.$$

Since $\mathcal{V} \cap \mathcal{S}$ does not cover a, none of the sets S_j belong to $\mathcal{V}$. Therefore, using the maximality of $\mathcal{V}$, $\{S_j\} \cup \mathcal{V}$ already has a finite subcover for each j. There exists therefore finitely many sets in $\mathcal{V}$ such that their union A_j satisfies the equality $S_j \cup A_j = X$. Then

$$X = \Big(S_1 \cap \cdots \cap S_n\Big) \cup \Big(A_1 \cup \cdots \cup A_n\Big) \subset V \cup A_1 \cup \cdots \cup A_n,$$

contradicting our hypothesis that $\mathcal{V}$ has no finite subcover. □

**Proof of Theorem 2.24* Introducing the *projections* $\pi_i : X \to X_i$, $\pi_i(x) := x_i$, by the definition of the product topology the family

$$\mathcal{S} := \left\{\pi_i^{-1}(U_i) \ : \ U_i \quad \text{is open in} \quad X_i, \quad i \in I\right\}$$

is a subbase of X. Therefore it suffices to prove that every cover $\mathcal{U} \subset \mathcal{S}$ has a finite subcover.

For each $i \in I$ we denote by $\mathcal{U}_i$ the family of open sets U_i in X_i satisfying $\pi_i^{-1}(U_i) \in \mathcal{U}$. There exists an index $k \in I$ such that $\mathcal{U}_k$ covers X_k. For otherwise there would exist a point $x \in X$ such that $\pi_i(x)$ is not covered by $\mathcal{U}_i$ for any i. Then $\mathcal{U}$ would not cover x, although it covers the whole space X by assumption.

If $\mathcal{U}_k$ covers X_k, then the compactness of the latter implies the existence of a finite subcover

$$X_k = U_1 \cup \cdots \cup U_n.$$

This implies the equality

$$X = \pi_k^{-1}(U_1) \cup \cdots \cup \pi_k^{-1}(U_n);$$

we have thus found a finite subcover of $\mathcal{U}$. □

2.5 * Convergence of Nets

Due to their countable character, sequences are less useful in topological spaces than in metric spaces.

Example Let X be an uncountable set (for example $X = \mathbb{R}$). The empty set and the complements of the countable sets form a topology on X. Although it is different from the discrete topology, the convergent sequences are the same in both of them: the eventually constant sequences.

The following notion may replace sequences in all topological spaces.

Definitions

- By a *net* in a set X[13] we mean a function $x : I \to X$ defined on a *directed set*, i.e., a non-empty set I endowed with a partial order having the following properties:

  – $i \ge i$ for all $i \in I$;
  – if $i \ge j$ and $j \ge k$, then $i \ge k$;
  – for all $i, j \in I$ there exists a $k \in I$ such that $k \ge i$ and $k \ge j$.

  Usually we write x_i instead of $x(i)$, and (x_i) or $(x_i)_{i \in I}$ instead of x.
- We say that a net $(x_i)_{i \in I}$ *eventually* satisfies some condition (or satisfies this condition *for all sufficiently large* i) if there exists a $j \in I$ such that x_i satisfies this condition for all $i \ge j$.
- A net $(x_i)_{i \in I}$ in X *converges* to $a \in X$ if for each neighborhood U of a, $x_i \in U$ for all sufficiently large i. We also say in this case that a is a *limit* of (x_i), and we write $x_i \to a$, $\lim x_i = a$ or $\lim\, (x_i)_{i \in I} = a$.
- We say that a net $(x_i)_{i \in I}$ *often* satisfies some condition if for each $j \in I$ there exists an $i \ge j$ such that x_i satisfies this condition.
- A net $(x_i)_{i \in I}$ in a *metric space* (X, d) is a *Cauchy net* if

$$\operatorname{diam} \left\{x_j \ : \ j \ge i\right\} \to 0$$

  in $\mathbb{R}$. Equivalently, this means that the net $\big(d(x_i, x_j)\big)_{(i,j) \in I \times I}$ converges to 0, where $I \times I$ is endowed with the following order relation:

$$(i, j) \ge (k, \ell) \quad \text{if} \quad i \ge k \text{ and } j \ge \ell.$$

Examples

- Sequences correspond to the case $I = \mathbb{N}$ with the usual ordering.
- Consider the following partial ordering of the set $\mathcal{U}$ of neighborhoods of a point a of a topological space:

$$U \ge V \iff U \subset V.$$

  If we choose a point $x_U \in U$ for each $U \in \mathcal{U}$, then we obtain a net $(x_U)_{U \in \mathcal{U}}$ converging to a.
- Every Cauchy sequence is also a Cauchy net.

Now we generalize some results on sequences:

Proposition 2.27 *Let X and Y be topological spaces, $a \in X$ and $D, A \subset X$.*

(a) *X is separated if and only if every net in X has at most one limit.*
(b) *$a \in \overline{D} \iff D$ contains a net converging to a.*

[13] This notion is equivalent to the more elegant but less transparent notion of a *filter*. See Császár [118].

(c) *A is closed in X if and only if*

$$(x_i) \subset A \quad \textit{and} \quad x_i \to a \Longrightarrow a \in A.$$

(d) *A function $f : X \to Y$ is continuous at $a \in X$ if and only if*

$$x_i \to a \quad \textit{in} \quad X \Longrightarrow f(x_i) \to f(a) \quad \textit{in} \quad Y.$$

(e) *A metric space is complete $\Longleftrightarrow$ every Cauchy net is convergent.*

Proof

(a) Let X be separated and $x_i \to a$ in X. We show that $x_i \not\to b$ if $b \neq a$. For this we separate a and b by two disjoint neighborhoods U and V. Since $x_i \to a$, we have $x_i \in U$ for all sufficiently large i, say $i \geq j$. Then $x_i \notin V$ for all $i \geq j$, so that $x_i \not\to b$.

On the other hand, if X is not separated, then there exist two points a and b such that each neighborhood U of a meets each neighborhood V of b. Let us consider the set I of all such pairs (U, V) of neighborhoods with the following partial ordering:

$$(U, V) \geq (U_0, V_0) \Longleftrightarrow U \subset U_0 \quad \text{and} \quad V \subset V_0.$$

Choosing a point $x_i \in U \cap V$ for each $i := (U, V) \in I$ we have $x_i \to a$ and $x_i \to b$.

(b) If a net $(x_i) \subset D$ converges to a, then each neighborhood U of a contains x_i if i is sufficiently large, so that U meets D. Hence $a \in \overline{D}$ by definition. Conversely, if $a \in \overline{D}$, then choosing for each neighborhood U of a a point $x_U \in U \cap D$, we get a net $x_U \to a$.

(c) If A is closed, $(x_i) \subset A$ and $x_i \to a$, then $a \in \overline{A} = A$ by (b). If A is not closed, then there exists a point $a \in \overline{A} \setminus A$, and then by (b) there exists a net $(x_i) \subset A$ satisfying $x_i \to a \notin A$.

(d) Let $f : X \to Y$ be continuous at a, and $x_i \to a$ in X. We have to show that $f(x_i) \to f(a)$ in Y. If V is a neighborhood of $f(a)$, then $U := f^{-1}(V)$ is a neighborhood of a by continuity, hence $x_i \in U$ for all sufficiently large i. Then $f(x_i) \in V$ for all sufficiently large i. This proves that $f(x_i)$ converges to $f(a)$.

If f is not continuous at a, then there exists a neighborhood V of $f(a)$ for which $f^{-1}(V)$ is not a neighborhood of a. Choose for each neighborhood U of a a point $x_U \in U \setminus f^{-1}(V)$. Then $x_U \to a$, but $f(x_U) \not\to f(a)$, because $f(x_U) \notin V$ for all U.

(e) If $x_i \to a$ is a convergent net in a metric space (X, d), then for each fixed $\varepsilon > 0$ there exists an index k such that $d(x_i, a) < \varepsilon/2$ for all $i \geq k$. Applying the triangle inequality we obtain that $d(x_i, x_j) < \varepsilon$ for all $i, j \geq k$, so that (x_i) is a Cauchy net.

If every Cauchy net is convergent in X, then in particular every Cauchy sequence is convergent, so that X is complete.

It remains to show that if X is complete, and (x_i) is a Cauchy net in X, then (x_i) is convergent. For each positive integer n we may choose an index $i(n)$ such that

$$d(x_i, x_j) < 1/n \quad \text{for all} \quad i, j \geq i(n).$$

We may assume that $i(1) \leq i(2) \leq \cdots$.[14] Then $x_{i(1)}, x_{i(2)}, \ldots$ is a Cauchy sequence, hence it converges to some point a.

We prove that $x_i \to a$. For any given $\varepsilon > 0$ there exists a positive integer n such that $1/n < \varepsilon$. Then

$$d(x_i, x_{i(m)}) < 1/n \quad \text{for all} \quad i \geq i(n) \quad \text{and} \quad m \geq n.$$

Since $x_{i(m)} \to a$, letting $m \to \infty$ and using the continuity of the metric we conclude that

$$d(x_i, a) \leq 1/n < \varepsilon \quad \text{for all} \quad i \geq i(n),$$

i.e., $x_i \to a$. □

The following examples show that the existence of convergent *subsequences* does not characterize the compact *topological* spaces.

Examples

- We consider the set X of functions $f : \mathbb{R} \to [0, 1]$ as the product $X := \prod_{t \in \mathbb{R}} I_t$ of the compact intervals $I_t = [0, 1]$. Convergence in X is pointwise convergence on $\mathbb{R}$. By Tychonoff's theorem (p. 53) X is compact.

  Let us denote by $f_n(x)$ the nth binary digit of x; if possible we consider finite expansions. Then the sequence (f_n) has no convergent subsequence. Indeed, for any given subsequence (f_{n_k}) consider a number $x \in [0, 1]$ whose n_kth binary digit is 0 or 1 according to whether k is even or odd. Then (f_{n_k}) is divergent at x.
- The functions $f : \mathbb{R} \to [0, 1]$ vanishing outside a countable set (that may depend on f) form a proper dense subset Y of X, so that Y is not compact. Nevertheless, in Y every sequence (f_n) has a convergent subsequence.

  Indeed, since the union of countably many countable sets is still countable, there exists a countable set A outside which all function f_n vanish. Hence they belong to the compact subspace $Z := \prod_{t \in \mathbb{R}} J_t$ of Y, where $J_t = [0, 1]$ if $t \in A$, and $J_t = \{0\}$ otherwise. One may readily check[15] that the topology of Z is metrizable with the metric

$$d(f, g) := \sum 2^{-n} \left| f(t_n) - g(t_n) \right|$$

[14] Use the last property in the definition of directed sets.

[15] Adapt the solution of Exercise 1.7, pp. 32 and 301.

where (t_n) is an arbitrary enumeration of the elements of A. Hence (f_n) has a convergent subsequence in the compact metric space Z, and then the convergence holds in X as well.

We may eliminate the above pathological situations by generalizing subsequences and cluster points.

Definition By a *subnet* of a net (x_i) we mean a net of the form $(x_{f(j)})_{j\in J}$ where the function $f : J \to I$ satisfies the following condition: for each $i \in I$ there exists a $j \in J$ such that

$$k \ge j \Longrightarrow f(k) \ge i.$$

Remark Subsequences correspond to the special case where $I = J = \mathbb{N}$ and f is an increasing function.

The following variant of Lemma 1.20 (p. 24) holds:

Lemma 2.28 *Given a net $(x_i)_{i\in I}$ and a point a in a topological space, the following conditions are equivalent:*

(a) *x_i often belongs to each neighborhood V of a;*
(b) *x_i often belongs to each open set U containing a;*
(c) *there exists a subnet $x_{f(j)} \to a$.*

Proof

(a) ⇒ (b) because U is a neighborhood of a.
(b) ⇒ (c) Let us denote by $\mathcal{U}$ the family of open sets containing a, and consider the set E of all pairs (i, U) in $I \times \mathcal{U}$ for which $x_i \in U$. Then E is a directed set for the partial order

$$(i, U) \ge (j, V) \Longleftrightarrow i \ge j \quad \text{and} \quad U \subset V.$$

Indeed, the reflexivity and transitivity relations are obvious. Furthermore, if $(i, U), (j, V) \in E$, then by (b) there exists a $k \in I$ satisfying $k \ge i$, $k \ge j$ and $x_k \in U \cap V$ because $U \cap V \in \mathcal{U}$. We conclude that

$$(k, U \cap V) \ge (i, U), \quad (k, U \cap V) \ge (j, V), \quad \text{and} \quad (k, U \cap V) \in E.$$

Setting $f(i, U) := i$ we obtain a subnet $(x_{f(i,U)})_{(i,U)\in E}$ of $(x_i)_{i\in I}$. Indeed, fixing $V \in \mathcal{U}$ arbitrarily, for each $i \in I$ we have the implication

$$(k, W) \ge (i, V) \Longrightarrow k \ge i \Longrightarrow f(k, W) \ge i$$

by the definition of f.

Finally, for any fixed $U \in \mathcal{U}$ we have to find $(i, V) \in E$ such that $x_{f(k,W)} \in U$ for all $(k, W) \geq (i, V)$. It suffices to choose $V := U$ and $i \in I$ satisfying $x_i \in U$. Indeed, then we have the implication

$$(k, W) \geq (i, V) \Longrightarrow x_{f(k,W)} = x_k \in W \subset U.$$

(c) $\Rightarrow$ (a) There exists a $k_1 \in I$ such that $x_{f(j)} \in V$ for all $j \geq k_1$. Furthermore, there exists a $k_2 \in I$ such $f(j) \geq i$ for all $j \geq k_2$. Choose $k \in I$ satisfying $k \geq k_1$ and $k \geq k_2$, then $f(k) \geq i$ and $x_{f(k)} \in V$. □

In view of Lemmas 1.20 (p. 24) and 2.28 the following definition is consistent with the former one given for metric spaces:

Definition In a topological space a is a *cluster point* of a net if the equivalent conditions of Lemma 2.28 are satisfied.

Remark Similarly to the last remark on p. 24, the limits and cluster points are the same for Cauchy nets.

Proposition 2.29 *Let X be a topological space.*

(a) *If $x_i \to a$, then every subnet of (x_i) converges to a.*
(b) *If $x_i \not\to a$, then (x_i) has a subnet of which a is not even a cluster point.*
(c) *A set K in X is compact $\Longleftrightarrow$ every net $(x_i) \subset K$ has a cluster point in K.*
(d) *Let X be a compact Hausdorff space. A net in X is convergent $\Longleftrightarrow$ it has a unique cluster point.*

Proof

(a) Let $(x_{f(j)})$ be a subnet of (x_i) and U a neighborhood of a. Since $x_i \to a$, there exists an index i_0 such that $x_i \in U$ for all $i \geq i_0$. By the definition of a subnet there exists an index $j_0 \in J$ such that $f(j) \geq i_0$ for all $j \geq j_0$. Then $x_{f(j)} \in U$ for all $j \geq j_0$. This proves that $x_{f(j)} \to a$.
(b) If $x_i \not\to a$, then a has a neighborhood U such that for each index i_0 there exists an $i \geq i_0$ satisfying $x_i \notin U$. Then considering the identical map $f : J \to I$ on the subset $J := \{i \in I \ : \ x_i \notin U\}$ of I we obtain a subnet $(x_j)_{j \in J}$ of (x_i). Since no element of this subnet belongs to U, a cannot be its cluster point.
(c) First we assume that K is not compact, and we fix an open cover $\{U_\alpha \ : \ \alpha \in A\}$ of K without any finite subcover. Consider the family I of the finite subsets of A with the partial ordering $i \geq j \Longleftrightarrow i \supset j$. By our assumption we may choose a point

$$x_i \in K \setminus \left(U_{\alpha_1} \cup \cdots \cup U_{\alpha_n}\right)$$

for each $i := \{\alpha_1, \ldots, \alpha_n\} \in I$. No subnet of $(x_i) \subset K$ converges to any $a \in K$ because a belongs to some U_α, and then $x_i \notin U_\alpha$ for all $i \geq \{\alpha\}$.

Now assume that no subnet of the net $(x_i) \subset K$ converges to any point of K. Then for each $a \in K$ there is a neighborhood U_a of a and an index i_a such that

$x_i \notin U_a$ for all $i \ge i_a$. We prove that the open cover

$$K \subset \bigcup_{a\in K} U_a$$

has no open subcover.

Indeed, choosing arbitrarily a finite number of points $a_1, \dots, a_n \in K$, there exists an index i such that $i \ge i_{a_1}, \dots, i \ge i_{a_n}$, and then

$$x_i \notin U_{a_1} \cup \cdots \cup U_{a_n}.$$

Hence K is not compact.

(d) We already know that every net (x_i) has at least one cluster point a. If (x_i) does not converge to a, then (x_i) has by (b) a subnet $(x_{f(j)})$ of which a is not even a cluster point. By (c) this subnet also has at least one cluster point b, necessarily different from a. But then b is also a cluster point of (x_i) by (a), so that (x_i) has at least two cluster points.

If $x_i \to a$, then (x_i) cannot have a cluster point different from a. Indeed, if $x_{f(j)} \to b$ for some subnet, then $x_{f(j)} \to a$ by (a), and therefore $a = b$ by Proposition 2.27 (a). □

2.6 Exercises

Exercise 2.1 Let A be a set in a topological space X. Prove the following equalities:

$$\begin{aligned}
&\operatorname{int}(X \setminus A) = \operatorname{ext} A;\\
&\operatorname{ext}(X \setminus A) = \operatorname{int} A;\\
&\partial(X \setminus A) = \partial A;\\
&\overline{A} = (\operatorname{int} A) \cup \partial A = A \cup \partial A = X \setminus \operatorname{ext} A.
\end{aligned}$$

Exercise 2.2 Let $A \subset X$ and $a \in X$. Prove the following:

$$\begin{aligned}
a \in \operatorname{int} A \quad &\Longleftrightarrow \quad A \quad \text{is a neighborhood of } a;\\
a \in \operatorname{ext} A \quad &\Longleftrightarrow \quad X \setminus A \quad \text{is a neighborhood of } a;\\
a \in \overline{A} \quad &\Longleftrightarrow \quad \text{each neighborhood of } a \text{ meets } A;\\
a \in \partial A \quad &\Longleftrightarrow \quad \text{each neighborhood of } a \text{ meets both } A \text{ and } X \setminus A.
\end{aligned}$$

Exercise 2.3 Prove the following statements:

(i) any two non-empty open intervals are homeomorphic;
(ii) any two non-degenerate bounded closed intervals are homeomorphic;
(iii) any two non-empty half-closed intervals are homeomorphic;
(iv) $(0, 1)$, $[0, 1]$ and $[0, 1)$ are pairwise non-homeomorphic;
(v) no interval is homeomorphic to a circle of the plane.

Exercise 2.4 (Accumulation Points) Given a point a and a set D in a topological space X, prove that the following properties are equivalent:

(i) every neighborhood of a meets D;
(ii) every open set containing a meets D;
(iii) there exists a net in D that converges to a.

We say that a is an *accumulation point* of a set $A \subset X$ if the above conditions are satisfied with $D := A \setminus \{a\}$. Show that this definition is compatible with the former one in metric spaces.

Show that in a compact topological space every infinite set has an accumulation point, but the converse statement may fail.

Exercise 2.5 (New Characterization of Compact Sets) We say that a family of sets has the *finite intersection property* if any finite number of sets of the family has a non-empty intersection.

Given a set K in a topological space, prove that the following properties are equivalent:

(i) K is compact;
(ii) if a family of closed sets in K has the finite intersection property, then the whole family has a non-empty intersection.

Exercise 2.6 (Initial Topology)

(i) Let Y be a non-empty set in a topological space X, and consider the embedding $i : Y \to X$, $i(x) := x$. Prove that the subspace topology on Y is the weakest topology on Y for which i is continuous.
(ii) Let $(X_i)_{i\in I}$ be an arbitrary non-empty family of topological spaces, and X the direct product of the *sets* X_i. Prove that the product topology on X is the weakest topology on X for which all projections $\pi_i : X \to X_i$ are continuous.
(iii) Given a non-empty set X, a family $(Y_i)_{i\in I}$ of topological spaces and a corresponding family of functions $f_i : X \to Y_i$, prove that there exists a weakest topology on X for which all functions f_i are continuous. It is called an *initial topology* on X.

Exercise 2.7 (Quotient Topology and Final Topology)

(i) Consider a function $f : X \to Y$ where X is a topological space. Prove that there exists a strongest topology on Y for which f is continuous.

Henceforth we consider this topology on Y. A map $f : X \to Y$ is called *open* if it sends open sets into open sets, and *closed* if it sends closed sets into closed sets.

(ii) f is open $\Longleftrightarrow f^{-1}(f(U)) \subset X$ is open for each open set $U \subset X$.
(iii) f is closed $\Longleftrightarrow f^{-1}(f(F)) \subset X$ is closed for each closed set $F \subset X$.

If there is an equivalence relation on X and $f(x)$ denotes the equivalence class of x for each $x \in X$, then it is called the *quotient topology* on $Y := f(X)$.

(iv) Given a non-empty set Y, a family $(X_i)_{i\in I}$ of topological spaces and a corresponding family of functions $f_i : X_i \to Y$, prove that there exists a finest topology on Y for which all functions f_i are continuous. It is called a *final topology* on Y.

Exercise 2.8 (Alexandroff's One-Point Compactification) If X is a non-compact topological space, then add the new symbol ∞ to X, and consider the larger set $\overline{X} := X \cup \{\infty\}$. Prove the following properties[16]:

(i) The open subsets of X and the sets $\overline{X}\setminus K$, where K runs over the closed compact subsets of X, form a topology $\mathcal{T}$ on $\overline{X}$.
(ii) This topology is compact.
(iii) X is a dense subspace of $\overline{X}$.

Exercise 2.9 Prove that an arbitrary (finite or infinite) product of connected spaces is connected.

Exercise 2.10 (Peano Curve) We recall from Exercise 1.8 (p. 32) that Cantor's ternary set C consists of those points $t \in [0, 1]$ which can be written in the form

$$t = 2\Big(\frac{t_1}{3} + \frac{t_2}{3^2} + \cdots + \frac{t_n}{3^n} + \cdots\Big)$$

with suitable integers $t_n \in \{0, 1\}$.

(i) Prove that the formula

$$t \mapsto \Big(\frac{t_1}{2} + \frac{t_3}{2^2} + \cdots + \frac{t_{2n-1}}{2^n} + \cdots , \frac{t_2}{2} + \frac{t_4}{2^2} + \cdots + \frac{t_{2n}}{2^n} + \cdots\Big)$$

defines a continuous function f of C onto $[0, 1] \times [0, 1]$.
(ii) Extend f to a continuous function of $[0, 1]$ onto $[0, 1] \times [0, 1]$.
(iii) For each positive integer N, construct a continuous function of $[0, 1]$ onto $[0, 1]^N$.

[16] This a generalization of the definition of the complex sphere in complex analysis.

Exercise 2.11 (Semi-continuity) A function $f : X \to \mathbb{R}$ on a topological space X is called *upper semi-continuous* if $\{x \in X : f(x) < \alpha\}$ is an open set for every $\alpha \in \mathbb{R}$, and *lower semi-continuous* if $\{x \in X : f(x) > \alpha\}$ is an open set for every $\alpha \in \mathbb{R}$. Prove the following results:

(i) f is upper semi-continuous $\Longleftrightarrow$ $-f$ is lower semi-continuous.
(ii) f is continuous $\Longleftrightarrow$ it is both upper and lower semi-continuous.
(iii) Let $\{f_i : i \in I\}$ be a non-empty family of upper semi-continuous functions on X. If $f := \inf_{i \in I} f_i$ is finite-valued everywhere, then f is upper semi-continuous.
(iv) If X is compact and f is upper semi-continuous, then f has a maximal value.

Chapter 3
Normed Spaces

Normed spaces are vector spaces endowed with a special metric that is compatible with its linear structure. They are very useful in Differential Calculus and in Applied Mathematics, among other subjects.

In this chapter we give an introduction to these spaces. For simplicity we consider only real vector spaces.[1]

3.1 Definitions and Examples

Definitions Let X be a real vector space.

- By a *norm* on X we mean a function $\|\cdot\| : X \to \mathbb{R}$ satisfying the following conditions for all *vectors* $x, y, z \in X$ and *scalars* (i.e., real numbers) λ:

  - $\|x\| \geq 0,$
  - $\|x\| = 0 \iff x = 0,$
  - $\|\lambda x\| = |\lambda| \cdot \|x\|,$
  - $\|x + y\| \leq \|x\| + \|y\|.$

  The last property is called the *triangle inequality*.
- By a *normed space* we mean a vector space X endowed with a norm on X.
- A set A is *bounded* in a normed space if there exists a constant M such that $\|x\| \leq M$ for all $x \in A$.
- A function $f : K \to X$, where X is a normed space, is *bounded* if its range is a bounded set in X.

[1] The complex case will be considered briefly in Sect. 3.6.

V. Komornik, *Topology, Calculus and Approximation*, Springer
Undergraduate Mathematics Series, DOI 10.1007/978-1-4471-7316-8_3

Examples

- $X = \mathbb{R}$, $\|x\| = |x|$. Henceforth we always consider this norm on $\mathbb{R}$.
- If K is a non-empty set and X a vector space, then the functions $f : K \to X$ form a vector space $\mathcal{F}(K, X)$ for the pointwise operations of addition and multiplication by scalars.
- If K is a non-empty set and $(X, \|\cdot\|)$ a normed space, then the *bounded* functions $f : K \to X$ form a linear subspace of $\mathcal{F}(K, X)$, denoted by $\mathcal{B}(K, X)$.[2] Furthermore, the formula

  $$\|f\|_\infty := \sup_{t \in K} \|f(t)\|_X$$

  defines a norm on $\mathcal{B}(K, X)$. Unless stated otherwise, henceforth $\mathcal{B}(K, X)$ will be endowed with this norm.

  For $X = \mathbb{R}$ we simply write $\mathcal{B}(K)$ instead of $\mathcal{B}(K, \mathbb{R})$.

We may define subspaces and products of normed spaces:

Definitions

- By a *(normed) subspace* of a normed space $(X, \|\cdot\|)$ we mean a linear subspace Y of X, endowed with the restriction of the norm $\|\cdot\|$ to Y.
- By the *product of the normed spaces* $(X_1, \|\cdot\|_1), \dots, (X_m, \|\cdot\|_m)$ we mean the vector space $X := X_1 \times \cdots \times X_m$ with the norm

  $$\|x\| := \|x_1\|_1 + \cdots + \|x_m\|_m, \quad x = (x_1, \dots, x_m) \in X.$$

The norm properties are easily verified in both cases.

Examples

- If K is a topological space, then the bounded continuous functions $f : K \to \mathbb{R}$ form a linear and hence normed subspace $C_b(K)$ of $\mathcal{B}(K)$ for the norm $\|\cdot\|_\infty$.
- The formula

  $$\|x\|_1 = |x_1| + \cdots + |x_m|, \quad x = (x_1, \dots, x_m) \in \mathbb{R}^m$$

  defines a norm on $\mathbb{R}^m$.

Now we introduce an important class of normed spaces.

[2] We will show in the next section that the definition of the function spaces $\mathcal{B}(K, X)$ and $C_b(K)$ on this page are consistent with the definition of these spaces in the preceding chapter.

Definitions

- By a *scalar product* on a vector space X we mean a function $(\cdot,\cdot) : X \times X \to \mathbb{R}$ satisfying for all vectors $x, y, z \in X$ and scalars α, β the following conditions:

  - $(x,x) \geq 0,$
  - $(x,x) = 0 \quad \Longleftrightarrow \quad x = 0,$
  - $(\alpha x + \beta y, z) = \alpha(x,z) + \beta(y,z),$
  - $(x,y) = (y,x).$

- By a *Euclidean* or *prehilbert space*[3] we mean a vector space endowed with a scalar product.

Examples

- $X = \mathbb{R}^m$, $(x,y) := x_1y_1 + x_2y_2 + \cdots + x_my_m$.
- I is a compact interval, $X = C(I)$ and $(f,g) := \int_I fg\ dt$.

Every Euclidean space has a natural norm:

Proposition 3.1 *If* $(\cdot,\cdot) : X \times X \to \mathbb{R}$ *is a scalar product, then the formula* $\|x\| := (x,x)^{1/2}$ *defines a norm on* X.

This norm satisfies the Cauchy–Schwarz inequality[4]

$$|(x,y)| \leq \|x\| \cdot \|y\|$$

and the parallelogram identity

$$\|x+y\|^2 + \|x-y\|^2 = 2\,\|x\|^2 + 2\,\|y\|^2$$

(see Fig. 3.1) for all $x, y \in X$.[5]

Proof The parallelogram identity follows by a direct computation:

$$\begin{aligned}
\|x+y\|^2 + \|x-y\|^2 &= (x+y, x+y) + (x-y, x-y) \\
&= (x,x) + (x,y) + (y,x) + (y,y) \\
&\qquad + (x,x) - (x,y) - (y,x) + (y,y) \\
&= 2(x,x) + 2(y,y) \\
&= 2\,\|x\|^2 + 2\,\|y\|^2 .
\end{aligned}$$

[3] In our terminology a Euclidean space is not necessarily finite-dimensional.

[4] The proof below also shows that equality holds if and only if x and y are proportional.

[5] Conversely, every norm satisfying the parallelogram identity may be defined by a suitable, unique scalar product. See Exercise 3.11, p. 91.

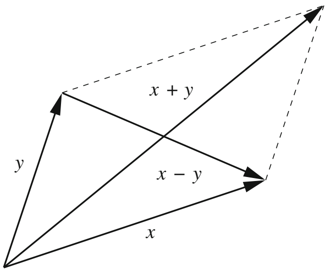


Fig. 3.1 The parallelogram identity

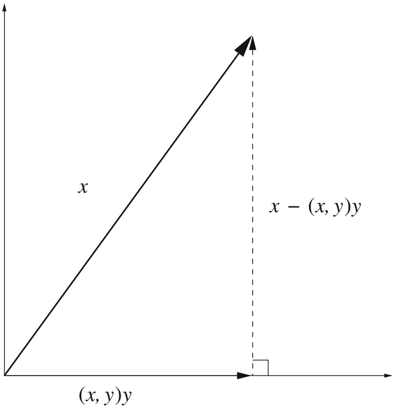


Fig. 3.2 The Cauchy–Schwarz inequality

The norm properties are also easily checked, except for the triangle inequality.

For $\|y\| = 1$ the Cauchy–Schwarz inequality is obtained by another direct computation[6]:

$$\begin{aligned} 0 &\le \|x - (x, y)y\|^2 = \big(x - (x, y)y, x - (x, y)y\big) \\ &= (x, x) - (x, y)(x, y) - (x, y)(y, x) + (x, y)(x, y)(y, y) \\ &= \|x\|^2 - |(x, y)|^2 \\ &= \|x\|^2 \, \|y\|^2 - |(x, y)|^2 . \end{aligned}$$

The general case follows by homogeneity. The case $y = 0$ is obvious. For $y \ne 0$ we write $y = ty'$ with $t = \|y\|$; then $\|y'\| = 1$, and hence

$$|(x, y)| = t \left|(x, y')\right| \le t \, \|x\| \, \left\|y'\right\| = \|x\| \cdot \|y\| .$$

[6] The result is essentially a consequence of Pythagoras' theorem: see Fig. 3.2.

Now the triangle inequality follows easily:

$$\begin{aligned}\|x+y\|^2 &= (x+y,x+y) = \|x\|^2+\|y\|^2+(x,y)+(y,x)\\ &\le \|x\|^2+\|y\|^2+2\,\|x\|\cdot\|y\| = \big(\,\|x\|+\|y\|\,\big)^2.\end{aligned}$$

□

Henceforth every Euclidean space will be automatically endowed with this norm.

Until now we have introduced three different norms on $\mathbb{R}^m$:

$$\begin{aligned}\|x\|_1 &= |x_1|+\cdots+|x_m|\,,\\ \|x\|_2 &= \big(\,|x_1|^2+\cdots+|x_m|^2\,\big)^{1/2},\\ \|x\|_\infty &= \max\{|x_1|\,,\ldots,|x_m|\}\,.\end{aligned}$$

The second norm of $\mathbb{R}^m$ comes from the natural scalar product defined above, while the third one is a special case of the norm of $\mathcal{B}(K,\mathbb{R})$ when $K=\{1,\ldots,m\}$. Figure 3.3 shows the "unit balls" of $\mathbb{R}^2$ for these norms.

Generalizing the first two examples we will prove that the formula

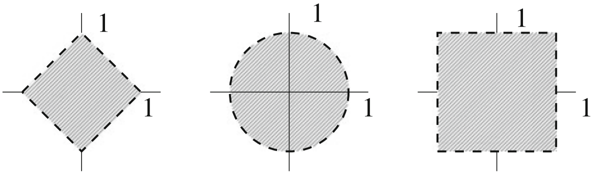
$$\|x\|_p := \big(\,|x_1|^p+\cdots+|x_m|^p\,\big)^{1/p}$$

defines a norm on $\mathbb{R}^m$ for each $1\le p<\infty$.

Definition The numbers $p,q\in[1,\infty]$ are called *conjugate exponents* if $p^{-1}+q^{-1}=1$.

Remark We have $p\in(1,\infty)$ if and only if $q\in(1,\infty)$. In this case the relation $p^{-1}+q^{-1}=1$ is equivalent to each of the following:

$$q=\frac{p}{p-1},\quad p=\frac{q}{q-1},\quad (p-1)(q-1)=1.$$



$p=1$ $p=2$ $p=\infty$

Fig. 3.3 "Unit balls"

Proposition 3.2 *Let p and q be conjugate exponents.*

(a) *(Young's inequality) If $x, y \geq 0$ and p, q are finite, then*

$$xy \leq \frac{x^p}{p} + \frac{y^q}{q}.$$

(b) *(Hölder's inequality) If $x, y \in \mathbb{R}^m$, then*

$$\left| \sum_{i=1}^{m} x_i y_i \right| \leq \|x\|_p \cdot \|y\|_q \, .$$

(c) *(Minkowski's inequality) If $x, y \in \mathbb{R}^m$, then*

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

(d) $\|\cdot\|_p$ *is a norm on* $\mathbb{R}^m$.

Proof

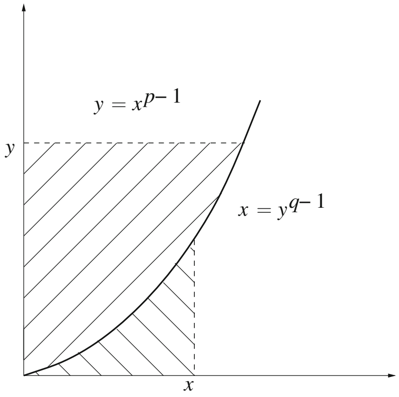
(a) We assume by symmetry that $p \geq 2$, and we consider the graph of the function $y = x^{p-1}$ or $x = y^{q-1}$; see Fig. 3.4. The shaded region contains the rectangle of sides x, y, hence its area is at least xy. Therefore

$$xy \leq \int_0^x s^{p-1} \, ds + \int_0^y t^{q-1} \, dt = \frac{x^p}{p} + \frac{y^q}{q}.$$

(b) The cases $x = 0$ and $y = 0$ being trivial, we assume that they are non-zero. First we assume that $p, q \in (1, \infty)$. If $\|x\|_p = \|y\|_q = 1$, then using Young's inequality we obtain the following estimate:

$$\left| \sum_{i=1}^{m} x_i y_i \right| \leq \sum_{i=1}^{m} |x_i| \cdot |y_i| \leq \sum_{i=1}^{m} \frac{x_i^p}{p} + \frac{y_i^q}{q} = p^{-1} + q^{-1} = 1.$$

Fig. 3.4 Young's inequality



In the general case we write $x = sx'$ and $y = ty'$ with $s := \|x\|_p$, $t := \|y\|_q$. Then $\|x'\|_p = \|y'\|_q = 1$, and hence

$$\Big|\sum_{i=1}^{m} x_i y_i\Big| = st\Big|\sum_{i=1}^{m} x_i' y_i'\Big| \le st\,\|x'\|_p \cdot \|y'\|_q = \|x\|_p \cdot \|y\|_q\,.$$

The case $p = 1$ follows from the definition of the norms:

$$\Big|\sum_{i=1}^{m} x_i y_i\Big| \le \sum_{i=1}^{m} |x_i| \cdot |y_i| \le \sum_{i=1}^{m} |x_i| \cdot \max_j |y_j| = \|x\|_1 \cdot \|y\|_\infty\,,$$

and then the case $p = \infty$ by exchanging the role of x and y.

(c) The cases $p = 1, \infty$ are already known, so we assume that $p, q \in (1, \infty)$. The case $x + y = 0$ being obvious, we also assume that $x + y \neq 0$. Furthermore, we may also assume by homogeneity that $\|x + y\|_p = 1$.[7] Then the required relation follows by applying Hölder's inequality and the equality $(p-1)q = p$:

$$\begin{aligned}
\|x+y\|_p &= \|x+y\|_p^p = \sum_{i=1}^{m} |x_i + y_i|^p \\
&\le \sum_{i=1}^{m} |x_i| \cdot |x_i + y_i|^{p-1} + \sum_{i=1}^{m} |y_i| \cdot |x_i + y_i|^{p-1} \\
&\le \|x\|_p \Big(\sum_{i=1}^{m} |x_i + y_i|^{(p-1)q}\Big)^{1/q} + \|y\|_p \Big(\sum_{i=1}^{m} |x_i + y_i|^{(p-1)q}\Big)^{1/q} \\
&= \big(\|x\|_p + \|y\|_p\big)\, \|x+y\|_p^{p/q} \\
&= \|x\|_p + \|y\|_p\,.
\end{aligned}$$

(d) We have just proved the triangle inequality. The other properties are obvious.

□

Similarly to the above, if I is a compact interval, then the formula

$$\|f\|_p := \Big(\int_I |f(t)|^p\ dt\Big)^{1/p}$$

defines a norm on $C(I)$ for each $1 \le p < \infty$[8]:

[7] We multiply both x and y by the same positive constant.

[8] The norm $\|\cdot\|_\infty$ on $C(I)$ has already been defined earlier.

Proposition 3.3 *(Riesz) Let p and q be two conjugate exponents.*

(a) *If $f, g \in C(I)$, then*

$$\left| \int_I fg \; dt \right| \le \|f\|_p \cdot \|g\|_q \, .$$

(b) *If $f, g \in C(I)$, then*

$$\|f + g\|_p \le \|f\|_p + \|g\|_p \, .$$

(c) $\|\cdot\|_p$ *is a norm on $C(I)$.*

Proof (a) The case $p = 1$ is obvious:

$$\left| \int_I fg \; dt \right| \le \int_I |f| \cdot |g| \; dt \le \int_I |f| \cdot \|g\|_\infty \; dt = \|f\|_1 \, \|g\|_\infty \, ,$$

and the case $p = \infty$ follows by symmetry.

For $p, q \in (1, \infty)$ it suffices to consider, as before, the case $\|f\|_p = \|g\|_q = 1$. Then the desired relation follows by applying Young's inequality:

$$\left| \int_I fg \; dt \right| \le \int_I |f| \cdot |g| \; dt \le \int_I \frac{|f|^p}{p} + \frac{|g|^q}{q} \; dt = p^{-1} + q^{-1} = 1 = \|f\|_p \cdot \|g\|_q \, .$$

(b) The case $p = \infty$ is already known, while the case $p = 1$ is straightforward:

$$\|f + g\|_1 = \int_I |f + g| \; dt \le \int_I |f| + |g| \; dt = \|f\|_1 + \|g\|_1 \, .$$

For $p \in (1, \infty)$ it suffices to consider, as in the preceding proposition, the case $\|f + g\|_p = 1$. Applying Hölder's inequality we get

$$\begin{aligned} \|f + g\|_p &= \|f + g\|_p^p = \int_I |f + g|^p \; dt \\ &\le \int_I |f| \cdot |f + g|^{p-1} \; dt + \int_I |g| \cdot |f + g|^{p-1} \; dt \\ &\le \|f\|_p \cdot \left\| \, |f + g|^{p-1} \right\|_q + \|g\|_p \cdot \left\| \, |f + g|^{p-1} \right\|_q \\ &= \|f\|_p + \|g\|_p \end{aligned}$$

because

$$\left\| \, |f + g|^{p-1} \right\|_q = \|f + g\|_{(p-1)q}^{p-1} = \|f + g\|_p^{p-1} = 1.$$

(c) We have just established the triangle inequality. The other properties are straightforward. □

3.2 Metric and Topological Properties

If $\|\cdot\|$ is a norm on a vector space X, then

$$d(x,y) := \|f-g\|$$

is a metric on X. (Sometimes we write $\|\cdot\|_X$ instead of $\|\cdot\|$ to avoid confusion.) We will therefore consider every normed space as a metric, and hence also as a topological space.

Remark Boundedness in the metric and normed sense are equivalent. Indeed, if a set A is bounded for the norm: $\|x\| < r$ for all $x \in A$, then $A \subset B_r(0)$. Conversely, if $A \subset B_r(y)$ for some ball, then $\|x\| < \|y\| + r$ for all $x \in A$.

In the rest of this chapter the letters X, Y, Z denote normed spaces.

Proposition 3.4

(a) *A sequence (x_n) converges to x in X if and only if $\|x_n - x\| \to 0$.*

(b) *A function $f : D \to Y$ ($D \subset X$) is continuous at a point $a \in D$ if and only if for each $\varepsilon > 0$ there exists a $\delta > 0$ such that*

$$x \in D \quad \textit{and} \quad \|x-a\|_X < \delta \Longrightarrow \|f(x)-f(a)\|_Y < \varepsilon.$$

(c) *A function $f : D \to Y$ ($D \subset X$) is uniformly continuous if and only if for each $\varepsilon > 0$ there exists a $\delta > 0$ such that*

$$x_1, x_2 \in D \quad \textit{and} \quad \|x_1 - x_2\|_X < \delta \Longrightarrow \|f(x_1)-f(x_2)\|_Y < \varepsilon.$$

(d) *The norm $\|\cdot\| : X \to \mathbb{R}$ is Lipschitz (and hence uniformly) continuous; more precisely,*

$$\big|\, \|x\| - \|y\| \,\big| \le \|x-y\|$$

for all $x, y \in X$.

(e) *Let $(x_n), (y_n) \subset X$ and $(\lambda_n) \subset \mathbb{R}$ be three sequences satisfying $x_n \to x$, $y_n \to y$ and $\lambda_n \to \lambda$. Then*

$$x_n + y_n \to x + y \quad \textit{and} \quad \lambda_n x_n \to \lambda x.$$

(f) *If the functions $f, g : X \hookrightarrow Y$ are continuous at a, then $f + g : X \hookrightarrow Y$ is continuous at a.*[9]

[9] The notation $\hookrightarrow$ was introduced on p. 2.

(g) *If the functions $h : X \hookrightarrow \mathbb{R}$ and $g : X \hookrightarrow Y$ are continuous at a, then $hg : X \hookrightarrow Y$ is continuous at a.*

(h) *The scalar product $(\cdot,\cdot) : X \times X \to \mathbb{R}$ is continuous: if $x_n \to x$ and $y_n \to y$, then $(x_n, y_n) \to (x, y)$.*

Proof (a), (b) and (c) are just reformulations of the definitions given for metric spaces.

(d) follows from the triangle inequality of the norm.

(e) and (h) are consequences of the following three estimates:

$$\begin{aligned}
\|(x_n + y_n) - (x + y)\| &\le \|x_n - x\| + \|y_n - y\| ;\\
\|\lambda_n x_n - \lambda x\| &= \|(\lambda_n - \lambda)(x_n - x) + (\lambda_n - \lambda)x + \lambda(x_n - x)\|\\
&\le |\lambda_n - \lambda| \cdot \|x_n - x\| + |\lambda_n - \lambda| \cdot \|x\| + |\lambda| \cdot \|x_n - x\| ;\\
|(x_n, y_n) - (x, y)| &= |(x_n - x, y_n - y) + (x_n - x, y) + (x, y_n - y)|\\
&\le \|x_n - x\| \cdot \|y_n - y\| + \|x_n - x\| \cdot \|y\| + \|x\| \cdot \|y_n - y\| ,
\end{aligned}$$

because the expressions on the right-hand sides tend to zero by assumption.

(f) and (g). If $x_n \to a$, then $f(x_n) \to f(a)$, $g(x_n) \to g(a)$ and $h(x_n) \to h(a)$. Using (e) we deduce that

$$(f + g)(x_n) = f(x_n) + g(x_n) \to f(a) + g(a) = (f + g)(a)$$

and

$$(hg)(x_n) = h(x_n)g(x_n) \to h(a)g(a) = (hg)(a).$$

□

Examples

- Let us consider in $X = \mathbb{R}^m$ a sequence (x_n) and a point a, and write $x_n = (x_{n1}, \dots, x_{nm})$, $a = (a_1, \dots, a_m)$. For each norm $\|\cdot\|_p$, $1 \le p \le \infty$, the norm convergence $x_n \to a$ is equivalent to the *component-wise convergence*: $x_{ni} \to a_i$ in $\mathbb{R}$ for $i = 1, \dots, m$. In particular, the convergence is the same for each p.
- Let us endow $Y = \mathbb{R}^m$ with the norm $\|\cdot\|_p$ for some $1 \le p \le \infty$. Consider a function $f : X \hookrightarrow \mathbb{R}^m$, and write $f = (f_1, \dots, f_m)$. Then f is continuous at a point a if and only if each component function $f_j : X \hookrightarrow \mathbb{R}$ is continuous at a.
- The metrics associated with the norms of $\mathcal{B}(K)$ and $\mathcal{B}(K, X)$ coincide with the metrics introduced in Chap. 1 (p. 4). In particular, the norm convergence is the uniform convergence on K.
- We may generalize the normed spaces $C_b(K)$ introduced on p. 66. If K is a topological space and $(X, \|\cdot\|)$ a normed space, then the bounded continuous functions $f : K \to X$ form a linear and hence normed subspace $C_b(K, X)$ of

$\mathcal{B}(K, X)$ for the norm $\|\cdot\|_\infty$.[10] If K is compact, then we may write $C(K, X)$ instead of $C_b(K, X)$ by Weierstrass' theorem (p. 25) and we may write

$$\|f\|_\infty = \max_{t \in K} \|f(t)\|$$

instead of using sup for every $f \in C(K, X)$.

Definition A complete normed space is called a *Banach space*. A complete Euclidean space is called a *Hilbert space*.

Examples

- We will prove in Theorem 3.10 below (p. 79) that every *finite-dimensional* normed space is complete.
- If X is a Banach space, then the spaces $\mathcal{B}(K, X)$ are also Banach spaces by Proposition 1.8, p. 14. In particular, the spaces $\mathcal{B}(K)$ are Banach spaces.
- If K is a topological space and X a Banach space, then $C_b(K, X)$ is also a Banach space, because it is a *closed* subspace of the Banach space $\mathcal{B}(K, X)$ by the example on p. 46. In particular, the spaces $C_b(K)$ are Banach spaces.
- If I is a compact interval, then the norms $\|\cdot\|_p$ for $p < \infty$ are *not* complete on the *vector space* $C(I)$: see Exercise 3.7, p. 90.

Remark A version of Proposition 1.16 (p. 21) states that every normed space may be considered as a dense subspace of some Banach space, and every Euclidean space may be considered as a dense subspace of some Hilbert space.[11]

There is a transparent characterization of connected *open* sets in normed spaces.

Definitions By a *segment* in a vector space X we mean a set of the form

$$[a, b] := \{(1-t)a + tb \ : \ 0 \le t \le 1\}$$

where $a, b \in X$.[12] The points a, b are called the *endpoints* of the segment.

A set A in a vector space X is *convex* if $[a, b] \subset A$ whenever $a, b \in A$.

By a *broken line* in a vector space X we mean a set of the form

$$L = [x_0, x_1] \cup [x_1, x_2] \cup \cdots \cup [x_{n-1}, x_n] \tag{3.1}$$

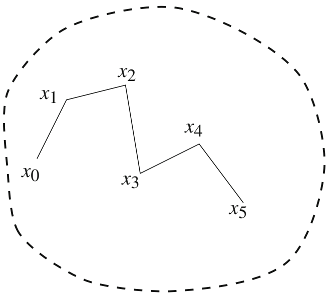
where $x_0, \dots, x_n \in X$. (See Fig. 3.5.) We say that L *connects* the points x_0 and x_n.

[10] The associated metric is the same as that introduced on p. 46.

[11] The proof of Proposition 1.16 allows us to define a suitable linear structure on the completion. See also [285, Corollary 2.21 (b)] for a direct proof using dual normed spaces.

[12] We do not exclude the case $a = b$.

Fig. 3.5 A broken line



Lemma 3.5 *The balls of normed spaces are convex.*

Proof We have to show that if $a, b \in B_r(y)$ and $x \in [a, b]$, then $\|x - y\| < r$. Since $x = (1-t)a + tb$ for some $0 \le t \le 1$, using the norm axioms we have[13]

$$\begin{aligned}\|x-y\| &= \|(1-t)(a-y) + t(b-y)\| \le (1-t)\|a-y\| + t\|b-y\| \\ &< (1-t)r + tr = r.\end{aligned}$$

□

Proposition 3.6 *For an* open *set U in a normed space the following three properties are equivalent:*

(a) *any two points of U may be connected by a broken line lying in U;*
(b) *any two points of U may be connected by a path lying in U;*
(c) *U is connected.*

Proof

$(a) \Rightarrow (b)$ It is sufficient to observe that every broken line $L \subset U$ is the range of a suitable path f in U. For example, if L is given by (3.1), then the formula

$$f(t) = (t-k+1)x_k + (k-t)x_{k-1}, \quad t \in [k-1, k], \quad k = 1, \ldots, n$$

defines such a path $f : [0, n] \to U$.

$(b) \Rightarrow (c)$ This is a special case of Proposition 2.17 (p. 49).

$(c) \Rightarrow (a)$ There is nothing to prove if U is empty. Otherwise fix a point $x \in U$, and consider the set A of points that may be connected to x by a broken line lying in U. We have to show that $A = U$. Since U is connected and A is non-empty (it contains x), it suffices to prove that A is both open and closed.

First we prove that A is open. If $y \in A$, then the open set U contains a small ball $B_r(y)$. This implies that $B_r(y) \subset A$. Indeed, if $L \subset U$ is a broken line connecting x and y, and $z \in B_r(y)$, then the broken line $L' := L \cup [y, z]$ connects x and z.

[13] The last inequality is strict because $(1-t)\|a-y\| < (1-t)r$ if $t < 1$, and $t\|b-y\| < tr$ if $t > 0$. As a matter of fact, it is sufficient to prove these relations for $0 < t < 1$.

Furthermore, this broken line still lies in U because $[y, z] \subset B_r(y) \subset U$ by the preceding lemma.

Next we prove that A is closed in U, i.e., $U \setminus A$ is open. If $z \in U \setminus A$, then U again contains a small ball $B_r(z)$. It suffices to show that $B_r(z) \subset U \setminus A$. Assume on the contrary that there exists a point $y \in B_r(z) \cap A$, and consider a broken line $L \subset U$ connecting x and y. Then $L' := L \cup [y, z]$ is a broken line connecting x and z in U, contradicting the choice of z. □

We say that two norms on a vector space X are *equivalent* if they define the same topology on X.

Proposition 3.7 *Two norms $\|\cdot\|$, $\|\cdot\|'$ on a vector space X are equivalent if and only if there exist two positive constants c_1, c_2 such that*

$$c_1 \|x\|' \le \|x\| \le c_2 \|x\|' \tag{3.2}$$

for all $x \in X$.

Furthermore, equivalent norms have the same Cauchy sequences, and therefore they are complete or non-complete at the same time.

Proof If the condition (3.2) is satisfied, then the same sequences converge to the same limits for both norms:

$$\|x_n - a\|' \to 0 \Longleftrightarrow \|x_n - a\| \to 0.$$

Hence the same sets are closed for both induced topologies, and therefore the open sets are also the same.

If the condition (3.2) is satisfied, then the same sequences have the Cauchy property for both induced metrics.

If there is no constant c_2 satisfying the corresponding inequality in (3.2), then there exists a sequence (x_n) satisfying $\|x_n\| = 1$ for all n, and $\|x_n\|' \to 0$. Then $x_n \to 0$ for the second norm, but not for the first, so that the induced topologies are different.

If there is no constant c_1 satisfying the corresponding inequality in (3.2), then we may repeat the preceding proof by exchanging the norms $\|\cdot\|$ and $\|\cdot\|'$. □

Remarks

- It is interesting to compare the proposition with our earlier remarks on the equivalence of *metrics* on p. 38 concerning the inequality (2.1).
- It follows from the proposition that changing a norm to an equivalent one the open, closed, compact, separable, dense, bounded or totally bounded sets, and the convergent and Cauchy sequences remain the same. Furthermore, the continuity, uniform continuity or Lipschitz continuity of a function $f : X \to Y$ is preserved if we change the norm of X and/or Y to equivalent ones.

3.3 Finite-Dimensional Normed Spaces

We recall that a set of *real numbers* is compact if and only if it is bounded and closed. We will extend this useful characterization to *all* finite-dimensional normed spaces.

We start with a lemma concerning the norm $\|\cdot\|_\infty$ on $\mathbb{R}^m$.

Lemma 3.8 *The normed space* $(\mathbb{R}^m, \|\cdot\|_\infty)$ *is complete, separable, and every bounded set is in fact* totally *bounded.*

Proof The space $(\mathbb{R}^m, \|\cdot\|_\infty)$ is complete because it is isomorphic to the Banach space $\mathcal{B}(\{1, \dots, m\})$.

It is separable because $\mathbb{Q}^m$ is a countable dense subset.

Finally, let K be a bounded set, and fix a number M such that $\|x\| < M$ for all $x \in K$. Given $r > 0$ arbitrarily, choose a large integer N such that $Nr > M$, and consider the balls $B_r(a_k)$ with

$$a_k = (k_1 r, \dots, k_m r), \quad k_1, \dots, k_m \in \{-N, \dots, N\}.$$

If $x = (x_1, \dots, x_m) \in K$, then one of these balls satisfies the inequalities $|x_j - k_j r| \le r/2 < r$ for all $j = 1, \dots, m$, i.e., $x \in B_r(a_k)$. □

A fundamental result of this section is the following

Theorem 3.9 (Tychonoff) *On a* finite-dimensional *vector space X all norms are equivalent.*

Proof of Theorem 3.9 Since every finite, say m-dimensional, vector space is isomorphic to $\mathbb{R}^m$ as a vector space, we may assume that $X = \mathbb{R}^m$.

Fix an arbitrary norm $\|\cdot\|$ on $\mathbb{R}^m$. It suffices to prove its equivalence with $\|\cdot\|' := \|\cdot\|_\infty$.

Consider the function

$$f : K \to \mathbb{R}, \quad f(x) = \|x\|$$

defined on the *unit sphere*

$$K := \{x \in \mathbb{R}^m \ : \ \|x\|_\infty = 1\}$$

of $(\mathbb{R}^m, \|\cdot\|_\infty)$. The set K is closed (by the continuity of the norm) and bounded, hence compact by Lemma 3.8 and Corollary 1.29 (p. 29).

The function f is continuous, even Lipschitz continuous. Indeed, introducing the canonical basis

$$e_1 = (1, 0, \dots, 0), \dots, e_m = (0, \dots, 0, 1)$$

of $\mathbb{R}^m$ and setting

$$M = \max\{\|e_1\|, \dots, \|e_m\|\},$$

the obvious inequality

$$\|x\| = \|x_1 e_1 + \cdots + x_m e_m\| \le M(|x_1| + \cdots + |x_m|) \le Mm\,\|x\|_\infty$$

yields the estimate

$$|f(x) - f(y)| = \Big| \|x\| - \|y\| \Big| \le \|x - y\| \le Mm\,\|x - y\|_\infty$$

for all $x, y \in \mathbb{R}^m$.

Applying Theorem 1.24, p. 25 (or Theorem 2.22, p. 52), there exist two points $a, b \in K$ satisfying

$$\|a\| \le \|x'\| \le \|b\|$$

for all $x' \in K$.

Every $x \in \mathbb{R}^m$ may be written in the form $x = tx'$ with $t = \|x\|_\infty$ and $x' \in K$. Multiplying the above inequalities by t we obtain that

$$\|a\| \cdot \|x\|_\infty \le \|x\| \le \|b\| \cdot \|x\|_\infty$$

for all $x \in \mathbb{R}^m$. We conclude by observing that $a, b \in K$, so that these vectors are non-zero, and therefore $c_1 := \|a\| > 0$ and $c_2 := \|b\| > 0$. □

The above theorem has important consequences:

Theorem 3.10 *Let $(X, \|\cdot\|)$ be a finite-dimensional normed space. Then*

(a) *X is complete;*
(b) *X is separable;*
(c) *a set in X is totally bounded if and only if it is bounded;*
(d) *a set in X is compact if and only if it is bounded and closed;*
(e) *every bounded sequence in X has a convergent subsequence.*

Proof We may assume again that $X = \mathbb{R}^m$ as a vector space. Furthermore, in view of the preceding theorem it suffices to consider the space $(\mathbb{R}^m, \|\cdot\|_\infty)$.

Properties (a), (b) and (c) have been proved in Lemma 3.8.

Property (d) follows from (a) and (c) by applying Corollary 1.29 (p. 29).

Property (e) follows by observing that a bounded sequence belongs to a sufficiently large closed ball, and that this ball is compact by (d) as a metric space. □

Remark We emphasize again that *all finite-dimensional normed spaces are Banach spaces, and all finite-dimensional Euclidean spaces are Hilbert spaces.*

The following consequence of Theorem 3.10 is of great importance in *Approximation theory*.

Proposition 3.11 *If Y is a finite-dimensional (linear) subspace of a normed space $(X, \|\cdot\|)$, then for each $x \in X$ there exists in Y a closest point a to x:*

$$\|x - a\| \le \|x - y\| \quad \textit{for all} \quad y \in Y. \tag{3.3}$$

Proof Choose a *minimizing* sequence $(y_n) \subset Y$:

$$\|x - y_n\| \to \inf_{y \in Y} \|x - y\| =: M.$$

Then (y_n) is a bounded sequence in a finite-dimensional space Y, hence it has a convergent subsequence: $y_{n_k} \to a \in Y$. Then

$$\|x - y_{n_k}\| \to M, \quad \text{and} \quad \|x - y_{n_k}\| \to \|x - a\|$$

by the continuity of the norm. Hence $\|x - a\| = M$ by the uniqueness of the limit. □

The closest point a is not always unique:

Examples (See Fig. 3.6.)

- Let

  $$Y = \{y \in \mathbb{R}^2 \ : \ y_1 + y_2 = 0\} \quad \text{and} \quad x = (1, 1)$$

  in $(\mathbb{R}^2, \|\cdot\|_1)$. Then for $a = (t, -t) \in Y$ we have

  $$\|x - a\| = |1 - t| + |1 + t| = 2$$

  if $|t| \le 1$, and $\|x - a\| > 2$ otherwise, so that a satisfies (3.3) for all $-1 \le t \le 1$.
- Considering the same Y and x in $(\mathbb{R}^2, \|\cdot\|_2)$, now only $a = (0, 0)$ satisfies (3.3) because

  $$\|x - a\|_2^2 = (1 - t)^2 + (1 + t)^2 = 2 + 2t^2$$

  has a unique minimum for $t = 0$.

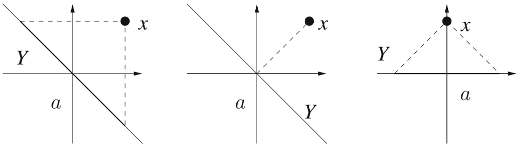


Fig. 3.6 Closest points

- If

$$Y = \{y \in \mathbb{R}^2 \ : \ y_2 = 0\} \quad \text{and} \quad x = (0, 1)$$

in $(\mathbb{R}^2, \|\cdot\|_\infty)$, then for $a = (t, 0) \in Y$ we have

$$\|x - a\|_\infty = \max\{|t|, 1\} = 1$$

if $|t| \le 1$, and $\|x - a\| > 1$ otherwise, so that a satisfies (3.3) for all $-1 \le t \le 1$.

Let us generalize the second example:

Proposition 3.12 (Legendre–Gauss) *If Y is a finite-dimensional (linear) subspace of a Euclidean space $(X, (\cdot, \cdot))$, then for each $x \in X$ there exists in Y a unique closest point a to x:*

$$\|x - a\| \le \|x - y\| \quad \textit{for all} \quad y \in Y. \tag{3.4}$$

This is the orthogonal projection *of x onto Y in the following sense*[14]*:*

$$a \in Y, \quad \textit{and} \quad (x - a, z) = 0 \quad \textit{for all} \quad z \in Y. \tag{3.5}$$

Finally, the map $x \mapsto a$ is linear, and $\|a\| \le \|x\|$ for all $x \in X$.

Remark The *method of least squares* in *Approximation theory* is based on the above proposition.

Proof By the preceding proposition there exists at least one point $a \in Y$ satisfying (3.4). This point satisfies (3.5). Indeed, for each $t > 0$ we deduce from (3.4) that

$$0 \ge t^{-1}\big(\|x - a\|^2 - \|x - (a + tz)\|^2\big) = 2(x - a, z) - t\|z\|^2.$$

Letting $t \to 0$ we get $(x - a, z) \le 0$. Applying the result to $-z$ instead of z we get the converse inequality, so that in fact $(x - a, z) = 0$.

In order to complete the proof of the uniqueness and the characterization it remains to show that if a point $a \in Y$ satisfies (3.5), then $\|x - y\| > \|x - a\|$ for each $y \in Y \setminus \{a\}$. Since $z = a - y \in Y \setminus \{0\}$, this may be shown as follows:

$$\begin{aligned}\|x - y\|^2 &= \|(x - a) + (a - y)\|^2 \\ &= \|x - a\|^2 + \|a - y\|^2 + 2(x - a, a - y) \\ &= \|x - a\|^2 + \|a - y\|^2 > \|x - a\|^2.\end{aligned}$$

[14] The proof will show that a is the only point satisfying (3.5).

If a and b are the orthogonal projections of x and y onto Y, then for any scalars α, β we deduce from (3.5) by linearity that

$$((\alpha x + \beta y) - (\alpha a + \beta b), z) = 0$$

for all $z \in Y$. Hence $\alpha a + \beta b$ is the orthogonal projection of $\alpha x + \beta y$ onto Y. This proves the linearity of the orthogonal projection.

Finally, applying (3.5) with $z = a$ and using the Cauchy–Schwarz inequality we obtain

$$\|a\|^2 = (a, a) = (x, a) \le \|x\| \cdot \|a\| ;$$

hence $\|a\| \le \|x\|$. □

3.4 Continuous Linear Maps

Motivated by matrix theory, the values $A(x)$ of a linear map $A : X \to Y$ are usually denoted by Ax. As before, the letters X, Y, Z stand for normed spaces.

For a linear map continuity at merely one point implies uniform and even Lipschitz continuity:

Lemma 3.13 *If a linear map $A : X \to Y$ is continuous at some point a, then there exists a number $L \ge 0$ such that*

$$\|Ax\| \le L\,\|x\| \quad \textit{for all} \quad x \in X. \tag{3.6}$$

Hence A is Lipschitz continuous with the constant L.

Proof If A is continuous at a, then there exists a $\delta > 0$ such that

$$z \in X \quad \text{and} \quad \|z - a\| \le \delta \Longrightarrow \|A(z-a)\| = \|Az - Aa\| \le 1.$$

Applying this with $z := a + x$ we obtain (3.6) with $L = 1/\delta$ for all $x \in X$ satisfying $\|x\| \le \delta$, and then by homogeneity for all $x \in X$.

Furthermore, (3.6) implies Lipschitz continuity:

$$\|Ax - Ay\| = \|A(x-y)\| \le L\,\|x - y\|$$

for all $x, y \in X$. □

We denote by $L(X, Y)$ the set of *continuous* linear maps $A : X \to Y$. If $A \in L(X, Y)$, then

$$\|A\| := \sup_{\|x\| \le 1} \|Ax\| < \infty$$

by the lemma.

Remark Excluding the trivial case $X = \{0\}$, $\|A\|$ is the smallest constant L satisfying (3.6), and the equalities

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} \quad \text{and} \quad \|A\| = \sup_{\|x\|=1} \|Ax\|$$

also hold.

The notation $\|A\|$ is justified:

Proposition 3.14 *$(L(X,Y), \|\cdot\|)$ is a normed space.*

Proof As a linear subspace of the vector space $\mathcal{F}(X,Y)$ of all functions $f : X \to Y$, $L(X,Y)$ is a vector space.

Among the norm axioms only the triangle inequality is not obvious. Using the triangle inequality in Y we have for each $x \in X$ the estimate

$$\|(A+B)x\| \le \|Ax\| + \|Bx\| \le \|A\| \cdot \|x\| + \|B\| \cdot \|x\| = (\|A\| + \|B\|) \cdot \|x\| ,$$

and hence $\|A+B\| \le \|A\| + \|B\|$ by the definition of $\|A+B\|$. □

Example We recall that every matrix (a_{ik}) of size $n \times m$ defines a linear map $A : \mathbb{R}^m \to \mathbb{R}^n$ by matrix multiplication if we represent the elements of $\mathbb{R}^m$ and $\mathbb{R}^n$ as column vectors. If we endow $\mathbb{R}^m$ with the $\|\cdot\|_1$ norm and $\mathbb{R}^n$ with the $\|\cdot\|_\infty$ norm, then

$$\|A\| = \max_{i,k} |a_{ik}|$$

by a simple computation.

The following result greatly simplifies the study of finite-dimensional linear maps.

Theorem 3.15 *If X is finite-dimensional, then every linear map $A : X \to Y$ is continuous.*

Proof By Theorem 3.9 (p. 78) we may assume that $X = (\mathbb{R}^m, \|\cdot\|_1)$. Introducing the canonical basis $e_1, \ldots, e_m$ of $\mathbb{R}^m$ again, setting

$$L := \max\{\|Ae_1\|, \ldots, \|Ae_m\|\}$$

we have

$$\begin{aligned} \|Ax\| &= \|x_1 Ae_1 + \cdots + x_m Ae_m\| \\ &\le |x_1| \cdot \|Ae_1\| + \cdots + |x_m| \cdot \|Ae_m\| \\ &\le L \|x\|_1 \end{aligned}$$

for all $x \in X$. □

Proposition 3.16

(a) *If $B \in L(X, Y)$ and $A \in L(Y, Z)$, then*

$$AB \in L(X, Z) \quad \textit{and} \quad \|AB\| \le \|A\| \cdot \|B\| .$$

(b) *The formula $\varphi(A, B) := AB$ defines a continuous bilinear map*

$$\varphi : L(Y, Z) \times L(X, Y) \to L(X, Z).$$

Proof

(a) The linearity of AB is obvious. It remains to observe that

$$\|ABx\| \le \|A\| \cdot \|Bx\| \le \|A\| \cdot \|B\| \cdot \|x\|$$

for all $x \in X$.

(b) Only the continuity is not obvious. Consider the following estimate:

$$\begin{aligned}
&\left\|\varphi(A', B') - \varphi(A, B)\right\| \\
&\quad = \left\|A'B' - AB\right\| \\
&\quad = \left\|(A' - A)(B' - B) + A(B' - B) + (A' - A)B\right\| \\
&\quad \le \left\|A' - A\right\| \cdot \left\|B' - B\right\| + \|A\| \cdot \left\|B' - B\right\| + \left\|A' - A\right\| \cdot \|B\| .
\end{aligned}$$

If $A' \to A$ and $B' \to B$, then each term on the right-hand side tends to zero. □

***Remark** If $X = Y$, then $L(X, X)$ is not only a normed space, but also a *normed algebra*. This means that the product of the operators is also defined, and it is compatible with the normed space structure.[15] More precisely, if $A, B, C \in L(X, X)$ and $\alpha \in \mathbb{R}$, then

- $\alpha(AB) = (\alpha A)B = A(\alpha B)$,
- $(A + B)C = AC + BC$,
- $A(B + C) = AB + AC$,
- $\|AB\| \le \|A\| \cdot \|B\|$.

The product is not commutative in general: $AB \neq BA$. See, e.g., Neumark [365] or Rudin [431] for an introduction to this theory.

Different normed spaces may have the same structure.

[15] The product is by definition the composition of the operators.

Definition X and Y are *isometrically isomorphic* if there exists a linear bijection $g : X \to Y$ such that

$$\|g(x)\| = \|x\|$$

for all $x \in X$.

It is often useful to identify isometrically isomorphic spaces. For example, every linear map $A : \mathbb{R} \to \mathbb{R}$ has the form $Ax = ax$ for a suitable real number a. This enables us to identify A with a. More generally, we have the following

Proposition 3.17 *The normed space $L(\mathbb{R}, X)$ is isometrically isomorphic with X via the function*

$$g : L(\mathbb{R}, X) \to X, \quad g(\varphi) := \varphi(1).$$

Proof The map g is clearly linear, and is invertible with inverse

$$f : X \to L(\mathbb{R}, X), \quad f(a)t := ta, \quad t \in \mathbb{R},$$

because $g(f(a)) = a$ for all $a \in X$, and $f(g(\varphi)) = \varphi$ for all $\varphi \in L(\mathbb{R}, X)$.

Finally, $\|\varphi\| = |\varphi(1)| = |g(\varphi)|$, so that g is an isometry. □

We end this section by proving a simple extension theorem that is widely applied in classical analysis.[16]

Proposition 3.18 *Let $f : M \to Y$ be a continuous linear map defined on a (linear) subspace M of a normed space X. If Y is a Banach space, then f may be uniquely extended to a continuous linear map $F : \overline{M} \to Y$ defined on the closure of M. Furthermore, $\|F\| = \|f\|$.*

Proof Changing X to $\overline{M}$ we may assume that M is dense in X. Then one may readily verify that the unique continuous extension $F : X \to Y$ of f, given by Proposition 1.14 (p. 20), is linear and satisfies the equality $\|F\| = \|f\|$. □

3.5 Continuous Linear Functionals

Definitions

- A real-valued linear map is also called a *linear functional*.
- The Banach space $L(X, \mathbb{R})$ of *continuous* linear functionals is called the *dual space* of the normed space X, and is often denoted by X'.

[16] See, e.g., the proof of Theorem 3.20 below in the separable case, Exercise 3.6 and the proof of Proposition 6.1 (a), pp. 88, 90, 142.

Normed spaces have "many" continuous linear functionals:

Proposition 3.19 *Let $(X, \|\cdot\|)$ be a normed space.*

(a) *For each $c \in X$ there exists a $\varphi \in X'$ such that $\|\varphi\| = \|c\|$ and $\varphi(c) = \|c\|^2$.*
(b) *For any two distinct $x_1, x_2 \in X$ there exists a $\varphi \in X'$ such that $\varphi(x_1) \neq \varphi(x_2)$:* the continuous linear functionals separate the points of X.

Proof for Euclidean spaces $(X, (\cdot,\cdot))$

(a) The formula $\varphi(x) := (c, x)$ defines a linear functional satisfying the equality $\varphi(c) = \|c\|^2$. This implies that $\|\varphi\| \geq \|c\|$. The continuity of φ and the converse inequality $\|\varphi\| \leq \|c\|$ follow from the Cauchy–Schwarz inequality:

$$|\varphi(x)| = |(c, x)| \leq \|c\| \cdot \|x\|$$

for all $x \in X$.
(b) Apply (a) with $c := x_1 - x_2$. □

Part (a) of the proposition for arbitrary normed spaces is deeper. Let us assume temporarily the following very important result:

***Theorem 3.20 (Helly–Hahn–Banach[17])** *If $f : M \to \mathbb{R}$ is a continuous linear functional on a (linear) subspace $M \subset X$, then f may be extended, preserving the norm, to a continuous linear functional $\varphi : X \to \mathbb{R}$.*

Remark Let us compare this theorem with Proposition 3.18. The theorem allows us to extend the linear map f to the whole space X, but only if $Y = \mathbb{R}$, and without ensuring the uniqueness of the extension.[18]

**Proof of Proposition 3.19 (a) in the general case* The formula

$$f(tc) := t\,\|c\|^2\,, \quad t \in \mathbb{R}$$

defines a linear functional of norm $\|c\|$ on the linear subspace M generated by the vector c. Since $f(c) = \|c\|^2$, the preceding theorem yields a suitable extension φ. □

The crucial step in the proof of Theorem 3.20 is the following

Lemma 3.21 (Helly) *Under the assumptions of Theorem 3.20, if $a \in X \setminus M$, then f may be extended, preserving its norm, to a continuous linear functional ψ defined on the linear subspace Y generated by M and a.*

[17] See p. 341 on Helly's essential contribution to this theorem.

[18] The extension is unique if X is a Euclidean space. Taylor [480] and Foguel [171] have characterized the normed spaces having the unique extension property; see, e.g., [109] or [285, Proposition 2.12].

Proof If $f = 0$, then we may choose $\varphi := 0$. Otherwise multiplying f by some constant we may assume that $\|f\| = 1$.

Fix a real number α (to be chosen later), and extend f linearly to a functional $\psi : Y \to \mathbb{R}$ by the formula[19]

$$\psi(x + ta) := f(x) + t\alpha, \quad x \in M, \quad t \in \mathbb{R}.$$

Since ψ is an extension of f, we have obviously $\|\psi\| \geq 1$. It remains to show that α may be chosen so as to satisfy the converse inequality $\|\psi\| \leq 1$.

Since $\psi(-y) = -\psi(y)$, it suffices to find α satisfying

$$\psi(x \pm ta) \leq \|x \pm ta\|$$

for all $x \in M$ and $t \geq 0$. Since ψ is an extension of f, the inequality is satisfied for $t = 0$. In the remaining cases dividing by $t > 0$ we obtain the equivalent condition

$$\psi(x' \pm a) \leq \left\|x' \pm a\right\| \quad \text{for all} \quad x' \in M;$$

this may be rewritten in the form

$$f(x') - \alpha \leq \left\|x' - a\right\| \quad \text{and} \quad f(x') + \alpha \leq \left\|x' + a\right\|,$$

or equivalently

$$f(x') - \left\|x' - a\right\| \leq \alpha \leq \left\|x' + a\right\| - f(x')$$

for all $x' \in M$.

Hence a suitable α exists if and only if

$$f(x') - \left\|x' - a\right\| \leq \left\|x'' + a\right\| - f(x'')$$

for all $x', x'' \in M$. This inequality may be proved as follows:

$$\begin{aligned} f(x') + f(x'') &= f(x' + x'') \leq \left\|x' + x''\right\| \\ &= \left\|(x' - a) + (x'' + a)\right\| \leq \left\|x' - a\right\| + \left\|x'' + a\right\|. \end{aligned}$$

□

**Proof of Theorem 3.20* If M has finite co-dimension m,[20] then the theorem follows by m successive applications of Helly's lemma.

[19] The uniqueness of the decomposition $x + ta$ implies that the definition is correct, and that ψ is linear.

[20] For example, if X is finite-dimensional.

In the general case we consider the family of all linear, norm-preserving extensions of f. If we identify the linear functionals with their graphs, then this family satisfies the hypotheses of Zorn's lemma (p. 54), so that f has at least one *maximal* linear, norm-preserving extension φ. By Helly's lemma φ is defined on the whole space X. □

Remark The application of Zorn's lemma may be avoided if X is separable: first we extend f to a dense subspace of X by repeating the first step countably many times, and then we apply Proposition 3.18 (p. 85). This was exactly the procedure used by Helly for $X = C(I)$ where I is a compact interval.

3.6 Complex Normed Spaces

Most results of this chapter may be easily adapted to the complex case. Let us briefly indicate the necessary modifications. We recall that every complex vector space may also be considered as a real vector space, by allowing only multiplication by real numbers. For example, $\mathbb{C}^m$ is isomorphic to $\mathbb{R}^{2m}$ as a real vector space.

Let X and Y be complex vector spaces. We say that the map $A : X \to Y$ is *linear* if

$$A(x+y) = A(x) + A(y) \quad \text{and} \quad A(\lambda x) = \lambda A(x)$$

for all $x, y \in X$ and $\lambda \in \mathbb{C}$, and *antilinear* if

$$A(x+y) = A(x) + A(y) \quad \text{and} \quad A(\lambda x) = \overline{\lambda} A(x)$$

for all $x, y \in X$ and $\lambda \in \mathbb{C}$.

By a *norm* defined on a complex vector space X we mean a real-valued function $\|\cdot\|$ satisfying the same axioms as in the real case for all $x, y, z \in X$ and for all $\lambda \in \mathbb{C}$ (instead of $\lambda \in \mathbb{R}$). A *normed space* is a vector space endowed with a norm. A norm induces a metric in the usual way, and the norm function is continuous for the corresponding topology.

A complex-valued function $(\cdot,\cdot) : X \times X \to \mathbb{C}$ defined on a complex vector space X is called a *scalar product* if it satisfies for all $x, y, z \in X$ and $\alpha, \beta \in \mathbb{C}$ (instead of $\alpha, \beta \in \mathbb{R}$) the same axioms as in the real case, with one modification: we change the symmetry relation $(x, y) = (y, x)$ to $(x, y) = \overline{(y, x)}$.

A *Euclidean space* is a vector space endowed with a scalar product. Since (x, x) is still real for all $x \in X$, the scalar product induces a norm in the usual way, which satisfies the Cauchy–Schwarz inequality and the parallelogram identity. The scalar product is continuous with respect to the norm topology.

A complete Euclidean space is called a *Hilbert space*. For example, $\mathbb{C}^m$ is a Hilbert space for the scalar product

$$(x, y) := x_1\overline{y_1} + x_2\overline{y_2} + \cdots + x_m\overline{y_m}.$$

Since a complex normed space induces the same metric as the associated real normed space (of double dimension), all metric and topological results of Sects. 3.1–3.4 remain valid with slight changes in some proofs. In condition (3.5) we have to replace the relation $(x-a,z)=0$ by $\Re(x-a,z)=0$,[21] and in the proof we have to write

$$(x-a,z)+(z,x-a)=2\Re(x-a,z)$$

instead of $2(x-a,z)$, and

$$(x-a,a-y)+(a-y,x-a)=2\Re(x-a,a-y)$$

instead of $2(x-a,a-y)$.

The only delicate problem is the extension of the Helly–Hahn–Banach theorem 3.20 to the complex case. In its statement we change $\mathbb{R}$ to $\mathbb{C}$. For the proof we first extend the real part of f by using the real case theorem, and then we *complexify* the extended functional $\psi : X \to \mathbb{R}$ by setting

$$\varphi(x) := \psi(x) - i\psi(ix).$$

Then $\Re\varphi = \psi$, $\|f\| = \|\psi\| \le \|\varphi\|$, and φ is a complex linear functional because

$$\varphi(ix) := \psi(ix) - i\psi(-x) = i\psi(x) + \psi(ix) = i\varphi(x).$$

It remains to show the converse inequality $\|\varphi\| \le \|\psi\|$. For each $x \in X$ there exists a $\lambda \in \mathbb{C}$ such that $|\lambda| = 1$ and $\lambda\varphi(x) = |\varphi(x)|$. Then

$$|\varphi(x)| = \varphi(\lambda x) = \psi(\lambda x) \le \|\psi\| \cdot |\lambda x| = \|\psi\| \cdot |x|,$$

i.e., $\|\varphi\| \le \|\psi\|$ indeed.

3.7 Exercises

Exercise 3.1 Prove that $\|x\|_p \to \|x\|_\infty$ for each $x \in \mathbb{R}^m$ as $p \to \infty$.

Exercise 3.2 Let A, B be two sets in a normed space and set

$$A + B := \{x + y \ : \ x \in A, \quad y \in B\}.$$

Prove the following:

(i) If A is open, then $A + B$ is open.
(ii) If A is compact and B is closed, then $A + B$ is closed.
(iii) If A and B are compact, then $A + B$ is compact.
(iv) If A and B are closed, then $A + B$ is not necessarily closed.

[21] The symbol $\Re$ stands for the real part.

Exercise 3.3 Consider the vector space of matrices of order n endowed with an arbitrary norm. Prove the following:

(i) The orthogonal matrices form a compact set.
(ii) The invertible matrices form an open set.
(iii) The symmetric matrices form a closed set.

Are these sets connected?

Exercise 3.4 Let X, Y be normed spaces, $K \subset X$ a convex compact set, and $f : K \to Y$ a *locally Lipschitz continuous* function: for each point $a \in K$ there exist a neighborhood V_a of a in K and a constant L_a such that

$$\|f(x_1) - f(x_2)\| \le L_a \|x_1 - x_2\|$$

for all $x_1, x_2 \in V_a$.

Prove that f is (globally) Lipschitz continuous.

Exercise 3.5 (Dini's Theorem) Let K be compact topological space and $(f_n) \subset C(K)$ a monotone sequence of continuous functions, converging *pointwise* to some $f \in C(K)$. Then the convergence is in fact uniform.

Exercise 3.6 (Integral of Continuous Functions) Given a Banach space X, we consider the Banach space $C([a,b],X)$ of the continuous functions $f : [a,b] \to X$ with the supremum norm.

A function $f \in C([a,b],X)$ is called *piecewise linear* if there exists a finite subdivision

$$a = x_0 < \cdots < x_n = b$$

such that f is affine in each subinterval $[x_{k-1}, x_k]$. Its integral is defined by

$$I(f) := \sum_{k=1}^{n} \frac{f(x_{k-1}) + f(x_k)}{2} (x_k - x_{k-1}).$$

Prove the following assertions:

(i) the piecewise linear functions form a linear subspace M;
(ii) M is dense in $C([a,b],X)$;
(iii) $I : M \to \mathbb{R}$ is a continuous linear functional;
(iv) I extends to a unique continuous linear functional $\tilde{I}$ on $C([a,b],X)$.

We usually write $\int_a^b f(x)\,dx$ instead of $\tilde{I}(f)$.

Exercise 3.7 Prove that $C([a,b])$ is not complete for any norm $\|\cdot\|_p$ with $1 \le p < \infty$.

Exercise 3.8 (Carathéodory's Theorem) If $x \in \mathbb{R}^n$ belongs to the convex hull[22] of a set $A \subset \mathbb{R}^n$, then there exists a subset $B \subset A$ of at most $n+1$ elements such that x already belongs to the convex hull of B.

Exercise 3.9 (Radon's Theorem) Any set of at least $n+2$ points in $\mathbb{R}^n$ can be partitioned into two disjoint sets whose convex hulls intersect.

Exercise 3.10 (Helly's Intersection Theorem)

(i) Let $C_1, \dots, C_k$ be a finite collection of convex subsets of $\mathbb{R}^n$, with $k > n$. If the intersection of every $n+1$ of these sets is non-empty, then the whole collection has a non-empty intersection.
(ii) Let $(C_i)_{i \in I}$ be an infinite collection of convex compact subsets of $\mathbb{R}^n$. If the intersection of every $n+1$ of these sets is non-empty, then the whole collection has a non-empty intersection.

Exercise 3.11 (Jordan–von Neumann Theorem) Consider a norm on a vector space X that satisfies the parallelogram identity. We are going to prove that it is associated with a uniquely defined scalar product.

(i) Show that the only possible scalar product is given by the formula

$$(x, y) = \frac{1}{4}\left(\|x+y\|^2 - \|x-y\|^2\right).$$

The next assertions, where $x, y, z \in X$ are arbitrary vectors, show that this formula does indeed define a scalar product, and that the induced norm is the given one:

(ii) $(x, z) + (y, z) = 2\left(\frac{x+y}{2}, z\right)$;
(iii) $(x, z) = 2\left(\frac{x}{2}, z\right)$;
(iv) $(x, z) + (y, z) = (x+y, z)$;
(v) $(nx, y) = n(x, y)$ for all $n \in \mathbb{Z}$;
(vi) $(\alpha x, y) = \alpha(x, y)$ for all $\alpha \in \mathbb{Q}$;
(vii) the maps $\alpha \mapsto \|\alpha x + y\|$ and $\alpha \mapsto \|\alpha x - y\|$ are continuous;
(viii) $(\alpha x, y) = \alpha(x, y)$ for all $\alpha \in \mathbb{R}$;
(ix) (x, y) is a scalar product associated with our norm.

[22] We recall from linear algebra that the *convex hull* of a set A in a vector space is the set of all convex combinations of the vectors in A. Equivalently, it is the intersection of all convex sets containing A as a subset.

Exercise 3.12 (A Fixed Point Theorem of Joó) Let X, Y be non-empty convex sets in normed spaces, and $y \mapsto K(y)$ a mapping of Y into the family of non-empty convex compact subsets of X. Assume that

- $K(y) \subset K(y_1) \cup K(y_2)$ whenever $y \in [y_1, y_2]$;
- $x \in K(y)$ whenever $x_n \to x$, $y_n \to y$, and $x_n \in K(y_n)$ for all n.

The goal of this exercise is to show that the sets $K(y)$ have a non-empty intersection.

First we consider the intersection of two sets. We assume on the contrary that there exist $y_1, y_2 \in Y$ such that $K(y_1) \cap K(y_2) = \varnothing$.

(i) Prove that for each $y \in [y_1, y_2]$ we have either $K(y) \subset K(y_1)$ or $K(y) \subset K(y_2)$.

(ii) Prove that there exists $y_3 \in [y_1, y_2]$ such that $K(y) \subset K(y_1)$ for all $y \in [y_1, y_3)$ and $K(y) \subset K(y_2)$ for all $y \in (y_3, y_2]$.

(iii) Assuming by symmetry that $K(y_3) \subset K(y_1)$, show that $K(y_4) \subset K(y_5)$ for any $y_5 \in (y_3, y_2]$ and $y_4 \in (y_3, y_5)$,[23] and hence

$$\cap_{y \in (y_3, y_2]} K(y) \neq \varnothing.$$

(iv) Show that

$$\cap_{y \in (y_3, y_2]} K(y) \subset K(y_3),$$

and hence $K(y_1) \cap K(y_2) \neq \varnothing$, contradicting our assumption.

Now we consider intersections of more than two sets.

(v) Assume by induction that any n sets $K(y)$ have a non-empty intersection, but there exist $n + 1$ sets $K(y_1), \dots, K(y_{n+1})$ with an empty intersection. Find a contradiction by applying the above results for the sets

$$K^*(y) := \cap_{i=3}^{n+1} K(y_i) \cap K(y), \quad y \in Y$$

instead of $K(y)$.

(vi) Conclude that

$$\cap_{y \in Y} K(y) \neq \varnothing.$$

Exercise 3.13 (von Neumann's Minimax Theorem) Let X, Y be non-empty convex compact sets in normed spaces, and $f : X \times Y \to \mathbb{R}$ a continuous function satisfying the following conditions:

- the subfunctions $x \mapsto f(x, y)$ are concave for each $y \in Y$;
- the subfunctions $y \mapsto f(x, y)$ are convex for each $x \in X$.

[23] Considering a linear order relation on $[y_1, y_2]$ we may express these conditions by the inequalities $y_1 < y_3 < y_4 < y_5 \leq y_2$.

The goal of this exercise is to prove the equality

$$\max_x \min_y f(x,y) = \min_y \max_x f(x,y). \tag{3.7}$$

(i) Show that both sides of (3.7) are well-defined, and

$$\max_x \min_y f(x,y) \le \min_y \max_x f(x,y).$$

(ii) Denoting by α the right-hand side of (3.7), show that the sets

$$K(y) := \{x \in X \ : \ f(x,y) \ge \alpha\}, \quad y \in Y$$

satisfy the conditions of the preceding exercise.

(iii) Show that the relation

$$\cap_{y \in Y} K(y) \ne \varnothing$$

implies the inequality

$$\max_x \min_y f(x,y) \ge \min_y \max_x f(x,y).$$

(iv) Show that if the maximum on the left-hand side of (3.7) is attained in x^* and the minimum on the right-hand side of (3.7) is attained in y^*, then (x^*, y^*) is a *saddle point*, i.e.,

$$f(x,y^*) \le f(x^*,y^*) \le f(x^*,y) \quad \text{for all} \quad x \in X \quad \text{and} \quad y \in Y. \tag{3.8}$$

(v) Conversely, show that the existence of a saddle point implies (3.7).

Part II
Differential Calculus

Kepler [274] and Fermat [167] noticed that (using today's terminology) the derivative of a function vanishes at points of minima and maxima. Fermat [167] and Descartes [125] solved geometrical problems using differential calculus.

Partial derivatives first appeared in the works of Newton and Leibniz around 1670–1680. Newton [368] solved problems of mechanics by applying power series to integrate differential equations like $x' = 1-3t+x+t^2+tx$. This stimulated many subsequent works by Leibniz, the Bernoulli brothers, Euler, Lagrange, Laplace and others.

Euler [151] and Lagrange [302, 308] extended the results of Kepler and Fermat to several variables and for conditional extrema.

Lagrange [310] and Cauchy [87] generalized Taylor's formula [481] for functions of several variables.

Using Euler's approximate solutions [156], Cauchy [89] established the existence of unique solutions for a large class of differential equations. His unpublished results were rediscovered by Lipschitz [336].

Peano [380] gave a new proof by using successive approximations. This method, going back at least to Liouville [335] or maybe even Cauchy [89], became very popular after the works of Picard [389], Bendixson [35] and Lindelöf [332].

Peano [379, 382] proved the *existence* of solutions under much weaker assumptions.

In Weierstrass' era of rigor, Dini [134] published the first proof of Descartes' implicit function theorem [125].

Peano [191] gave a new form of the remainder term in Taylor's formula, which is very useful when computing limits.

The *total derivative* was introduced by Weierstrass [508] and Stolz [470], and became widely accepted after the works of Fréchet [177]. Clarifying some earlier results of Euler [147, 155], Schwarz [441] and Young [515] proved that the higher-order derivatives are usually symmetric functions.

The following works contain detailed historical accounts: [60–62, 72, 81, 120, 137, 216, 217, 277] and [475]. Numerous exercises and further results may be found in the books: [15, 63, 105, 118, 124, 130, 170, 224, 278, 398, 430, 450, 451].

In this chapter the letters X, Y, Z denote arbitrary normed spaces. However, the reader may first consider that they are equal to $\mathbb{R}^N$, $N = 1, 2, \ldots$.

Chapter 4
The Derivative

4.1 Definitions and Elementary Properties

The classical definition (see Fig. 4.1)

$$f'(a) := \lim_{x\to a} \frac{f(x)-f(a)}{x-a}$$

of the derivative remains meaningful for *vector-valued* functions $f : \mathbb{R} \hookrightarrow Y$.[1] For functions $f : X \hookrightarrow \mathbb{R}$ of vector *variables*, however, the fraction is undefined if $\dim X > 1$. But there exist two equivalent definitions that may be adapted to these cases:

Let us observe that for $A \in \mathbb{R}$ the relation

$$\frac{f(x)-f(a)}{x-a} \to A \quad \text{as} \quad x \to a \tag{4.1}$$

is equivalent to

$$\frac{|f(a+h)-f(a)-Ah|}{|h|} \to 0 \quad \text{as} \quad h \to 0. \tag{4.2}$$

The relation (4.1) means that the function

$$x \mapsto \frac{f(x)-f(a)}{x-a}$$

[1] The notation $\hookrightarrow$ was introduced on p. 2.

V. Komornik, *Topology, Calculus and Approximation*, Springer
Undergraduate Mathematics Series, DOI 10.1007/978-1-4471-7316-8_4

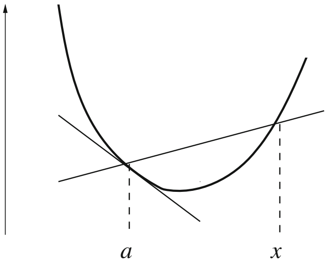


Fig. 4.1 The notion of the derivative

has a *continuous* extension to a. Equivalently, there exists a function $u : D(f) \to \mathbb{R}$ that is continuous at a, and satisfies the identity[2]

$$f(x) - f(a) \equiv u(x)(x-a). \tag{4.3}$$

By Proposition 3.17 (p. 85) the number A in (4.2) may be identified with the linear map $h \mapsto Ah$ in $L(\mathbb{R}, \mathbb{R})$, and the function u in (4.3) may be considered as a function $u : \mathbb{R} \hookrightarrow L(\mathbb{R}, \mathbb{R})$. Using this interpretation and changing the absolute values to norms in (4.2), both relations are meaningful for more general functions $f : X \hookrightarrow Y$ as well, and they are still equivalent:

Lemma 4.1 *Given a function $f : X \hookrightarrow Y$, a point $a \in D(f)$ and a number $r > 0$, the following properties are equivalent:*

(a) *there exists a continuous linear map $A \in L(X, Y)$ such that*

$$\frac{\|f(a+h) - f(a) - Ah\|}{\|h\|} \to 0 \quad \textit{as} \quad \|h\| \to 0; \tag{4.4}$$

(b) *there exists a function $u : D(f) \to L(X, Y)$ satisfying (4.3), and continuous at a.*

(c) *there exists a function $u : D(f) \cap B_r(a) \to L(X, Y)$ satisfying (4.3), and continuous at a.*

Furthermore, (a) implies the existence of u satisfying $u(a) = A$, and (b), (c) imply (a) with $A = u(a)$.

Proof

(b) $\Longrightarrow$ (c) is obvious.

(c) $\Longrightarrow$ (a) The relation

$$\frac{\|f(a+h) - f(a) - u(a)h\|}{\|h\|} = \frac{\|\big(u(a+h) - u(a)\big)h\|}{\|h\|} \le \|u(a+h) - u(a)\|$$

[2] The symbol $\equiv$ means in this book that equality holds for all points where both sides are defined.

implies (4.4) with $A := u(a)$ because for $h \to 0$ the right-hand side tends to zero by the continuity of u at a.
(a) $\Longrightarrow$ (b) For each $x \in D(f) \setminus \{a\}$ there exists a $\varphi_x \in X'$ such that[3]

$$\|\varphi_x\| = 1/\|x-a\| \quad \text{and} \quad \varphi_x(x-a) = 1.$$

Then the formula

$$u(x)h := Ah + \varphi_x(h)\big(f(x) - f(a) - A(x-a)\big), \quad h \in X$$

defines a continuous linear map $u(x) \in L(X,Y)$ satisfying (4.3).
It remains to show that, setting $u(a) := A$, the function $u : D(f) \to L(X,Y)$ is continuous at a. We have for each $h \in X$ the estimate

$$\begin{aligned}\big\|\big(u(x) - u(a)\big)h\big\| &= \big\|\varphi_x(h)\big(f(x) - f(a) - A(x-a)\big)\big\| \\ &\le \|\varphi_x\| \cdot \|h\| \cdot \|f(x) - f(a) - A(x-a)\|\,;\end{aligned}$$

and since $\|\varphi_x\| = 1/\|x-a\|$, we have

$$\|u(x) - u(a)\| \le \frac{\|f(x) - f(a) - A(x-a)\|}{\|x-a\|}.$$

If $x \to a$, then the fraction tends to zero by (a), so that $u(x) \to u(a)$. □

The relation (4.4) is often written in the more convenient form

$$f(a+h) = f(a) + Ah + o(h), \quad h \to 0.$$

This expresses transparently that the derivative yields a good *linear approximation* of f in a neighborhood of a.

Remark If $\dim X > 1$ the function u is not unique: a natural example will occur later (p. 115).

Definitions Let $f : X \hookrightarrow Y$.

- f is *differentiable* at $a \in D(f)$ if it is defined in a neighborhood of a, and satisfies one of the equivalent properties in Lemma 4.1. The map A is called the *(Fréchet or total) derivative* of f at a, and is denoted by $f'(a)$.
- The *derivative function* of f is the function $f' : D_1 \to L(X,Y)$ defined by the formula $a \mapsto f'(a)$ on the set D_1 of points where f is differentiable.
- f is *differentiable* if $D_1 = D(f)$, i.e., if it is defined on an open set, and is differentiable at every point of its domain of definition.

[3] We apply Proposition 3.19 on p. 86 with $c = x - a$, and we divide the resulting functional by $\|c\|^2$.

- f is *continuously differentiable* or *belongs to the class* C^1 if it is differentiable, and if its derivative $f' : D(f) \to L(X, Y)$ is continuous (everywhere). We write $f \in C^1$ in this case.

Remarks

- Neither the differentiability nor the derivative of f changes if we change the norm of X or Y to an equivalent one.
- We define the derivative only in interior points in order to simplify several important theorems in the sequel.
- Property (a) is usually easier to check, while property (b) allows us to simplify many proofs.

Examples

- Every constant function is differentiable, and its derivative is identically zero because the numerator in (4.4) vanishes.
- Every continuous linear map $f \in L(X, Y)$ is differentiable, and its derivative is the constant function $f'(a) := f$ for every $a \in X$ because the numerator in (4.4) vanishes again.
- Let $\varphi : X \times Y \to Z$ be a *bilinear* map satisfying for some constant M the estimate

$$\|\varphi(x, y)\| \le M \|x\| \cdot \|y\|$$

for all $x \in X$ and $y \in Y$.[4]

Then φ is differentiable, and $\varphi'(x, y) \in L(X \times Y, Z)$ is given by the formula

$$\varphi'(x, y)(h, k) = \varphi(x, k) + \varphi(h, y), \quad (h, k) \in X \times Y.$$

Using this formula we see that the derivative function is a continuous *linear* map

$$\varphi' \in L(X \times Y, L(X \times Y, Z)).$$

Indeed, the linearity of φ' follows from the bilinearity of φ. Furthermore, we infer from the estimate

$$\left\|\varphi'(x, y)(h, k)\right\| \le M \|x\| \cdot \|k\| + M \|h\| \cdot \|y\|$$

and from the definitions

$$\|(x, y)\| = \|x\| + \|y\| \quad \text{and} \quad \|(h, k)\| = \|h\| + \|k\|$$

[4] This is equivalent to the continuity of φ, see Lemma 5.1 below, p. 118.

that

$$\left\|\varphi'(x,y)(h,k)\right\| \le M \left\|(x,y)\right\| \cdot \left\|(h,k)\right\|$$

for all $(x,y), (h,k) \in X \times Y$. Hence

$$\left\|\varphi'(x,y)\right\| \le M \left\|(x,y)\right\| \quad \text{for all} \quad (x,y) \in X \times Y,$$

i.e., φ' is continuous with $\|\varphi'\| \le M$.

Proposition 4.2

(a) *If f is differentiable at a, then it is continuous at a.*

(b) *If $f : X \hookrightarrow Y$ is differentiable at a, then*

$$\lim_{t \to 0} \frac{f(a+th) - f(a)}{t} = f'(a)h \tag{4.5}$$

for all $h \in X$. Hence the derivative $f'(a)$ is unique.

(c) *Differentiation is a linear operation: if $f, g : X \hookrightarrow Y$ are differentiable at a and $\alpha, \beta \in \mathbb{R}$, then $\alpha f + \beta g : X \hookrightarrow Y$ is differentiable at a, and*

$$(\alpha f + \beta g)'(a) = \alpha f'(a) + \beta g'(a).$$

(d) *If $g : X \hookrightarrow Y$ is differentiable at a and $f : Y \hookrightarrow Z$ is differentiable at $g(a)$, then $f \circ g : X \hookrightarrow Z$ is differentiable at a, and*

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

(e) *The preceding implication remains valid if we replace the words "differentiable at a" by one of the following: "continuously differentiable at a", "differentiable" or "continuously differentiable".*

Proof

(a) It follows from (4.3) that

$$\|f(x) - f(a)\| \le \|u(x)\| \cdot \|x - a\|.$$

If $x_n \to a$, then $u(x_n) \to u(a)$, so that $\|u(x_n)\| \cdot \|x_n - a\| \to 0$, and therefore $f(x_n) \to f(a)$ by the preceding inequality.

(b) Apply (4.3) with $x = a + th$ and use the equality $u(a) = f'(a)$.

(c) As the intersection of two neighborhoods,

$$D := D(\alpha f + \beta g) = D(f) \cap D(g)$$

is a neighborhood of a. By our assumptions there exist two functions $u, v : D \to L(X,Y)$, continuous at a and satisfying the identities

$$f(x) - f(a) \equiv u(x)(x-a) \quad \text{and} \quad g(x) - g(a) \equiv v(x)(x-a).$$

Then the identity

$$(\alpha f + \beta g)(x) - (\alpha f + \beta g)(a) \equiv (\alpha u(x) + \beta v(x))(x - a)$$

is also satisfied. Since $\alpha u + \beta v : D \to L(X, Y)$ is continuous at a by Proposition 3.4 (p. 73), $\alpha f + \beta g$ is differentiable at a, and

$$(\alpha f + \beta g)'(a) = (\alpha u + \beta v)(a) = \alpha u(a) + \beta v(a) = \alpha f'(a) + \beta g'(a).$$

(d) Since g is continuous at a, $D := D(f \circ g) = D(g) \cap g^{-1}(D(f))$ is a neighborhood of a. By our assumptions there exists a function $v : D(g) \to L(X, Y)$, continuous at a and satisfying the identity

$$g(x) - g(a) \equiv v(x)(x - a),$$

and there exists a function $u : D(f) \to L(Y, Z)$, continuous at $g(a)$ and satisfying the identity

$$f(y) - f(g(a)) \equiv u(y)(y - g(a)).$$

Then we have

$$f(g(x)) - f(g(a)) = u(g(x))(g(x) - g(a)) = u(g(x))v(x)(x - a)$$

for all $x \in D$, i.e.,

$$(f \circ g)(x) - (f \circ g)(a) \equiv w(x)(x - a)$$

in D with

$$w(x) := u(g(x))v(x) \in L(X, Z).$$

Since the composite function[5] $w = \varphi \circ (u \circ g, v)$ is continuous at a by Proposition 1.6 (p. 11), we conclude that $f \circ g$ is differentiable at a, and

$$(f \circ g)'(a) = w(a) = u(g(a))v(a) = f'(g(a))g'(a).$$

(e) If g' is continuous at a and f' is continuous at $g(a)$, then $(f \circ g)'$ is continuous at a by the preceding formula. The other two versions are obtained by applying the obtained results to every $a \in D(g)$. □

[5] Here φ denotes the continuous bilinear map introduced in Proposition 3.16, p. 84.

Remarks

- The limit (4.5) is called the *derivative of f at a in the direction h*. Thus a (Fréchet) differentiable function is differentiable in every direction. The last example on p. 104 shows that the converse implication may fail.[6]
- The derivative of a composite function $f \circ g$ takes a simpler form when f is linear: since $f'(g(a)) = f$, we have

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

This remark will frequently be used in the sequel.

Examples

- If $f : X \hookrightarrow \mathbb{R}$ and $g : X \hookrightarrow Y$ are differentiable at a, then $fg : X \hookrightarrow Y$ is differentiable at a, and

$$(fg)'(a)h = (f'(a)h)g(a) + f(a)g'(a)h \tag{4.6}$$

for all $h \in X$.

For the proof we introduce the function $F : X \hookrightarrow \mathbb{R} \times X$ defined by $F(x) := (f(x), g(x))$, $x \in D(f) \cap D(g)$ and the bilinear function $\varphi : \mathbb{R} \times X \to X$ defined by $\varphi(c, x) := cx$. Then $fg = \varphi \circ F$.

If $\|h\| \to 0$, then

$$\begin{aligned} &\frac{\|F(a+h) - F(a) - (f'(a)h, g'(a)h)\|}{\|h\|} \\ &\qquad = \frac{\|f(a+h) - f(a) - f'(a)h\|}{\|h\|} \\ &\qquad\qquad + \frac{\|g(a+h) - g(a) - g'(a)h\|}{\|h\|} \to 0 \end{aligned}$$

by the differentiability of f and g at a. Hence F is differentiable at a and $F'(a)h = (f'(a)h, g'(a)h)$ for all $h \in X$.

Next we observe that

$$\|\varphi(c, x)\| = \|cx\| = |c| \cdot \|x\| \quad \text{for all} \quad (c, x) \in \mathbb{R} \times X,$$

so that φ is differentiable by the example on page 100, and

$$\varphi'(c, x)(t, k) = \varphi(c, k) + \varphi(t, x) = ck + tx$$

[6] We return to this question in Proposition 4.12, p. 113.

for all $(c, x), (t, k) \in \mathbb{R} \times X$.

Applying Proposition 4.2 (d) we conclude that $fg = \varphi \circ F$ is differentiable at a, and

$$\begin{aligned}(fg)'(a)h &= (\varphi \circ F)'(a)h = \varphi'(F(a))F'(a)h\\ &= \varphi'(f(a), g(a))(f'(a)h, g'(a)h) = f(a)g'(a)h + (f'(a)h)g(a)\end{aligned}$$

for all $h \in X$.

- We show that if $A \in L(H, H)$ is a continuous linear map on a Euclidean space, then the *quadratic form* $f : H \to \mathbb{R}$ defined by $f(x) := (Ax, x)$ is differentiable.

  First we compute the directional derivatives. For any fixed $a, h \in H$ we have

  $$\begin{aligned}\lim_{t\to 0} \frac{f(a+th) - f(a)}{t} &= \lim_{t\to 0} [(Aa, h) + (Ah, a) + t(Ah, h)]\\ &= (Aa, h) + (Ah, a).\end{aligned}$$

  Hence the derivative, if it exists, is given by the formula

  $$f'(a)h = (Aa, h) + (Ah, a), \quad a, h \in H.$$

  Now we check the relation (4.4): since

  $$\frac{|f(a+h) - f(a) - (Aa, h) - (Ah, a)|}{\|h\|} = \frac{|(Ah, h)|}{\|h\|} \le \|Ah\| \le \|A\| \cdot \|h\|,$$

  it suffices to observe that as $h \to 0$ the last expression tends to zero.
- The function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by the formulas

  $$f(0,0) = 0 \quad \text{and} \quad f(x, y) = \frac{2xy}{x^2 + y^2} \quad \text{otherwise}$$

  is differentiable in every direction at 0, but it is not totally differentiable because it is not even continuous at 0. Indeed, it takes any value between -1 and 1 in every neighborhood of $(0, 0)$.

As in the case of functions of a real variable, the derivative is very helpful when finding extremal values of functions. By symmetry we consider only the case of *minima*.

Definition A function $f : X \hookrightarrow \mathbb{R}$ has a *local minimum* at $a \in D(f)$ if there exists an $r > 0$ such that $f(x) \ge f(a)$ for all $x \in D(f) \cap B_r(a)$.

Proposition 4.3 (Fermat) *If f has a local minimum at a and is differentiable at a, then $f'(a) = 0$.*

Proof Fix $h \in X$ arbitrarily. By our assumptions we have

$$f(a) \le f(a+th) = f(a) + tf'(a)h + o(t)$$

as $t \to 0$ (see Fig. 4.2). Dividing by $t > 0$ we get

$$0 \le f'(a)h + o(1);$$

letting $t \searrow 0$ this yields $f'(a)h \ge 0$. Changing h to $-h$ we obtain the converse inequality $f'(a)h \le 0$, so that finally $f'(a)h = 0$ for all $h \in X$. □

4.2 Mean Value Theorems

First we recall some classical theorems:

Proposition 4.4 *Let $f, g, h : [a,b] \to \mathbb{R}$ be continuous functions on a non-degenerate interval, differentiable on (a,b).*

(a) *(Rolle) If $h(a) = h(b)$, then there exists a $c \in (a,b)$ such that $h'(c) = 0$.*
(b) *(Lagrange) There exists a $c \in (a,b)$ such that $f(b) - f(a) = f'(c)(b-a)$.*
(c) *(Cauchy) If $g' \neq 0$ in (a,b), then there exists a $c \in (a,b)$ such that*

$$\frac{f(b)-f(a)}{g(b)-g(a)} = \frac{f'(c)}{g'(c)}.$$

See Fig. 4.3 for the geometric meaning of Lagrange's theorem.

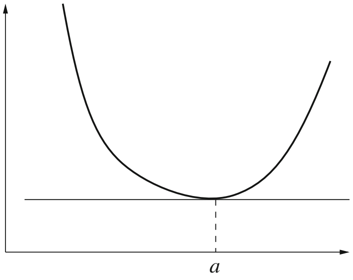


Fig. 4.2 Minimum

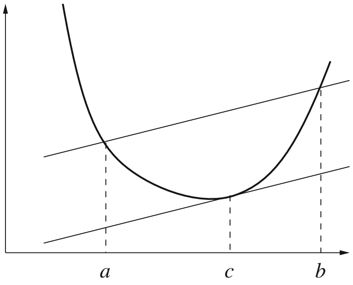


Fig. 4.3 The mean value theorem

Proof

(a) The function h has maximal and minimal values by Weierstrass' theorem. Since $h(a) = h(b)$, at least one of them is attained at some point $c \in (a, b)$, and then $h'(c) = 0$ by Fermat's theorem.

(b) We apply (a) to the function

$$h(t) := f(t) - \frac{f(b) - f(a)}{b - a}(t - a).$$

(c) Since $g' \neq 0$ in (a, b), applying (b) to g we see that the fraction on the left-hand side is well defined. The existence of c follows by applying Rolle's theorem to

$$h(t) := f(t) - \frac{f(b) - f(a)}{g(b) - g(a)}(g(t) - g(a)).$$

□

□

The last result allows us to establish a useful method of finding limits.

Corollary 4.5 (L'Hospital) *Let $f, g : (a, b) \to \mathbb{R}$ be two differentiable functions with $g' \neq 0$ in (a, b). If*

$$\lim_b f = \lim_b g = 0 \quad \textit{and} \quad \lim_b \frac{f'}{g'} = A,$$

then

$$\lim_b \frac{f}{g} = A.$$

Proof We have to prove that

$$\lim_b \frac{f(t_n)}{g(t_n)} = A$$

for every sequence $t_n \nearrow b$ in (a, b).

Fix an arbitrary closed convex neighborhood V of A.[7] If $n > m$, then

$$\frac{f(t_n) - f(t_m)}{g(t_n) - g(t_m)} = \frac{f'(c_{n,m})}{g'(c_{n,m})}$$

[7] A bounded closed interval if A is finite, and a closed halfline if A is infinite.

for some $c_{n,m} \in (t_m, t_n)$ by Cauchy's mean value theorem. If m is sufficiently large, then the right-hand side belongs to V by our assumption. Then letting $n \to \infty$ we obtain that $f(t_m)/g(t_m)$ also belongs to V. □

Lagrange's theorem remains valid for functions of a *vector variable*. In what follows a segment $[a, b]$ in a vector space is called *non-degenerate* if $a \neq b$, and we use the natural notation $(a, b) := [a, b] \setminus \{a, b\}$.

Proposition 4.6 (Lagrange) *If $f : X \hookrightarrow \mathbb{R}$ is continuous on a non-degenerate segment $[a, b] \subset X$ and differentiable on (a, b), then there exists a $c \in (a, b)$ satisfying*

$$f(b) - f(a) = f'(c)(b - a). \tag{4.7}$$

Proof Setting $\ell(t) := (1 - t)a + tb$, the composite function $g := f \circ \ell : [0, 1] \to \mathbb{R}$ satisfies the conditions of the preceding proposition. There exists therefore $0 < t < 1$ such that

$$(f \circ \ell)(1) - (f \circ \ell)(0) = (f \circ \ell)'(t).$$

Equivalently, we have

$$f(b) - f(a) = f'(\ell(t))(b - a),$$

i.e., (4.7) is satisfied with $c := \ell(t)$. □

Example The proposition may fail for *vector-valued* functions. For example, the function $f(x) := (\cos x, \ \sin x)$ satisfies

$$|f'(x)| = |(-\sin x, \cos x)| = 1$$

for every $x \in \mathbb{R}$. Since $f(2\pi) - f(0) = 0$, we have $f(2\pi) - f(0) \neq 2\pi f'(c)$ for all $c \in \mathbb{R}$.

For *vector-valued* functions a weaker, but still useful result holds:

Theorem 4.7 *If $f : X \hookrightarrow Y$ is continuous on a non-degenerate segment $[a, b] \subset X$ and differentiable on (a, b), then there exist $c_1, c_2 \in (a, b)$ such that*

$$\|f(b) - f(a)\| \le \left\| f'(c_1)(b - a) \right\| \tag{4.8}$$

and

$$\left\| f(b) - f(a) - f'(a)(b - a) \right\| \le \left\| (f'(c_2) - f'(a))(b - a) \right\| . \tag{4.9}$$

Proof Fix $\varphi \in Y'$ (to be chosen later) and apply the preceding proposition to $\varphi \circ f : X \hookrightarrow \mathbb{R}$ instead of f: there exists a $c_1 \in (a,b)$ such that

$$\varphi\,(f(b) - f(a)) = \varphi(f(b)) - \varphi(f(a)) = \varphi f'(c_1)(b-a). \tag{4.10}$$

This implies (4.8) if we choose φ satisfying

$$\varphi\,(f(b) - f(a)) = \|f(b) - f(a)\| \quad \text{and} \quad \|\varphi\| \le 1$$

(this is possible by Proposition 3.19, p. 86), because

$$\varphi f'(c_1)(b-a) \le \|\varphi\| \cdot \left\|f'(c_1)(b-a)\right\| \le \left\|f'(c_1)(b-a)\right\|.$$

For the proof of (4.9) we apply the preceding proposition to $\varphi \circ g : X \hookrightarrow \mathbb{R}$ with $g(x) := f(x) - f'(a)(x-a)$ and with $\varphi \in Y'$ satisfying $\|\varphi\| \le 1$ and

$$\varphi\,\big(f(b) - f(a) - f'(a)(b-a)\big) = \left\|f(b) - f(a) - f'(a)(b-a)\right\|.$$

There exists a $c_2 \in (a,b)$ such that

$$\varphi(g(b)) - \varphi(g(a)) = \varphi g'(c_2)(b-a),$$

i.e.,

$$\varphi\,\big(f(b) - f(a) - f'(a)(b-a)\big) = \varphi(f'(c_2) - f'(a))(b-a),$$

and hence

$$\begin{aligned}\left\|f(b) - f(a) - f'(a)(b-a)\right\| &= \varphi(f'(c_2) - f'(a))(b-a)\\ &\le \left\|(f'(c_2) - f'(a))(b-a)\right\|.\end{aligned}$$

□

The above theorem has important consequences.

Corollary 4.8 *Let $f : D \to Y$ be a differentiable function. If D is connected and f' vanishes on D, then f is constant.*

Proof For any given $a, b \in D$ there exists a broken line

$$L = \cup_{i=1}^{n} [x_{i-1}, x_i] \subset D$$

connecting $x_0 = a$ and $x_n = b$ by Proposition 3.6 (p. 76). Since $f'(x) \equiv 0$, applying (4.8) on each segment $[x_{i-1}, x_i]$ we obtain the equalities

$$f(x_0) = f(x_1) = \cdots = f(x_n),$$

whence $f(a) = f(b)$. □

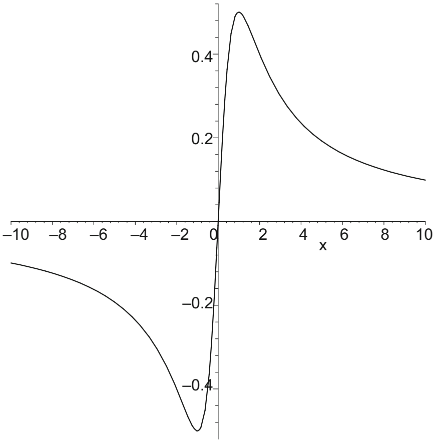


Fig. 4.4 Graph of f

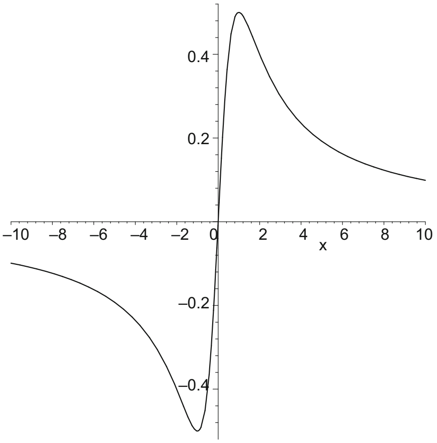


Fig. 4.5 Graph of f'

Now we consider a sequence of differentiable functions $f_n : U \to Y$. If $f_n \to f$ and $f_n' \to g$ pointwise on U, then we cannot conclude in general that f is differentiable and $f' = g$. We recall a counterexample, and then we generalize a classical sufficient condition to normed spaces.

Example Let $f(x) = x/(1 + x^2)$, $x \in \mathbb{R}$; see the graphs of f and f' in Figs. 4.4 and 4.5. Then the sequence of functions $f_n(x) := n^{-1}f(nx)$ tends to zero uniformly in $\mathbb{R}$, and

$$f_n'(x) = f'(nx) \to \begin{cases} 0 & \text{if } x \neq 0, \\ 1 & \text{if } x = 0. \end{cases}$$

Proposition 4.9 *If the convergence $f_n' \to g$ is* uniform, *then f is differentiable and $f' = g$.*

Proof Since the functions f_n are differentiable, U is an open set. For any given $a \in U$ we fix a ball $B_r(a) \subset U$. By Theorem 4.7 the following inequalities hold for all $x \in B_r(a)$ and $m, n = 1, 2, \dots$:

$$\|(f_m - f_n)(x) - (f_m - f_n)(a)\| \le \sup_{y \in B_r(a)} \left\|(f_m' - f_n')(y)\right\| \cdot \|x - a\| .$$

Letting $m \to \infty$ this yields

$$\|f(x) - f(a) - (f_n(x) - f_n(a))\| \le \sup_{y \in B_r(a)} \left\|(g - f_n')(y)\right\| \cdot \|x - a\| .$$

For any given $\varepsilon > 0$ we fix n such that

$$\sup_{y \in B_r(a)} \left\|(g - f_n')(y)\right\| \le \frac{\varepsilon}{3};$$

then

$$\|f(x) - f(a) - (f_n(x) - f_n(a))\| \le \frac{\varepsilon}{3} \|x - a\|$$

for all $x \in B_r(a)$.

Since f_n is differentiable at a, there exists $0 < r' \le r$ such that

$$\left\|f_n(x) - f_n(a) - f_n'(a)(x - a)\right\| \le \frac{\varepsilon}{3} \|x - a\|$$

for all $x \in B_{r'}(a)$. Since

$$\left\|g(a) - f_n'(a)\right\| \le \frac{\varepsilon}{3}$$

by the choice of n, using the triangle inequality we deduce from the preceding three estimates that

$$\|f(x) - f(a) - g(a)(x - a)\| \le \varepsilon \|x - a\|$$

for all $x \in B_{r'}(a)$. Hence f is differentiable at a, and $f'(a) = g(a)$. □

Example Let U be a non-empty open set in X, and denote by $C_b^1(U, Y)$ the vector space of all continuously differentiable functions $f : U \to Y$ for which f and f' are bounded. If Y is complete, then $C_b^1(U, Y)$ is a Banach space for the norm

$$\|f\| := \sup_{x \in U} \|f(x)\| + \sup_{x \in U} \left\|f'(x)\right\| .$$

Indeed, if (f_n) is a Cauchy sequence for this norm, then (f_n) and (f_n') are Cauchy sequences in $C_b(U, Y)$. Since $C_b(U, Y)$ is a Banach space there exist $f, g \in C_b(U, Y)$ such that $f_n \to f$ and $f_n' \to g$ uniformly on U. By the preceding proposition we have $g = f'$; hence $f \in C_b^1(U, Y)$, and $f_n \to f$ $C_b^1(U, Y)$.

4.3 The Functions $\mathbb{R}^m \hookrightarrow \mathbb{R}^n$

The differentiability of vector-valued functions may be reduced to that of real-valued functions:

Proposition 4.10 *Let $f = (f_1, \dots, f_n) : X \hookrightarrow \mathbb{R}^n$ and $a \in D(f)$.*

(a) *f is differentiable at a if and only if all components $f_j : \mathbb{R}^m \hookrightarrow \mathbb{R}$ are differentiable at a. Then the equality*

$$f'(a)h = (f_1'(a)h, \dots, f_n'(a)h) \tag{4.11}$$

holds for all $h \in X$.

(b) *The preceding implication remains valid if we replace the words "differentiable at a" by one of the following: "continuously differentiable at a", "differentiable" or "continuously differentiable".*

Proof Assume first that f is differentiable at a. We have $f_j = P_j \circ f$, where $P_j(y_1, \dots, y_n) := y_j$ is a continuous linear *projection*. Applying Proposition 4.2 (d) (p. 101) we conclude that the functions f_j are differentiable at a, and

$$f_j'(a) = P_j'(f(a))f'(a) = P_j f'(a), \quad 1 \le j \le n.$$

If, moreover, f' is continuous at a, then the composite functions $f_j' = P_j \circ f'$ are also continuous at a by Proposition 1.7 (a) (p. 12).

Now assume that the functions f_j are differentiable in a. Introducing the continuous linear *embeddings*

$$B_j y := (0, \dots, 0, y, 0, \dots, 0)$$

of $\mathbb{R}$ into $\mathbb{R}^n$, where y stands in the jth position, we have

$$f = \sum_{j=1}^n B_j \circ f_j.$$

Applying Proposition 4.2 (c) and (d) we conclude that f is differentiable at a, and

$$f'(a) = \sum_{j=1}^n B_j'(f_j(a))f_j'(a) = \sum_{j=1}^n B_j f_j'(a).$$

The last equality is equivalent to (4.11). If, moreover, the functions f_j' are continuous at a, then $f' = \sum B_j \circ f_j'$ is also continuous at a by Proposition 1.7 (a).

Finally, the global statements follow by applying the above results for each $a \in D(f)$. □

Next we study real-valued functions $f : \mathbb{R}^m \hookrightarrow \mathbb{R}$. We introduce the usual orthonormal basis $e_1, \ldots, e_m$ of $\mathbb{R}^m$.

Definition By the ith *partial derivative* of a function $f : \mathbb{R}^m \hookrightarrow \mathbb{R}$ at a we mean its derivative in the direction e_i at a; it is denoted by

$$D_i f(a), \quad \frac{\partial f}{\partial x_i}(a) \quad \text{or} \quad \partial_i f(a).$$

Equivalently, $D_i f(a)$ is the derivative of the function

$$\mathbb{R} \ni x_i \mapsto f(a_1, \ldots, a_{i-1}, x_i, a_{i+1}, \ldots, a_m) \in \mathbb{R}$$

at a_i (if it exists).

We also introduce the *partial derivative function* $D_i f : \mathbb{R}^m \hookrightarrow \mathbb{R}$.

Remark If $f : \mathbb{R}^m \hookrightarrow \mathbb{R}$ is differentiable at a, then the partial derivatives $D_i f(a)$ also exist, and $D_i f(a) = f'(a)e_i$ for each i by Proposition 4.2 (b). Hence

$$f'(a)h = \sum_{i=1}^{m} D_i f(a)h_i \tag{4.12}$$

for all $h \in \mathbb{R}^m$. Defining the *gradient vector* of f by the formula[8]

$$\nabla f(a) := (D_1 f(a), \ldots, D_m f(a)),$$

by (4.12) the relation

$$f(a+h) = f(a) + \nabla f(a) \cdot h + o(h), \quad h \to 0$$

holds, where the dot stands for the usual scalar product of $\mathbb{R}^m$. This reveals the geometrical interpretation of $\nabla f(a)$: starting from a it is the direction of the largest growth of f.

It is often convenient to identify f' with ∇f, and to consider the partial derivatives as the components of f'.

If we represent the elements of $\mathbb{R}^m$ and $\mathbb{R}^n$ as column vectors, then combining (4.11) and (4.12) we obtain the

[8] It is pronounced "nabla f".

Proposition 4.11 *If $f : \mathbb{R}^m \hookrightarrow \mathbb{R}^n$ is differentiable at $a \in \mathbb{R}^m$, then its derivative may be computed by the following formula:*

$$f'(a)h = \begin{pmatrix} D_1f_1(a) \cdots D_mf_1(a) \\ \vdots \qquad\qquad \vdots \\ D_1f_n(a) \cdots D_mf_n(a) \end{pmatrix} \begin{pmatrix} h_1 \\ \vdots \\ h_m \end{pmatrix}, \quad h \in \mathbb{R}^m. \tag{4.13}$$

Generalizing the identification of f' and ∇f for real-valued functions, it is often convenient to identify $f'(a)$ with the matrix in (4.13).

The example on p. 104 shows that the existence of the partial derivatives does not imply total differentiability, and hence the validity of the formula (4.13). We have, however, the following useful result:

Proposition 4.12 *A function $f : U \to \mathbb{R}$, $U \subset \mathbb{R}^m$, is continuously differentiable if and only if its partial derivatives exist and are continuous in U.*

Proof We identify f' with ∇f for convenience.

If f is continuously differentiable, then we already know that the partial derivatives exist, and $f' = (D_1f, \dots, D_mf)$. Since f' is continuous, its components D_if are also continuous.

Now assume that the partial derivatives exist and are continuous in U, and consider the norm $\|\cdot\|_\infty$ on $\mathbb{R}^m$.

First we show that f is differentiable at each point $a = (a_1, \dots, a_m) \in U$. Fix a ball $B_r(a) \subset U$.[9] If $x = (x_1, \dots, x_m) \in B_r(a)$, then the broken line

$$L = [y_0, y_1] \cup \cdots \cup [y_{m-1}, y_m]$$

defined by the points

$$\begin{aligned} y_0 &:= (a_1, a_2, \dots, a_{m-1}, a_m), \\ y_1 &:= (x_1, a_2, \dots, a_{m-1}, a_m), \\ &\vdots \\ y_{m-1} &:= (x_1, x_2, \dots, x_{m-1}, a_m), \\ y_m &:= (x_1, x_2, \dots, x_{m-1}, x_m) \end{aligned}$$

connects a and x, and lies in $B_r(a)$ by the definition of the norm. (See the right-hand side of Fig. 4.6 for $m = 2$.)

[9] We recall that a differentiable function is defined on an open set by definition.

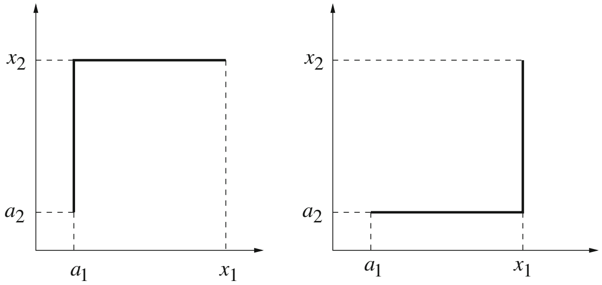


Fig. 4.6 Proof of Proposition 4.12

Applying the classical mean value theorem to the differences on the right-hand side of the equality

$$f(x)-f(a)=\sum_{j=1}^{m} f(y_j)-f(y_{j-1}),$$

we obtain the relation

$$f(x)-f(a)=\sum_{j=1}^{m} D_jf(z_j)(x_j-a_j)$$

with suitable points $z_j \in [y_{j-1}, y_j]$.[10]

Now the formula

$$u(x)h := \sum_{j=1}^{m} D_jf(z_j)h_j$$

defines a continuous linear map $u(x) \in L(\mathbb{R}^m, \mathbb{R})$ satisfying the equality

$$f(x)-f(a)=u(x)(x-a).$$

We claim that u is continuous at a. It follows from the trivial estimate

$$\left|\bigl(u(x)-u(a)\bigr)h\right| \le \sum_{j=1}^{m} \left|D_jf(z_j)-D_jf(a)\right| \cdot |h_j|$$

[10] The points z_j depend on x. If $y_{j-1}=y_j$ we may take $z_j=y_j$.

and the inequalities $|h_j| \leq \|h\|$ that

$$\|u(x) - u(a)\| \leq \sum_{j=1}^{m} \left|D_j f(z_j) - D_j f(a)\right| .$$

If $x \to a$, then $z_j \to a$ for $j = 1, \dots, m$, so that the right-hand side tends to zero by the continuity of the partial derivatives. Hence $u(x) \to u(a)$.

We have proved that f is differentiable and $f' = (D_1 f, \dots, D_m f)$. Since the components $D_j f$ of f' are continuous by assumption, f' is also continuous. □

***Remark** There are $m!$ different ways to introduce broken lines L for which the above proof remains valid (see the left-hand side of Fig. 4.6 for $m = 2$). They usually lead to different functions u. This shows that the function u in the definition of the total derivative is not always unique.

4.4 Exercises

Exercise 4.1 Prove that the function

$$f : \mathbb{R}^3 \to \mathbb{R}^2, \quad f(x, y, z) := (\sin(x - e^z), x^2 + y^2)$$

is differentiable, and compute its derivatives.

Exercise 4.2 Study the continuity and differentiability of the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(0,0) = 0 \quad \text{and} \quad f(x, y) = \frac{2xy^2}{x^2 + y^4} \quad \text{otherwise.}$$

Exercise 4.3 Prove the following statements:

(i) The function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by the formulas $f(0,0) = 0$ and

$$f(x, y) = \frac{xy}{\sqrt{x^2 + y^2}} \quad \text{otherwise}$$

is continuous and has bounded partial derivatives $D_1 f$ and $D_2 f$ in a neighborhood of $(0,0)$, but it is not differentiable at $(0,0)$.

(ii) The function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by the formulas $f(0,0) = 0$ and

$$f(x, y) := (x^2 + y^2) \sin \frac{1}{x^2 + y^2}$$

has unbounded partial derivatives $D_1 f$ and $D_2 f$ in every neighborhood of $(0,0)$, but it is differentiable at $(0,0)$.

Exercise 4.4 Let E be a Euclidean space and $\|\cdot\|$ the corresponding norm. Prove that the *inversion*

$$f(x) := \frac{x}{\|x\|^2}, \quad x \in E \setminus \{0\}$$

is differentiable, and compute its derivative.

Exercise 4.5 Let $f : X \hookrightarrow Y$ be differentiable in $B_r(a) \setminus \{a\}$ and continuous at a. Prove that if f' has a limit A at a, then f is differentiable at a, and $f'(a) = A$.

Exercise 4.6 Let $f : X \hookrightarrow Y$ be a differentiable function on a convex open set U. Prove the following:

(i) f is (globally) Lipschitz continuous $\Longleftrightarrow f'$ is bounded.
(ii) If $f \in C^1$, then f is locally Lipschitz continuous.[11]

Exercise 4.7 (Euler's Identity) A function $f : \mathbb{R}^n \setminus \{0\} \to \mathbb{R}$ is said to be *homogeneous of order m* for some real constant m if $f(tx) = t^m f(x)$ for all $x \in \mathbb{R}^n \setminus \{0\}$ and $t > 0$.

(i) Are the functions

$$f(x, y, z) := (x - 2y + 3z)^2 \quad \text{and} \quad g(x, y, z) := \left(\frac{x}{y}\right)^{y/z}$$

homogeneous? If yes, of what order?
(ii) Prove that every differentiable homogeneous function satisfies Euler's identity $x \cdot \nabla f(x) \equiv mf(x)$.
(iii) Prove that, conversely, if a differentiable function satisfies Euler's identity, then it is homogeneous.

Exercise 4.8 (Differentiation with Respect to a Parameter) Let $f : (-r, r) \times [a, b] \to \mathbb{R}$ be a continuous function, and set

$$F(x) = \int_a^b f(x, y)\, dy, \quad x \in (-r, r).$$

Assume that $D_1 f$ exists and is continuous in $(-r, r) \times [a, b]$. Prove that $F : (-r, r) \to \mathbb{R}$ is differentiable, and

$$F'(x) = \int_a^b D_1 f(x, y)\, dy, \quad x \in (-r, r).$$

[11] See the definition in Exercise 3.4, p. 90.

Chapter 5
Higher-order derivatives

As in the preceding chapter, the letters X, Y, Z always denote normed spaces.

5.1 Continuous multilinear maps

Let $(X, \|\cdot\|)$ be the product of the normed spaces

$$(X_1, \|\cdot\|_1), \ldots, (X_m, \|\cdot\|_m);$$

we recall that

$$\|x\| := \|(x_1, \ldots, x_m)\| := \|x_1\|_1 + \cdots + \|x_m\|_m .$$

Definition A map $A : X \to Y$ is *m-linear* if the functions

$$X_j \ni x_j \mapsto A(x_1, \ldots, x_m) \in Y$$

are linear for any fixed $j \in \{1, \ldots, m\}$ and $x_i \in X_i$, $i \in \{1, \ldots, m\} \setminus \{j\}$.

Examples

- For $m = 1$ we get the usual linear maps.
- Every matrix (a_{ij}) of size $m \times n$ defines a bilinear map $A : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ by the formula

$$A(x, y) := \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} x_i y_j .$$

- A determinant of order m is an m-linear map $\mathbb{R}^m \times \cdots \times \mathbb{R}^m \to \mathbb{R}$.
- The scalar product of a Euclidean space H is a bilinear map $H \times H \to \mathbb{R}$.

V. Komornik, *Topology, Calculus and Approximation*, Springer
Undergraduate Mathematics Series, DOI 10.1007/978-1-4471-7316-8_5

Lemma 5.1 *An m-linear map $A : X \to Y$ is continuous if and only if there exists a constant $M \geq 0$ such that*

$$\|A(x_1, \ldots, x_m)\|_Y \leq M \, \|x_1\|_1 \cdots \|x_m\|_m \tag{5.1}$$

for all $x_i \in X_i$, $i = 1, \ldots, m$.

Proof If A is continuous at 0, then there exists an $r > 0$ such that

$$\|z\|_X \leq r \Longrightarrow \|Az\|_Y \leq 1.$$

Setting $M := (m/r)^m$, (5.1) follows. Indeed, writing the vectors x_i in the form

$$x_i = t_i z_i \quad t_i := (m/r) \, \|x_i\|_i \, ,$$

we have

$$\|z\|_X = \|z_1\|_1 + \cdots + \|z_m\|_m = mr/m = r$$

and therefore

$$\|Ax\|_Y = t_1 \cdots t_m \, \|Az\|_Y \leq t_1 \cdots t_m = M \, \|x_1\|_1 \cdots \|x_m\|_m \, .$$

Conversely, if (5.1) is satisfied, then for any $x, h \in X$ we have

$$A(x+h) = \sum_{k=0}^{m} A(x_1 + h_1, \ldots, x_{k-1} + h_{k-1}, h_k, x_{k+1}, \ldots, x_m),$$

and hence

$$\|A(x+h) - Ax\|_Y \leq M \sum_{k=1}^{m} \|x_1 + h_1\| \cdots \|x_{k-1} + h_{k-1}\| \cdot \|h_k\| \cdot \|x_{k+1}\| \cdots \|x_m\| \, .$$

For any fixed x the right-hand side tends to zero as $h \to 0$, so that $A(x+h) \to Ax$. □

We denote by $L^m(X, Y)$ the set of continuous m-linear maps $A : X \to Y$. Similarly to the earlier case $m = 1$, for $A \in L^m(X, Y)$ we have

$$\|A\| := \sup_{\|x_1\|_1 \leq 1} \cdots \sup_{\|x_m\|_m \leq 1} \|A(x_1, \ldots, x_m)\|_Y < \infty$$

by the lemma, and $\|A\|$ is the smallest nonnegative constant M satisfying (5.1).

Proposition 5.2 $(L^m(X,Y), \|\cdot\|)$ *is a normed space.*

Proof A simple adaptation of the proof of Proposition 3.14 (p. 83). □

The study of multilinear maps may be reduced to that of linear maps. Given $A \in L^m(X_1 \times \cdots \times X_m, Y)$, the formula

$$f(A)(x_1)(x_2,\dots,x_m) := A(x_1,\dots,x_m)$$

defines a map $f(A)$ of X_1 into the set of functions $X_2 \times \cdots \times X_m \to Y$.

It follows from the multilinearity of A and from the estimate

$$\|f(A)(x_1)(x_2,\dots,x_m)\| = \|A(x_1,\dots,x_m)\| \le \|A\| \cdot \|x_1\| \cdots \|x_m\|$$

that $f(A)(x_1) \in L^{m-1}(X_2 \times \cdots \times X_m, Y)$ for each $x_1 \in X_1$ and, moreover,

$$f(A) \in L(X_1, L^{m-1}(X_2 \times \cdots \times X_m, Y))$$

with $\|f(A)\| \le \|A\|$.

Even more is true:

Proposition 5.3 *The map*

$$f : L^m(X_1 \times \cdots \times X_m, Y) \to L(X_1, L^{m-1}(X_2 \times \cdots \times X_m, Y))$$

is an isometric isomorphism.

Proof If $B \in L(X_1, L^{m-1}(X_2 \times \cdots \times X_m, Y))$, then the formula

$$g(B)(x_1,\dots,x_m) := (Bx_1)(x_2,\dots,x_m)$$

defines a continuous m-linear map

$$g(B) \in L^m(X_1 \times \cdots \times X_m, Y).$$

The m-linearity is obvious. Furthermore, the estimate

$$\begin{aligned}\|g(B)(x_1,\dots,x_m)\| &= \|B(x_1)(x_2,\dots,x_m)\| \\ &\le \|Bx_1\| \cdot \|x_2\| \cdots \|x_m\| \\ &\le \|B\| \cdot \|x_1\| \cdots \|x_m\|\end{aligned}$$

implies that $g(B)$ is continuous, and $\|g(B)\| \le \|B\|$.

It follows from the definitions that f, g are linear,

$$f(g(B)) = B \quad \text{for all} \quad B \in L(X_1, L^{m-1}(X_2 \times \cdots \times X_m, Y))$$

and

$$g(f(A)) = A \quad \text{for all} \quad A \in L^m(X_1 \times \cdots \times X_m, Y).$$

Hence f is a (linear) bijection between the two vector spaces with $f^{-1} = g$. Finally, f is an isometry because

$$\|A\| = \|g(f(A))\| \le \|f(A)\| \le \|A\|$$

for all A. □

The above proposition has important consequences:

Corollary 5.4

(a) *The normed spaces*

$$L(X_1, L(X_2, \dots L(X_m, Y) \dots) \quad \textit{and} \quad L^m(X, Y)$$

are isometrically isomorphic.

(b) *If* $\dim X < \infty$*, then every* m*-linear map* $A : X \to Y$ *is continuous.*

Proof

(a) This follows from the preceding proposition by induction on m.

(b) Combine (a) with Theorem 3.15 (p. 83). □

Remark Henceforth we identify the spaces in (a). If $X_1 = \cdots = X_m = Z$, then we write $X = Z^m$, so that we identify

$$L(Z, L(Z, \dots L(Z, Y) \dots) \quad \text{and} \quad L^m(Z^m, Y).$$

If, moreover, $Z = \mathbb{R}$, then these spaces may be identified with Y by Proposition 3.17 (p. 85).

These identifications simplify the manipulation of higher-order derivatives.

5.2 Higher-order derivatives

Higher-order derivatives are defined recursively:

Definitions Let $f : X \hookrightarrow Y$, $a \in D(f)$ and $k \ge 2$.

- f is *k times differentiable at a* if it is differentiable in a neighborhood of a, and $f' : X \hookrightarrow L(X, Y)$ is $(k-1)$ times differentiable at a. Then $(f')^{(k-1)}(a)$ is called the *kth derivative* of f at a, and is denoted by $f^{(k)}(a)$. Thus we have

$$f^{(k)}(a) \in L^k(X^k, Y) \quad \text{and} \quad f^{(k)} : X \hookrightarrow L^k(X^k, Y).$$

- f is *k times differentiable* if (its domain of definition is open, and) f is *k times differentiable* at each $a \in D(f)$.
- f is *k times continuously differentiable at a* if it is k times differentiable at a, and $f^{(k)}$ is continuous at a.
- f is *k times continuously differentiable* or *belongs to the class C^k* if it is k times differentiable, and $f^{(k)}$ is continuous. We write $f \in C^k$ in this case.
- f is *infinitely many times differentiable* or *belongs to the class C^∞*, if $f \in C^k$ for all $k = 1, 2, \ldots$. We express this property by writing $f \in C^\infty$.

Examples

- Every constant function is differentiable, and its derivative is identically zero, hence also a constant function. Consequently, the constant functions belong to the class C^∞.
- Every continuous linear map belongs to the class C^∞ because it is differentiable, and its derivative is a constant function.
- Every continuous bilinear map belongs to the class C^∞ because it is differentiable, and its derivative is a continuous linear map: see the last example preceding Proposition 4.2, p. 101.
- Given a non-empty open set U in X, we denote by $C_b^k(U, Y)$ the set of functions $f : U \to Y$ of class C^k such that all functions $f, f', \ldots, f^{(k)}$ are bounded. If Y is complete, then using Proposition 4.9 (p. 109) it follows that $C_b^k(U, Y)$ is a Banach space for the norm

$$\|f\| := \sup_{x\in U} \|f(x)\| + \sup_{x\in U} \left\| f'(x) \right\| + \cdots + \sup_{x\in U} \left\| f^{(k)}(x) \right\| .$$

We may extend Proposition 4.2 (p. 101):

Proposition 5.5 *Consider two functions $g : X \hookrightarrow Y$, $f : Y \hookrightarrow Z$ and their composition $f \circ g : X \hookrightarrow Z$.*

(a) *The map $f \mapsto f^{(k)}(a)$ is linear.*
(b) *If g is k times differentiable at a and f is k times differentiable at $g(a)$, then $f \circ g$ is k times differentiable at a.*
(c) *The preceding implication remains valid if we replace the words "differentiable at a" by one of the following: "continuously differentiable at a", "differentiable" or "continuously differentiable".*

Proof The case $k = 1$ is contained in Proposition 4.2. Let $k \ge 2$, and assume that the results are already known for $k - 1$.

(a) Each linear combination $\alpha f + \beta g$ satisfies the equality

$$\begin{aligned}
(\alpha f + \beta g)^{(k)}(a) &= \left((\alpha f + \beta g)'\right)^{(k-1)}(a) \\
&= \left(\alpha f' + \beta g'\right)^{(k-1)}(a) \\
&= \alpha f^{(k)}(a) + \beta g^{(k)}(a).
\end{aligned}$$

(b) Since $k \geq 2$, g' is defined in a neighborhood of a, and f' is defined in a neighborhood of $g(a)$. By Proposition 4.2 the function $f \circ g$ is differentiable in a neighborhood of a, and

$$(f \circ g)' = \varphi \circ (f' \circ g, g'), \tag{5.2}$$

where φ denotes the continuous bilinear map of Proposition 3.16 (p. 84). On the right-hand side the functions g and g' are $k-1$ times differentiable at a, and f' is $k-1$ times differentiable at $g(a)$. Since $\varphi \in C^\infty$, by the induction hypothesis the function $(f \circ g)'$ is $k-1$ times differentiable at a. Consequently $f \circ g$ is k times differentiable at a.

(c) If $g^{(k)}$ is continuous at a and $f^{(k)}$ is continuous at $g(a)$, then the preceding proof yields that $(f \circ g)^{(k)}$ is continuous at a. The other two versions are obtained by applying (b) and this result to every point of $D(g)$. □

In the rest of this section we study a symmetry property of higher-order derivatives.

Example We recall again that every continuous bilinear functional $\varphi : X \times Y \to Z$ is differentiable, and

$$\varphi'(x,y)(h,k) = \varphi(x,k) + \varphi(h,y).$$

Since $\varphi' : X \times Y \to L(X \times Y, Z)$ is a continuous linear map, it follows that

$$\varphi''(x,y)\big((h_1,k_1),(h_2,k_2)\big) = \varphi(h_1,k_2) + \varphi(h_2,k_1) :$$

the bilinear functional $\varphi''(x,y)$ is symmetric.

Remarks

- The above example is a special case of a result of Euler, stating that the mixed partial derivatives $D_1D_2f(a)$ and $D_2D_1f(a)$ are usually equal. Schwarz proved this under the assumption that D_1D_2f and D_2D_1f exist in a neighborhood of a, and are continuous at a.
- Schwarz also gave an example of a function $f : \mathbb{R}^2 \to \mathbb{R}$ such that $D_1D_2f(0,0)$ and $D_2D_1f(0,0)$ exist, but are not equal. A simpler counterexample was given by Peano: $f(0,0) := 0$, and

$$f(x,y) := xy\frac{x^2-y^2}{x^2+y^2} \quad \text{if} \quad x^2+y^2 > 0.$$

We will show that total differentiability also implies the equality $D_1D_2f = D_2D_1f$.

Definition $A \in L^k(X^k, Y)$ is *symmetric* if $A(x_1,\dots,x_k)$ is invariant under permutations of the vectors $x_1,\dots,x_k \in X$.

Theorem 5.6 (Young) *If $f : X \hookrightarrow Y$ is k times differentiable at a for some $k \geq 2$, then the derivative $f^{(k)}(a) \in L^k(X^k, Y)$ is symmetric.*

Proof for C^2 functions $f : \mathbb{R}^2 \hookrightarrow \mathbb{R}$ We assume by translation that $a = (0,0)$. Since

$$f''(a)(h,k) = D_1D_1f(a)h_1k_1 + D_1D_2f(a)h_1k_2 \\ + D_2D_1f(a)h_2k_1 + D_2D_2f(a)h_2k_2,$$

it suffices to show that $D_1D_2f(a) = D_2D_1f(a)$. Choose a small ball $B_r(a)$ on which f is defined. Then the functions

$$F(x) := \int_0^x \int_0^x D_1D_2f(s,t)\; ds\; dt$$

and

$$G(x) := \int_0^x \int_0^x D_2D_1f(s,t)\; dt\; ds$$

are defined for all $0 < x < r$ and

$$D_1D_2f(0,0) = \lim_{x \to a} \frac{F(x)}{x^2}, \quad D_2D_1f(0,0) = \lim_{x \to a} \frac{G(x)}{x^2}$$

by the continuity of the functions under the integral sign. This yields the required result because $F(x) \equiv G(x)$. Indeed, applying the Newton–Leibniz formula we have

$$\begin{aligned} F(x) &= \int_0^x D_2f(x,t) - D_2f(0,t)\; dt \\ &= f(x,x) - f(0,x) - f(x,0) + f(0,0) \end{aligned}$$

and

$$\begin{aligned} G(x) &= \int_0^x D_1f(s,x) - D_1f(s,0)\; ds \\ &= f(x,x) - f(x,0) - f(0,x) + f(0,0). \end{aligned}$$

□

**Proof for twice differentiable functions* Let $g : \mathbb{R}^2 \hookrightarrow \mathbb{R}$ be twice differentiable at $(0,0)$. We claim that

$$D_1D_2g(0,0) = D_2D_1g(0,0). \tag{5.3}$$

Fixing a sufficiently small number $\ell > 0$, the formula

$$u(x) := g(x,\ell) - g(x,0)$$

defines a function $u : \mathbb{R} \hookrightarrow \mathbb{R}$ that is differentiable at each point of the segment $[0,\ell]$. Applying the classical mean value theorem there exists $0 < t < \ell$ such that

$$u(\ell) - u(0) = u'(t)\ell = (D_1g(t,\ell) - D_1g(t,0))\ell. \tag{5.4}$$

Since D_1g differentiable at 0 by our assumption, for $\ell \to 0$ we have

$$D_1g(t,\ell) = D_1g(0,0) + D_1D_1g(0,0)t + D_2D_1g(0,0)\ell + o(t+\ell)$$

and

$$D_1g(t,0) = D_1g(0,0) + D_1D_1g(0,0)t + o(t),$$

and hence, since $0 < t < \ell$,

$$u(\ell) - u(0) = D_2D_1g(0,0)\ell^2 + o(\ell^2). \tag{5.5}$$

Similarly, the function $v : \mathbb{R} \hookrightarrow \mathbb{R}$ defined by the formula

$$v(y) := g(\ell,y) - g(0,y)$$

satisfies

$$v(\ell) - v(0) = D_1D_2g(0,0)\ell^2 + o(\ell^2). \tag{5.6}$$

Now we observe that

$$u(\ell) - u(0) = v(\ell) - v(0).$$

Therefore (5.5) and (5.6) yield the equality

$$D_2D_1g(0,0) - D_1D_2g(0,0) = \frac{o(\ell^2)}{\ell^2}.$$

Letting $\ell \to 0$, (5.3) follows. □

Remark Using the above functions u and v, Schwarz's theorem may be proved as follows. Since D_2D_1g and D_1D_2g exist in a neighborhood of $(0,0)$, we deduce from (5.4) for each sufficiently small $\ell > 0$ that

$$u(\ell) - u(0) = D_2D_1g(t,s))\ell^2$$

for some $t, s \in (0, \ell)$. Similarly, we have

$$v(\ell) - v(0) = D_1D_2g(t', s'))\ell^2$$

for some $t', s' \in (0, \ell)$. Since $u(\ell) - u(0) = v(\ell) - v(0)$, and D_2D_1g, D_1D_2g are continuous at $(0, 0)$, dividing by ℓ^2 and then letting $\ell \to 0$ we get (5.3).

To prepare the proof of the general case we present a method to reduce the study of derivatives of functions $X \to L^k(X^k, Y)$ to that of simpler functions $X \to Y$:

Lemma 5.7 *Let $f : X \hookrightarrow Y$ be k times differentiable at a for some $k \geq 2$. For any fixed vectors $h_1, \dots, h_k \in X$ the function $\psi : X \hookrightarrow Y$ defined by*

$$\psi(x) := f^{(k-1)}(x)(h_2, \dots, h_k)$$

is differentiable at a, and

$$\psi'(a)h_1 := f^{(k)}(a)(h_1, \dots, h_k).$$

Proof We may write ψ in the form $\psi = A \circ f^{(k-1)}$ with the evaluation map $A \in L(L^{k-1}(X^{k-1}, Y), Y)$ defined by the formula $AL := L(h_2, \dots, h_k)$. Using this formula we obtain that ψ is differentiable at a, and

$$\psi'(a)h_1 = Af^{(k)}(a)h_1 = f^{(k)}(a)(h_1, \dots, h_k).$$

□

Proof of Theorem 5.6 in the general case For any fixed vectors

$$h_1, \dots, h_k \in X$$

we have to prove the equalities

$$f^{(k)}(a)(h_1, \dots, h_k) = f^{(k)}(a)(h_{i_1}, \dots, h_{i_k})$$

for all permutations $i_1, \dots, i_k$ of $1, \dots, k$. Since each permutation is composed of finitely many transpositions of consecutive elements, it suffices to show that

$$f^{(k)}(a)(h_1, \dots, h_k) = f^{(k)}(a)(h_1, \dots, h_{j-1}, h_{j+1}, h_j, h_{j+2}, \dots, h_k) \tag{5.7}$$

for $j = 1, \dots, k - 1$.

We start with the case $j = 1$. For any given functional $\varphi \in Y'$ the formula

$$g(s, t) := \varphi(f^{(k-2)}(a + sh_1 + th_2)(h_3, \dots, h_k))$$

defines a function $g : \mathbb{R}^2 \hookrightarrow \mathbb{R}$ that is twice differentiable at $(0,0)$. A simple computation shows that

$$D_2 g(s,0) = \varphi(f^{(k-1)}(a + sh_1)(h_2, h_3, \dots, h_k)),$$
$$D_1 D_2 g(0,0) = \varphi(f^{(k)}(a)(h_1, h_2, \dots, h_k)),$$

and

$$D_1 g(0,t) = \varphi(f^{(k-1)}(a + th_2)(h_1, h_3, \dots, h_k)),$$
$$D_2 D_1 g(0,0) = \varphi(f^{(k)}(a)(h_2, h_1, \dots, h_k)).$$

By what we already know, it follows that

$$\varphi(f^{(k)}(a)(h_1, h_2, \dots, h_k)) = \varphi(f^{(k)}(a)(h_2, h_1, \dots, h_k)).$$

Since this equality is satisfied for all $\varphi \in Y'$, applying Proposition 3.19 (p. 86) the relation (5.7) follows.

Now let $j \geq 2$. Since $k - j + 1 < k$, the function f is $k - j + 1$ times differentiable in a neighborhood of a. Applying the preceding case we have

$$f^{(k-j+1)}(x)(h_j, h_{j+1}, \dots, h_k) = f^{(k-j+1)}(x)(h_{j+1}, h_j, \dots, h_k)$$

in this neighborhood. Differentiating this equality $j - 1$ times at a, and using the preceding lemma, (5.7) follows. □

5.3 Taylor's formula

The purpose of this section is to establish several variants of Taylor's formula, providing good local approximations of sufficiently smooth functions.

For brevity we denote by h^k the vector $(h, \dots, h) \in X^k$.

Theorem 5.8 (Peano) *If $f : X \hookrightarrow Y$ is n times differentiable at a, then*

$$f(a+h) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!} h^k + o(\|h\|^n), \quad h \to 0. \tag{5.8}$$

Proof For $n = 1$ this is the definition of the derivative. Let $n \geq 2$, and assume by induction that

$$f'(a+h_1) = \sum_{i=0}^{n-1} \frac{(f')^{(i)}(a) h_1^i}{i!} + o(\|h_1\|^{n-1}), \quad h_1 \to 0. \tag{5.9}$$

Choose a ball $B_r(a)$ in which f is defined and is $n-1$ times differentiable. For each fixed $h \in B_r(0)$ the formula

$$g(t) := f(a+th) - \sum_{k=1}^{n} \frac{f^{(k)}(a)h^k}{k!} t^k$$

defines a function $g : [0,1] \to Y$ satisfying the assumptions of Theorem 4.7 (p. 107). Therefore there exists a $t \in (0,1)$ such that $\|g(1) - g(0)\| \le \|g'(t)\|$, i.e.,

$$\left\| f(a+h) - \sum_{k=0}^{n} \frac{f^{(k)}(a)h^k}{k!} \right\| \le \left\| f'(a+th)h - \sum_{k=1}^{n} \frac{f^{(k)}(a)h^k}{(k-1)!} t^{k-1} \right\|.$$

Applying (5.9) with $h_1 = th$, the right-hand side may be estimated as follows:

$$\left\| f'(a+th)h - \sum_{k=1}^{n} \frac{f^{(k)}(a)h^k}{(k-1)!} t^{k-1} \right\| \le o(\|th\|^{n-1})\,\|h\| = o(\|h\|^n).$$

□

Next we generalize the Lagrange mean value theorem.[1]

Proposition 5.9 (Lagrange) *If $f : X \hookrightarrow \mathbb{R}$ is n times differentiable on a non-degenerate segment $[a,b] \subset X$, then there exists a $c \in (a,b)$ such that*

$$f(b) = \sum_{k=0}^{n-1} \frac{f^{(k)}(a)}{k!}(b-a)^k + \frac{f^{(n)}(c)}{n!}(b-a)^n.$$

***Remark** The proposition and the following theorem remain valid, with the same proof, under the weaker assumptions that f is continuous on $[a,b]$, n times differentiable on (a,b), and $n-1$ times continuously differentiable at a. See also the comment on p. 342.

Proof Similarly to the proof of Proposition 4.6 we may reduce the problem to the case $X = \mathbb{R}$ and $[a,b] = [0,1]$ with the help of the affine function $\ell(t) = (1-t)a + tb$. Then we repeatedly apply the Cauchy mean value theorem (Proposition 4.4 (c), p. 105) to the functions

$$g(t) := f(t) - \sum_{k=0}^{n-1} \frac{f^{(k)}(0)}{k!} t^k \quad \text{and} \quad h(t) := t^n$$

as follows.

[1] Proposition 4.6 (p. 107) corresponds to the case $n = 1$.

Since $g^{(k)}(0) = h^{(k)}(0) = 0$ for $k = 0, \dots, n-1$, we have

$$\begin{aligned}\frac{g(1)}{h(1)} &= \frac{g(1)-g(0)}{h(1)-h(0)}\\ &=\frac{g'(t_1)}{h'(t_1)} = \frac{g'(t_1)-g'(0)}{h'(t_1)-h'(0)}\\ &\dots\\ &=\frac{g^{(n)}(t_n)}{h^{(n)}(t_n)}\end{aligned}$$

for suitable points $0 < t_n < t_{n-1} < \cdots < t_1 < 1$.

Since $h(1) = 1$, $g^{(n)}(t_n) = f^{(n)}(t_n)$ and $h^{(n)}(t_n) = n!$, the proposition follows with $c = t_n$. □

Now we generalize Theorem 4.7 (p. 107):

Theorem 5.10 *If $f : X \hookrightarrow Y$ is n times differentiable on a non-degenerate segment $[a,b] \subset X$, then there exist $c_1, c_2 \in (a,b)$ such that*

$$\Big\| f(b) - \sum_{k=0}^{n-1} \frac{f^{(k)}(a)(b-a)^k}{k!} \Big\| \le \frac{\big\| f^{(n)}(c_1)(b-a)^n \big\|}{n!} \tag{5.10}$$

and

$$\Big\| f(b) - \sum_{k=0}^{n} \frac{f^{(k)}(a)(b-a)^k}{k!} \Big\| \le \frac{\big\| (f^{(n)}(c_2) - f^{(n)}(a))(b-a)^n \big\|}{n!}. \tag{5.11}$$

Remark The assumptions are stronger than those of Theorem 5.8, but the error estimates are also stronger.

Proof Fix a functional $\varphi \in Y'$ (to be chosen later). Applying the preceding proposition to the composite function $\varphi \circ f : X \hookrightarrow \mathbb{R}$, there exists a $c \in (a,b)$ (depending on φ) such that

$$(\varphi \circ f)(b) = \sum_{k=0}^{n-1} \frac{(\varphi \circ f)^{(k)}(a)(b-a)^k}{k!} + \frac{(\varphi \circ f)^{(n)}(c)(b-a)^n}{n!}.$$

Since $(\varphi \circ f)^{(k)} = \varphi \circ f^{(k)}$ for all k, this may be rewritten as

$$\varphi(f(b)) = \varphi\Big(\sum_{k=0}^{n-1} \frac{f^{(k)}(a)(b-a)^k}{k!} + \frac{f^{(n)}(c)(b-a)^n}{n!} \Big),$$

and this equality implies the estimates

$$\varphi\Big(f(b)-\sum_{k=0}^{n-1}\frac{f^{(k)}(a)(b-a)^k}{k!}\Big)\le\|\varphi\|\,\frac{\big\|f^{(n)}(c)(b-a)^n\big\|}{n!}$$

and

$$\varphi\Big(f(b)-\sum_{k=0}^{n}\frac{f^{(k)}(a)(b-a)^k}{k!}\Big)\le\|\varphi\|\,\frac{\big\|(f^{(n)}(c)-f^{(n)}(a))(b-a)^n\big\|}{n!}.$$

The estimates (5.10) and (5.11) hence follow by applying Proposition 3.19 (p. 86) to choose φ such that $\|\varphi\|\le 1$, and

$$\varphi\Big(f(b)-\sum_{k=0}^{n-1}\frac{f^{(k)}(a)(b-a)^k}{k!}\Big)=\Big\|f(b)-\sum_{k=0}^{n-1}\frac{f^{(k)}(a)(b-a)^k}{k!}\Big\|$$

or

$$\varphi\Big(f(b)-\sum_{k=0}^{n}\frac{f^{(k)}(a)(b-a)^k}{k!}\Big)=\Big\|f(b)-\sum_{k=0}^{n}\frac{f^{(k)}(a)(b-a)^k}{k!}\Big\|,$$

respectively. □

Our last version of Taylor's formula uses the integral of continuous, Banach space-valued functions that we will introduce in Section 6.1 (p. 141). However, in the most important finite-dimensional case we may simply write $f=\sum_{j=1}^m f_je_j$ in some fixed basis $e_1,\dots,e_m$ of Y, and define

$$\int_a^b f(t)\,dt:=\sum_{j=1}^m\Big(\int_a^b f_j(t)\,dt\Big)e_j.$$

The following theorem reduces to the *Newton–Leibniz formula* if $Y=\mathbb{R}$ and $n=1$.

Theorem 5.11 (Taylor's formula with integral remainder) *If $f:\mathbb{R}\hookrightarrow Y$ is n times continuously differentiable on a segment $[a,b]$, where Y is a Banach space, then*

$$f(b)=\sum_{k=0}^{n-1}\frac{f^{(k)}(a)}{k!}(b-a)^k+\int_a^b f^{(n)}(t)\frac{(b-t)^{n-1}}{(n-1)!}\,dt. \tag{5.12}$$

Proof If $Y = \mathbb{R}$, then for $n = 1$ this is the Newton–Leibniz formula, while the cases $n \ge 2$ follow by induction because

$$\begin{aligned}\int_a^b f^{(n-1)}(t)\frac{(b-t)^{n-2}}{(n-2)!}\,dt\\ &= \left[f^{(n-1)}(t)\frac{-(b-t)^{n-1}}{(n-1)!}\right]_a^b + \int_a^b f^{(n)}(t)\frac{(b-t)^{n-1}}{(n-1)!}\,dt\\ &= \frac{f^{(n-1)}(a)}{(n-1)!}(b-a)^{n-1} + \int_a^b f^{(n)}(t)\frac{(b-t)^{n-1}}{(n-1)!}\,dt.\end{aligned}$$

In the general case we fix a functional $\varphi \in Y'$, and we apply the just obtained result to the composite function $\varphi \circ f : X \hookrightarrow \mathbb{R}$:

$$(\varphi \circ f)(b) = \sum_{k=0}^{n-1} \frac{(\varphi \circ f)^{(k)}(a)}{k!}(b-a)^k + \int_a^b (\varphi \circ f)^{(n)}(t)\frac{(b-t)^{n-1}}{(n-1)!}\,dt.$$

Since $(\varphi \circ f)^{(k)} = \varphi \circ f^{(k)}$ for all k, and

$$\int_a^b (\varphi \circ g)(t)\,dt = \varphi \int_a^b g(t)\,dt$$

for every continuous function $g : [a,b] \to Y$,[2] we conclude that

$$\varphi(A) = \varphi(B)$$

for all $\varphi \in Y'$, where A and B denote the two sides of (5.12). Applying Proposition 3.19 we conclude that $A = B$. □

5.4 Local extrema

We complete Proposition 4.3 (p. 104) by using the second derivative. By symmetry we consider only the case of minima.

Definitions

- A function $f : X \hookrightarrow \mathbb{R}$ has a *strict local minimum* at $a \in D(f)$ if there exists an $r > 0$ such that $f(x) > f(a)$ for all $x \in D(f) \cap B_r(a)$, $x \neq a$.

[2] See Proposition 6.1 (d) below, p. 142.

- The quadratic form associated with the bilinear form $A \in L^2(X^2, \mathbb{R})$ is *positive semidefinite* if $A(h,h) \geq 0$ for all $h \in X$. We express this property by writing $A \geq 0$.
- The quadratic form associated with the bilinear form $A \in L^2(X^2, \mathbb{R})$ is *positive definite* if there exists a constant $c > 0$ such that $A(h,h) \geq c\,\|h\|^2$ for all $h \in X$. We write $A > 0$ in this case.[3]

Theorem 5.12 (Lagrange–Hesse) *Let $f : X \hookrightarrow \mathbb{R}$ be twice differentiable at a.*

(a) *If f has a local minimum at a, then $f'(a) = 0$ and $f''(a) \geq 0$.*
(b) *If $f'(a) = 0$ and $f''(a) > 0$, then f has a strict local minimum at a.*

Proof

(a) Fix an arbitrary vector $h \in X$. Using the minimality of $f(a)$ and applying Theorem 5.8 (p. 126) we obtain

$$f(a) \leq f(a+th) = f(a) + tf'(a)h + \frac{t^2}{2} f''(a)(h,h) + o(t^2), \quad t \to 0.$$

Since we already know from Proposition 4.3 that $f'(a) = 0$, hence

$$0 \leq f''(a)(h,h) + \frac{o(t^2)}{t^2}, \quad t \to 0.$$

Letting $t \to 0$ we conclude that $f''(a)(h,h) \geq 0$.

(b) Since $f'(a) = 0$, by Theorem 5.8 we have

$$f(a+h) = f(a) + \frac{1}{2} f''(a)(h,h) + o(\|h\|^2), \quad h \to 0. \tag{5.13}$$

Fix $c > 0$ such that

$$f''(a)(h,h) \geq 2c\,\|h\|^2$$

for all $h \in X$, then for $h \to 0$ (5.13) yields

$$f(a+h) - f(a) \geq (c + o(1))\,\|h\|^2 .$$

If $\|h\|$ is sufficiently small, but $h \neq 0$, then the right-hand side is positive, and therefore $f(a+h) > f(a)$. □

[3] If $\dim X < \infty$, then an equivalent definition is that $A(h,h) > 0$ for all non-zero vectors h.

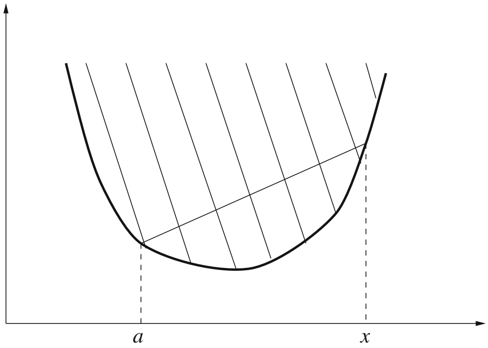


Fig. 5.1 A convex epigraph

Example (Peano) Given $p > q > 0$ we define a polynomial $f : \mathbb{R}^2 \to \mathbb{R}$ by the formula

$$f(x,y) := (y^2 - 2px)(y^2 - 2qx).$$

Then

$$f(0,0) = f'(0,0) = 0 \quad \text{and} \quad f''(0,0) \geq 0,$$

but f has no local minimum at $(0,0)$ because f takes both positive and negative values in each neighborhood of $(0,0)$. Nevertheless, the restriction of f to each line containing $(0,0)$ has a local minimum at $(0,0)$.

5.5 Convex functions

Definition A function $f : K \to \mathbb{R}$ is *convex* if K is a convex set in a vector space, and

$$f\left((1-t)x + ty\right) \leq (1-t)f(x) + tf(y)$$

for all $x, y \in K$ and $t \in [0,1]$.[4] Equivalently, f is convex if its *epigraph*

$$\operatorname{epi}(f) := \{(x,y) \in K \times \mathbb{R} \ : \ x \in K \text{ and } y \geq f(x)\}$$

is a convex set (see Figure 5.1).

[4] Since $x, y \in K$, it suffices to check this inequality for $x \neq y$ and $t \in (0,1)$.

Examples

- The constant functions defined on convex sets are convex.
- Every linear functional is convex.
- Every norm is convex.
- The sum of convex functions is convex.
- A positive multiple of a convex function is convex.
- If $g : K \to \mathbb{R}, f : \mathbb{R} \to \mathbb{R}$ are convex functions and f is non-decreasing, then $f \circ g$ is convex, because

$$f(g((1-t)x+ty)) \le f((1-t)g(x)+tg(y)) \le (1-t)f(g(x))+tf(g(y))$$

for all $x, y \in K$ and $t \in [0,1]$.

Proposition 5.13 (Jensen's inequality) *If $f : K \to \mathbb{R}$ is convex, then*

$$f(t_1x_1 + \cdots + t_nx_n) \le t_1 f(x_1) + \cdots + t_n f(x_n)$$

for all $x_1, \dots, x_n \in K$ and $t_1, \dots, t_n \in [0,1]$ with $t_1 + \cdots + t_n = 1$.

Proof This follows from the definition by induction on n. □

The following result simplifies the minimization of convex functions:

Proposition 5.14 *A local minimum of a convex function in a normed space is necessarily a global minimum.*

Proof Let $f : K \to \mathbb{R}$ be a convex function, $a \in K$, and U a neighborhood of a in K such that $f(a) \le f(y)$ for all $y \in U$. Given $x \in K$ arbitrarily, there exists a $t \in (0,1]$ (close to zero) such that

$$y := (1-t)a + tx = a + t(x-a) \in U.$$

Then

$$f(a) \le f(y) = f((1-t)a+tx) \le (1-t)f(a) + tf(x),$$

whence $f(a) \le f(x)$. □

There are two useful characterizations of *differentiable* convex functions:

Proposition 5.15 *Let $f : K \to \mathbb{R}$ be a* differentiable *function on a convex* open *set K of a normed space X. The following properties are equivalent (see Figure 5.2 for $K = \mathbb{R}$):*

$$f \text{ is convex}; \tag{a}$$

$$f(x) \ge f(a) + f'(a)(x-a) \text{ for all } x, a \in K; \tag{b}$$

$$f' \text{ is monotone, i.e., } (f'(x) - f'(a))(x-a) \ge 0 \text{ for all } x, a \in K. \tag{c}$$

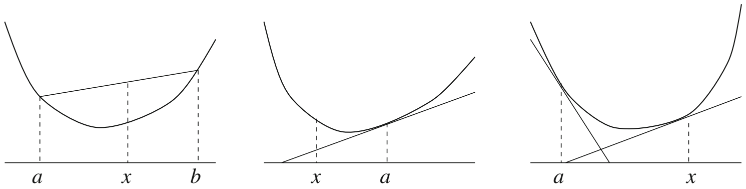


Fig. 5.2 Characterizations of convexity

Remark If $X = \mathbb{R}$, then (c) is equivalent to the monotonicity relation

$$a \le x \Longrightarrow f'(a) \le f'(x).$$

Proof

(a) $\Longrightarrow$ (b) We have

$$f(tx + (1-t)a) \le tf(x) + (1-t)f(a)$$

for all $t \in (0,1)$ by convexity, and hence

$$f(x) - f(a) \ge \frac{f(tx + (1-t)a) - f(a)}{t} = \frac{f(a + t(x-a)) - f(a)}{t}.$$

Letting $t \to 0$ we get (b).
(b) $\Longrightarrow$ (c) Exchanging the roles of a and x in (b) we have

$$f(a) \ge f(x) + f'(x)(a-x).$$

Summing the two inequalities we get (c).
(c) $\Longrightarrow$ (a) Let $x, y \in K$, $x \ne y$. Fix $0 < t < 1$ and consider the point $z = tx + (1-t)y$. Applying Proposition 4.6 (p. 107) there exist points $c \in (x,z)$ and $d \in (z,y)$ such that

$$\begin{aligned} A :&= \frac{tf(x) + (1-t)f(y) - f(z)}{t(1-t)} \\ &= \frac{f(x) - f(z)}{1-t} + \frac{f(y) - f(z)}{t} \\ &= \frac{f'(c)(x-z)}{1-t} + \frac{f'(d)(y-z)}{t} \\ &= \big(f'(c) - f'(d)\big)(x-y). \end{aligned}$$

Since $c \neq d$, we have $x - y = \alpha(c - d)$ with a suitable number $\alpha > 0$, and hence

$$A = \alpha\big(f'(c) - f'(d)\big)(c - d) \geq 0$$

by (c). □

We have seen that the local and global minima coincide for convex functions. Our next result further simplifies work with minima:

Corollary 5.16 *A differentiable convex function $f : K \to \mathbb{R}$ has a minimum at a if and only if $f'(a) = 0$.*

Proof The necessity of the condition $f'(a) = 0$ follows from Proposition 4.3 (p. 104). Conversely, if $f'(a) = 0$, then $f(x) \geq f(a)$ for all $x \in K$ by Proposition 5.15 (b). □

Proposition 5.17 *If $f : K \to \mathbb{R}$ is a* twice *differentiable function on a convex open set, then the properties (a), (b), (c) of Proposition 5.15 are also equivalent to the following:*

$$f''(a) \geq 0 \quad \textit{for all } a \in K. \tag{d}$$

Proof

(c) $\Longrightarrow$ (d) Fix $h \in X$ arbitrarily. If $t > 0$ is sufficiently small, then $a + th \in K$, and therefore

$$\big(f'(a + th) - f'(a)\big)(a + th - a) \geq 0$$

by (c). Dividing by t^2 we get

$$\frac{f'(a + th) - f'(a)}{t} h \geq 0,$$

and letting $t \to 0$ we obtain that $f''(a)(h, h) \geq 0$.
(d) $\Longrightarrow$ (b) Fix $a, x \in K$ arbitrarily. Applying Proposition 5.9 (p. 127) there exists a $b \in [a, x]$ such that

$$f(x) = f(a) + f'(a)(x - a) + \frac{1}{2} f''(b)(x - a, x - a).$$

Since the last term is nonnegative by assumption, (b) follows. □

In finite dimensions convexity implies continuity:

Proposition 5.18 *If U is a convex* open *set in a* finite-dimensional *normed space, then every convex function $f : U \to \mathbb{R}$ is* locally Lipschitz continuous:[5] *each point $a \in U$ has a neighborhood $V \subset U$ such that for some constant $L = L(V)$ we have the estimate*

$$\|f(x_1) - f(x_2)\| \le L\,\|x_1 - x_2\| \quad \textit{for all} \quad x_1, x_2 \in V$$

with some constant L depending only on V.

Proof By Theorem 3.9 (p. 78) we may assume that $X = (\mathbb{R}^n, \|\cdot\|_\infty)$. Fix a small $r > 0$ such that $\overline{B_{2r}(a)} \subset U$, denote by $b_1, \dots, b_{2^n}$ the vertices of the *cube* $\overline{B_{2r}(a)}$, and set

$$M := \max\{f(b_1), \dots, f(b_{2^n})\}.$$

First we prove that f is bounded in $\overline{B_{2r}(a)}$. Every point $x \in \overline{B_{2r}(a)}$ has a representation of the form

$$x = \sum_{i=1}^{2^n} t_i b_i, \quad t_i \ge 0, \quad \sum_{i=1}^{2^n} t_i = 1;$$

using Jensen's inequality this implies the upper estimate

$$f(x) \le \sum_{i=1}^{2^n} t_i f(b_i) \le M. \tag{5.14}$$

Since $2a - x \in \overline{B_{2r}(a)}$, applying the convexity inequality

$$f(a) \le \frac{f(x) + f(2a - x)}{2}$$

we obtain the lower estimate

$$f(x) \ge 2f(a) - f(2a - x) \ge 2f(a) - M. \tag{5.15}$$

If $x_1, x_2 \in B_r(a)$ are two distinct points, then (see Figure 5.3)

$$x_3 := x_2 + r\frac{x_2 - x_1}{\|x_2 - x_1\|} \in B_{2r}(a).$$

[5] We have already encountered this notion in Exercises 3.4 and 4.6, pp. 90, 116.

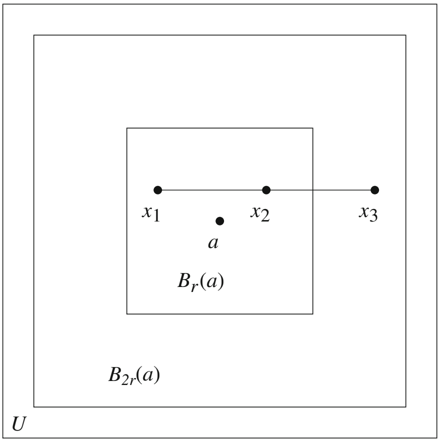


Fig. 5.3 Proof of Proposition 5.18

Then

$$rx_1 + \|x_2 - x_1\| x_3 = (\|x_2 - x_1\| + r)x_2,$$

whence

$$x_2 = tx_3 + (1-t)x_1 \quad \text{with} \quad t := \frac{\|x_2 - x_1\|}{\|x_2 - x_1\| + r} \in (0,1).$$

Since f is convex, we have

$$f(x_2) \le tf(x_3) + (1-t)f(x_1);$$

hence, using (5.14) and (5.15), we obtain the inequality

$$f(x_2) - f(x_1) \le t\big(f(x_3) - f(x_1)\big) \le \frac{2M - 2f(a)}{r} \|x_2 - x_1\| .$$

Since the roles of x_1 and x_2 are symmetric, we conclude that f is Lipschitz continuous on $B_r(a)$ with the constant $L := (2M - 2f(a))/r$. □

5.6 The functions $\mathbb{R}^m \hookrightarrow \mathbb{R}^n$

We generalize the results of Section 4.3 (p. 111).

Proposition 5.19 *Let* $f = (f_1, \dots, f_n) : \mathbb{R}^m \hookrightarrow \mathbb{R}^n$ *and* $k \ge 1$.

(a) *f is k times differentiable at a if and only if $f_1, \dots, f_n$ are k times differentiable at a.*
(b) *f is k times differentiable if and only if $f_1, \dots, f_n$ are k times differentiable.*
(c) *$f \in C^k$ if and only if $f_1, \dots, f_n \in C^k$.*

Proof We again use the continuous linear maps P_j and B_j introduced in the proof of Proposition 4.10 (p. 111). They belong to the class C^∞ by the remark on p. 121.

(a) If f is k times differentiable at a, then the functions $f_j = P_j \circ f$ are also k times differentiable at a by Proposition 5.5. Conversely, if $f_1, \dots, f_n$ are k times differentiable at a, then $f = B_1 \circ f_1 + \cdots + B_n \circ f_n$ is also k times differentiable at a by the same proposition.

The proofs of (b) and (c) are analogous. □

Now we generalize Proposition 4.12 (p. 113):

Proposition 5.20 *A function $f : \mathbb{R}^m \hookrightarrow \mathbb{R}$ is k times continuously differentiable $(k \ge 1)$ if and only if all partial derivatives $D_1 f, \dots, D_m$ exist and are $k-1$ times continuously differentiable at $D(f)$.*

Proof The case $k = 1$ is contained in Proposition 4.12. Assume henceforth that $k \ge 2$.

If $f \in C^k$, then by Proposition 4.2 (p. 101) the partial derivatives $D_i f$ exist, and they are given by the formula $D_i f = L_i \circ f'$ where L_i denotes the continuous linear map $L_i A := Ae_i$. As compositions of C^{k-1} functions, the functions $D_i f$ also belong to C^{k-1} by Proposition 5.5.

Conversely, if the partial derivatives $D_i f : D(f) \to \mathbb{R}$ exist and belong to C^{k-1}, then f belongs to C^1 by Proposition 4.12, and

$$f'(a)h = \sum_{i=1}^{m} D_i f(a) h_i \quad \text{or} \quad f'(a)h = \nabla f(a) \cdot h$$

for all $a \in D(f)$ and $h \in \mathbb{R}^m$ by Proposition 4.11 (p. 113).

Since the components $D_i f$ of ∇f belong to C^{k-1}, applying the preceding proposition we conclude that $f \in C^k$. □

Finally, we discuss the matrix representation of the second derivative:

Proposition 5.21 (Hesse) *If $f : \mathbb{R}^m \hookrightarrow \mathbb{R}$ is twice differentiable at a, then*

$$f''(a)(h,k) = (h_1, \dots, h_m) \begin{pmatrix} D_1 D_1 f(a) & \cdots & D_1 D_m f(a) \\ \vdots & & \vdots \\ D_m D_1 f(a) & \cdots & D_m D_m f(a) \end{pmatrix} \begin{pmatrix} k_1 \\ \vdots \\ k_m \end{pmatrix}$$

for all $h, k \in \mathbb{R}^m$.

Proof For any fixed $k \in \mathbb{R}^m$ we have

$$\varphi(x) := f'(x)k = \sum_{j=1}^{m} k_j D_j f(x)$$

by Proposition 4.11. Applying Lemma 5.7 (p. 125) and using Proposition 4.11 again the required equality follows:

$$f''(a)(h,k) = \varphi'(a)h = \sum_{j=1}^{m} k_j \sum_{i=1}^{m} D_i D_j f(x) h_i = \sum_{j=1}^{m} \sum_{i=1}^{m} D_i D_j f(a) h_i k_j.$$

□

Remark By Theorem 5.6 (p. 123) the *Hessian* matrix is symmetric, and hence diagonalizable.[6] It follows that $f''(a)$ is positive definite (resp. positive semidefinite) if and only if all eigenvalues of its Hessian are positive (resp. nonnegative).

5.7 Exercises

Exercise 5.1 Every continuous m-linear map belongs to C^∞ and its kth derivative vanishes for all $k > m$.

Exercise 5.2 Compute the first and second derivatives of the Euclidean norm in $\mathbb{R}^n$.

Exercise 5.3 Find the local and global extrema of

$$f(x,y) := x^4 + y^2 \quad \text{and} \quad g(x,y) := x^3 + y^2.$$

Exercise 5.4 (C^∞ functions of compact support)

(i) Prove that the formula

$$h(t) := \begin{cases} e^{-1/t} & \text{if } t > 0, \\ 0 & \text{if } t \le 0 \end{cases}$$

defines a function $h : \mathbb{R} \to \mathbb{R}$ of class C^∞, all of whose derivatives vanish at 0.

[6] We recall the proof in Section 7.3 below (p. 174).

(ii) Prove that the formula[7]

$$f(x) := \begin{cases} e^{-1/(1-\|x\|^2)}) & \text{if } \|x\| < 1, \\ 0 & \text{if } \|x\| \geq 1 \end{cases}$$

defines a function $f : \mathbb{R}^n \to \mathbb{R}$ of class C^∞, strictly positive in the unit ball, and vanishing outside this ball.

Exercise 5.5 (Laplace equation) Given a twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, we denote by $\Delta f(x_1, \dots, x_n)$ the trace of the Hessian matrix $f''(x_1, \dots, x_n)$, i.e.,

$$\Delta f := \sum_{i=1}^{n} D_i^2 f;$$

Δ is called the *Laplacian operator*.[8]

(i) Prove that the function $f(x) := \ln |x|$ satisfies the *Laplace equation* $\Delta f = 0$ in $\mathbb{R}^2 \setminus \{0\}$.
(ii) Prove that the function $f(x) := |x|^{-1}$ satisfies the *Laplace equation* $\Delta f = 0$ in $\mathbb{R}^n \setminus \{0\}$ for $n = 3, 4, \dots$.

Exercise 5.6 (Heat equation) Prove that for any non-zero real number a, the formula

$$f(t, x) := \frac{1}{2a\sqrt{\pi t}} e^{-\frac{x^2}{4a^2 t}}, \quad t > 0, x \in \mathbb{R}$$

defines a solution of Fourier's *heat equation*[9]

$$D_1 f = a^2 D_2^2 f.$$

Exercise 5.7 (Wave equation) Prove that if $f, g : \mathbb{R} \to \mathbb{R}$ are two arbitrary functions of class C^2 and a is a positive constant, then *d'Alembert's formula*

$$u(t, x) := f(x + at) + g(x - at)$$

defines a solution of the one-dimensional wave equation

$$D_1^2 u - a^2 D_2^2 u = 0.$$

[7] We consider the usual Euclidean norm.

[8] See also Section 12.3 below, p. 290.

[9] See Section 12.5 (p. 294) for a derivation of this equation.

Chapter 6
Ordinary Differential Equations

This chapter is an introduction to ordinary differential equations.[1] The letters X, Y, Z will always denote *Banach* spaces, i.e., *complete* normed spaces.

On the first reading Sect. 6.1 may be skipped, and the reader may assume that $X = \mathbb{R}$: this simple special case already suffices for many important applications.

6.1 Integrals of Vector-Valued Functions

Let $[a, b] \subset \mathbb{R}$ be a non-degenerate compact interval. We have seen in Sects. 3.1 and 3.2 (pp. 65 and 73) that the bounded functions $f : [a, b] \to Y$ form a Banach space $\mathcal{B}([a, b], Y)$ for the norm

$$\|f\|_\infty := \sup_{t \in [a,b]} \|f(t)\|$$

of uniform convergence. We could adapt the Riemann integral to vector-valued functions, but a simpler notion will suffice for our purposes.[2]

Definition $f : [a, b] \to Y$ is a *step function* if there exist points $a = t_0 < t_1 < \cdots < t_n = b$ and vectors $y_1, \ldots, y_n \in Y$ such that

$$f(t) = y_i \quad \text{for all} \quad t_{i-1} < t < t_i, \quad i = 1, \ldots, n.$$

[1] The books of Arnold [16, 17], Burkill [70], Coddington–Levinson [112], Hartman [224], Ince [252], Pontryagin [398] contain many more results.

[2] In Exercise 3.6 (p. 90) a slightly less general notion was introduced by the same method.

V. Komornik, *Topology, Calculus and Approximation*, Springer
Undergraduate Mathematics Series, DOI 10.1007/978-1-4471-7316-8_6

The step functions form a linear subspace $\mathcal{E}$ of $\mathcal{B}([a,b],Y)$, and the formula

$$I(f) := \sum_{i=1}^{n}(t_i - t_{i-1})y_i$$

defines a linear map $I : \mathcal{E} \to Y$. The obvious estimate

$$\|I(f)\| \le \sum_{i=1}^{n}(t_i - t_{i-1})\,\|y_i\| \le (b-a)\,\|f\|_\infty$$

shows that I is continuous.

Definition $f \in \mathcal{B}([a,b],Y)$ is *integrable* if there exists a sequence (f_n) of step functions converging to f uniformly on $[a,b]$. Then the *integral* of f is defined by the formula

$$\int_a^b f(t)\,dt := \lim I(f_n).$$

Remark The case when $Y = \mathbb{R}^m$ is particularly simple: $f : [a,b] \to \mathbb{R}^m$ is integrable if and only if its components are integrable, and then

$$\int_a^b f(t)\,dt = \Big(\int_a^b f_1(t)\,dt, \dots, \int_a^b f_m(t)\,dt\Big).$$

Proposition 6.1

(a) *The integrable functions $f : [a,b] \to Y$ form a Banach space $\overline{\mathcal{E}}$, and the integral is a continuous linear map of $\overline{\mathcal{E}}$ into Y.*

(b) *Every continuous function $f : [a,b] \to Y$ is integrable.*

(c) *If $f : [a,b] \to Y$ is integrable, then $\|f\| : [a,b] \to \mathbb{R}$ is also integrable, and*

$$\Big\|\int_a^b f(t)\,dt\Big\| \le \int_a^b \|f(t)\|\;dt \le (b-a)\,\|f\|_\infty\,. \tag{6.1}$$

(d) *If $f : [a,b] \to Y$ is integrable and $A \in L(Y,Z)$, then $A \circ f : [a,b] \to Z$ is integrable, and*

$$\int_a^b Af(t)\,dt = A\int_a^b f(t)\,dt.$$

(e) *Let $a < c < b$. A function $f : [a,b] \to Y$ is integrable if and only if its restrictions $f|_{[a,c]}$ and $f|_{[c,b]}$ are integrable. Then*

$$\int_a^b f(t)\,dt = \int_a^c f(t)\,dt + \int_c^b f(t)\,dt.$$

Proof (a) By definition the set of integrable functions is the closure of $\mathcal{E}$ in $\mathcal{B}([a,b],Y)$. The existence and uniqueness of the integral is a special case of Proposition 3.18 (p. 85).

(b) For each positive integer n we define $f_n(a)=f(a)$, $h=(b-a)/n$ and

$$f_n(t)=f(a+kh) \quad \text{if} \quad a+(k-1)h<t\le a+kh, \quad k=1,\dots,n.$$

Then f_n is a step function, and $\|f-f_n\|_\infty \to 0$ by Heine's theorem (p. 27).

(c), (d) and (e) The results are elementary for step functions. Hence the general case follows by continuity because if a sequence (f_n) of step functions converges uniformly to f, then $(\|f_n\|)$, (Af_n), $(f_n\chi_{[a,c]})$[3] and $(f_n\chi_{[c,b]})$ are sequences of step functions, converging uniformly to $\|f\|$, Af, $f\chi_{[a,c]}$ and $f\chi_{[c,b]}$, respectively. □

6.2 Definitions and Examples

Let X be a Banach space, $D\subset\mathbb{R}\times X$ be an open set, $(\tau,\xi)\in D$, and $f:D\to X$ a continuous function. We consider the *initial value problem*

$$x'=f(t,x), \quad x(\tau)=\xi \tag{6.2}$$

consisting of the *differential equation* $x'=f(t,x)$ and the *initial condition* $x(\tau)=\xi$.

Definition By a *solution* of (6.2) we mean a differentiable function $x:I\to\mathbb{R}^N$, defined on some open interval $I\subset\mathbb{R}$, and satisfying the following conditions:

$$\begin{aligned}&(t,x(t))\in D \quad \text{and} \quad x'(t)=f(t,x(t)) \quad \text{for all} \quad t\in I,\\ &\tau\in I \quad \text{and} \quad x(\tau)=\xi.\end{aligned}$$

Geometrically the solution is a function whose graph is a curve in D, passing through the point (τ,ξ). If we draw a small segment of slope $f(t,x)$ at each point $(t,x)\in D$ (see Fig. 6.1), then the graph is tangent to the corresponding segment at each point.

Remark Since f is continuous, the right-hand side of (6.2) is continuous for every solution x, and hence $x\in C^1$. It follows by induction[4] that if $f\in C^k$ for some $k=1,2,\dots$, then the solutions belong to C^{k+1}.

[3] We denote by χ_A the *characteristic function* of a set A, i.e., $\chi_A(x)=1$ if $x\in A$, and $\chi_A(x)=0$ otherwise.

[4] This is a so-called "bootstrap argument".

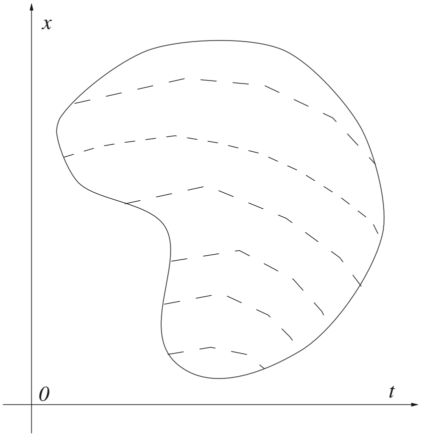


Fig. 6.1 Geometric interpretation of the solution

Examples In the first five examples we set $X = \mathbb{R}$, $D = \mathbb{R}^2$ and $\tau = 0$.

(i) If f is continuous and $f(t, x)$ does not depend on x, then our problem reduces to that of finding a primitive function: the solution is given by the formula[5]

$$x(t) = \xi + \int_0^t f(s)\, ds, \quad t \in \mathbb{R}.$$

(ii) If $f(t, x) = x$ the formula

$$x(t) = \xi e^t, \quad t \in \mathbb{R}$$

gives an obvious solution; see Fig. 6.2.[6]

(iii) For $f(t, x) = x^2$ the formula

$$x(t) = \begin{cases} \xi/(1-\xi t), & t \in (-\infty, 1/\xi) \quad \text{if} \quad \xi > 0; \\ 0, & t \in \mathbb{R} \quad \text{if} \quad \xi = 0; \\ \xi/(1-\xi t), & t \in (1/\xi, \infty) \quad \text{if} \quad \xi < 0 \end{cases}$$

gives a solution; see Fig. 6.3.

(iv) (Peano) If $f(t, x) = 3x^{2/3}$, then the formula

$$x(t) := (\xi^{1/3} + t)^3, \quad t \in \mathbb{R}$$

[5] When $f(t, x)$ does not depend on x, then changing the initial value ξ the graph of the solution is translated "vertically".

[6] When $f(t, x)$ does not depend on t, then changing the initial value ξ the graph of the solution is translated "horizontally".

Fig. 6.2 Example (ii)

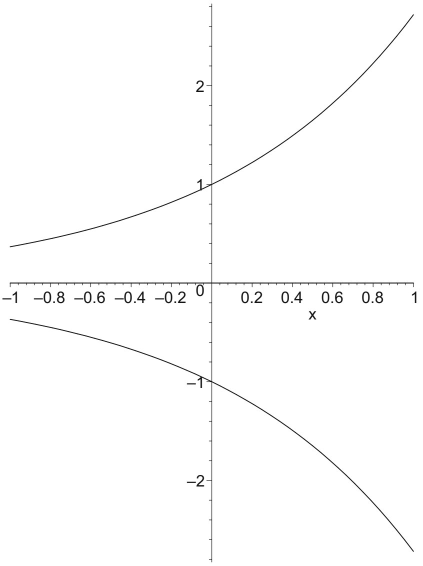
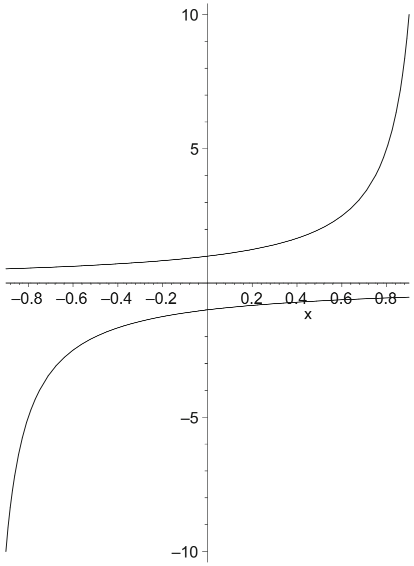


Fig. 6.3 Example (iii)



gives a solution. In particular, for $\xi = 0$ the function $x(t) = t^3$ is a solution; see Fig. 6.4. But the constant function $x(t) \equiv 0$ is also a solution!

(v) It is hopeless to try to find explicit solutions to differential equations like the following:

$$x' = \sin \sin e^{tx^3}, \quad x(e) = \pi.$$

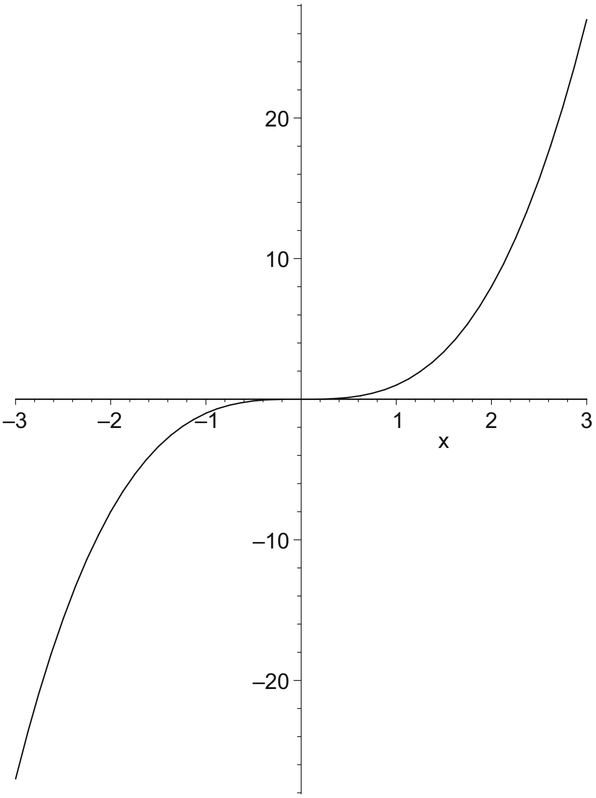


Fig. 6.4 Example (iv)

However, it will follow from the general results of the next section that it has a unique solution $x : \mathbb{R} \to \mathbb{R}$, and we will show in Chap. 12 that this solution may be approximated with arbitrary precision.

(vi) Finally, we consider an example with $X = \mathbb{R}^2$. Let

$$f : \mathbb{R}^3 \to \mathbb{R}^2, \quad f(t, x_1, x_2) = (-x_2, x_1), \quad \tau = 0 \quad \text{and} \quad \xi = (0, 1).$$

In other words, we consider the *system* of differential equations

$$x_1' = -x_2 \quad \text{and} \quad x_2' = x_1$$

with the initial conditions

$$x_1(0) = 1 \quad \text{and} \quad x_2(0) = 0.$$

One may readily check that a solution is given by the formula

$$x(t) := (\cos t, \sin t), \quad t \in \mathbb{R}.$$

Its graph is a spiral in $\mathbb{R}^3$.

If $x : I \to X$ is a solution of (6.2), then infinitely many other solutions are obtained by restricting x to open subintervals of I containing τ. This suggests the following notion:

Definition A solution of (6.2) is *maximal* if it cannot be extended to a solution defined on a larger interval.

For example, the solutions given in the above Examples (i), (ii), (iv), (vi) are maximal because they are defined on the whole real line. The solutions of Example (iii) are also maximal because for $\xi \neq 0$ they have no finite limits at $1/\xi$, and hence they have no continuous extensions to this point.

In Example (iv) we found two different maximal solutions. The reason behind this curious phenomenon will be revealed in the next section.

Remark We could also investigate differential equations of higher order:

$$x^{(n)} = f(t, x, x', \dots, x^{(n-1)}), \quad x^{(i)}(\tau) = \xi_i, \ i = 0, \dots, n-1,$$

where $n > 1$, $D \subset \mathbb{R} \times X^n$ is an open set, $f : D \to X$ is a continuous function, and $(\tau, \xi_0, \dots, \xi_{n-1}) \in D$. By definition a solution is an n times differentiable function $x : I \to X$, satisfying

$$(t, x(t), x'(t), \dots, x^{(n-1)}(t)) \in D$$

and

$$x^{(n)}(t) = f(t, x(t), x'(t), \dots, x^{(n-1)}(t))$$

for all $t \in I$,

$$\tau \in I, \quad \text{and} \quad x^{(i)}(\tau) = \xi_i, \ i = 0, \dots, n-1.$$

This problem may be reduced to the earlier one as follows. Set $\xi = (\xi_0, \dots, \xi_{n-1})$, and introduce a function $F : D \to X^n$ by the formula

$$F(t, y_0, \dots, y_{n-1}) := (y_1, \dots, y_{n-1}, f(t, y_0, \dots, y_{n-1})).$$

Then D is an open set in $\mathbb{R} \times X^n$, and $F : D \to X^n$ is continuous.

If x is a solution of the n-order problem, then $y := (x, x', \dots, x^{(n-1)})$ is a solution of the problem

$$y' = F(t, y), \quad y(\tau) = \xi.$$

Conversely, if y is a solution of the latter problem, then $x := y_0$ solves the n-order problem.

Example Applying this reduction to the second-order problem

$$x'' = -x, \quad x(\tau) = \xi_0, \quad x'(\tau) = \xi_1$$

we obtain Example (vi) above.

6.3 The Cauchy–Lipschitz Theorem

We consider the problem

$$x' = f(t, x), \quad x(\tau) = \xi, \tag{6.2}$$

where X is a Banach space, $D \subset \mathbb{R} \times X$ is an open set, $f : D \to X$ is a continuous function and $(\tau, \xi) \in D$. Let us generalize the partial derivatives[7]:

Definition Let $(t_0, x_0) \in D$. If the function

$$X \ni x \mapsto f(t_0, \cdot) \in X$$

has a (total) derivative at x_0, then it is called the *second partial derivative* of f at (t_0, x_0), and is denoted by $D_2f(t_0, x_0)$. This defines a function $D_2f : D \hookrightarrow L(X, X)$.

The following theorem is of the highest importance:

Theorem 6.2 (Cauchy–Lipschitz) *If f and D_2f exist and are continuous in D, then the problem (6.2) has a unique maximal solution.*

Remarks

- The assumptions of the theorem are satisfied in all examples of the preceding section, except Example (iv). In Examples (ii), (iii), (v), (vi) the function f even belongs to C^∞.
- Peano proved that if X is *finite-dimensional*, then the mere *continuity* of f already ensures the *existence* of at least one maximal solution. Example (iv) of the

[7] For $X = \mathbb{R}$ the following definition reduces to the earlier definition of D_2f.

preceding section shows that the maximal solution is not necessarily unique in this case.

See p. 343 for further comments on Peano's theorem.

- Instead of using the partial derivative D_2f, Lipschitz assumed that each point $(t,x) \in D$ has a small neighborhood V such that all points $(t_1,x_1),(t_1,x_2) \in V$ satisfy the estimate

$$\|f(t_1,x_1)-f(t_1,x_2)\| \le L\,\|x_1-x_2\| \tag{6.3}$$

with some constant L depending only on V.[8]

It follows from the mean value theorem 4.7 (p. 107) that Cauchy's assumption implies that of Lipschitz with any $L > \|D_2f(t,x)\|$ in a sufficiently small ball V centered at any given $(t,x) \in D$. Our proof below will use this (somewhat weaker) condition (6.3).

The first proofs of the theorem were based on Euler's piecewise linear approximating solutions.[9] Here we apply the method of successive approximations.

We introduce an equivalent *integral equation*:

Lemma 6.3 *Given an open interval I containing τ, a function $x : I \to X$ is a solution of* (6.2) *if and only if it is* continuous, *and*

$$x(t) = \xi + \int_\tau^t f(s,x(s))\,ds \quad \textit{for all} \quad t \in I. \tag{6.4}$$

Proof Every solution $x : I \to X$ of (6.2) belongs to C^1, and therefore

$$x(t) = \xi + \int_\tau^t x'(s)\,ds = \xi + \int_\tau^t f(s,x(s))\,ds$$

for all $t \in I$ by the Newton–Leibniz formula.

Conversely, if a *continuous* function $x : I \to X$ satisfies (6.4), then $x(\tau) = \xi$. Furthermore, since the right-hand side of (6.4) is a primitive of a continuous function, $x \in C^1$; differentiating (6.4) we obtain the differential equation in (6.2). □

The equation (6.4) suggests a natural recursive approximation of the solution by defining $x_0(t) := \xi$ and

$$x_{n+1}(t) := \xi + \int_0^t f(s,x_n(s))\,ds, \quad n = 0,1,\ldots.$$

[8] Lipschitz was not aware of Cauchy's unpublished work, see p. 343. The condition (6.3) is weaker than the local Lipschitz property of Proposition 5.18 (p. 136).

[9] We will also use Euler's method in Chap. 12.

Example Applying this method to the problem

$$x' = x, \qquad x(0) = 1$$

we get the recursive relations $x_0(t) := 1$ and

$$x_{n+1}(t) := 1 + \int_0^t x_n(s)\,ds, \quad n = 0, 1, \ldots.$$

We obtain by induction that

$$x_n(t) = \sum_{k=0}^{n} \frac{t^k}{k!},$$

and

$$x(t) := \lim x_n(t) = e^t$$

is indeed a solution.

Remarks

- In the above example the sequence (x_n) converges on the whole real line. In general the convergence holds only in some small neighborhood of τ.
- The following proof of the theorem shows that the integral equation (6.4) may also be used to obtain good approximating solutions.

Proof of Theorem 6.2 First step: existence of a "cylinder of security". Fix $M > \|f(\tau, \xi)\|$, $L > \|D_2 f(\tau, \xi)\|$ and then a small $r > 0$ such that

$$|t - \tau| \le r \quad \text{and} \quad \|x - \xi\| \le Mr \Longrightarrow$$
$$(t, x) \in D, \quad \|f(t, x)\| \le M \quad \text{and} \quad \|D_2 f(t, x)\| \le L.$$

It follows that f satisfies the Lipschitz condition (6.3) in this cylinder.[10]

We claim that each solution $x : I \to X$ of (6.4) satisfies the inequality

$$\|x(t) - \xi\| \le Mr \quad \text{for all} \quad t \in I \cap [\tau - r, \tau + r].$$

See Fig. 6.5 for $X = \mathbb{R}$: the graph of the solution stays in a "cylinder of security".

Assume on the contrary that $\|x(t) - \xi\| > Mr$ for some $t \in I \cap [\tau - r, \tau + r]$. We may assume by symmetry that $t \ge \tau$. Since $x(\tau) = \xi$, we have $t > \tau$, and by

[10] The condition (6.3) will be used only in Step 2. If the local Lipschitz condition is assumed instead of the continuity of $D_2 f$ in the theorem, then (6.3) is satisfied for sufficiently small r.

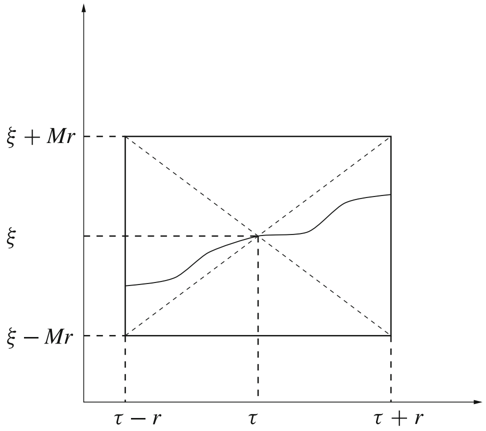


Fig. 6.5 Rectangle and cone of security

continuity there exists a $t' \in (\tau, t)$ such that $\|x(t') - \xi\| = Mr$, and $\|x(s) - \xi\| < Mr$ for all $s \in [\tau, t')$. Then

$$\left\|x(t') - \xi\right\| \le \int_\tau^{t'} \|f(s, x(s))\| \; ds \le M(t' - \tau) < Mr,$$

contradicting the choice of t'.

Remark Applying the above estimate for the endpoints of $[\tau - r, \tau + r]$ we obtain that $\|x(\tau \pm r) - \xi\| \le Mr$. Since we may repeat the proof for any $r' \in (0, r)$ instead of r, we have the stronger conclusion

$$\|x(t) - \xi\| \le M\,|t - \tau| \quad \text{for all} \quad t \in I \cap [\tau - r, \tau + r].$$

This means that the graph of the solution stays in a smaller "cone of security".[11]

Second step: local existence and uniqueness. We prove that (6.4) has a unique solution defined on $J = [\tau - r, \tau + r]$. Since every solution $x : J \to X$ belongs to the closed ball

$$\overline{B_{Mr}(\xi)} := \{y \in C_b(J; X) \; : \; \|y(t) - \xi\| \le Mr \quad \text{for all} \quad t \in J\}$$

by the preceding step, it is sufficient to prove that the map F defined by

$$(Fy)(t) := \xi + \int_\tau^t f(s, y(s)) \; ds, \quad y \in \overline{B_{Mr}(\xi)}, \; t \in J$$

has a unique fixed point.

[11] We do not need this stronger property in the sequel.

If $y \in \overline{B_{Mr}(\xi)}$, then using the choice of r and J we have

$$\|(Fy)(t) - \xi\| = \left\| \int_\tau^t f(s, y(s))\, ds \right\| \le M\,|t - \tau| \le Mr$$

for all $t \in J$. Hence F is a well-defined map of $\overline{B_{Mr}(\xi)}$ into itself.

Next we observe that the formula[12]

$$\|y\|_L := \sup_{t \in J} \|y(t)\|\, e^{-L|t-\tau|}$$

defines an equivalent norm on $C_b(J; X)$ because

$$\|y\|_L \le \|y\|_\infty \le e^{Lr}\, \|y\|_L$$

for all y. Since, as a closed subset of the Banach space $C_b(J; X)$, $\overline{B_{Mr}(\xi)}$ is a complete metric space, it remains to show that F is a contraction for the new metric of $\overline{B_{Mr}(\xi)}$.

If $x, y \in \overline{B_{Mr}(\xi)}$ and $t \in J$, then using (6.3) we have

$$\begin{aligned} \|(Fx)(t) - (Fy)(t)\| &= \left\| \int_\tau^t \Big(f(s, x(s)) - f(s, y(s)) \Big)\, ds \right\| \\ &\le \left| \int_\tau^t L\, \|x(s) - y(s)\|\ ds \right|, \end{aligned}$$

and hence

$$\begin{aligned} \|(Fx)(t) - (Fy)(t)\|\, e^{-L|t-\tau|} &\le \left| \int_\tau^t L\, \|x(s) - y(s)\|\, e^{-L|t-\tau|}\, ds \right| \\ &= \left| \int_\tau^t L e^{-L|t-s|}\, \|x(s) - y(s)\|\, e^{-L|s-\tau|}\, ds \right| \\ &\le \left| \int_\tau^t L e^{-L|t-s|}\, \|x - y\|_L\ ds \right| \\ &= \Big(1 - e^{-L|t-\tau|}\Big)\, \|x - y\|_L \\ &\le \big(1 - e^{-Lr}\big)\, \|x - y\|_L\,. \end{aligned}$$

Since this holds for all t, we have

$$\|Fx - Fy\|_L \le \big(1 - e^{-Lr}\big)\, \|x - y\|_L$$

for all $x, y \in \overline{B_{Mr}(\xi)}$, i.e., F is indeed a contraction.

[12] Observe that the choice of L was not needed before.

Third step: existence and uniqueness of a maximal solution. We consider the set $\{x_\alpha : I_\alpha \to X\}$ of all solutions of (6.4). By the local existence this set is non-empty.

We claim that any two solutions are equal on the intersection of their domains. Assume on the contrary that there exist two solutions $x_\alpha : I_\alpha \to X$ and $x_\beta : I_\beta \to X$ such that $x_\alpha(t') \neq x_\beta(t')$ in some $t' \in I_\alpha \cap I_\beta$. Assume by symmetry that $t' > \tau$, and let us denote by τ' the greatest lower bound of these values t'. Then $\xi' := x_\alpha(\tau') = x_\beta(\tau')$: this is obvious if $\tau' = \tau$, while for $\tau' > \tau$ this follows by continuity from the equality $x_\alpha = x_\beta$ on $[\tau, \tau')$.

It follows that the problem

$$x' = f(t,x), \quad x(\tau') = \xi'$$

has at least two solutions in every arbitrarily small neighborhood $(\tau' - r, \tau' + r)$ of τ'. This contradicts the result of the preceding step.[13]

Now we introduce the *open interval*[14] $I := \cup_\alpha I_\alpha$, and for each $t \in I$ we define $x(t) := x_\alpha(t)$ where α an arbitrary index satisfying $t \in I_\alpha$. By the just proven uniqueness property the definition does not depend on the particular choice of α, and hence $x : I \to X$ is a solution of (6.4) that is an extension of every other solution. □

6.4 Additional Results and Linear Equations

The conclusion of the Cauchy–Lipschitz theorem may be strengthened, and its proof simplified under a somewhat stronger assumption. Let $f : I \times X \to X$ be a continuous function, where I is an arbitrary non-degenerate interval and X a Banach space. Assume that there exists a continuous[15] function $L : I \to \mathbb{R}$ such that

$$\|f(t,x) - f(t,y)\| \le L(t)\, \|x - y\| \quad \text{for all} \quad t \in I \quad \text{and} \quad x, y \in X.$$

We denote by $C(I,X)$ the *vector space* of continuous functions $x : I \to X$.[16]

Proposition 6.4 *For any given $\tau \in I$ and $h \in C(I,X)$ the integral equation*

$$x(t) = h(t) + \int_\tau^t f(s, x(s))\, ds, \quad t \in I$$

has a unique solution $x \in C(I;X)$.

[13] The reasoning of the preceding step is valid for any other point $(\tau', \xi') \in D$ instead of (τ, ξ).

[14] This is an interval because $\tau \in \cap I_\alpha$.

[15] The local boundedness is already sufficient.

[16] If I is compact, then this reduces to the former notation.

Proof It is sufficient to prove that for each compact subinterval $K \subset I$ satisfying $\tau \in K$ the map $F : C(K,X) \to C(K,X)$ defined by the formula

$$(Fy)(t) = h(t) + \int_{\tau}^{t} f(s, y(s))\, ds, \quad t \in K$$

has a unique fixed point.

This is obtained by repeating the second step of the proof of the Cauchy–Lipschitz theorem with $C(K,X)$ instead of $\overline{B_{Mr}(\xi)}$, and with the Lipschitz constant $L := \max_{t\in K} L(t)$.[17] □

The assumptions of the proposition are satisfied for the *linear* problems

$$x' = A(t)x + c(t), \quad x(\tau) = \xi, \tag{6.5}$$

where $A : I \to L(X,X)$ and $c : I \to X$ are given continuous functions on an open interval $I = (\alpha, \beta)$. This is a special case of (6.2) with

$$D := I \times X \quad \text{and} \quad f(t,x) = A(t)x + c(t).$$

Indeed, since $D_2 f(t,x) = A(t), f$ and $D_2 f$ are continuous.

An important feature of linear equations[18] is the following:

Corollary 6.5 *The maximal solution of the linear problem* (6.5) *is defined on the whole interval I.*

Proof We apply the preceding proposition with $L(t) := \|A(t)\|$. □

In the rest of this section we return to the general assumptions of the Cauchy–Lipschitz theorem.

Proposition 6.6 *Let* $x : (a,b) \to X$ *be the maximal solution of* (6.2) *and* $K \subset D$. *If* K *is compact, then there exist two points* t_1, t_2 *such that* $a < t_1 < \tau < t_2 < b$ *and* $(t_i, x(t_i)) \notin K$, $i = 1, 2$.

Remark

- If $\dim X < \infty$, then the intuitive meaning of the proposition is that the maximal solutions are defined up to the boundary of D.

  In particular, if $D = \mathbb{R} \times X$ and $b < \infty$ for the maximal solution $x : (a,b) \to X$ of (6.2), then $\|x(t)\|$ is unbounded as $t \to b$. (A similar property holds if $a > -\infty$.)

  The last property does not hold in *any* infinite-dimensional Banach space, even if $f(t,x)$ is independent of t: see the comments on p. 343.

[17] The first part of that proof is obvious here: F maps $C(K,X)$ into itself.

[18] Linear differential equations are studied in detail in Pontryagin [398], Burkill [70], Coddington [111], and Walter [505].

- The proof will show that for the existence of t_2 (resp. t_1) it is sufficient to assume instead of the compactness that K is closed, f is bounded on K, and $b < \infty$ (resp. $a > -\infty$).

Proof of Proposition 6.6 If, for example, $(t, x(t)) \in K$ for all $t \in [\tau, b)$, then $b < \infty$ by the compactness of K.

We claim that the limit $x(b-0) \in X$ exists. Indeed, if a sequence $(t_n) \subset (\tau, b)$ converges to b, then setting $M := \sup_K \|f\|$ we have

$$\|x(t_m) - x(t_n)\| = \left\| \int_{t_n}^{t_m} f(s, x(s))\, ds \right\| \le M|t_m - t_n| \to 0$$

as $m, n \to \infty$, so that $(x(t_n))$ is a Cauchy sequence. Since X is complete, $x(t_n)$ converges to some point $x_b \in X$. As in the proof of Proposition 1.14 (p. 20), the limit x_b does not depend on the particular choice of the sequence (t_n). Therefore $\lim_{b-0} x = x_b$.

Since $(b, x_b) \in K \subset D$, applying Theorem 6.2 in (b, x_b) we obtain an extension of the solution x beyond b, contradicting the maximality of the solution. □

Next we prove that the solutions depend continuously on the initial data. This is important for applications because the initial data usually come from measurements, and hence they are only approximately known.

***Proposition 6.7 (Niccoletti)** *Assume that the conditions of Theorem 6.2 are satisfied. Consider a solution $x : J \to X$ of (6.2) and fix a compact subinterval $[c, d] \subset J$ such that $c < \tau < d$. For each $\varepsilon > 0$ there exists a $\delta > 0$ such that if $\|\xi - \eta\| < \delta$, then the maximal solution of the problem*

$$y' = f(t, y), \quad y(\tau) = \eta$$

is defined on an interval (a, b) containing $[c, d]$ as a subset, and

$$\|x(t) - y(t)\| < \varepsilon \quad \textit{for all} \quad t \in [c, d].$$

For the proof we need a lemma:

Lemma 6.8 (Gronwall) *Let $\varphi : [\tau, b) \to \mathbb{R}$ be a continuous, nonnegative function. If there exist two constants $C, L > 0$ such that*

$$\varphi(t) \le C + L \int_\tau^t \varphi(s)\, ds \quad \textit{for all} \quad \tau \in [\tau, b)$$

then

$$\varphi(t) \le Ce^{L(t-\tau)} \quad \textit{for all} \quad \tau \in [\tau, b).$$

Proof Integrating the inequality

$$\frac{d}{dt}\Big(e^{-Lt}\int_\tau^t \varphi(s)\,ds\Big) = e^{-Lt}\Big(\varphi(t) - L\int_\tau^t \varphi(s)\,ds\Big) \le Ce^{-Lt}$$

between τ and t we obtain that

$$e^{-Lt}\int_\tau^t \varphi(s)\,ds \le \frac{C}{L}\big(e^{-L\tau} - e^{-Lt}\big).$$

Hence

$$\varphi(t) \le C + L\int_\tau^t \varphi(s)\,ds \le C + Ce^{Lt}\big(e^{-L\tau} - e^{-Lt}\big) = Ce^{L(t-\tau)}.$$

□

Proof of Proposition 6.7 First step. Let us introduce for each $\varepsilon' \ge 0$ the set

$$K_{\varepsilon'} := \big\{(t,y) \in \mathbb{R}\times X \ : \ t \in [c,d] \quad \text{and} \quad \|x(t) - y\| \le \varepsilon'\big\}.$$

Since $K_0 \subset D$ is compact and f, D_2f are continuous, there exist two positive constants M, L such that $M > \|f\|$ and $L > \|D_2f\|$ on K_0. If $\varepsilon' > 0$ is sufficiently small, then $K_{\varepsilon'} \subset D$ because D is open, and the inequalities $M > \|f\|$ and $L > \|D_2f\|$ still hold on $K_{\varepsilon'}$ by continuity. We may assume that $\varepsilon' < \varepsilon$. Applying the Lagrange inequality it follows that

$$\|f(t,x(t)) - f(t,y)\| \le L\,\|x(t) - y\|$$

for all $(t,y) \in K_{\varepsilon'}$.

Second step. Now choose $0 < \delta < e^{-L(d-c)}\varepsilon'$, and assume on the contrary that $\tau < b \le d$ (the case $c \le a < \tau$ is similar). Then by Proposition 6.6 (and the second remark following its statement) there exists a $t' \in (\tau, b)$ such that $\|x(t') - y(t')\| = \varepsilon'$ and $\|x(t) - y(t)\| < \varepsilon'$ for all $t \in [\tau, t')$.

Then we deduce from the integral equation for all $t \in [\tau, t')$ the estimates

$$\begin{aligned}\|x(t) - y(t)\| &\le \|\xi - \eta\| + \Big\|\int_\tau^t \Big(f(s,x(s)) - f(s,y(s))\Big)\,ds\Big\| \\ &< \delta + L\int_\tau^t \|x(s) - y(s)\|\,ds.\end{aligned}$$

Applying Gronwall's lemma we obtain that

$$\|x(t) - y(t)\| \le \delta e^{L(t-\tau)} \le \delta e^{L(d-c)}$$

for all $t \in [\tau, t')$. By continuity this holds for $t = t'$, too. This yields $\|x(t') - y(t')\| < \varepsilon'$ by our choice of δ, contradicting our initial assumption on t'. □

***Remark** If we assume the local Lipschitz condition instead of the continuity of $D_2 f$, then the first step of the above proof may be modified as follows. For each $t \in [c, d]$ there exist positive numbers δ_t, ε_t, M_t, L_t such that if

$$|t - t'| < \delta_t \quad \text{and} \quad \|x(t) - y\| \le \varepsilon_t,$$

then

$$(t', x(t')), (t', y) \in D, \quad \left\|f(t', y)\right\| \le M_t$$

and

$$\left\|f(t', x(t')) - f(t', y)\right\| \le L_t \left\|x(t') - y\right\|.$$

We may cover $[c, d]$ with finitely many intervals $(t_i - \delta_{t_i}, t_i + \delta_{t_i})$. Then the required property follows by choosing

$$\varepsilon' := \min \varepsilon_i, \quad M := \max M_i \quad \text{and} \quad L := \max L_i.$$

6.5 Explicit Solutions

Few differential equations may be solved explicitly. We briefly indicate two methods here.[19]

Separable equations We consider the problem

$$x' = g(t)h(x), \quad x(\tau) = \xi$$

where $g, h : \mathbb{R} \hookrightarrow \mathbb{R}$ are continuous functions defined on open intervals. If $\tau \in D(g)$, $\xi \in D(h)$ and $h(\xi) \ne 0$, then a maximal solution is given by the formula

$$\int_\xi^x \frac{1}{h(y)}\, dy = \int_\tau^t g(s)\, ds.$$

Formally we integrate the equality

$$\frac{1}{h(x)}\, dx = g(t)\, dt \quad \text{coming from} \quad \frac{dx}{dt} = g(t)h(x).$$

[19] Many more are given in Kamke [269]. See also Exercises 6.1–6.5, pp. 160–162.

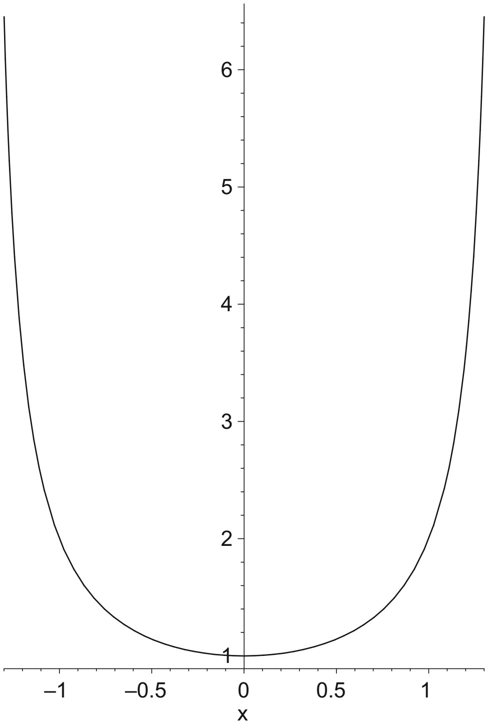


Fig. 6.6 Separable equation: solution of the example

If h is locally Lipschitz continuous, then Theorem 6.2 applies, so that the maximal solution is unique.[20]

Examples

- The maximal solution of the problem

$$x' = tx^2, \qquad x(0) = 1$$

is given by the formula

$$\int_1^x \frac{1}{y^2}\,dy = \int_0^t s\,ds;$$

hence (see Fig. 6.6)

$$x(t) = 2/(2 - t^2), \quad -\sqrt{2} < t < \sqrt{2}.$$

[20] The uniqueness holds without this extra hypothesis: see, e.g., Walter [505, pp. 16–17].

- Consider the *homogeneous* linear differential equation

  $$x' = a(t)x, \quad x(\tau) = \xi,$$

  where $a : I \to \mathbb{R}$ is a given continuous function. If $\xi \neq 0$, then the formula

  $$\int_{\xi}^{x} \frac{1}{y}\,dy = \int_{\tau}^{t} a(s)\,ds =: A(t)$$

  yields the maximal solution

  $$x(t) = \xi e^{A(t)}, \quad t \in I. \tag{6.6}$$

  The last formula is also valid for $\xi = 0$.

Introduction of a new unknown function

Examples

- The problem

  $$x' = 1 + (x-t)^2, \quad x(0) = 1$$

  becomes separable by setting $y(t) := x(t) - t$:

  $$y' = y^2, \quad y(0) = 1.$$

  Applying the preceding method we obtain the maximal solution

  $$x(t) = t + \frac{1}{1-t}, \qquad -\infty < t < 1.$$

  See Fig. 6.7.
- Finally we consider the *non-homogeneous* linear differential equation

  $$x' = a(t)x + b(t), \quad x(\tau) = \xi,$$

  where $a, b : I \to \mathbb{R}$ are given continuous functions. Motivated by the formula (6.6) we seek the solutions in the form[21]

  $$x(t) = \xi(t)e^{A(t)}.$$

  Then we get for ξ the *homogeneous* equation

  $$\xi'(t) = e^{-A(t)}b(t).$$

[21] This is the simplest case of the *method of variation of constants,* due to Euler and Lagrange.

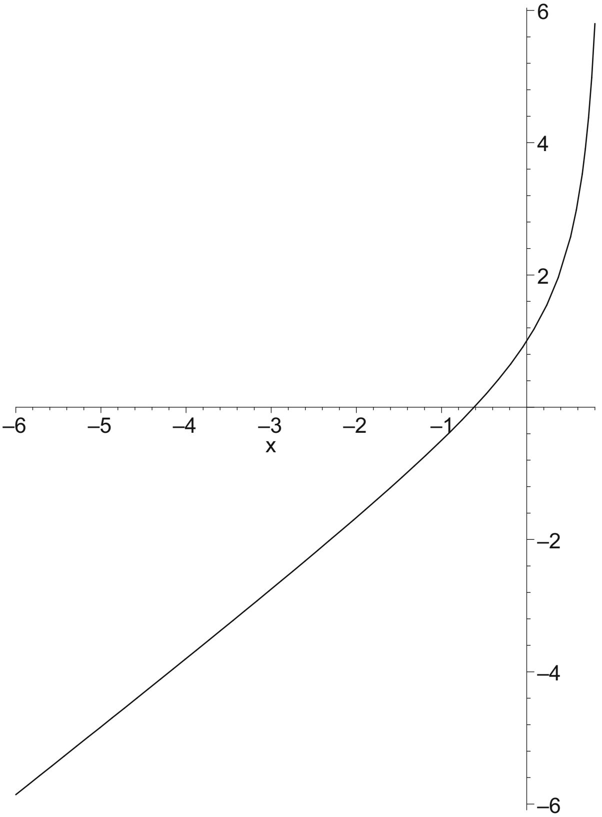


Fig. 6.7 New unknown function: solution of the example

Solving this we obtain at last the formula

$$x(t) = e^{A(t)}\xi + \int_{\tau}^{t} e^{A(t)-A(s)}b(s)\,ds, \quad t \in I$$

for the maximal solution of the original problem.

6.6 Exercises

Exercise 6.1 Solve the following problem:

$$(t^2 - 1)x' + 2tx^2 = 0, \quad x(0) = 1.$$

Exercise 6.2 (Homogeneous Equations) Prove that a differential equation of the form

$$x' = f\Big(\frac{x}{t}\Big),$$

where $f : I \to \mathbb{R}$ is a continuous function on some interval I, may be transformed into a separable equation by introducing the new unknown function $y := x/t$.

Apply this to the equation

$$x' = \frac{2tx - x^2}{t^2}.$$

Exercise 6.3 (Exact Equations) Given two continuous functions $g, h : D \to \mathbb{R}$ defined on a non-empty open set $D \subset \mathbb{R}^2$, the formal expression

$$g(t,x)\ dt + h(t,x)\ dx = 0 \tag{6.7}$$

means either one of the differential equations

$$\frac{dx}{dt} = -\frac{g(t,x)}{h(t,x)} \quad \text{and} \quad \frac{dt}{dx} = -\frac{h(t,x)}{g(t,x)};$$

i.e., we may consider either x as a function of t, or t as a function of x.[22]

Prove that if there exists a function $F : D \to \mathbb{R}$ of class C^1 satisfying[23]

$$D_1F := \frac{dF}{dt} = g \quad \text{and} \quad D_2F := \frac{dF}{dx} = h$$

in D, then the solutions of the algebraic equations

$$F(t,x) = c, \quad c \in \mathbb{R}$$

solve the differential equation (6.7).

Apply this method to the equation

$$x' = \frac{2t + 3t^2x}{3x^2 - t^3}.$$

[22] Sometimes it is easier to solve the second problem than the first one.

[23] Then F is called a *potential function* for the *vector field* (g, h).

Exercise 6.4 (Characterization of Exact Equations) Assume that the functions $g, h : D \to \mathbb{R}$ in (6.7) are of class C^1.

(i) Prove that if (6.7) is exact and $g, h : D \to \mathbb{R}$ are of class C^1, then $D_2 g = D_1 h$ on D.
(ii) Conversely, prove that if D is a rectangle and $D_2 g = D_1 h$ on D, then (6.7) is exact.

Exercise 6.5 (Multipliers) Sometimes an equation of the form (6.7) is not exact, but there exists a non-zero function $m : D \to \mathbb{R}$ such that the equivalent equation

$$m(t,x)g(t,x)\ dt + m(t,x)h(t,x)\ dx = 0$$

is exact.

Solve the differential equation

$$x\ dt - (4t^2 x + t)\ dx = 0$$

by finding a suitable multiplier $m(t)$, depending only on t.

Exercise 6.6 (Liouville's Formula) Let $A : I \to \mathbb{R}^{n \times n}$ be a continuous matrix function on an open interval I, and $W : I \to \mathbb{R}^{n \times n}$ a differentiable matrix function satisfying the *matrix differential equation* $W' = AW$. Show that $w := \det W : I \to \mathbb{R}$ satisfies the differential equation

$$w' = (\operatorname{tr} A)w.$$

Exercise 6.7 Consider the differential equation $x'(t) = f(t,x)$ where $f : \mathbb{R}^2 \to \mathbb{R}$ is given by the formula

$$f(t,x) := \begin{cases} 2t & \text{if } x \geq t^2, \\ 2x/t & \text{if } |x| < t^2, \\ -2t & \text{if } x \leq -t^2. \end{cases}$$

For which initial conditions does this equation have a unique solution?

Exercise 6.8 We consider the initial value problem

$$x' = 2t - 2\sqrt{\max\{x, 0\}}, \quad x(0) = 0.$$

Prove the following statements:

(i) The problem has a unique solution $x : [0, \infty) \to \mathbb{R}$.[24]

[24] We consider a one-sided derivative at 0.

(ii) The method of successive approximations does not converge if we start with $x_0(t) = 0$.

(iii) The method of successive approximations converges to the solution if we start with $x_0(t) = \alpha t^2$ for some $\alpha > 0$.

Exercise 6.9 Show that the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(0,0) = 0, \quad \text{and} \quad f(t,x) = \frac{4t^3 x}{t^4 + x^2} \quad \text{otherwise}$$

is continuous.

Does the problem $x'(t) = f(t,x)$, $x(0) = 0$ have a unique solution?

Exercise 6.10 (A Mean-Value Formula) Let u solve the differential equation

$$-u'' + qu = \lambda u$$

in an open interval I, where $q : I \to \mathbb{R}$ is a continuous function and λ is a nonnegative number.

Prove that

$$u(x-t) + u(x+t) - 2u(x)\cos\sqrt{\lambda}t = \int_{x-t}^{x+t} q(s)u(s)\frac{\sin\sqrt{\lambda}(t-|x-s|)}{\sqrt{\lambda}}\,ds$$

whenever $x \pm t \in I$.

Exercise 6.11 (An "anti-Gronwall" Inequality) Prove that if a non-increasing function $E : [0,\infty) \to [0,\infty)$ satisfies for some $\alpha > 0$ the condition

$$\alpha \int_t^\infty E(s)\,ds \le E(t)$$

for all $t \ge 0$, then

$$E(t) \le E(0)e^{1-\alpha t}$$

for all $t \ge 0$.

Chapter 7
Implicit Functions and Their Applications

In this chapter we continue to use the letters X, Y, Z to denote *Banach* spaces, i.e, *complete* normed spaces.[1]

7.1 Implicit Functions

We consider the equation $f(x,y) = 0$ for some given function $f : \mathbb{R}^2 \hookrightarrow \mathbb{R}$. When we may solve it explicitly to get a function $y = g(x)$, then the *level curve*

$$\Gamma := \{(x,y) \in D(f) \ : \ f(x,y) = 0\}$$

is equal to the graph

$$\{(x, g(x)) \ : \ x \in D(g)\}$$

of g.

Such a representation does not always exist. For example, the level curve of $f(x,y) = x^2 + y^2 - 1$, the unit circle, is not the graph of any function $g : \mathbb{R} \hookrightarrow \mathbb{R}$, because it has different points having the same abscissa. (See Fig. 7.1.)

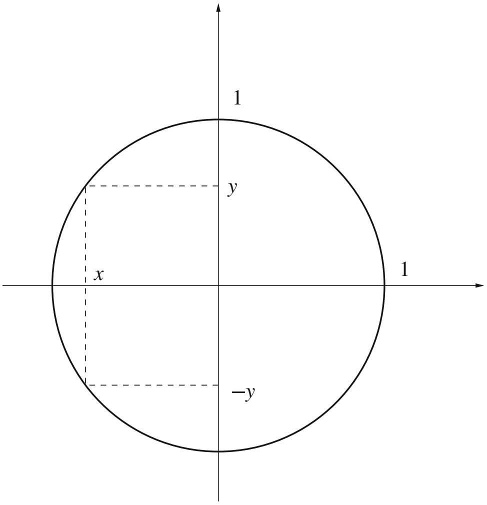
However, if $(x_0, y_0) \in \Gamma$ and $y_0 \neq 0$, then in a small neighborhood of (x_0, y_0) the level curve Γ *is* the graph of the function

$$g(x) = \begin{cases} \sqrt{1-x^2} & \text{if } y_0 > 0, \\ -\sqrt{1-x^2} & \text{if } y_0 < 0. \end{cases}$$

[1] We recall that all finite-dimensional normed spaces are complete.

V. Komornik, *Topology, Calculus and Approximation*, Springer
Undergraduate Mathematics Series, DOI 10.1007/978-1-4471-7316-8_7

Fig. 7.1 Implicit functions



This example may be generalized:

Proposition 7.1 *Let* $f : \mathbb{R}^n \times \mathbb{R} \hookrightarrow \mathbb{R}$ *be a* continuous *function in a neighborhood of* (x_0, y_0)*, satisfying* $f(x_0, y_0) = 0$*. Assume that for some neighborhood* $D \times [y_0 - d, y_0 + d]$ *of* (x_0, y_0) *the function*

$$y \mapsto f(x, y)$$

is (strictly) increasing in $[y_0 - d, y_0 + d]$ *for each fixed* $x \in D$*.*

Then there exist an open neighborhood $D' \subset D$ *of* x_0 *and a* continuous *function* $g : D' \to \mathbb{R}$ *such that*

$$\left\{(x, y) \in D' \times [y_0 - d, y_0 + d] \ : \ f(x, y) = 0\right\} = \left\{(x, g(x)) \ : \ x \in D'\right\}. \tag{7.1}$$

The proposition remains valid if we change the adjective “increasing” to “decreasing”: it suffices to consider $-f$ instead of f.

Proof Since $f(x_0, y_0) = 0$, and $f(x_0, \cdot)$ is increasing in $[y_0 - d, y_0 + d]$, we have[2]

$$f(x_0, y_0 - d) < 0 < f(x_0, y_0 + d).$$

[2] We indicate in Fig. 7.2 the sign of f at some points in the case $n = 1$.

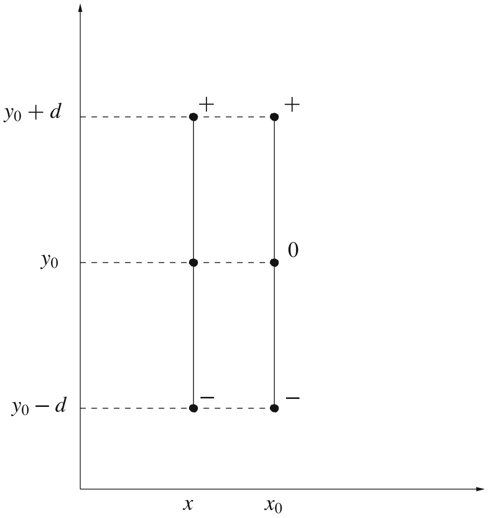


Fig. 7.2 Proof of Proposition 7.1

Since f is continuous at the points $(x_0, y_0 \pm d)$, there exists an open neighborhood $D' \subset D$ of x_0 such that

$$f(x, y_0 - d) < 0 < f(x, y_0 + d)$$

for all $x \in D'$. By Bolzano's theorem (p. 47) there is a point $y_0 - d < y < y_0 + d$ such that $f(x, y) = 0$. Since the function $f(x, \cdot)$ is increasing, this point is unique. Setting $g(x) := y$ we obtain therefore a function $g : D' \to \mathbb{R}$ satisfying (7.1), and the inequalities

$$y_0 - d < g(x) < y_0 + d \tag{7.2}$$

for all $x \in D'$.

It remains to show that g is continuous at each $x \in D'$. For any fixed $\varepsilon > 0$, using (7.2) we may choose $0 < \varepsilon' \leq \varepsilon$ such that

$$[g(x) - \varepsilon', g(x) + \varepsilon'] \subset [y_0 - d, y_0 + d].$$

Repeating the above reasoning with

$$(x, g(x)) \quad \text{and} \quad D' \times [g(x) - \varepsilon', g(x) + \varepsilon']$$

instead of

$$(x_0, y_0) \quad \text{and} \quad D \times [y_0 - d, y_0 + d],$$

we obtain a neighborhood D'' of x such that

$$|g(\tilde{x}) - g(x)| < \varepsilon' \leq \varepsilon$$

for all $\tilde{x} \in D''$. □

Next we give a more practical sufficient condition that also ensures the differentiability of the implicit function:

Proposition 7.2 (Descartes–Dini) *Let* $f : \mathbb{R}^n \times \mathbb{R} \hookrightarrow \mathbb{R}$ *be a* C^k *function in a neighborhood of* (x_0, y_0) *for some* $k \geq 1$*. Assume that*

$$f(x_0, y_0) = 0 \quad \textit{and} \quad D_2 f(x_0, y_0) \neq 0.$$

Then there exist a neighborhood $V' \subset D(f)$ *of* (x_0, y_0) *and a* C^k *function* $g : \mathbb{R}^n \hookrightarrow \mathbb{R}$ *such that*

$$\left\{(x, y) \in V' \; : \; f(x, y) = 0\right\} = \{(x, g(x)) \; : \; x \in D(g)\}. \tag{7.3}$$

Proof Assume, for example, that $D_2 f(x_0, y_0) > 0$, and fix a neighborhood $D \times [y_0 - d, y_0 + d]$ of (x_0, y_0) in which $D_2 f(x, y) > 0$. Then the hypotheses of the preceding proposition are satisfied. Therefore there exist an open neighborhood $D' \subset D$ of x_0 and a *continuous* function $g : D' \to \mathbb{R}$ satisfying (7.3) with $V' : = D' \times [y_0 - d, y_0 + d]$.

We prove that g is differentiable. Given $(x_1, y_1) \in D(f)$ arbitrarily, by the differentiability of f there is a function $u : D(f) \to L(\mathbb{R}^n \times \mathbb{R}, \mathbb{R})$, continuous at (x_1, y_1) and satisfying the relations

$$f(x, y) - f(x_1, y_1) \equiv u(x, y)(x - x_1, y - y_1)$$

and $u(x_1, y_1) = f'(x_1, y_1)$. Equivalently, there exist two functions

$$u_1 : D(f) \to L(\mathbb{R}^n, \mathbb{R}) \quad \text{and} \quad u_2 : D(f) \to \mathbb{R},$$

continuous at (x_1, y_1) and satisfying the relations

$$f(x, y) - f(x_1, y_1) \equiv u_1(x, y)(x - x_1) + u_2(x, y)(y - y_1)$$

and[3]

$$u_1(x_1, y_1) = D_1 f(x_1, y_1), \quad u_2(x_1, y_1) = D_2 f(x_1, y_1).$$

[3] For $n > 1$ the partial derivative $D_1 f(x_1, y_1)$ is defined in a generalized sense, as $D_2 f(t_0, x_0)$ in the Cauchy–Lipschitz theorem, p. 148.

Choosing $y = g(x)$ and $y_1 = g(x_1)$ we infer that

$$0 \equiv u_1(x, g(x))(x - x_1) + u_2(x, g(x))(g(x) - g(x_1)).$$

Since the composite function

$$x \mapsto u_2(x, g(x)) = D_2 f(x, g(x))$$

is continuous and positive at x_1, $u_2(x, g(x)) > 0$ in a suitable neighborhood $D'' \subset D'$ of x_1, and then

$$g(x) - g(x_1) \equiv -\frac{u_1(x, g(x))}{u_2(x, g(x))}(x - x_1)$$

in D''. Since the fraction is continuous at x_1, we conclude that g is differentiable at x_1, and

$$g'(x_1) = -\frac{u_1(x_1, g(x_1))}{u_2(x_1, g(x_1))} = -\frac{D_1 f(x_1, g(x_1))}{D_2 f(x_1, g(x_1))}.$$

Since x_1 is arbitrary, g is differentiable, and

$$g'(x) = -\frac{D_1 f(x, g(x))}{D_2 f(x, g(x))} \quad \text{for all} \quad x \in D'. \tag{7.4}$$

Moreover, the fraction on the right-hand side is composed of continuous functions, so that $g \in C^1$.

If $f \in C^2$, then the right-hand side of (7.4) is composed of C^1 functions (because we already know that $g \in C^1$), so that $g \in C^2$. If $f \in C^k$ for some $k > 2$, then by a "bootstrap" argument (p. 143) we obtain that $g \in C^k$. □

Remark In general we cannot express g explicitly from the defining equation $f(x, g(x)) = 0$. However, its *derivative* may be computed at any given point $(x, y) \in V'$ as follows. Differentiating the equation $f(x, g(x)) = 0$ we get

$$D_1 f(x, g(x)) + D_2 f(x, g(x)) g'(x) = 0,$$

whence[4]

$$g'(x) = -\frac{D_1 f(x, g(x))}{D_2 f(x, g(x))} = -\frac{D_1 f(x, y)}{D_2 f(x, y)}.$$

For $f : \mathbb{R}^2 \hookrightarrow \mathbb{R}$ this formula often enables us to draw the graph of g and hence Γ.

[4] Compare to (7.4).

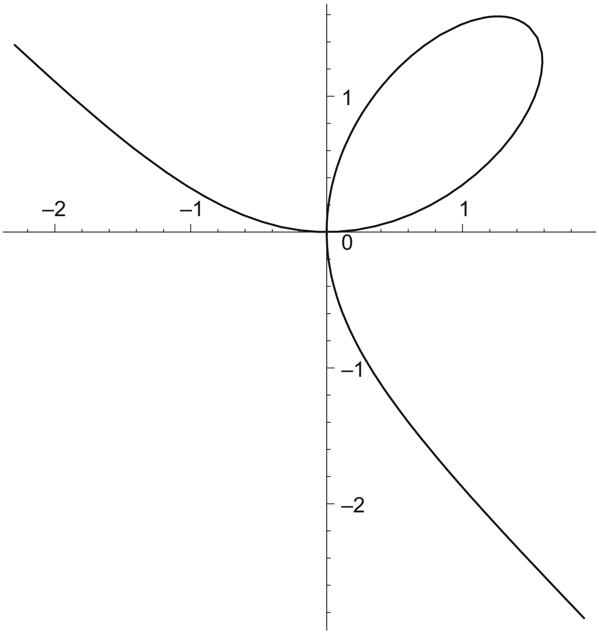


Fig. 7.3 Folium of Descartes

Examples

- Let us return to the example of the circle Γ. Applying Proposition 7.2 to the function

$$f(x,y) := x^2 + y^2 - 1$$

at a point $(x_0, y_0) \in \Gamma$ with $y_0 \neq 0$, we obtain the well-known fact that in a small neighborhood of this point Γ is the graph of a C^∞ function $g : \mathbb{R} \hookrightarrow \mathbb{R}$.

- Descartes applied his theorem and the preceding remark to construct the tangents to the curve defined by the equation $x^3 + y^3 = 3xy$. See Fig. 7.3.
- Let us consider the C^∞ function $f : \mathbb{R}^3 \to \mathbb{R}$ given by the formula

$$f(u,v,y) := u^2 + v^2 + y^2 - 3uvy.$$

Since

$$\frac{\partial f}{\partial y}(1,1,1) = -1 \neq 0,$$

writing $x := (u, v)$ and choosing $x_0 = (1, 1)$, $y_0 = 1$ we may apply Proposition 7.2. We obtain that in a suitable neighborhood of $(1, 1, 1)$ the set

$$\Gamma := \left\{(u, v, y) \in \mathbb{R}^3 \ : \ u^2 + v^2 + y^2 = 3uvy\right\}$$

is the graph of a C^∞ function $g : \mathbb{R}^2 \hookrightarrow \mathbb{R}$. Geometrically this is a *surface*.

In Sects. 7.4–7.5 we will generalize Proposition 7.2 for *vector-valued* functions f by using a different method.

7.2 Lagrange Multipliers

Proposition 7.3 (Lagrange) *Let $f_0, f : \mathbb{R}^m \hookrightarrow \mathbb{R}$ be C^1 functions, $m \geq 2$. Assume that the restriction of f_0 to*

$$\Gamma := \{x \in D(f_0) \cap D(f) \ : \ f(x) = 0\}$$

has a local extremum at some point $a \in \Gamma$. Then

$$\lambda_0 f_0'(a) + \lambda f'(a) = 0 \tag{7.5}$$

for suitable real numbers λ_0 and λ, at least one of which is different from zero.

Remarks

- Apart from some pathological cases, the set Γ is not open in $\mathbb{R}^m$, so that Fermat's theorem (p. 104) does not apply.
- The *Lagrange multipliers* λ_0, λ have a geometric interpretation. First let $m = 2$, and consider a path on Γ passing through a, i.e., a function $\gamma : (\alpha, \beta) \to \mathbb{R}^2$ satisfying $\gamma(t_0) = a$, $\gamma'(t_0) \neq 0$, and $\gamma(t) \in \Gamma$ for all t. Then $f \circ \gamma$ vanishes identically, while $f_0 \circ \gamma$ has a local extremum at t_0. Hence

$$(f \circ \gamma)'(t_0) = (f_0 \circ \gamma)'(t_0) = 0,$$

i.e.,

$$f'(a)\gamma'(t_0) = f_0'(a)\gamma'(t_0) = 0.$$

Identifying $\gamma'(t_0) \in L(\mathbb{R}, \mathbb{R}^2)$ with a non-zero vector $v \in \mathbb{R}^2$ and using gradient vectors (p. 112) we may rewrite them in the form

$$\nabla f(a) \cdot v = \nabla f_0(a) \cdot v = 0.$$

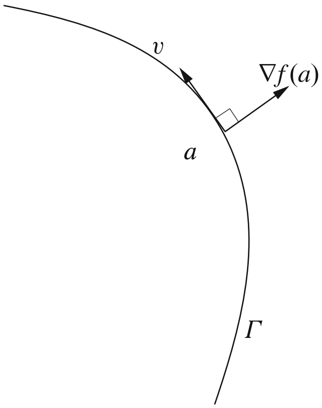


Fig. 7.4 Orthogonality of the gradient

Being orthogonal to the same non-zero vector, $\nabla f(a)$ and $\nabla f_0(a)$ are parallel, and the relation (7.5) follows. Geometrically v is the tangent vector at a to Γ, so that $\nabla f(a)$ and $\nabla f_0(a)$ are orthogonal to Γ at a. See Fig. 7.4.

If $m > 2$, then we apply the above argument for *every* tangent vector v at a.

- Writing $a = (a_1, \ldots, a_m)$ the proposition reduces to the problem of finding the conditional extrema of the solutions of the (usually nonlinear) system of algebraic equations

$$f(a_1, \ldots, a_m) = 0,$$
$$\lambda_0 D_k f_0(a) + \lambda D_k f(a) = 0, \quad k = 1, \ldots, m,$$

having $m + 1$ equations and $m + 2$ unknowns: $a_1, \ldots, a_m, \lambda_0, \lambda$. The solutions with $\lambda_0 = \lambda = 0$ are of no interest.

In applications one can often show the implication

$$\lambda_0 = 0 \Longrightarrow \lambda = 0$$

for all solutions. Then we may assume by a homogeneity argument that $\lambda_0 = 1$, and the number of unknowns becomes equal to that of the equations. This new system often has only finitely many solutions.

We give an important application in the next section.

- Sometimes (following Lagrange…) the theorem is formulated with $\lambda_0 = 1$. The examples

$$f_0(x_1, x_2) := x_1 \quad \text{and} \quad f(x_1, x_2) := x_1^2 + x_2^2, \quad (x_1, x_2) \in \mathbb{R}^2$$

and

$$f_0(x_1, x_2) := x_1 \quad \text{and} \quad f(x_1, x_2) := x_1^3 - x_2^2, \quad (x_1, x_2) \in \mathbb{R}^2$$

show that without further assumptions this formulation is incorrect.

Proof If $f'(a) = 0$, then we may choose $\lambda_0 = 0$ and $\lambda = 1$. Otherwise there exists a k such that $D_k f(a) \neq 0$. To simplify the notation we assume henceforth that $D_m f(a) \neq 0$.

By the results of the preceding section a has a neighborhood $U \subset \mathbb{R}^m$ such that $U \cap \Gamma$ is the graph of a suitable C^1 function $g : D' \to \mathbb{R}$, defined on an open set D' of $\mathbb{R}^{m-1}$.

Let $(c, g(c)) = a$, then the function

$$h(y) := f_0(y, g(y)), \quad y \in D'$$

or more explicitly

$$h(y_1, \dots, y_{m-1}) := f_0(y_1, \dots, y_{m-1}, g(y_1, \dots, y_{m-1})), \quad (y_1, \dots, y_{m-1}) \in D'$$

has a local extremum at c, and hence $h'(c) = 0$, because D' is open. Equivalently,[5]

$$D_k f_0(a) + D_m f_0(a) D_k g(c) = 0, \quad k = 1, \dots, m-1.$$

On the other hand, differentiating the identity

$$f(y, g(y)) \equiv 0$$

at c we get the equality

$$D_k f(a) + D_m f(a) D_k g(c) = 0, \quad k = 1, \dots, m-1.$$

Eliminating the terms $D_k g(c)$ from these equations we obtain that

$$D_m f(a) D_k f_0(a) - D_m f_0(a) D_k f(a) = 0, \quad k = 1, \dots, m-1.$$

The last equality obviously holds for $k = m$, too, so that

$$D_m f(a) f_0'(a) - D_m f_0(a) f'(a) = 0,$$

i.e., (7.5) holds with

$$\lambda_0 = D_m f(a) \neq 0 \quad \text{and} \quad \lambda = -D_m f_0(a). \qquad \square$$

[5] We differentiate $f_0(y_1, \dots, y_{m-1}, g(y_1, \dots, y_{m-1}))$ with respect to y_k.

7.3 The Spectral Theorem

We use Lagrange multipliers in the proof of an important theorem of linear algebra.

Definition A linear map $A \in L(H, H)$ on a Euclidean space H is *symmetric* if

$$(Ax, y) = (x, Ay)$$

for all $x, y \in H$.

Example Consider the usual scalar product on $H = \mathbb{R}^m$, and represent the elements of $H = \mathbb{R}^m$ as column vectors. If (a_{ij}) is a symmetric matrix of order m, then the formula

$$Ax := \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mm} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}$$

defines a symmetric map $A \in L(H, H)$ because

$$(Ax, y) = \sum_{i,j=1}^{m} a_{ij} x_i y_j = \sum_{i,j=1}^{m} a_{ji} x_i y_j = (x, Ay)$$

for all $x, y \in H$.

Theorem 7.4 (Cauchy) *If $A \in L(H, H)$ is a symmetric map on a finite-dimensional Euclidean space $H \neq \{0\}$, then H has an orthonormal basis formed by eigenvectors of A.*

The key to the proof is the following lemma:

Lemma 7.5 *A has at least one eigenvalue.*

Proof The example on p. 104 shows that the formulas

$$f_0(x) := (Ax, x) \quad \text{and} \quad f(x) := (x, x) - 1$$

define two differentiable functions, and that

$$f_0'(a)h = (Aa, h) + (Ah, a) = (2Aa, h),$$
$$f'(a)h = (a, h) + (h, a) = (2a, h)$$

for all $a, h \in H$. Since A is continuous, the linear maps $f_0', f' : H \to L(H, \mathbb{R})$ are continuous, and hence belong to C^∞.

Since H is finite-dimensional, its unit sphere

$$\Gamma := \{x \in H \ : \ f(x) = 0\}$$

is compact (see Theorem 3.10, p. 79), so that the continuous function $f_0|_\Gamma$ has a maximal value $f_0(e)$. We claim that $e \in \Gamma$ is an eigenvector of A.

By Proposition 7.3 we have

$$\lambda_0 f_0'(e) + \lambda f'(e) = 0,$$

i.e.,

$$\lambda_0 Ae + \lambda e = 0$$

for suitable real numbers λ_0, λ, at least one of which is non-zero. Since $e \in \Gamma$ implies that $e \neq 0$, the equality $\lambda_0 = 0$ would imply that $\lambda = 0$. Since this case is excluded, we have $\lambda_0 \neq 0$, and then $Ae = (-\lambda/\lambda_0)e$. □

Remark Here we may avoid the use of Lagrange multipliers as follows. By the definition of e, and since $e \neq 0$, for any given $h \in H$ we have

$$\Bigl(A\frac{e+th}{\|e+th\|}, \frac{e+th}{\|e+th\|}\Bigr) \le (Ae, e)$$

or equivalently

$$(A(e+th), e+th) \le (Ae, e)(e+th, e+th)$$

for all t sufficiently close to zero. Since $(e, e) = 1$ and $(Ae, h) = (Ah, e)$, this yields

$$2t(Ae, h) + o(t) \le \bigl(2t(e, h) + o(t)\bigr)(Ae, e).$$

Dividing by $2t$ and then letting $t \to \pm 0$ we conclude that $(Ae, h) = (Ae, e)(e, h)$ for all h. Hence $Ae = (Ae, e)e$.

Proof of Theorem 7.4 We seek $n := \dim H$ vectors $e_1, \dots, e_n \in H$ and real numbers $\lambda_1, \dots, \lambda_n$ satisfying the relations

$$(e_i, e_j) = \delta_{ij} \quad \text{and} \quad Ae_j = \lambda_j e_j \tag{7.6}$$

for all $i, j = 1, \dots, n$. Here δ_{ij} denotes the usual *Kronecker symbol*:

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

By the preceding lemma there exists a vector $e_1 \in H$ and a real number λ_1 such that

$$(e_1, e_1) = 1 \quad \text{and} \quad Ae_1 = \lambda_1 e_1.$$

Let $1 \le k < n$, and assume by induction that there exist vectors $e_1, \dots, e_k \in H$ and real numbers $\lambda_1, \dots, \lambda_k$ satisfying (7.6) for $i, j = 1, \dots, k$.

Let us introduce the linear subspace

$$H_k := \{x \in H \ : \ (x, e_1) = \cdots = (x, e_k) = 0\}$$

and the restriction A_k of A to H_k. If $x \in H_k$, then $Ax \in H_k$ because

$$(Ax, e_i) = (x, Ae_i) = (x, \lambda_i e_i) = \lambda_i (x, e_i) = 0 \quad \text{for all} \quad i = 1, \dots, k$$

by the symmetry of A.

Applying the preceding lemma for $A_k \in L(H_k, H_k)$ instead of A, there exist a unit vector $e_{k+1} \in H_k$ and $\lambda_{k+1} \in \mathbb{R}$ such that $Ae_{k+1} = \lambda_{k+1} e_{k+1}$. Using the definition of H_k, the relations (7.6) are now satisfied for all $i, j = 1, \dots, k+1$.

After n steps we obtain a basis having the required properties. □

Remark The spectral theorem has countless generalizations to linear operators in Banach spaces of arbitrary dimension. They play an important role in physical applications.[6]

7.4 * The Inverse Function Theorem

It follows from the continuity of the determinant function that an invertible matrix remains invertible if we slightly change its elements. The following proposition extends this property to arbitrary Banach spaces.

Definition A continuous linear map $A \in L(X, Y)$ is *continuously invertible* if it is bijective and if its inverse A^{-1} is also continuous: $A^{-1} \in L(Y, X)$.

Let us denote by Ω (resp. Ω') the set of continuously invertible linear maps in $L(X, Y)$ (resp. in $L(Y, X)$).

[6] See, e.g., von Neumann [362] and Reed–Simon [406].

Proposition 7.6

(a) *Ω is an open set in $L(X, Y)$. More precisely, if $A \in \Omega$, $B \in L(X, Y)$ and*

$$\|A - B\| < 1/\left\|A^{-1}\right\|, \tag{7.7}$$

then $B \in \Omega$.

(b) *The formula $\Phi(A) := A^{-1}$ defines a homeomorphism between Ω and Ω'.*

(c) *The functions $\Phi : \Omega \to \Omega'$ and $\Phi^{-1} : \Omega' \to \Omega$ belong to the class C^∞.*

Remark The set Ω may be empty. This is the case when X and Y have different finite dimensions. In parts (b) and (c) we assume that Ω is non-empty.

Proof

(a) First we prove that B is a bijection, i.e., for each $y \in Y$ the equation $Bx = y$ has a unique solution. It suffices to show that the function $\varphi : X \to X$ defined by the formula

$$\varphi(x) := x + A^{-1}(y - Bx)$$

has a unique fixed point, because the equality $\varphi(x) = x$ is equivalent to $y = Bx$. By Theorem 1.10 (p. 16) it is sufficient to show that φ is a contraction. This follows from the hypothesis $\left\|A^{-1}\right\| \cdot \|A - B\| < 1$ because

$$\begin{aligned}\|\varphi(x_1) - \varphi(x_2)\| &= \left\|A^{-1}(A - B)(x_1 - x_2)\right\| \\ &\le \left\|A^{-1}\right\| \cdot \|A - B\| \cdot \|x_1 - x_2\|\end{aligned}$$

for all $x_1, x_2 \in X$.

To prove the continuity of B^{-1} first we observe that

$$\|x\| = \left\|A^{-1}Ax\right\| \le \left\|A^{-1}\right\| \cdot \|Ax\|$$

and

$$\|Ax\| \le \|(A - B)x\| + \|Bx\| \le \|A - B\| \cdot \|x\| + \|Bx\|.$$

Consequently,

$$\|x\| \le \left\|A^{-1}\right\| \cdot \|Ax\| \le \left\|A^{-1}\right\| (\|A - B\| \cdot \|x\| + \|Bx\|),$$

and hence

$$(1 - \left\|A^{-1}\right\| \cdot \|A - B\|)\,\|x\| \le \left\|A^{-1}\right\| \cdot \|Bx\|$$

for all $x \in X$. Using (7.7) we conclude that $B \in \Omega$, and

$$\left\|B^{-1}\right\| \le \frac{\left\|A^{-1}\right\|}{1 - \|A^{-1}\| \cdot \|A-B\|}. \tag{7.8}$$

(b) The relation $(A^{-1})^{-1} = A$ implies that $\Phi : \Omega \to \Omega'$ is a bijection.
If $B \in L(X,Y)$ and $\|A-B\| < 1/\left\|A^{-1}\right\|$, then using (7.8) we have

$$\begin{aligned}\left\|B^{-1} - A^{-1}\right\| &= \left\|B^{-1}(A-B)A^{-1}\right\| \\ &\le \left\|B^{-1}\right\| \cdot \|A-B\| \cdot \left\|A^{-1}\right\| \\ &\le \frac{\left\|A^{-1}\right\|^2 \|A-B\|}{1 - \|A^{-1}\| \cdot \|A-B\|}.\end{aligned}$$

If $B \to A$, then the right-hand side tends to zero, so that Φ is continuous. Exchanging the roles of X and Y we obtain the continuity of Φ^{-1} as well.

(c) By symmetry it suffices to show that $\Phi \in C^\infty$. First we show that Φ is differentiable at each $A \in \Omega$, and

$$\Phi'(A)H = -A^{-1}HA^{-1} \quad \text{for all} \quad H \in L(X,Y). \tag{7.9}$$

Indeed, the linear map $H \mapsto -A^{-1}HA^{-1}$ is continuous from $L(X,Y)$ into $L(Y,X)$ because

$$\left\|-A^{-1}HA^{-1}\right\| \le \left\|A^{-1}\right\|^2 \|H\|$$

for all $H \in L(X,Y)$. Furthermore,

$$\begin{aligned}\Phi(A+H) - \Phi(A) + A^{-1}HA^{-1}& \\ &= (A+H)^{-1} - A^{-1} + A^{-1}HA^{-1} \\ &= (A+H)^{-1}[A - (A+H) + (A+H)A^{-1}H]A^{-1} \\ &= (A+H)^{-1}HA^{-1}HA^{-1} \\ &= o(H)\end{aligned}$$

as $H \to 0$. Indeed,

$$\left\|(A+H)^{-1}HA^{-1}HA^{-1}\right\| \le \left\|(A+H)^{-1}\right\| \cdot \left\|A^{-1}\right\|^2 \|H\|^2,$$

and

$$\left\|(A+H)^{-1}\right\| \cdot \left\|A^{-1}\right\|^2 \|H\| \to 0.$$

To prove that $\Phi \in C^\infty$, we write (7.9) in the form $\Phi' = \varphi \circ (\Phi, \Phi)$ with the bilinear map

$$\varphi(C, D)H := -CHD.$$

The trivial estimate

$$\|-CHD\| \le \|C\| \cdot \|H\| \cdot \|D\|$$

shows that

$$\varphi : L(Y, X)^2 \to L(L(X, Y), L(Y, X))$$

is a continuous bilinear map (with norm ≤ 1). Hence $\varphi \in C^\infty$. Since Φ is differentiable, using the formula $\Phi' = \varphi \circ (\Phi, \Phi)$ a bootstrap argument (p. 143) shows that $\Phi \in C^\infty$. □

Definition Let $U \subset X$ and $V \subset Y$ be two open sets. A bijection f between U and V is called a C^k-*diffeomorphism* if $f \in C^k$ and $f^{-1} \in C^k$.

Example The map $A \mapsto A^{-1}$ of the preceding proposition is a C^∞-diffeomorphism of Ω onto Ω'.

Now we are ready to prove the following fundamental theorem:

Theorem 7.7 (Inverse Function Theorem) *Let X, Y be Banach spaces and $f : X \hookrightarrow Y$ a C^k function for some $k \ge 1$. If $f'(a) \in L(X, Y)$ is invertible at some point $a \in D(f)$, then there exist an open neighborhood U of a and an open neighborhood V of $f(a)$ such that $f|_U$ is a C^k-diffeomorphism of U onto V.*

Proof We proceed in five steps.

(i) Changing f to $[f'(a)]^{-1} \circ f$ we may assume that $Y = X$ and $f'(a) = I$ (the identity map of X). Since f' is continuous at a, there exists an open ball U centered at a such that

$$\|I - f'(x)\| < \frac{1}{2} \quad \text{for all} \quad x \in U. \tag{7.10}$$

Then $f'(x)$ is invertible for each $x \in U$ by the preceding proposition.

(ii) Since U is convex, it follows from (7.10) by applying the mean value theorem that

$$\|(x_1 - f(x_1)) - (x_2 - f(x_2))\| \le \frac{1}{2} \|x_1 - x_2\| \tag{7.11}$$

for all $x_1, x_2 \in U$. Since

$$\|x_1 - x_2\| - \|f(x_1) - f(x_2)\| \le \|(x_1 - f(x_1)) - (x_2 - f(x_2))\|,$$

this yields the estimate

$$\|f(x_1) - f(x_2)\| \ge \frac{1}{2}\|x_1 - x_2\| \quad \text{for all} \quad x_1, x_2 \in U. \tag{7.12}$$

The inequality (7.12) implies that $f|_U$ has a continuous inverse g.

(iii) We show that $V := f(U)$ is an open set. Fix $y_0 = f|_U(x_0) \in V$ arbitrarily, and then fix $r > 0$ such that

$$K := \overline{B_{2r}(x_0)} \subset U.$$

It suffices to show that $B_r(y_0) \subset V$.

Choose $y \in B_r(y_0)$ arbitrarily and consider the map $\psi : K \to X$ defined by the formula

$$\psi(x) := y + x - f(x).$$

The estimate (7.11) shows that ψ is a contraction. Furthermore, $\psi(K) \subset K$, because for $x \in K$ we have

$$\begin{aligned}\|\psi(x) - x_0\| &\le \|\psi(x) - \psi(x_0)\| + \|\psi(x_0) - x_0\|\\ &\le \frac{1}{2}\|x - x_0\| + \|y - y_0\|\\ &< r + r = 2r.\end{aligned}$$

Since K is closed and hence complete, ψ has a fixed point $x \in K$ by Theorem 1.10 (p. 16). Then $x \in U$ and $y = f(x) \in V$ by the definition of ψ.

(iv) We show that g is differentiable and $g' = \Phi \circ f' \circ g$, where $\Phi : \Omega \to \Omega$ denotes the C^∞-diffeomorphism of Proposition 7.6. (Now we have $\Omega' = \Omega$ because $Y = X$.)

Fix $y_0 \in V$ arbitrarily. Since f is differentiable in $x_0 := g(y_0) \in U$, there exists a function $u : U \to L(X, X)$, continuous at x_0 and satisfying

$$f(x) - f(x_0) = u(x)(x - x_0)$$

for all $x \in U$. Writing $y = f(x)$ this implies that

$$y - y_0 = u(g(y))(g(y) - g(y_0))$$

for all $y \in V$.

Since $(u \circ g)(y_0) = u(x_0) = f'(x_0)$ is invertible and $u \circ g$ is continuous at y_0, by Proposition 7.6 there exists a neighborhood $V' \subset V$ of y_0 such that $u(g(y))$ is invertible for all $y \in V'$. Hence the formula

$$v := \Phi \circ u \circ g$$

defines a map $v : V' \to L(X, X)$ that is continuous at y_0 and satisfies the equality

$$g(y) - g(y_0) = v(y)(y - y_0)$$

for all $y \in V'$. We conclude that g is differentiable at y_0, and

$$g'(y_0) = v(y_0) = \Phi(f'(g(y_0))).$$

(v) In the just proven equality

$$g' = \Phi \circ f' \circ g$$

we have $f' \in C^{k-1}$ by assumption. Since $\Phi \in C^\infty$ and g is continuous, g' is also continuous, i.e., $g \in C^1$. By the usual bootstrap reasoning (p. 143) we obtain that $g \in C^k$. If $f \in C^\infty$, then we get $g \in C^k$ for all finite k, i.e., $g \in C^\infty$. □

Remark If X is finite-dimensional, then we may avoid the application of the fixed point theorem as follows.[7] Without loss of generality we may assume that the norm is Euclidean. Choose $y_0 = f|_U(x_0) \in V$ and $r > 0$ as in part (iii) of the preceding proof. Given $y \in B_{r/2}(y_0)$ arbitrarily, we minimize the differentiable, and hence continuous, function $\rho(x) := \|f(x) - y\|^2$ on the *compact* ball $\overline{B_{2r}(x_0)}$. Since

$$\|f(x_0) - y\| = \|y_0 - y\| < \frac{r}{2},$$

while on the boundary of this ball using (7.12) we have

$$\|f(x) - y\| > \|f(x) - y_0\| - \frac{r}{2} \ge \frac{1}{2}\|x - x_0\| - \frac{r}{2} = \frac{r}{2},$$

the minimum is attained at some *interior* point x of the ball, and therefore $\rho'(x) = 0$. The required equality $f(x) - y = 0$ follows because

$$0 = \rho'(x)h = 2(f(x) - y, f'(x)h)$$

for all $h \in X$, and $f'(x)$ is onto.

[7] It may be adapted to infinite-dimensional Hilbert spaces, too.

7.5 * The Implicit Function Theorem

Using the inverse function theorem we may greatly generalize Proposition 7.2 (p. 168).

Definition The Banach space X is the *direct sum* of its *closed* linear subspaces R and S if

$$R \cap S = \{0\} \quad \text{and} \quad R + S = X.$$

Theorem 7.8 (Implicit Function Theorem) *Let $f : X \hookrightarrow Y$ be a C^k function where X, Y are Banach spaces and $k \geq 1$. Let $a \in D(f)$ and consider the set*

$$\Gamma := \{x \in D(f) \; : \; f(x) = f(a)\}.$$

Assume that X has a direct sum decomposition $X = R + S$ such that $f'(a)|_S \in L(S, Y)$ is invertible. Then there exists a neighborhood U of a and a C^k function $g : R \hookrightarrow S$ such that

$$U \cap \Gamma = \{(r, g(r)) \; : \; r \in D(g)\}.$$

Proof We proceed in four steps.

(i) The formula

$$F(r, s) = (r, f(r, s))$$

defines a C^k function $F : D(f) \to R \times Y$. Setting

$$A = f'(a)|_R \quad \text{and} \quad B = f'(a)|_S$$

we have

$$f'(a)x = Ar + Bs \quad \text{for all} \quad x = r + s \quad \text{with} \quad r \in R, \quad s \in S,$$

i.e.,

$$F'(a)x = \begin{pmatrix} I & 0 \\ A & B \end{pmatrix} \begin{pmatrix} r \\ s \end{pmatrix}.$$

Since B is invertible by assumption, a matrix computation shows that $F'(a)$ is invertible, and

$$F'(a)^{-1}(r,y) = \begin{pmatrix} I & 0 \\ -B^{-1}A & B^{-1} \end{pmatrix} \begin{pmatrix} r \\ y \end{pmatrix}.$$

Applying the inverse function theorem there exist open neighborhoods U, V of $a \in X$ and $F(a) \in R \times Y$ such that $F|_U$ is a C^k-diffeomorphism of U onto V.

(ii) We claim that $U \cap \Gamma$ is the graph of a suitable function $g : R \hookrightarrow S$. It is sufficient to show that if $(r,s), (r,\tilde{s}) \in U$ and $f(r,s) = f(r,\tilde{s})$, then $s = \tilde{s}$. This follows from the injectivity of $F|_U$ because

$$F(r,s) = (r, f(r,s)) = (r, f(r,\tilde{s})) = F(r,\tilde{s}).$$

(iii) We show that g is defined on an open set. We observe that

$$\begin{aligned} D(g) &= \{r \in R \ : \ \text{for some } s \in S \text{ we have } (r,s) \in U \text{ and } f(r,s) = f(a)\} \\ &= \{r \in R \ : \ \text{for some } s \in S \text{ we have } (r,s) \in U \text{ and } F(r,s) = (r, f(a))\} \\ &= \{r \in R \ : \ (r, f(a)) \in V\}. \end{aligned}$$

Since V is open, the last expression shows that $D(g)$ is open, too.

(iv) If $r \in D(g)$, then

$$F(r, g(r)) = (r, f(r, g(r))) = (r, f(a)),$$

so that

$$(F|_U)^{-1}(r, f(a)) = (r, g(r)).$$

Since $(F|_U)^{-1} \in C^k$, its component g is also of class C^k. □

Example We consider the function

$$f : \mathbb{R}^3 \to \mathbb{R}^2, \quad f(x_1,x_2,x_3) := (x_1 + x_2 + x_3, x_1^2 + x_2^2 + x_3^2)$$

in a neighborhood of $a = (0,1,0)$. Since

$$f'(a) = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 2 & 0 \end{pmatrix},$$

representing $\mathbb{R}^3$ as the direct sum of

$$R = \{(r, 0, 0) \ : \ r \in \mathbb{R}\} \quad \text{and} \quad S = \{(0, s_2, s_3) \ : \ s_2, s_3 \in \mathbb{R}\},$$

$f'(a)|_S$ is invertible: its matrix is composed of the last two columns of the above matrix.

Applying the theorem near $(0, 1, 0)$ we obtain that

$$\Gamma = \left\{(x_1, x_2, x_3) \in \mathbb{R}^3 \ : \ x_1 + x_2 + x_3 = x_1^2 + x_2^2 + x_3^2 = 1\right\}$$

is the graph of a suitable C^∞ function $g : \mathbb{R} \hookrightarrow \mathbb{R}^2$. Geometrically this is a *space curve*.

Remark More generally, if the theorem may be applied to a function $f : \mathbb{R}^m \hookrightarrow \mathbb{R}^n$ in some point $a \in \mathbb{R}^m$, then Γ is an $(m-n)$-dimensional *manifold*[8] in a neighborhood of a.

7.6 * Lagrange Multipliers: General Case

Let D be an open set in a Banach space X, and $f_0 : D \to \mathbb{R}, f : D \to \mathbb{R}^n$ two C^1 functions. We seek the local extrema of f_0 under the condition $f = 0$. Writing $f = (f_1, \dots, f_n)$ we have

$$\Gamma := \{x \in D \ : \ f_1(x) = \cdots = f_n(x) = 0\}.$$

Theorem 7.9 (Lagrange) *If $f_0|_\Gamma$ has a local extremum at a, then*

$$\lambda_0 f_0'(a) + \cdots + \lambda_n f_n'(a) = 0 \tag{7.13}$$

for suitable real numbers $\lambda_0, \dots, \lambda_n$, at least one of which is non-zero.

Remarks

- The existence of non-trivial Lagrange multipliers with $\lambda_0 = 0$ means that the derivatives $f_1'(a), \dots, f_n'(a)$ of the constraint functions are linearly dependent, but it does not provide any information concerning the possible extremal values.
- Lyusternik generalized this theorem for functions $f : D \to Y$ with arbitrary Banach spaces Y.
- Kuhn and Tucker proved sharper results when the functions f_i are *convex*.[9]

[8] See, e.g., Milnor [348] for more details.

[9] See, e.g., [285, Theorem 1.10].

Proof We consider the C^1 function

$$F := (f_0, \dots, f_n) : D \to \mathbb{R}^{n+1}$$

where $\mathbb{R}^{n+1}$ is endowed with the usual scalar product. It is sufficient to show that the linear map $F'(a) \in L(X, \mathbb{R}^{n+1})$ is *not* onto. Indeed, then there exists a non-zero vector $(\lambda_0, \dots, \lambda_n)$ that is orthogonal to its range M in $\mathbb{R}^{n+1}$, and hence (7.13) follows.[10]

Assume on the contrary that $M = \mathbb{R}^{n+1}$, and choose $e_0, \dots, e_n \in X$ such that

$$F'(a)e_0, \dots, F'(a)e_n$$

is a basis of $\mathbb{R}^{n+1}$. We claim that the restriction of F to $Y := \operatorname{vect}\{e_0, \dots, e_n\}$ satisfies at a the assumptions of the inverse function theorem.

Indeed, $F \in C^1$ in a neighborhood of a, and $F'(a) \in L(Y, \mathbb{R}^{n+1})$ is *onto* by our assumption. Moreover, since $\dim Y = \dim \mathbb{R}^{n+1}$, $F'(a)$ is an isomorphism.

By Theorem 7.7 the function $F : Y \hookrightarrow \mathbb{R}^{n+1}$ is a local diffeomorphism at a. Consequently, the formula[11]

$$x_n^{\pm} := F^{-1}\Big(f_0(a) \pm \frac{1}{n}, 0, \dots, 0\Big)$$

defines two sequences in Γ, converging to a and satisfying the inequalities

$$f_0(x_n^-) < f_0(a) < f_0(x_n^+)$$

if n is sufficiently large. But then $f_0|_\Gamma$ cannot have a local extremum at a. □

7.7 * Differential Equations: Dependence on the Initial Data

We return to the initial value problem

$$x' = f(t,x), \quad x(\tau) = \xi \tag{7.14}$$

investigated in Chap. 6. We assume that $f : D \to X$ satisfies the conditions of the Cauchy–Lipschitz theorem (p. 148). We fix a solution $x : J \to X$ (7.14) and a *compact* subinterval $I = [c, d]$ of J satisfying $c < \tau < d$.

[10] We may take, for example, $(\lambda_0, \dots, \lambda_n) := x - y$, where y is the orthogonal projection of an arbitrarily chosen point $x \in \mathbb{R}^{n+1} \setminus M$ in the sense of Proposition 3.12, p. 81. The subspace M is finite-dimensional, hence closed.

[11] The formula is meaningful if n is sufficiently large.

By Proposition 6.7 (p. 155) there exists a continuous function $g : U \to C(I,X)$ defined in some neighborhood U of ξ such that for each $\eta \in U$, $g(\eta)$ is the restriction to I of the maximal solution of the problem

$$y' = f(t,y), \quad y(\tau) = \eta.$$

In fact, g is more regular[12]:

Proposition 7.10 *The map g is continuously differentiable. If, moreover, $D_2^k f$ exists and is continuous for some $k \geq 2$, then $g \in C^k$.*

Proof

(i) Since the graph of $x|_I$ is compact, there exists an $r > 0$ such that $s \in I$ and $\|z - x(s)\| < r$ imply $(s,z) \in D$. Therefore the formula

$$F(\eta,y)(t) := y(t) - \eta - \int_\tau^t f(s,y(s))\, ds$$

defines a function $F : X \times V \to C(I,X)$ for some open neighborhood V of $g(\xi) = x|_I$ in $C(I,X)$. Since $F(\eta, g(\eta)) = 0$ for all $\eta \in U$, it remains to prove that F satisfies the conditions of the implicit function theorem (p. 182).

(ii) The map

$$V \ni u \mapsto D_2 f(\cdot, u(\cdot)) \in C(I,X)$$

is continuous. For otherwise there exist $u \in V$, $\varepsilon > 0$ and two sequences $(u_k) \subset V$ and $(t_k) \subset I$ such that $\|u_k - u\|_\infty \to 0$, and

$$\|D_2 f(t_k, u_k(t_k)) - D_2 f(t_k, u(t_k))\| \geq \varepsilon \tag{7.15}$$

for all k. Since the interval I is compact, we may assume by taking a subsequence that $t_k \to t \in I$. Then $u(t_k) \to u(t)$ and $u_k(t_k) \to u(t)$, so that the left-hand side of (7.15) tends to zero by the continuity of $D_2 f$ at $(t, u(t))$. This contradicts the choice of ε.

(iii) We claim that $F \in C^1$ in $X \times V$, and

$$F'(\eta,y)(\theta,z)(t) = z(t) - \theta - \int_\tau^t D_2 f(s,y(s))z(s)\, ds \tag{7.16}$$

for all

$$(\eta,y) \in X \times V, \quad (\theta,z) \in X \times C(I,X) \quad \text{and} \quad t \in I.$$

[12] Further results are given, for example, in Coddington–Levinson [112] and Pontryagin [398].

For each fixed $(\eta, y) \in X \times V$, denoting temporarily by $A(\theta, z)(t)$ the right-hand side of (7.16), the trivial estimate

$$\|A(\theta, z)\|_\infty \le \Big(1 + \int_I |D_2 f(s, y(s))| \; ds\Big) \cdot (\|\theta\| + \|z\|_\infty)$$

shows that $A \in L(X \times C(I, X), C(I, X))$.

Furthermore, if (θ, z) is sufficiently close to 0, then the following equality holds for all $t \in I$:

$$\begin{aligned} h(t) :&= F(\eta + \theta, y + z)(t) - F(\eta, y)(t) - A(\theta, z)(t) \\ &= -\int_\tau^t f(s, y(s) + z(s)) - f(s, y(s)) - D_2 f(s, y(s))z(s) \; ds \\ &= -\int_\tau^t \int_0^1 \big[D_2 f(s, y(s) + rz(s)) - D_2 f(s, y(s))\big] z(s) \; dr \; ds. \end{aligned}$$

Using (ii) we conclude for each fixed $(\eta, y) \in X \times V$ the relation

$$\|h\|_\infty = o(\|z\|_\infty) \quad \text{as} \quad \|z\|_\infty \to 0.$$

This estimate is even stronger than the estimate

$$\|h\|_\infty = o(|\theta| + \|z\|_\infty) \quad \text{as} \quad |\theta| + \|z\|_\infty \to 0$$

ensuring the differentiability.

Using (ii) again, the equality (7.16) also implies that F' is continuous at every point $(\eta, y) \in X \times V$.

(iv) We show that $D_2 F(\xi, g(\xi)) = D_2 F(\xi, x|_I)$ is invertible. For any fixed $h \in C(I, X)$ the equation $D_2 F(\xi, x|_I) z = h$ is equivalent to the linear integral equation

$$z(t) - \int_\tau^t D_2 f(s, x(s)) z(s) \; ds = h(t), \quad t \in I.$$

This equation has a unique solution by Proposition 6.4 (p. 153).[13]

(v) If $D_2^k f$ exists and is continuous for some $k \ge 2$, then a straightforward adaptation of the computation in (iii) above shows that $F \in C^k$ and

$$F^{(j)}(\eta, y)\big((\theta_1, z_1), \dots, (\theta_j, z_j)\big)(t) = -\int_\tau^t D_2^j f(s, y(s)) z_1(s) \cdots z_j(s) \; ds$$

for $j = 2, \dots, k$ and for all

$$(\eta, y) \in X \times V, \quad (\theta_1, z_1), \dots, (\theta_j, z_j) \in X \times C(I, X) \quad \text{and} \quad t \in I. \qquad \square$$

[13] We may choose $L(t) = \|D_2 f(t, x(t))\|$.

7.8 Exercises

Exercise 7.1 (Cayley–Hamilton Theorem) Let A be a real or complex matrix of order n with characteristic polynomial

$$\det(tI - A) = c_0 t^n + c_1 t^{n-1} + \cdots + c_n.$$

(i) Show that

$$\det(I - tA) = c_0 + c_1 t + \cdots + c_n t^n$$

for all $t \neq 0$.

(ii) We recall the identity

$$\det(I - tA)I = (I - tA)\,\mathrm{adj}(I - tA)^*$$

where $\mathrm{adj}\, B$ denotes the determinant formed by the algebraic adjugates of the matrix B.

Show that if t is sufficiently close to zero, then $I - tA$ has an inverse equal to $I + tA + t^2A^2 + \cdots$, so that

$$(I + tA + t^2A^2 + \cdots)(c_0 + c_1 t + \cdots + c_n t^n) = \mathrm{adj}(I - tA).$$

(iii) Show that the right-hand side is a matrix-coefficient polynomial of degree $\leq n - 1$, so that after development and simplification the left-hand side does not contain the power t^n.

(iv) Infer from the preceding assertion that

$$c_0 A^n + c_1 A^{n-1} + \cdots + c_n I = 0.$$

Exercise 7.2 (Tangent Lines) Determine the equations of the *tangent lines* at a given point (x_0, y_0) of the following curves:

(i) the *ellipse*

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \quad (a > b > 0);$$

(ii) the *hyperbola*

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1 \quad (a > b > 0);$$

(iii) the *parabola*

$$x^2 = 2py \quad (p > 0).$$

Exercise 7.3 (Folium of Descartes) Consider the equation[14]

$$x^3 + y^3 - 3xy = 0.$$

Assuming that y is a function of x, find the maxima and minima of $y(x)$.

Exercise 7.4

(i) Prove that the function $f : \mathbb{R} \to \mathbb{R}$ defined by the formula $f(0) := 0$ and

$$f(x) := x + 2x^2 \sin \frac{1}{x} \quad \text{otherwise}$$

has a bounded derivative at $(-1, 1)$, $f'(0) \neq 0$, but f is not monotone in any neighborhood of 0.

(ii) Prove that the formula

$$f(x, y) := (e^x \cos y, e^x \sin y)$$

defines a C^∞ function $f : \mathbb{R}^2 \to \mathbb{R}^2$ whose derivative does not vanish anywhere, but f is not one-to-one.

Why don't these examples contradict the inverse function theorem?

Exercise 7.5 (Inequalities) Prove the following inequalities by using Lagrange multipliers:

(i) (Euclid, Cauchy) The arithmetic and geometric means of any positive numbers $x_1, \dots, x_n$ satisfy the inequality

$$\sqrt[n]{x_1 x_2 \cdots x_n} \leq \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

(ii) (Young) If $p, q > 1$ are *conjugate exponents*, i.e., if $p^{-1} + q^{-1} = 1$, then

$$xy \leq \frac{x^p}{p} + \frac{x^q}{q}$$

for all $x, y > 0$.

(iii) (Hölder) Given two *conjugate exponents* $p, q > 1$, any positive numbers $x_1, \dots, x_n$ and $y_1, \dots, y_n$ satisfy the inequality

$$x_1 y_1 + x_2 y_2 + \cdots + x_n y_n \leq (x_1^p + x_2^p + \cdots + x_n^p)^{1/p} (y_1^q + y_2^q + \cdots + y_n^q)^{1/q}.$$

When do we have equality in the above inequalities?

[14] See Fig. 7.3.

Exercise 7.6 (Power Means) Given positive real numbers $a_1, \dots, a_n$, set $f(0) := \sqrt[n]{a_1 a_2 \dots a_n}$, and

$$f(x) := \left(\frac{a_1^x + a_2^x + \cdots + a_n^x}{n}\right)^{1/x} \quad \text{if} \quad x \in \mathbb{R} \setminus \{0\}\,.$$

Prove that f is non-decreasing and continuous.

Exercise 7.7 Solve the following problems:

(i) Among all right parallelepipeds of a given volume, which has the smallest surface area?

(ii) Among all triangles of a given perimeter, which has the largest area?

Part III
Approximation Methods

The computational needs of celestial mechanics and navigation led Napier [358] and Bürgi [68] to the discovery of the logarithm. Briggs [66] improved Napier's tables; in his work he employed the divided differences of Harriot [223] and interpolation.

Descartes's analytical geometry [125] increased the importance of the solution of algebraic equations.

Independently of some similar results of Briggs [66] and Gregory [208], Newton [367] generalized the binomial formula by expanding the function $(1+x)^{1/2}$ into a "Taylor series".

In his differential calculus, Newton [369] developed a general method for the numerical solution of not necessarily algebraic equations. It was subsequently studied by Raphson [403].

Newton [369] also obtained integral estimates by interpolation. His method was completed by Cotes [113].

Euler [146, 148] and Maclaurin [340] found a powerful way to estimate the sum of series of the form $\sum f(n)$ for regular functions f.

Euler [156] invented a classical method for the numerical solution of ordinary differential equations. It was improved by Runge [433]. Generalizing his idea, Heun [235] and Kutta [299] constructed efficient algorithms.

Gauss [187] improved the Newton–Cotes formulas by a clever choice of the interpolating abscissas. His results led to the rich theory of orthogonal polynomials, developed by Legendre [325], Jacobi [255], Chebyshev [95, 96], Hermite [232], Posse [399], Laguerre [313], Stieltjes [463], Markov [342, 343] and many others.

Completing the unpublished work of Fourier, Sturm [473] devised an elegant algorithm for the localization of the real roots of polynomials. Sturm-type polynomial sequences also appear in a natural way in the theory of orthogonal polynomials and when finding eigenvalues of symmetric matrices.

The following works contain rich historical accounts on this subject: [62, 72, 81, 100, 137, 197, 216, 217, 246, 277, 475, 493].

In this chapter we can only treat some selected topics. Many other questions are studied in the following books: [100, 108, 110, 121, 246, 254, 298, 316, 404, 476].

Chapter 8
Interpolation

In order to shorten tedious astronomical computations, for centuries multiplication was transformed into addition by using the trigonometrical identity

$$2\cos\alpha\cos\beta = \cos(\alpha+\beta) + \cos(\alpha-\beta).$$

The increasing pressure imposed by navigation led Napier and Bürgi to the independent invention of the logarithmic function, which yielded the more tractable identity

$$\log(ab) = \log a + \log b.$$

After twenty years of hard work, Napier published his logarithmic tables.

Briggs, greatly impressed by his achievement, with the agreement of the already old Napier, constructed much more precise logarithmic tables, in base 10. Thanks to his clever ideas on accelerating the computations, it took him much less time than before. As we will see in this chapter, his interpolation technique proved to be extremely fruitful in the later development of mathematics.

In this chapter we denote

- by $\mathcal{P}$ the vector space of real algebraic polynomials,
- by $\deg p$ the degree of a polynomial p, and
- by $\mathcal{P}_n$ the subspace $\{p \in \mathcal{P} \ : \ \deg p \le n\}$ of dimension $n+1$ for $n = 0, 1, \ldots.$

V. Komornik, *Topology, Calculus and Approximation*, Springer
Undergraduate Mathematics Series, DOI 10.1007/978-1-4471-7316-8_8

We recall that, counted with multiplicity, a non-zero polynomial has at most as many real roots as its degree.[1] It follows that if a polynomial $p \in \mathcal{P}_n$ has more than n real roots, then $p \equiv 0$.[2]

Until now the spaces $C^n(I)$ have only been defined for *open* intervals. It is useful to generalize them as follows:

Definition Given an arbitrary interval I and a nonnegative integer n, we denote by $C^n(I)$ the set of functions $f : I \to \mathbb{R}$ that may be extended to functions of class C^n, defined on some open interval containing I.

Equivalently, f belongs to C^n in the interior of I, and $f, f', \ldots, f^{(n)}$ may be extended continuously to I.[3]

We write simply $C(I)$ instead of $C^0(I)$.

In this chapter we assume that $I = [a, b]$ is a *non-degenerate compact* interval, i.e., $-\infty < a < b < \infty$.

8.1 Lagrange Interpolation

Given finitely many points $x_1, \ldots, x_n \in I$ and corresponding real numbers $y_1, \ldots, y_n$, $n \geq 1$, we seek a polynomial p of minimal degree, satisfying the equalities (see Fig. 8.1)

$$p(x_k) = y_k, \quad k = 1, \ldots, n. \tag{8.1}$$

Proposition 8.1 (Lagrange)

(a) *There exists a unique polynomial* $p \in \mathcal{P}_{n-1}$ *satisfying* (8.1).
(b) *For every* $f \in C(I)$ *there exists a unique polynomial* $p \in \mathcal{P}_{n-1}$ *satisfying*

$$p(x_k) = f(x_k), \quad k = 1, \ldots, n.$$

Definition The polynomial p is called the *Lagrange interpolating polynomial* of f associated with the points or *nodes* $x_1, \ldots, x_n$.

[1] We say that a is a root of multiplicity m of a function f if

$$f^{(j)}(a) = 0 \quad \text{for} \quad j = 0, \ldots, m-1, \quad \text{and} \quad f^{(m)}(a) \neq 0.$$

[2] We recall that in this book the symbol $\equiv$ means that equality holds for all points where both sides are defined.

[3] In higher dimensions this equivalence holds only for domains having a sufficiently regular boundary. See, e.g., Grisvard [211].

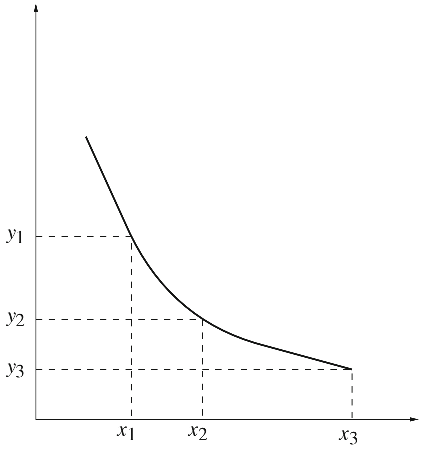


Fig. 8.1 Lagrange interpolation

Proof It is sufficient to prove (a): (b) then follows by choosing $y_k := f(x_k)$.

Existence. Introducing the *Lagrange basis polynomials*

$$\ell_k(x) := \prod_{\substack{1\le j\le n\\ j\ne k}} \frac{x-x_j}{x_k-x_j}, \quad k=1,\dots,n,$$

we have

$$\ell_k \in \mathcal{P}_{n-1} \quad \text{and} \quad \ell_k(x_j) = \delta_{kj} := \begin{cases} 1 & \text{if } k=j; \\ 0 & \text{if } k\ne j \end{cases}$$

for all k. Then $p := \sum_{k=1}^{n} y_k\ell_k$ belongs to $\mathcal{P}_{n-1}$ and satisfies (8.1).

Uniqueness. If $q \in \mathcal{P}_{n-1}$ satisfies

$$q(x_k) = y_k, \quad k=1,\dots,n,$$

then $p-q$ has more roots than its degree, because $\deg(p-q) \le n-1$ and $p-q$ vanishes at $x_1,\dots,x_n$. Hence $p-q=0$ and thus $p=q$. □

Remark Setting $\omega(x) := (x-x_1)\cdots(x-x_n)$ we have the more compact formula

$$\ell_k(x) = \frac{\omega(x)}{\omega'(x_k)(x-x_k)}$$

for each k and for every $x \ne x_k$.

The formula $Lf := p$ defines a linear *projection* $L : C(I) \to \mathcal{P}_{n-1}$, because $Lf = f$ for every $f \in \mathcal{P}_{n-1}$.

If f is sufficiently smooth, then Lf is a good approximation of f:

Theorem 8.2 (Cauchy) *Let $f \in C^n(I)$. For each $x \in I$ there exists a $\xi = \xi(x) \in I$ such that*

$$f(x) - (Lf)(x) = \frac{\omega(x)}{n!} f^{(n)}(\xi). \tag{8.2}$$

Consequently,

$$\|f - Lf\|_\infty \le \frac{\|\omega\|_\infty}{n!} \left\| f^{(n)} \right\|_\infty . \tag{8.3}$$

Proof It suffices to prove (8.2). If $x \in \{x_1, \dots, x_n\}$, then both sides vanish for all ξ. Henceforth we assume that $x \notin \{x_1, \dots, x_n\}$. Then the formula

$$g(y) := f(y) - (Lf)(y) - \frac{f(x) - (Lf)(x)}{\omega(x)} \omega(y) \tag{8.4}$$

defines a function $g \in C^n(I)$ that has at least $n + 1$ different roots in I: $g(x) = 0$, and $g(x_k) = 0$ for all $k = 1, \dots, n$.

A repeated application of Rolle's theorem yields that g' has at least n different roots, g'' has at least $n - 1$ different roots, and so on. Finally, $g^{(n)}(\xi) = 0$ for at least one point $\xi \in I$.

Differentiating (8.4) n times we obtain the identity

$$g^{(n)}(y) \equiv f^{(n)}(y) - \frac{f(x) - (Lf)(x)}{\omega(x)} n!$$

because $(Lf)^{(n)} \equiv 0$ and $\omega^{(n)} \equiv n!$. Choosing $y = \xi$ we obtain the required equality. □

Remark The estimate (8.3) is optimal in the sense that we have equality for every *polynomial* $f \in \mathcal{P}_n$ (and both sides are different from zero if $\deg f = n$), because $f - Lf$ is a multiple of ω by (8.2).

8.2 Minimization of Errors: Chebyshev Polynomials

The estimate (8.3) suggests the problem of seeking $x_1, \dots, x_n \in I$ so as to minimize the constant $\|\omega\|_\infty$. To simplify the notation we assume that $I = [-1, 1]$.[4]

[4] The general case follows by an affine change of variable.

Theorem 8.3 (Chebyshev) *We have*

$$\|\omega\|_\infty \ge 2^{1-n}$$

for every choice of points $x_1 > \cdots > x_n$ *in* $I := [-1, 1]$, *with equality if*

$$x_k = \cos\left(\frac{2k-1}{2n}\pi\right), \quad k = 1, \ldots, n. \tag{8.5}$$

For the proof we introduce the functions

$$T_n(x) := \cos(n \arccos x), \quad x \in [-1, 1], \quad n = 0, 1, \ldots; \tag{8.6}$$

see Figs. 8.2–8.7.

Lemma 8.4 *We have* $T_0(x) = 1$, $T_1(x) = x$, *and*

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \quad n = 1, 2, \ldots. \tag{8.7}$$

Consequently, T_n *is a polynomial of degree n. Finally, the equality*

$$T_n(x) = 2^{n-1} \prod_{k=1}^{n} \left(x - \cos\left(\frac{2k-1}{2n}\pi\right)\right) \tag{8.8}$$

holds for all $n \ge 1$ *and* $x \in [-1, 1]$.

Fig. 8.2 $T_1(x)$

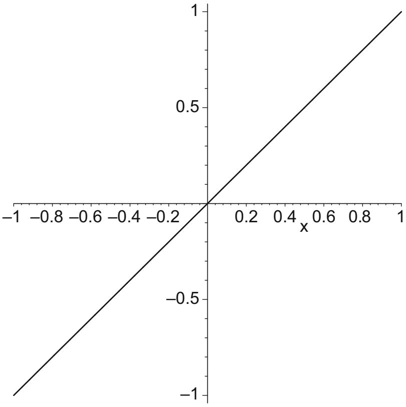


Fig. 8.3 $T_2(x)$

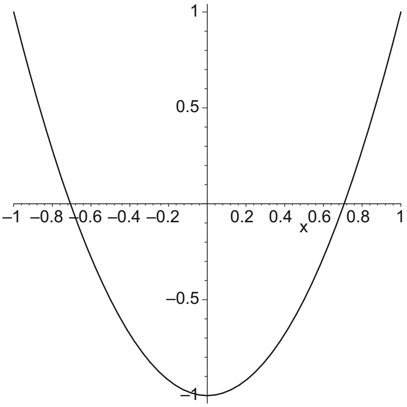


Fig. 8.4 $T_3(x)$

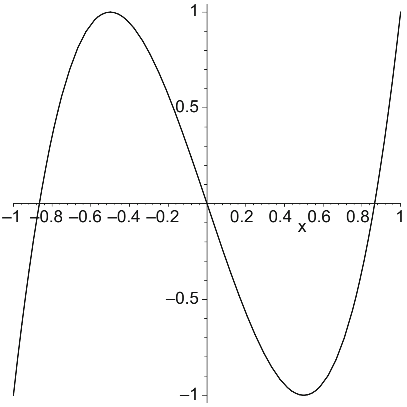


Fig. 8.5 $T_4(x)$

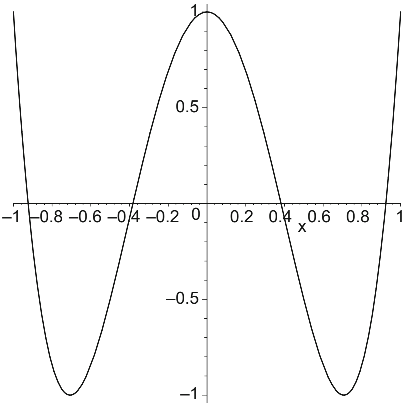


Fig. 8.6 $T_5(x)$

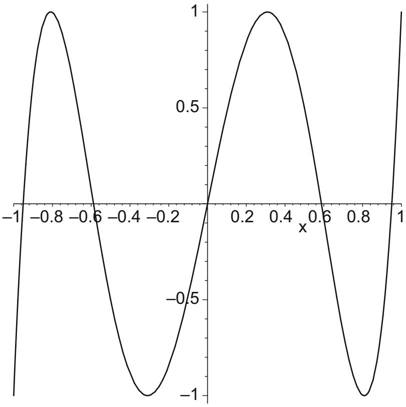
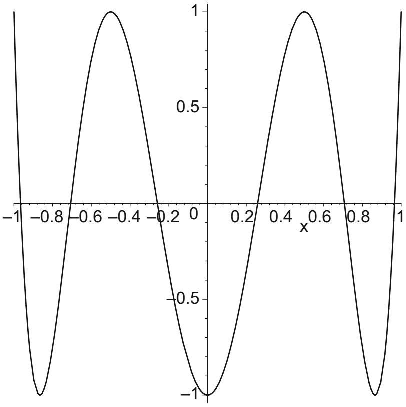


Fig. 8.7 $T_6(x)$



Proof The equalities $T_0(x) = 1$ and $T_1(x) = x$ are obvious. The relation (8.7) follows from the trigonometric identity

$$\cos(n+1)t = 2\cos t \cos(nt) - \cos(n-1)t$$

with the substitution $t := \arccos x$.

Using (8.7) we obtain by induction on n that T_n is a polynomial of degree n, and that its main coefficient is equal to 2^{n-1} for $n = 1, 2, \dots$. For (8.8) it remains to show that T_n vanishes at the n points

$$(1 >) \cos\frac{\pi}{2n} > \cos\frac{3\pi}{2n} > \cdots > \cos\Big(\frac{2n-1}{2n}\pi\Big)(> -1).$$

This can be checked by a direct computation: using (8.6) we have

$$T_n\Big(\cos\Big(\frac{2k-1}{2n}\pi\Big)\Big) = \cos\Big(\frac{2k-1}{2}\pi\Big) = (-1)^k \cos\frac{\pi}{2} = 0$$

for all $k = 1, \dots, n$. □

Proof of Theorem 8.3 It follows from the definition (8.6) of the Chebyshev polynomials that the polynomial $p := 2^{1-n}T_n$ satisfies $\|p\|_\infty = 2^{1-n}$. If we choose the points x_k according to (8.5), then $\omega = p$ by (8.8), so that $\|\omega\|_\infty = 2^{1-n}$. It remains to establish the inequality $\|\omega\|_\infty \ge \|p\|_\infty$ in the general case.

Assume on the contrary that $\|\omega\|_\infty < \|p\|_\infty$ for some point system (x_k), and consider the points

$$y_k = \cos\frac{k\pi}{n}, \quad k = 0, \dots, n.$$

Observe that

$$1 = y_0 > y_1 > \cdots > y_n = -1,$$

and that $p(y_k) = (-1)^k \|p\|_\infty$ for $k = 0, \dots, n$ by (8.6). Using the assumption $\|\omega\|_\infty < \|p\|_\infty$ we conclude that

$$\operatorname{sign}\,(p-\omega)(y_k) = (-1)^k, \quad k = 0, \dots, n.$$

Hence the polynomial[5] $p - \omega \in \mathcal{P}_{n-1}$ has at least one root in each of the pairwise disjoint intervals $(y_n, y_{n-1}), \dots, (y_1, y_0)$. Since it has more roots than its degree, we conclude that $p \equiv \omega$, contradicting our hypothesis $\|\omega\|_\infty < \|p\|_\infty$. □

8.3 Divided Differences: Newton's Interpolating Formula

The Lagrange interpolation formula has a drawback: when we add a new point, we have to recompute all the coefficients. For practical computation it is better to follow Newton's original suggestion to seek an interpolating polynomial of the form

$$(Lf)(x) = A_1 + A_2(x - x_1) + \cdots + A_n(x - x_1)\cdots(x - x_{n-1}). \tag{8.9}$$

Choosing $x = x_1, \dots, x_n$ we may determine consecutively the coefficients $A_1, A_2, \dots, A_n$.

We may also give an explicit formula for the coefficients as follows. Given k distinct points $x_1, \dots, x_k$, we introduce the polynomial

$$\omega_k(x) = (x - x_1)\cdots(x - x_k)$$

[5] Observe that the main terms of p and ω eliminate each other.

and the symmetric function

$$f(x_1,\dots,x_k) := \sum_{j=1}^{k} \frac{f(x_j)}{\omega_k'(x_j)}.$$

For $k = 1$ the latter coincides with the usual meaning of $f(x_1)$, while for $n = 2$ we have

$$f(x_1,x_2) = \frac{f(x_2) - f(x_1)}{x_2 - x_1}.$$

Proposition 8.5 (Newton)

(a) *The Lagrange interpolating polynomial is given by the formula*

$$(Lf)(x) = \sum_{k=1}^{n} f(x_1,\dots,x_k) \prod_{j=1}^{k-1} (x - x_j). \tag{8.10}$$

(b) *The coefficients in* (8.10) *may also be computed by the formulas of* divided differences*:*

$$f(x_1,\dots,x_k) = \frac{f(x_2,\dots,x_k) - f(x_1,\dots,x_{k-1})}{x_k - x_1}, \quad k = 2,3,\dots. \tag{8.11}$$

Proof

(a) Define the numbers A_k by (8.9), and introduce for each $m = 1,\dots,n$ the polynomial

$$p_m(x) := \sum_{k=1}^{m} A_k \prod_{j=1}^{k-1} (x - x_j). \tag{8.12}$$

Then $p_m \in \mathcal{P}_{m-1}$, and $p_m(x_k) = (Lf)(x_k) = f(x_k)$ whenever $k \le m$, so that p_m is the Lagrange interpolating polynomial of f associated with $x_1,\dots,x_m$. Hence, using the remark preceding Theorem 8.2,

$$p_m(x) = \sum_{k=1}^{m} f(x_k) \frac{\omega_m(x)}{\omega_m'(x_k)(x - x_k)} \tag{8.13}$$

for every $x \in \mathbb{R} \setminus \{x_1,\dots,x_m\}$. The equality $A_m = f(x_1,\dots,x_m)$ follows by comparing the main coefficients on the right-hand side of (8.12) and (8.13).

(b) We introduce the Lagrange interpolating polynomials $p, q \in \mathcal{P}_{k-2}$ defined by the conditions

$$p(x_j) = f(x_j),\ 1 \le j \le k-1 \quad \text{and} \quad q(x_j) = f(x_j),\ 2 \le j \le k,$$

and we define $r \in \mathcal{P}_{k-1}$ by the formula

$$r(x) := \frac{x - x_1}{x_k - x_1} q(x) + \frac{x_k - x}{x_k - x_1} p(x). \tag{8.14}$$

It is easy to check that $r(x_j) = f(x_j)$ for every $1 \le j \le k$, so that r is the Lagrange interpolating polynomial of f associated with $x_1, \dots, x_k$.

Using (a) the main coefficients[6] of p, q, r are equal to

$$f(x_1, \dots, x_{k-1}), \quad f(x_2, \dots, x_k) \quad \text{and} \quad f(x_1, \dots, x_k),$$

respectively. Using (8.14), (8.11) follows. □

Example Let $f(x) = x^3$, $n = 3$, $x_1 = 1$, $x_2 = 2$, $x_3 = 4$. The scheme

$$\begin{array}{c|lll}
x_1 & f(x_1) & & \\
 & & f(x_1, x_2) & \\
x_2 & f(x_2) & & f(x_1, x_2, x_3) \\
 & & f(x_2, x_3) & \\
x_3 & f(x_3) & &
\end{array}$$

makes it easy to determine the polynomial

$$(Lf)(x) = f(x_1) + f(x_1, x_2)(x - x_1) + f(x_1, x_2, x_3)(x - x_1)(x - x_2);$$

in the present case we may deduce from the scheme

$$\begin{array}{c|ccc}
1 & 1 & & \\
 & & 7 & \\
2 & 8 & & 7 \\
 & & 28 & \\
4 & 64 & &
\end{array}$$

that

$$(Lf)(x) = 1 + 7(x - 1) + 7(x - 1)(x - 2).$$

[6] More precisely, we take the coefficients of x^{k-2} in p, q and of x^{k-1} in r.

8.4 Hermite Interpolation

Consider n distinct points $x_1, \dots, x_n \in I$ and corresponding positive integers $m_1, \dots, m_n$. Given a set

$$\left\{ y_k^{(j)} \; : \; k = 1, \dots, n, \quad j = 0, \dots, m_k - 1 \right\}$$

of real numbers, we seek a polynomial p of minimal degree, satisfying the following conditions:

$$p^{(j)}(x_k) = y_k^{(j)}, \quad k = 1, \dots, n, \quad j = 0, \dots, m_k - 1. \tag{8.15}$$

The following generalization of Proposition 8.1 holds:

Proposition 8.6 (Hermite) *Put $m = m_1 + \cdots + m_n$.*

(a) *There exists a unique polynomial $p \in \mathcal{P}_{m-1}$ satisfying (8.15).*
(b) *If $f \in C^m(I)$, then there exists a unique polynomial $p \in \mathcal{P}_{m-1}$ such that*

$$p^{(j)}(x_k) = f^{(j)}(x_k), \quad k = 1, \dots, n, \quad j = 0, \dots, m_k - 1. \tag{8.16}$$

Proof As in Proposition 8.1, it is sufficient to prove (a). We have to show that the linear map $A : \mathcal{P}_{m-1} \to \mathbb{R}^m$, defined by the formula

$$Ap := (p(x_1), \dots, p^{(m_1-1)}(x_1), \dots, p(x_n), \dots, p^{(m_n-1)}(x_n)),$$

is bijective. Since

$$\dim \mathcal{P}_{m-1} = \dim \mathbb{R}^m < \infty,$$

it is sufficient to check that A is one-to-one, i.e., $Ap = 0 \Longrightarrow p = 0$.

If $p \in \mathcal{P}_{m-1}$ and $Ap = 0$, then x_k is a root of p with multiplicity $\geq m_k$. Hence p has at least $m_1 + \cdots + m_n = m$ roots, which is larger than its degree. We conclude that $p = 0$. □

Definition p is called the *Hermite interpolating polynomial* of f associated with $x_1, \dots, x_n$ and the multiplicities $m_1, \dots, m_n$.

The formula $Hf := p$ defines a linear projection $H : C^m(I) \to \mathcal{P}_{m-1}$.

Examples

- The case $m_1 = \cdots = m_n = 1$ corresponds to Lagrange interpolation.
- If $n = 1$, then Hf is the Taylor polynomial of order m_1 of f at x_1.

- We prove that if $m_1 = \cdots = m_n = 2$ then we have

$$p(x) = \sum_{k=1}^{n} y_k^{(0)} h_k^0(x) + y_k^{(1)} h_k^1(x),$$

where

$$\begin{aligned} h_k^1(x) &= (x - x_k)\ell_k(x)^2, \\ h_k^0(x) &= \big(1 - 2\ell_k'(x_k)(x - x_k)\big)\ell_k(x)^2 \\ &= \ell_k(x)^2 - 2\ell_k'(x_k)h_k^1(x), \end{aligned}$$

and $\ell_1, \ldots, \ell_k$ are the Lagrange basis polynomials:

$$\ell_k \in \mathcal{P}_{n-1}, \quad \ell_k(x_j) = \delta_{kj}.$$

Since $p \in \mathcal{P}_{2n-1}$, it is sufficient to check the relations

$$h_k^0(x_j) = (h_k^1)'(x_j) = \delta_{kj} \quad \text{and} \quad (h_k^0)'(x_j) = h_k^1(x_j) = 0.$$

A straightforward computation shows that

$$\begin{aligned} h_k^1(x_j) &= (x_j - x_k)\ell_k(x_j)^2 = (x_j - x_k)\delta_{kj} = 0, \\ h_k^0(x_j) &= \ell_k(x_j)^2 - 2\ell_k'(x_k)h_k^1(x_j) = \ell_k(x_j)^2 = \delta_{kj}, \\ (h_k^1)'(x_j) &= \ell_k^2(x_j) + 2(x_j - x_k)\ell_k(x_j)\ell_k'(x_j) \\ &= \delta_{kj} + 2(x_j - x_k)\delta_{kj}\ell_k'(x_j) = \delta_{kj}, \end{aligned}$$

and finally

$$\begin{aligned} (h_k^0)'(x_j) &= 2\ell_k(x_j)\ell_k'(x_j) - 2\ell_k'(x_k)(h_k^1)'(x_j) \\ &= 2\delta_{kj}\big(\ell_k'(x_j) - \ell_k'(x_k)\big) = 0. \end{aligned}$$

In order to generalize Theorem 8.2, we set

$$\Omega(x) := (x - x_1)^{m_1} \cdots (x - x_n)^{m_n}. \tag{8.17}$$

Theorem 8.7 (Stieltjes) *Let $f \in C^m(I)$. For each $x \in I$ there exists a point $\xi = \xi(x) \in I$ such that*

$$f(x) - (Hf)(x) = \frac{\Omega(x)}{m!} f^{(m)}(\xi). \tag{8.18}$$

Consequently,

$$\|f - Hf\|_\infty \le \frac{\|\Omega\|_\infty}{m!} \left\| f^{(m)} \right\|_\infty . \tag{8.19}$$

Remark The estimate (8.19) is again optimal: we have equality for every polynomial $f \in \mathcal{P}_m$ because $f - Hf$ is a multiple of Ω by (8.18), and the two sides are different from zero for polynomials of degree m.

For the proof we need a generalization of Rolle's theorem for multiple roots:

Lemma 8.8 *Counted with multiplicity,*[7] *if $g \in C^1(I)$ has at least $m + 1$ roots, then g' has at least m roots.*

Proof By hypothesis there exist points $x_1 < \cdots < x_n$ in I and natural numbers $m_1, \ldots, m_n$ such that $m_1 + \cdots + m_n = m + 1$, and x_k is a root of multiplicity $\ge m_k$ of g for each k. By Rolle's theorem g' has at least one root in each of the disjoint intervals $(x_1, x_2), \ldots, (x_{n-1}, x_n)$. Furthermore, each x_k is a root of multiplicity $\ge m_k - 1$ of g'. Summarizing, g' has at least

$$(n-1) + \sum_{k=1}^{n} (m_k - 1) = m_1 + \cdots + m_n - 1 = m$$

roots with multiplicity. □

Proof of Theorem 8.7 It suffices to prove (8.18). If $x \in \{x_1, \ldots, x_n\}$, then both sides of (8.18) vanish for any choice of ξ.

Assume henceforth that $x \notin \{x_1, \ldots, x_n\}$. Then $\Omega(x) \ne 0$, and we may define a function $g \in C^m(I)$ by setting

$$g(y) := f(y) - (Hf)(y) - \frac{f(x) - (Hf)(x)}{\Omega(x)} \Omega(y). \tag{8.20}$$

We claim that g has at least $m + 1$ roots (counted with multiplicity). Indeed, we have $g(x) = 0$; furthermore, each x_k is a root of multiplicity $\ge m_k$ of g by (8.16)

[7] See the definition of multiplicity in the footnote of page p. 193.

and (8.17). Since $g \in C^m(I)$, a repeated application of the preceding lemma yields $g^{(m)}(\xi) = 0$ for some $\xi \in I$.

On the other hand, differentiating (8.20) m times we get the equality

$$\begin{aligned} g^{(m)}(y) &\equiv f^{(m)}(y) - (Hf)^{(m)}(y) - \frac{f(x) - (Hf)(x)}{\Omega(x)} \Omega^{(m)}(y) \\ &\equiv f^{(m)}(y) - \frac{f(x) - (Hf)(x)}{\Omega(x)} m!. \end{aligned}$$

Choosing $y := \xi$ this yields

$$0 = g^{(m)}(\xi) = f^{(m)}(\xi) - \frac{f(x) - (Hf)(x)}{\Omega(x)} m!,$$

and this is equivalent to (8.18). □

Example Consider the following special case with $n = 3$ and $m = 4$:

$$I = [-r, r], \quad (x_1, x_2, x_3) = (-r, 0, r), \quad (m_1, m_2, m_3) = (1, 2, 1).$$

If $f \in C^4(I)$, then $Hf \in \mathcal{P}_3$ is defined by the conditions

$$(Hf)(x) = f(x) \quad \text{for} \quad x \in \{-r, 0, r\}, \quad \text{and} \quad (Hf)'(0) = f'(0). \tag{8.21}$$

By Theorem 8.7 we have

$$|(f - Hf)(x)| \le \frac{|\Omega(x)|}{4!} \left\| f^{(4)} \right\|_\infty$$

for every $x \in I$. Integrating this inequality in I and using the equality

$$\begin{aligned} \|\Omega\|_1 &= \int_{-r}^{r} (x + r)x^2(r - x)\, dx \\ &= \int_{-r}^{r} x^2(r^2 - x^2)\, dx \\ &= \frac{4r^5}{15}, \end{aligned}$$

we get the following estimate:

$$\left| \int_I f\, dx - \int_I Hf\, dx \right| \le \frac{r^5 \left\| f^{(4)} \right\|_\infty}{90}. \tag{8.22}$$

Let us compute the second integral. Since $Hf \in \mathcal{P}_3$, it can be written in the form

$$(Hf)(x) \equiv a + bx + cx(x+r) + dx(x+r)(x-r)$$

for suitable real numbers a, b, c, d. Then[8]

$$\int_{-r}^{r} Hf\, dx = \int_{-r}^{r} a + cx^2\, dx = 2ra + \frac{2r^3c}{3}, \tag{8.23}$$

and it remains to determine the coefficients a and c.

The first three conditions in (8.21) lead to the linear system

$$\begin{cases} f(-r) = (Hf)(-r) = a - br \\ f(0) = (Hf)(0) = a \\ f(r) = (Hf)(r) = a + br + 2cr^2; \end{cases}$$

solving it we get

$$a = f(0) \quad \text{and} \quad 2r^2c = f(-r) - 2f(0) + f(r).$$

Substituting these values into (8.23), and putting the result into (8.22), we obtain an error estimate of *Simpson's formula*:

$$\Bigl| \int_{-r}^{r} f\, dx - 2r \cdot \frac{f(-r) + 4f(0) + f(r)}{6} \Bigr| \le \frac{r^5 \, \|f^{(4)}\|_\infty}{90} \tag{8.24}$$

for all $f \in C^4([-r, r])$.

8.5 Theorems of Weierstrass and Fejér

One of the most important theorems in approximation theory is the following[9]:

Theorem 8.9 (Weierstrass) *For each $f \in C(I)$ there exists a sequence (p_n) of polynomials, converging uniformly to f in I.*

The theorem enables us to replace complicated functions by simpler polynomials in many investigations.

[8] The integrals of the odd powers of x vanish.

[9] We recall that in this chapter the letter I denotes always a *compact* interval.

There is a natural approach to proving this theorem. Choose n distinct points $x_1^n, \dots, x_n^n \in I$ for each $n = 1, 2, \dots$, and consider the corresponding Lagrange interpolation polynomials of f:

$$L_n f \in \mathcal{P}_{n-1}, \quad (L_n f)(x_k^n) = f(x_k^n), \quad 1 \le k \le n, \quad n = 1, 2, \dots.$$

We might hope that for some clever universal choice of the node system $\{x_k^n\}$, we have $\|f - L_n f\|_\infty \to 0$ as $n \to \infty$, for every $f \in C(I)$. After decades of fruitless efforts, this turned out to be impossible:

Theorem 8.10 (Faber) *For any given node system $\{x_k^n\}$ there exists an $f \in C(I)$ such that $\|f - L_n f\|_\infty \not\to 0$.*

***Remarks**

- Faber's proof was a new application of the method of *condensation of singularities,* introduced by Riemann, that eventually led to the fundamental Helly–Banach–Steinhaus theorem of Functional analysis. See, e.g., [285], Proposition 8.21 (b), for a proof based on this theorem.
- Erdős and Vértesi proved a far-reaching improvement of Faber's theorem: for any given node system $\{x_k^n\}$ there exists an $f \in C(I)$ such that

  $$\limsup |L_n f(x)| = \infty$$

  for almost all $x \in I$.[10] In particular, $(L_n f(x)) \subset \mathbb{R}$ is divergent for almost every $x \in I$.

In spite of all this, Fejér discovered that Weierstrass's theorem *may* be proved in this way, by using *Hermite* interpolation. Assume by an affine change of variable that $I = [-1, 1]$, and consider the Chebyshev nodes of Theorem 8.3 (p. 197):

$$x_k^n = \cos\left(\frac{2k-1}{2n}\pi\right), \quad 1 \le k \le n < \infty.$$

For any given $f \in C(I)$ define the Hermite interpolating polynomials by

$$h_n \in \mathcal{P}_{2n-1}, \quad h_n(x_k^n) = f(x_k^n), \quad h_n'(x_k^n) = 0, \ 1 \le k \le n$$

for each $n = 1, 2, \dots$; see Fig. 8.8.[11]

[10] We use the expressions "almost all" and "almost every" in Lebesgue's sense. See, e.g., Burkill [69], Halmos [220], Rudin [430], or [285].

[11] Fejér used the expression *step parabolas* to emphasize that their graphs have horizontal tangents at the nodes.

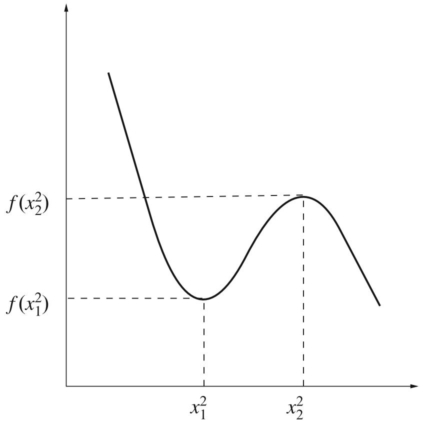


Fig. 8.8 Graph of h_2

Theorem 8.9 will readily follow from the following:

Theorem 8.11 (Fejér) *We have* $\|f - h_n\|_\infty \to 0$ *as* $n \to \infty$.

Proof From the example following Proposition 8.6 we have

$$\begin{aligned} h_n(x) &\equiv \sum_{k=1}^{n} f(x_k^n)\bigl(1 - 2(\ell_k^n)'(x_k^n)(x - x_k^n)\bigr)\ell_k^n(x)^2 \\ &=: \sum_{k=1}^{n} f(x_k^n)A_k^n(x). \end{aligned} \tag{8.25}$$

Assume temporarily the identity

$$A_k^n(x) \equiv \frac{(1 - x_k x)\cos^2(n \arccos x)}{n^2(x - x_k)^2}, \quad x \in I \setminus \{x_k\}. \tag{8.26}$$

Observe the following properties of the numbers $A_k^n(x)$ for all $x \in I$:

$$A_k^n(x) \ge 0, \qquad n^2(x - x_k^n)^2 A_k^n(x) \le 2 \qquad \text{and} \qquad \sum_{k=1}^{n} A_k^n(x) = 1.$$

Indeed, the first two follow from (8.26) and the inequalities $|x|, \left|x_k^n\right| \le 1$. For the third one we apply (8.25) with $f \equiv 1$, and we observe that $h_n \equiv f$ by the *definition* of Hermite interpolation.

For any fixed $f \in C(I)$ and $\varepsilon > 0$ we have to find a natural number N such that

$$\|f - h_n\|_\infty \le \varepsilon \tag{8.27}$$

for all $n \ge N$. Since f is uniformly continuous on the compact interval I, there exists a $\delta > 0$ such that

$$x, y \in I \quad \text{and} \quad |x - y| < \delta \Longrightarrow |f(x) - f(y)| \le \frac{\varepsilon}{2}.$$

Setting

$$K_{n,x} := \left\{1 \le k \le n : |x - x_k^n| < \delta\right\}$$

we have for every $x \in I$ the following estimate:

$$\begin{aligned}
|f(x) - h_n(x)| &= \Big|\sum_{k=1}^{n} A_k^n(x)\big(f(x) - f(x_k^n)\big)\Big| \\
&\le \sum_{k=1}^{n} A_k^n(x)|f(x) - f(x_k^n)| \\
&\le \sum_{k \in K_{n,x}} A_k^n(x)\frac{\varepsilon}{2} + \sum_{k \notin K_{n,x}} \frac{4}{n^2\delta^2}\,\|f\|_\infty \\
&\le \sum_{k=1}^{n} A_k^n(x)\frac{\varepsilon}{2} + \sum_{k=1}^{n} \frac{4}{n^2\delta^2}\,\|f\|_\infty \\
&= \frac{\varepsilon}{2} + \frac{4\,\|f\|_\infty}{n\delta^2}.
\end{aligned}$$

Hence

$$\|f - h_n\|_\infty \le \frac{\varepsilon}{2} + \frac{4\,\|f\|_\infty}{n\delta^2},$$

so that (8.27) holds by choosing

$$N \ge \frac{8\,\|f\|_\infty}{\varepsilon\delta^2}.$$

It remains to establish (8.26) for each fixed n. We omit for brevity the index n, and we write x_k and ℓ_k instead of x_k^n and ℓ_k^n. Using the notation

$$\omega(x) := (x - x_1)\cdots(x - x_n),$$

the equality

$$\ell_k(x) = \frac{\omega(x)}{\omega'(x_k)(x - x_k)}$$

holds for all $x \neq x_k$, whence

$$\ell_k'(x) = \frac{\omega'(x)(x - x_k) - \omega(x)}{\omega'(x_k)(x - x_k)^2}.$$

Applying L'Hospital's rule (Corollary 4.5, p. 106) we deduce that

$$\begin{aligned}\ell_k'(x_k) &= \lim_{x \to x_k} \frac{\omega'(x)(x - x_k) - \omega(x)}{\omega'(x_k)(x - x_k)^2} \\ &= \lim_{x \to x_k} \frac{\omega''(x)(x - x_k)}{2\omega'(x_k)(x - x_k)} = \frac{\omega''(x_k)}{2\omega'(x_k)}.\end{aligned}$$

We recall from Lemma 8.4 (p. 197) that

$$\omega(x) = 2^{1-n} \cos(n \arccos x)$$

for all $x \in [-1, 1]$. Hence

$$\omega'(x) = n2^{1-n} \sin(n \arccos x)(1 - x^2)^{-1/2},$$

for all $x \in (-1, 1)$, and therefore[12]

$$\omega'(x_k) = n2^{1-n}(-1)^{k-1}(1 - x_k^2)^{-1/2}$$

and

$$\omega''(x_k) = n2^{1-n} \sin(n \arccos x_k)(1 - x_k^2)^{-3/2} x_k.$$

Using these relations we conclude that

$$\ell_k'(x_k) = \frac{x_k}{2(1 - x_k^2)} \quad \text{and} \quad \ell_k(x)^2 = \frac{1 - x_k^2}{n^2(x - x_k)^2} \cos^2(n \arccos x).$$

[12] We use the relations $\arccos x_k = \frac{2k-1}{2n}\pi$.

Now (8.26) follows from (8.25):

$$\begin{aligned}\bigl(1-2\ell_k'(x_k)(x-x_k)\bigr)\ell_k(x)^2 \\ &= \Bigl(1-\frac{x_k(x-x_k)}{1-x_k^2}\Bigr)\frac{1-x_k^2}{n^2(x-x_k)^2}\cos^2(n\arccos x) \\ &= \frac{(1-x_kx)\cos^2(n\arccos x)}{n^2(x-x_k)^2}.\end{aligned}$$

□

There are many other proofs: see the comment on p. 346.

8.6 Spline Functions

We consider only a special case here.[13] Fix $n(\geq 2)$ real numbers $x_1 < \cdots < x_n$, and set $I = [x_1, x_n]$.

Definition (Schoenberg) By a *(cubic) spline* we mean a function $s \in C^2(I)$ satisfying the following conditions:

$$s|_{[x_j,x_{j+1}]} \in \mathcal{P}_3 \quad \text{for} \quad j = 1, \dots, n-1, \quad \text{and} \quad s''(x_1) = s''(x_n) = 0.$$

The splines have the Lagrange interpolation property:

Proposition 8.12 *For any given real numbers $y_1, \dots, y_n$ there exists a unique spline s such that*

$$s(x_j) = y_j, \quad j = 1, \dots, n. \tag{8.28}$$

First we prove a lemma:

Lemma 8.13 *Given a spline s, we denote by c_j the constant value of s''' in the open interval (x_j, x_{j+1}), $j = 1, \dots, n-1$. Then*

$$\int_I f''s''\,dx = \sum_{j=1}^{n-1} c_j\bigl(f(x_j) - f(x_{j+1})\bigr)$$

for all $f \in C^2(I)$.

[13] See, e.g., Laurent [316] for a general exposition.

Proof We integrate by parts and we use the conditions $s''(x_1) = s''(x_n) = 0$ as follows:

$$\begin{aligned}\int_I f''s''\,dx &= [f's'']_{x_1}^{x_n} - \int_I f's'''\,dx\\ &= -\int_{x_1}^{x_n} f's'''\,dx\\ &= -\sum_{j=1}^{n-1} c_j \int_{x_j}^{x_{j+1}} f'\,dx\\ &= \sum_{j=1}^{n-1} c_j\big(f(x_j) - f(x_{j+1})\big).\end{aligned}$$

□

Proof of Proposition 8.12 The splines are given by the formula

$$s(x) = a_i x^3 + b_i x^2 + c_i x + d_i, \quad x_i \le x \le x_{i+1},\ i = 1, \dots, n-1,$$

where the parameters a_i, b_i, c_i, d_i satisfy the compatibility conditions

$$s^{(k)}(x_i - 0) = s^{(k)}(x_i + 0), \quad i = 2, \dots, n-1, \quad k = 0, 1, 2 \tag{8.29}$$

and

$$s''(x_1) = s''(x_n) = 0. \tag{8.30}$$

Thus we have to show that the *linear* system (8.28), (8.29), (8.30) has a unique solution a_i, b_i, c_i, d_i.

Since the number $4n - 4$ of equations is equal to the number of unknowns, it is sufficient to show that the homogeneous system has only the trivial solution. Equivalently, we have to show that if a spline s satisfies

$$s(x_1) = \cdots = s(x_n) = 0, \tag{8.31}$$

then it vanishes identically.

For this we apply the preceding lemma with $f := s$. Using also (8.31) we obtain the equality

$$\int_I (s'')^2\,dx = \sum_{j=1}^{n-1} c_j\big(s(x_j) - s(x_{j+1})\big) = 0.$$

Hence $s'' \equiv 0$, i.e., s is an affine function in I. Since by (8.31) it has at least two roots, we have necessarily $s \equiv 0$. □

The importance of splines comes from the fact that they are the smoothest solutions of the interpolation problem

$$g(x_j) = y_j, \quad j = 1, \ldots, n \tag{8.32}$$

in the following sense:

Proposition 8.14 (Holladay) *Let s be a spline satisfying*

$$s(x_j) = y_j, \quad j = 1, \ldots, n, \tag{8.33}$$

and $g \in C^2(I)$ an arbitrary function satisfying (8.32). Then

$$\int_I |g''|^2 \, dx \ge \int_I |s''|^2 \, dx,$$

with equality only if $g = s$.

Proof Applying Lemma 8.13 with $f = g - s$ and using (8.32), (8.33), we obtain that $\int_I (g-s)'' s'' \, dx = 0$. Therefore

$$\int_I |g''|^2 \, dx = \int_I |(g-s)'' + s''|^2 \, dx = \int_I |(g-s)''|^2 + |s''|^2 \, dx.$$

This proves the inequality, and that equality holds only if $(g-s)'' \equiv 0$. In the last case $g - s$ is affine in I, and it has at least $n \ge 2$ roots by (8.32) and (8.33), so that $g - s \equiv 0$. □

8.7 Exercises

Exercise 8.1 Prove the formula for the Lagrange basis polynomials given in the remark on p. 195.

Exercise 8.2 If p is a polynomial of degree d, then $(p(k))$ is called a *generalized arithmetic sequence* or an *arithmetic sequence of order d*.

Given a sequence (a_k) we define the difference sequences $(\Delta_k^{(0)}), (\Delta_k^{(1)}), \ldots$ by the recurrence relations

$$\Delta_k^{(0)} := a_k, \quad \text{and} \quad \Delta_k^{(j+1)} := \Delta_{k+1}^{(j)} - \Delta_k^{(j)}, \quad j = 0, 1, \ldots .$$

Prove the following:

(i) A non-zero sequence (a_k) is an arithmetic sequence of order $\leq d \iff \Delta_k^{(d+1)} \equiv 0$.

(ii) If (a_k) is a generalized arithmetic sequence, then

$$a_k = \sum_{j=0}^{k} \binom{k}{j} \Delta_0^{(j)}, \quad k = 0, 1, \ldots .$$[14]

(iii) Apply this formula to compute $1^2 + \cdots + k^2$ and $1^3 + \cdots + k^3$.

Exercise 8.3 Generalize Lagrange interpolation to several variables as follows.[15] Fix two integers $n, k \geq 1$ and consider in $\mathbb{R}^n$ the *simplex*

$$K := \{x = (x_1, \ldots, x_n) \in \mathbb{R}^n \ : \ x_1 \geq 0, \ldots, x_n \geq 0, \ x_1 + \cdots + x_n \leq k\} .$$

We denote by Σ the set of points $x \in K$ with integer coordinates, and by $\mathcal{F}$ the vector space of functions $g : \Sigma \to \mathbb{R}$. Finally, we denote by $\mathcal{P} = \mathcal{P}_{n,k}$ the vector space of polynomials of variables $x_1, \ldots, x_n$, of total degree $\leq k$.

The purpose of this exercise is to show that each $g \in \mathcal{F}$ has a unique extension $p \in \mathcal{P}$.

(i) Given $g \in \mathcal{F}$ arbitrarily, show that the formula

$$p(x_1, \ldots, x_n) := \sum_{\ell \in \Sigma} g(x_1, \ldots, x_n) \prod_{j=0}^{n} \prod_{i=0}^{\ell_j - 1} \frac{x_j - i}{\ell_j - i}$$

with

$$x_0 := k - x_1 - \cdots - x_n, \quad \ell_0 := k - \ell_1 - \cdots - \ell_n$$

defines an extension $p \in \mathcal{P}$ of g.

(ii) Show that the map

$$(\ell_1, \ldots, \ell_n) \mapsto x_1^{\ell_1} \cdots x_n^{\ell_n}$$

is a bijection between Σ and a basis of $\mathcal{P}$.

(iii) Prove that $\mathcal{F}$ and $\mathcal{P}$ have the same (finite) dimension.

(iv) Conclude.

[14] If (a_k) is of order d, then for $j > d$ the binomial coefficients vanish.

[15] The result of this exercise is the starting point of the *finite element method* for the numerical solution of partial differential equations. See, e.g., Ciarlet [107], Raviart–Thomas [405].

Exercise 8.4 (Cauchy) Consider the divided differences for a function $f \in C^{n-1}(I)$.

(i) Prove that

$$f(x_1, \ldots, x_n) = \frac{f^{(n-1)}(\xi)}{(n-1)!}$$

for some $\xi \in (\min x_i, \max x_i)$.

(ii) Show that

$$f(x_1, \ldots, x_n) \to \frac{f^{(n-1)}(x)}{(n-1)!}$$

as $x_1, \ldots, x_n \to x$.

Exercise 8.5 (Theorems of Eudoxos and Archimedes)

(i) Consider a conical frustum created by slicing the top off a right circular cone (with the cut made parallel to the base). Prove that its volume may be computed by Simpson's formula

$$V = \frac{h}{6}(A + 4B + C) \tag{8.34}$$

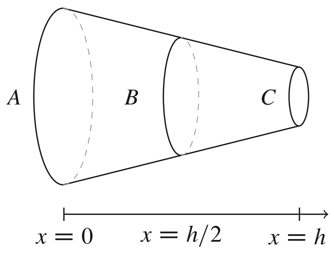
where h, A, C, B denote the height of the frustum, its base areas and and the area of the parallel middle section, respectively. See Fig. 8.9.

Simplify (8.34) to

$$V = \frac{h}{3}(A + \sqrt{AC} + C). \tag{8.35}$$

(ii) Consider a spherical segment obtained by cutting a ball with a pair of parallel planes. Prove that its volume may be computed by Simpson's formula (8.34) where h, A, C, B denote the width of the spherical segment, its base areas and and the area of the parallel middle section, respectively.

Fig. 8.9 Conical frustum



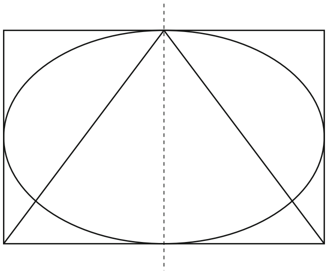


Fig. 8.10 Theorem of Archimedes

Show that for $A = 0$ the formula reduces to

$$V = \pi h^2 \left(R - \frac{h}{3} \right) \tag{8.36}$$

where R denotes the radius of the sphere.

(iii) Let us inscribe into a rectangle an isosceles triangle and an ellipse having the same symmetry axis as shown on Fig. 8.10. Determine the proportion of the corresponding volumes of revolution.

Exercise 8.6 (Painlevé) Generalize Weierstrass' theorem: if $f \in C^k(I)$ for some positive integer k, then there exists a sequence (p_n) of polynomials such that $p_n^{(j)}$ converges uniformly to $f^{(j)}$ for each $j = 0, 1, \dots, k$.

Chapter 9
Orthogonal Polynomials

The Chebyshev polynomials have an interesting orthogonality property. Namely, using the orthogonality of the trigonometrical functions $\cos nt$ on $(0, \pi)$, we obtain by the change of variable $t = \arccos x$ that $T_n(x) = \cos nt$, $dt/dx = -(1-x^2)^{-1/2}$, and therefore

$$\int_{-1}^{1} T_n(x)T_k(x)(1-x^2)^{-1/2}\,dx = \int_0^{\pi} \cos nt \cos kt\,dt = 0$$

for all $n \neq k$.

There are many similar orthogonal polynomial sequences for other *weight functions* $w(x)$ in place of $(1-x^2)^{-1/2}$. We give here an introduction to this theory.

In this chapter we denote by I an arbitrary non-degenerate interval.

9.1 Gram–Schmidt Orthogonalization

We recall that using orthogonal coordinates, i.e., orthogonal bases, we may simplify many computations in analytical geometry. We show that similar bases exist in infinite-dimensional Euclidean spaces, too.

In this section we denote by E a Euclidean space endowed with the scalar product (x, y) and the associated norm $\|x\| := (x, x)^{1/2}$.

Definitions Let $(x_n)_{n\geq 0}$ be a sequence of vectors in E.

- The sequence (x_n) *is linearly independent* if for any *finite* sequence $\alpha_0, \dots, \alpha_k$ of real numbers the following implication holds:

$$\alpha_0 x_0 + \cdots + \alpha_k x_k = 0 \Longrightarrow \alpha_0 = \cdots = \alpha_k = 0.$$

V. Komornik, *Topology, Calculus and Approximation*, Springer
Undergraduate Mathematics Series, DOI 10.1007/978-1-4471-7316-8_9

- The sequence (x_n) is *orthogonal* if

$$i \neq j \Longrightarrow (x_i, x_j) = 0.$$

Proposition 9.1 (Gram–Schmidt Orthogonalization) *Let* $x_0, x_1, \dots$ *be a sequence of* non-zero *vectors in E.*

(a) *If* (x_n) *is orthogonal, then it is also linearly independent.*
(b) *If* (x_n) *is linearly independent, then there exists a unique orthogonal sequence* (y_n) *of the form*

$$y_n = x_n - \sum_{k=0}^{n-1} \alpha_k^n x_k, \quad \alpha_k^n \in \mathbb{R}, \quad n = 0, 1, \dots. \tag{9.1}$$

Furthermore, $(y_n, x_k) = 0$ *whenever* $n > k$.

Proof

(a) If $x := \alpha_0 x_0 + \dots + \alpha_k x_k = 0$, then

$$0 = \|x\|^2 = (x, x) = \sum_{i,j=0}^{k} \alpha_i \alpha_j (x_i, x_j) = \sum_{i=0}^{k} \alpha_i^2 \|x_i\|^2 .$$

Hence $\alpha_0 = \dots = \alpha_k = 0$ because $\|x_i\| > 0$ for all i.
(b) Let z_{n-1} be the orthogonal projection[1] of x_n onto the finite-dimensional and hence closed subspace vect $\{x_0, \dots, x_{n-1}\}$. Then $y_n := x_n - z_{n-1}$ has the form (9.1), and satisfies $(y_n, x_k) = 0$ for $k = 0, \dots, n-1$. Since $y_k \in$ vect $\{x_0, \dots, x_{n-1}\}$ if $k < n$, it follows that $(y_n, y_k) = 0$.

Conversely, if (y_n) is an orthogonal sequence of the form (9.1), then we have necessarily $y_0 = x_0$, and then we see by induction that each x_k is a linear combination of $y_0, \dots, y_k$, so that $(y_n, x_k) = 0$ for all $k < n$. Hence y_n is orthogonal to vect $\{x_0, \dots, x_{n-1}\}$. The latter contains $x_n - y_n$, so that it is necessarily the (unique) orthogonal projection of x_n onto vect $\{x_0, \dots, x_{n-1}\}$. □

9.2 Orthogonal Polynomials

We introduce a class of Euclidean spaces.[2]

[1] See Proposition 3.12, p. 81.

[2] See Szegő [476] for a more general class.

Definition By a *weight function* we mean a continuous, positive function w defined on a non-degenerate interval I, for which all integrals

$$\int_I x^n w(x)\,dx, \quad n = 0, 1, \ldots$$

converge.[3]

Remark It follows from the definition that the integral $\int_I qw\,dx$ converges for all polynomials q, too. If I is *compact*, then this is obvious because $qw \in C(I)$.

For the rest of this section we fix a weight function $w : I \to \mathbb{R}$, and we introduce the scalar product

$$(p, q) := \int_I pqw\,dx$$

on the vector space $\mathcal{P}$ of polynomials.

Proposition 9.2 *There exists a unique orthogonal sequence of polynomials $p_0, p_1, \ldots$ such that the leading term of $p_n(t)$ is t^n.*

Furthermore, $(p_n, q) = 0$ for all polynomials q of degree $< n$.

Proof We apply Proposition 9.1 (b) to the sequence $x_n(t) := t^n$. The latter sequence is linearly independent because non-zero polynomials have only finitely many roots, and therefore do not vanish identically in I. □

Examples

- If $\alpha, \beta > -1$, then the formula

$$w(x) := (1 - x)^\alpha (1 + x)^\beta$$

  defines a weight function on the interval $I = (-1, 1)$. The corresponding polynomials given by Proposition 9.2 are called the *Jacobi polynomials* of indices (α, β). In the special cases $\alpha = \beta = 0, -\frac{1}{2}, \frac{1}{2}$ we get the *Legendre polynomials* and the *Chebyshev polynomials of the first and second kind, respectively*. We have already encountered the Chebyshev polynomials of the first kind during the proof of Theorem 8.3 (p. 197); the Chebyshev polynomials of the second kind play a similar role in Proposition 10.2 below (p. 233).
- If $\alpha > -1$, then the formula

$$w(x) := x^\alpha e^{-x}$$

[3] In other words, the integrals exist and are finite.

defines a weight function on the interval $I = (0, \infty)$. The corresponding polynomials given by Proposition 9.2 are called the *Laguerre polynomials*.

- The weight function

$$w(x) := e^{-x^2}$$

leads to the *Hermite polynomials* on $I = \mathbb{R}$ in a similar way.

Remark The above *classical orthogonal polynomials* play an important role in many problems of practical interest. See the exercises at the end of this chapter for some of their interesting properties.[4]

The following recursive formula simplifies the computation of orthogonal polynomials: each element is already determined by the two preceding ones.

Proposition 9.3 (Stieltjes) *The orthogonal polynomials satisfy the recurrence relations*

$$p_{n+1}(x) \equiv (x - \mu_n)p_n(x) - \lambda_n p_{n-1}(x), \quad n = 1, 2, \ldots$$

with

$$\mu_n = \frac{\int_I x p_n(x)^2 w(x)\, dx}{\int_I p_n(x)^2 w(x)\, dx} \quad \textit{and} \quad \lambda_n = \frac{\int_I p_n(x)^2 w(x)\, dx}{\int_I p_{n-1}(x)^2 w(x)\, dx}.$$

Remark We have $p_0(x) \equiv 1$ and $p_1(x) \equiv x - \mu_0$ with

$$\mu_0 := \frac{\int_I x w(x)\, dx}{\int_I w(x)\, dx}$$

by a straightforward computation. After this, $p_2, p_3, \ldots$ may be computed recursively by using the formula of the proposition.

Proof Since $p_0, \ldots, p_n$ form a basis of $\mathcal{P}_n$ and $xp_n(x) - p_{n+1}(x) \in \mathcal{P}_n$ (the leading terms eliminate each other), we have

$$xp_n(x) - p_{n+1}(x) \equiv \sum_{i=0}^{n} c_i p_i(x) \tag{9.2}$$

[4] Courant–Hilbert [114], Jackson [254], Natanson [359] and Szegő [476] contain many additional results.

for suitable real numbers $c_0, \dots, c_n$. It remains to show that $c_k = 0$ if $k \le n-2$, and that

$$c_n = \frac{\int_I xp_n(x)^2 w(x)\ dx}{\int_I p_n(x)^2 w(x)\ dx} \qquad \text{and} \qquad c_{n-1} = \frac{\int_I p_n(x)^2 w(x)\ dx}{\int_I p_{n-1}(x)^2 w(x)\ dx}.$$

Multiplying (9.2) by $p_k(x)w(x)$ for $k = 0, \dots, n$, and using the orthogonality of the polynomials p_j, we obtain the equalities

$$\int_I xp_n(x)p_k(x)w(x)\ dx = \sum_{i=0}^{n} c_i(p_i, p_k) = c_k \left\|p_k\right\|^2. \tag{9.3}$$

Choosing $k = n$ this yields $c_n = \mu_n$.

If $k = n-1$, then $xp_k(x) - p_n(x)$ belongs to $\mathcal{P}_{n-1}$, and hence it is orthogonal to p_n. Hence we can change $xp_k(x)$ to $p_n(x)$ on the left-hand side of (9.3); this yields $c_{n-1} = \lambda_n$.

Finally, for $k \le n-2$ the polynomial $xp_k(x)$ belongs to $\mathcal{P}_{n-1}$, and hence is orthogonal to p_n. Therefore the left-hand side of (9.3) vanishes. Since $\|p_k\| > 0$, we conclude that $c_k = 0$. □

9.3 Roots of Orthogonal Polynomials

The roots of orthogonal polynomials have surprising properties.

Proposition 9.4 (Stieltjes) *Let $p_0, p_1, \dots$ be a sequence of orthogonal polynomials.*

(a) *p_n has n distinct simple real roots, all lying in the interior of I.*
(b) *The roots of consecutive polynomials separate each other*[5]*: if*

$$x_1 < \cdots < x_n \quad \text{and} \quad y_1 < \cdots < y_{n-1}$$

are the roots of p_n and p_{n-1}, respectively, then

$$x_1 < y_1 < x_2 < \cdots < x_{n-1} < y_{n-1} < x_n. \tag{9.4}$$

Proof

(a) Since p_n has at most n roots, it suffices to prove that it changes sign at least n times in the interval I. Assume on the contrary that p_n changes sign only $k < n$

[5] See the graphs of some Legendre polynomials in see Figs. 9.1–9.4.

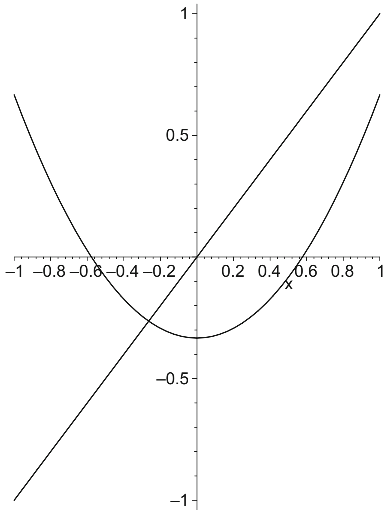


Fig. 9.1 $n = 1, 2$

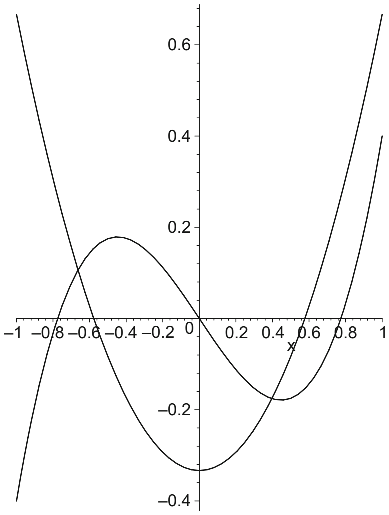


Fig. 9.2 $n = 2, 3$

times inside I (we do not exclude the case $k = 0$), and let $x_1 < \cdots < x_k$ be the corresponding roots of p_n. Then the polynomial

$$q(x) := (x - x_1) \cdots (x - x_k) \qquad (q(x) := 1 \quad \text{if} \quad k = 0)$$

has degree $k < n$, and therefore it is orthogonal to p_n. This is, however, impossible, because the non-zero product function $p_n q w$ does not change sign in I and therefore $\int_I p_n q w \, dx \neq 0$.

Fig. 9.3 $n = 3, 4$

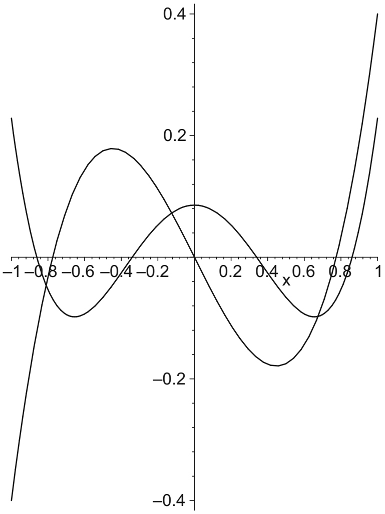
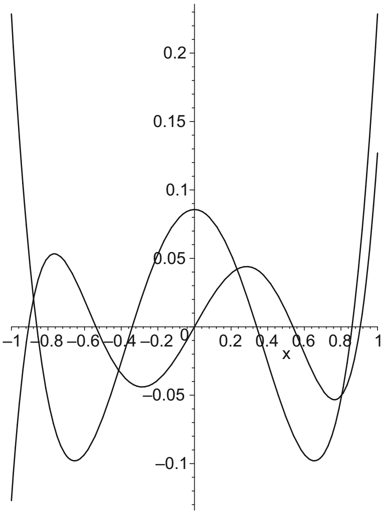


Fig. 9.4 $n = 4, 5$



(b) First we prove the following property:

$$p_k(\alpha) = 0 \Longrightarrow p_{k+1}(\alpha)p_{k-1}(\alpha) < 0. \tag{9.5}$$

In particular, consecutive orthogonal polynomials have no common roots.

Indeed, since p_0 has no root, $p_k(\alpha) = 0$ may hold only if $k \geq 1$. Then we deduce from the recurrence relations of Proposition 9.3 that

$$p_{k+1}(\alpha)p_{k-1}(\alpha) = -\lambda_k p_{k-1}(\alpha)^2 \leq 0,$$

because $\lambda_k > 0$ from the formula of Proposition 9.3.

For the strict inequality it remains to show that $p_{k-1}(\alpha) \neq 0$. If $p_k(\alpha) = p_{k-1}(\alpha) = 0$ then we would have $p_{k-2}(\alpha) = 0$ from the recursion formula, and we would obtain successively $p_j(\alpha) = 0$ for all $j < k$. This, however, is impossible because p_0 has no root.

Property (9.4) holds (formally) for $n = 1$. Assume by induction that it holds for some $n \geq 1$. Then

$$\operatorname{sign} p_{n-1}(x_k) = (-1)^{n-k}, \quad k = 1, \dots, n,$$

because p_{n-1} changes sign at the points y_j by (a), and $p_n(\infty) = \infty$ (the leading coefficient is positive). Using (9.5) this implies that

$$\operatorname{sign} p_{n+1}(x_k) = (-1)^{n+1-k}, \quad k = 1, \dots, n.$$

Since[6]

$$p_{n+1}(\infty) = \infty \quad \text{and} \quad \operatorname{sign} p_{n+1}(-\infty) = (-1)^{n+1},$$

p_{n+1} changes sign in each of the disjoint open intervals

$$(-\infty, x_1), (x_1, x_2), \dots, (x_{n-1}, x_n), (x_n, \infty), \tag{9.6}$$

and therefore it has a root in each of them. Since the number of intervals is equal to $\deg p_{n+1}$, none of the intervals may contain more than one root, and p_{n+1} cannot have other roots. □

Remarks

- The proof of (b) also provides a second proof of (a).
- In the proof of (b), besides (9.5) we only need that p_0 has no real root, and that $\deg p_n \leq n$ for all n.

Later we will need the following result:

Corollary 9.5 *Let (p_n) be an orthogonal sequence of polynomials. Then*

$$p_0 \quad \textit{has no real root}; \tag{a}$$

$$p_k(\alpha) = 0 \Longrightarrow p_{k+1}(\alpha) p_{k-1}(\alpha) < 0; \tag{b}$$

$$p_n(\alpha) = 0 \Longrightarrow (p_n p_{n-1})'(\alpha) > 0. \tag{c}$$

[6] This follows from (a), but it is sufficient to recall that the number of non-real roots of a real polynomial is even.

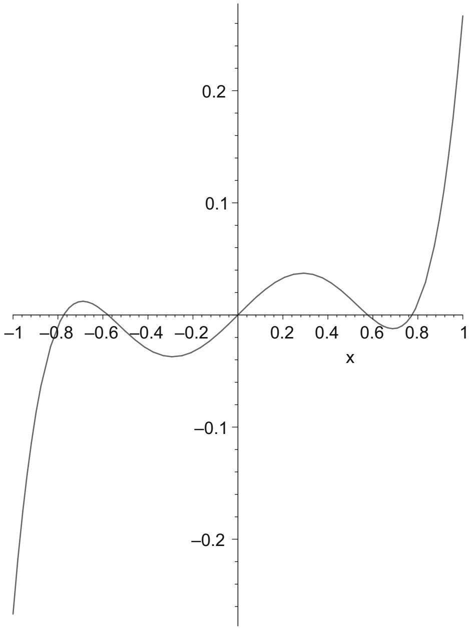


Fig. 9.5 Graph of p_2p_3

Proof

(a) is obvious because $p_0 \equiv 1$.

(b) was established during the above proof.

(c) The product polynomial p_np_{n-1} changes sign at each root, because the roots are simple by the preceding proposition, and therefore $(p_np_{n-1})'$ is alternately positive and negative at these points. Since the roots of p_n and p_{n-1} are alternating, using the notation (9.4) we conclude that $(p_np_{n-1})'(x_k)$ has the same sign for all k. (See Fig. 9.5 for the Legendre polynomials.) To finish the proof we observe that $(p_np_{n-1})'(x_n) > 0$, because the polynomial p_np_{n-1} is positive for all $x > x_n$. □

9.4 Exercises

Exercise 9.1 Let w be an *even* weight function on $I = (-1, 1)$ and consider the corresponding orthogonal polynomials. Show that p_n is even if n is even, and p_n is odd if n is odd.

Exercise 9.2 Is the sequence of polynomials $1, x, x^2, \dots$ orthogonal for a suitable weight function on some interval?

Exercise 9.3 Consider the Legendre polynomials $p_n(x) = x^n + \cdots$ corresponding to the weight function $w \equiv 1$ on $I = (-1, 1)$.

(i) Show that $q_n := \dfrac{d^n}{dx^n}(x^2 - 1)^n$ is a polynomial of degree n, and the sequence (q_n) is orthogonal.
(ii) Deduce from (i) that $p_n = a_n q_n$ for all n with suitable coefficients a_n.
(iii) Show that $y = q_n$, and hence $y = p_n$ satisfies the differential equation

$$(1 - x^2)y'' - 2xy' + n(n + 1)y = 0.$$

Exercise 9.4 Generalizing the preceding exercise, consider the Jacobi polynomials $p_n(x) = x^n + \cdots$ corresponding to the weight function $w(x) = (1 - x)^\alpha (1 + x)^\beta$ on $I = (-1, 1)$.

(i) Show that

$$q_n := (1 - x)^{-\alpha}(1 + x)^{-\beta} \frac{d^n}{dx^n}\left[(1 - x)^{n+\alpha}(1 + x)^{n+\beta}\right]$$

is a polynomial of degree n, and the sequence (q_n) is orthogonal with respect to w.
(ii) Deduce from (a) that $p_n = a_n q_n$ for all n with suitable non-zero coefficients a_n.
(iii) Show that $y = p_n$ satisfies the differential equation

$$(1 - x^2)y'' + (\beta - \alpha - (\alpha + \beta + 2)x)y' + n(n + \alpha + \beta + 1)y = 0.$$

Exercise 9.5

(i) Prove that for $\alpha = \beta = -1/2$ the Jacobi polynomials reduce to the Chebyshev polynomials[7]:

$$p_n(x) = 2^{1-n} \cos(n \arccos x), \quad n = 1, 2, \dots.$$

(ii) Prove that for $\alpha = \beta = 1/2$ the Jacobi polynomials are given by the formulas

$$p_n(x) = 2^{-n} \frac{\sin((n + 1) \arccos x)}{\sin(\arccos x)}, \quad n = 0, 1, \dots.$$

They are called the *Chebyshev polynomials of the second kind.*

[7] See the formula (8.6), p. 197.

Exercise 9.6 Consider the Laguerre polynomials $p_n(x) = x^n + \cdots$ corresponding to the weight function $w(x) = x^\alpha e^{-x}$ in $I = (0, \infty)$.

(i) Show that

$$q_n := x^{-\alpha} e^x \frac{d^n}{dx^n} \left(x^{n+\alpha} e^{-x}\right)$$

is a polynomial of degree n, and the sequence (q_n) is orthogonal with respect to w.

(ii) Deduce from (a) that $p_n = a_n q_n$ for all n with suitable non-zero coefficients a_n.

(iii) Show that $y = p_n$ satisfies the differential equation

$$xy'' + (\alpha + 1 - x)y' + ny = 0.$$

Exercise 9.7 Consider the Hermite polynomials $p_n(x) = x^n + \cdots$ corresponding to the weight function $w(x) := e^{-x^2}$ in $I = \mathbb{R}$.

(i) Show that

$$p_n(x) := \left(-\frac{1}{2}\right)^n e^{x^2} \frac{d^n}{dx^n} \left(e^{-x^2}\right).$$

(ii) Show that $y = p_n$ satisfies the differential equation

$$y'' - 2xy' + 2ny = 0.$$

(iii) (Generating function) Prove the identity

$$e^{-2xt-t^2} = \sum_{n=0}^{\infty} \frac{p_n(x)}{n!} (-2t)^n.$$

Exercise 9.8 Show that the weight functions w of all classical orthogonal polynomials satisfy *Pearson's differential equation*

$$\frac{w'(x)}{w(x)} = \frac{D + Ex}{A + Bx + Cx^2}$$

with suitable parameters A, B, C, D, E.

Exercise 9.9 Fix two electric charges of mass $1/2$ at ± 1, and place n unit charges at $x_1, \ldots, x_n \in (-1, 1)$. Assume that the system is in an electrostatic equilibrium. Prove that $x_1, \ldots, x_n$ are the zeros of the nth Legendre polynomial.

Chapter 10
Numerical Integration

For his celestial mechanics Newton had to evaluate complicated integrals. He and his followers generalized the trapezoidal rule

$$\int_a^b f(x)\,dx \approx \frac{f(a)+f(b)}{2}(b-a)$$

and Simpson's rule

$$\int_a^b f(x)\,dx \approx \frac{f(a)+4f(\frac{a+b}{2})+f(b)}{6}(b-a)$$

by seeking approximations of the form

$$\int_I f(x)\,dx \approx A_1 f(x_1)+\cdots+A_n f(x_n)$$

where the points x_k and coefficients A_k do not depend on the particular choice of the function f. We will present some of these results.[1]

As in the preceding section, w denotes an arbitrary weight function on some nondegenerate interval I. Besides the usual norms $\|\cdot\|_p$ on I we will also use the norms

$$\|f\|_{1,w} := \int_I |f|\,w\,dx, \quad \|f\|_{2,w} := \Big(\int_I |f|^2\,w\,dx\Big)^{1/2}$$

and the scalar product $(\cdot,\cdot)_w$ associated with the last norm.

[1] See, e.g., Ralston–Rabinowitz [404] for a more detailed exposition.

V. Komornik, *Topology, Calculus and Approximation*, Springer Undergraduate Mathematics Series, DOI 10.1007/978-1-4471-7316-8_10

10.1 Lagrange Formulas

Given n distinct points $x_1, \dots, x_n \in I$, $n \geq 1$, we seek real numbers $A_1, \dots, A_n$ such that the equality

$$\int_I f\, w\, dx = A_1 f(x_1) + \cdots + A_n f(x_n)$$

holds for *all* polynomials $p \in \mathcal{P}_m$, with the greatest possible m.

Let us introduce the usual polynomial

$$\omega(x) := (x - x_1) \cdots (x - x_n)$$

and the Lagrange basis polynomials

$$\ell_k \in \mathcal{P}_{n-1}, \quad \ell_k(x_j) = \delta_{kj}.$$

Proposition 10.1 *There exist real numbers $A_1, \dots, A_n$ such that*

$$\int_I f\, w\, dx = \sum_{k=1}^{n} A_k f(x_k) \quad \textit{for all} \quad f \in \mathcal{P}_{n-1}. \tag{10.1}$$

They are determined uniquely:

$$A_k = \int_I \ell_k w\, dx \quad \textit{for all} \quad k. \tag{10.2}$$

Furthermore, if $f \in C^n(I)$ and $\left\| f^{(n)} \right\|_\infty < \infty$, then

$$\left| \int_I f\, w\, dx - \sum_{k=1}^{n} A_k f(x_k) \right| \leq \frac{\|\omega\|_{1,w}}{n!} \left\| f^{(n)} \right\|_\infty. \tag{10.3}$$

Proof Uniqueness. If (10.1) is satisfied, then choosing $f = \ell_j$ the relations (10.2) follow:

$$\int_I \ell_j w\, dx = \sum_{k=1}^{n} A_k \ell_j(x_k) = \sum_{k=1}^{n} A_k \delta_{jk} = A_j.$$

Existence. Define the numbers A_k by (10.2). For any given $f \in C^n(I)$ consider the corresponding Lagrange interpolation polynomial:

$$p := \sum_{k=1}^{n} f(x_k) \ell_k. \tag{10.4}$$

By Theorem 8.2 (p. 196) we have

$$|f(x)-p(x)| \le \frac{|\omega(x)|}{n!}\|f^{(n)}\|_\infty$$

for all $x \in I$. By the definition of the weight function this implies that $(f-p)w$ is integrable. Then $fw=(f-p)w+pw$ is also integrable, and

$$\begin{aligned}\Big|\int_I f\,w\,dx - \int_I pw\,dx\Big| &= \Big|\int_I (f-p)w\,dx\Big| \\ &\le \int_I \frac{|\omega|}{n!} w\,dx\,\|f^{(n)}\|_\infty \\ &= \frac{\|\omega\|_{1,w}}{n!}\|f^{(n)}\|_\infty.\end{aligned}$$

This yields the estimate (10.3) because (10.2) and (10.4) imply that

$$\int_I pw\,dx = \int_I \sum_{k=1}^n f(x_k)\ell_k w\,dx = \sum_{k=1}^n A_k f(x_k).$$

Finally, (10.1) follows from (10.3) because if $f \in \mathcal{P}_{n-1}$, then $f^{(n)} \equiv 0$, so that the right-hand side of (10.3) vanishes. □

Similarly to Theorem 8.3 (p. 197) we may try to minimize the norm $\|\omega\|_{1,w}$ in (10.3) by an appropriate choice of the points x_i. For $w \equiv 1$ the answer is given by the following theorem[2]:

***Proposition 10.2 (Korkin–Zolotarev)** *If $I=[-1,1]$ and $w \equiv 1$, then*

$$\|\omega\|_1 \ge 2^{1-n}$$

for all choices of points $x_1 > \cdots > x_n$ in I. Equality holds if and only if

$$x_k = \cos\frac{k\pi}{n+1}, \qquad k=1,\dots,n.$$

Remark The optimal nodes x_k are the roots of the Chebyshev polynomials of the second kind[3]:

$$p_n(\cos\theta) = \frac{\sin((n+1)\theta))}{\sin\theta}, \qquad n=0,1,\dots.$$

[2] For the proof we refer to Achieser [3], Natanson [359] or Timan [491].

[3] See Exercise 9.5 (ii), p. 228.

10.2 Newton–Cotes Rules

Newton and Cotes investigated the case of equidistant nodes

$$x_k = a + \frac{k-1}{n-1}(b-a), \quad k = 1, \dots, n$$

for the weight function $w \equiv 1$ on a compact interval $[a, b]$.

Examples

- For $n = 2$ we have

  $$A_1 = A_2 = \frac{b-a}{2} \quad \text{and} \quad \|\omega\|_1 = \int_a^b (x-a)(b-x)\,dx = \frac{(b-a)^3}{6},$$

  and the estimate (10.3) of Proposition 10.1 takes the form

  $$\left| \int_a^b f\,dx - \frac{f(a)+f(b)}{2}(b-a) \right| \le \frac{(b-a)^3}{12} \, \|f''\|_\infty$$

  for all $f \in C^2(I)$. This is the error estimate of the *trapezoidal rule*.
- If $n = 3$, then an easy computation shows that

  $$A_1 = A_3 = \frac{b-a}{6}, \quad A_2 = \frac{4(b-a)}{6} \quad \text{and} \quad \|\omega\|_1 = \frac{(b-a)^4}{32}.$$

  Proposition 10.1 now leads to the error estimate of *Simpson's rule*[4]:

  $$\left| \int_a^b f\,dx - \frac{f(a) + 4f(\frac{a+b}{2}) + f(b)}{6}(b-a) \right| \le \frac{(b-a)^4}{192} \, \|f'''\|_\infty$$

  for all $f \in C^3(I)$.

It is natural to expect that by increasing n the Newton–Cotes formulas become more and more efficient, and that (with obvious notations)

$$\sum_{k=1}^{n} A_k^n f(x_k^n) \to \int_I f\,dx \quad \text{if} \quad n \to \infty$$

[4] Compare with the estimate (8.24), p. 207.

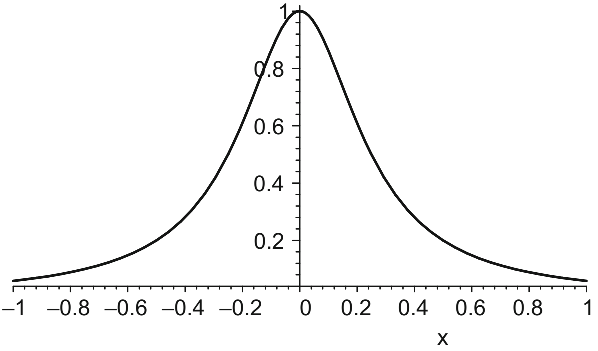


Fig. 10.1 Runge's example

for all $f \in C(I)$. It came as a surprise when Méray gave a counterexample for a similar problem. The following counterexample is due to Runge: for the integral

$$\int_{-1}^{1} \frac{1}{1+16x^2}\,dx = \frac{\arctan 4}{2} \approx 0.66291$$

(see Fig. 10.1) the Newton–Cotes formulas give the following values:

$$\begin{aligned} 0.11765 &\quad \text{for} \quad n = 2, \\ 1.37255 &\quad \text{for} \quad n = 3, \\ 0.59306 &\quad \text{for} \quad n = 6, \\ 0.69993 &\quad \text{for} \quad n = 8, \end{aligned}$$

but

$$-1.19379 \quad \text{for} \quad n = 21 \ldots.$$

10.3 Gauss Rules

Proposition 10.1 (p. 232) is valid for all node systems. Gauss asked and answered the following question: *is it possible to choose the nodes so as to make the equality* (10.1) *hold for polynomials of higher degree as well?* In order to formulate his result we consider the orthogonal polynomials p_n associated with a weight function $w : I \to \mathbb{R}$.

Theorem 10.3 *For each fixed $n \geq 1$ there exist points $x_1, \dots, x_n \in I$ and real numbers $A_1, \dots, A_n$ such that*

$$\int_I f\,w\,dx = \sum_{k=1}^{n} A_k f(x_k) \quad \text{for all} \quad f \in \mathcal{P}_{2n-1}. \tag{10.5}$$

There is a unique choice: the points x_k are the roots of p_n, and

$$A_k = \int_I \ell_k w\,dx, \quad k = 1, \dots, n. \tag{10.6}$$

Moreover, if $f \in C^{2n}(I)$ and $\left\| f^{(2n)} \right\|_\infty < \infty$, then

$$\left| \int_I f\,w\,dx - \sum_{k=1}^{n} A_k f(x_k) \right| \leq \frac{\|\omega\|_{2,w}^2}{(2n)!} \left\| f^{(2n)} \right\|_\infty. \tag{10.7}$$

Remark The estimate (10.7) is optimal: the proof given below shows that we have equality for all polynomials $f \in \mathcal{P}_{2n}$. Moreover, if $\deg f = 2n$ then the two sides of this equality are different from zero, so that equality (10.5) does not hold any more.

Proof Uniqueness. It is sufficient to show that (10.5) implies $\omega = p_n$: the uniqueness of the coefficients A_k and the formula (10.6) will then follow from Proposition 10.1. First we show that ω is orthogonal to $\mathcal{P}_{n-1}$. Indeed, if $q \in \mathcal{P}_{n-1}$, then applying (10.5) for the polynomial $f := \omega q \in \mathcal{P}_{2n-1}$ and using the equalities $\omega(x_1) = \cdots = \omega(x_n) = 0$ we obtain that

$$(\omega, q)_w = \int_I \omega q w\,dx = \sum_{k=1}^{n} A_k q(x_k)\omega(x_k) = 0.$$

Since p_n is also orthogonal to $\mathcal{P}_{n-1}$ and $\omega - p_n \in \mathcal{P}_{n-1}$, it follows that

$$\|\omega - p_n\|_{2,w}^2 = (\omega - p_n, \omega - p_n)_w = (\omega, \omega - p_n)_w - (p_n, \omega - p_n)_w = 0;$$

hence $\omega = p_n$.

Existence. Let $\omega := p_n$, i.e., let $x_1, \dots, x_n$ be the roots of p_n, and define the numbers A_k by (10.6). If $f \in \mathcal{P}_{2n-1}$, then we have $f = \omega q + r$ for suitable polynomials $q, r \in \mathcal{P}_{n-1}$. Applying the equality (10.1) of Proposition 10.1 for r, and taking into account that $f(x_k) = r(x_k)$ for all k (because $\omega(x_k) = 0$), we obtain the following equality:

$$\int_I r\,w\,dx = \sum_{k=1}^{n} A_k r(x_k) = \sum_{k=1}^{n} A_k f(x_k). \tag{10.8}$$

Furthermore, since $\omega = p_n$ is orthogonal to $q \in \mathcal{P}_{n-1}$, we have

$$\int_I f\,w\,dx = \int_I (\omega q + r)w\,dx = (\omega, q)_w + \int_I rw\,dx = \int_I rw\,dx. \tag{10.9}$$

The required equality (10.5) follows from (10.8) and (10.9).

The estimate (10.7). For any fixed $f \in C^{2n}(I)$ we introduce the Hermite interpolation polynomial $p \in \mathcal{P}_{2n-1}$ satisfying

$$p(x_k) = f(x_k) \quad \text{and} \quad p'(x_k) = f'(x_k), \quad k = 1, \dots, n.$$

It follows from Theorem 8.7 (p. 205) that[5]

$$|f(x) - p(x)| \le \frac{\omega(x)^2}{(2n)!} \| f^{(2n)} \|_\infty$$

for all $x \in I$. Consequently,

$$\Big| \int_I f\,w\,dx - \int_I pw\,dx \Big| = \Big| \int_I (f - p)w\,dx \Big| \le \frac{\|\omega\|_{2,w}^2}{(2n)!} \| f^{(2n)} \|_\infty.$$

This implies (10.7) because applying (10.5) for p and using the relations $p(x_k) = f(x_k)$ we have

$$\int_I pw\,dx = \sum_{k=1}^{n} A_k p(x_k) = \sum_{k=1}^{n} A_k f(x_k). \qquad \square$$

Example Let us reconsider Runge's integral

$$\int_{-1}^{1} \frac{1}{1 + 16x^2}\,dx \approx 0.66291.$$

The following table illustrates the superiority of the Gauss rules over the Newton–Cotes rules.

[5] Here we have $m = 2n$ and $\Omega = \omega^2$.

n	Newton–Cotes	Gauss
2	0.11765	0.31579
3	1.37255	0.99371
4	0.56942	0.51182
5	0.56941	0.77215
6	0.59306	0.60292
7	0.83220	0.70192
8	0.69993	0.64002
9	0.48527	0.67721
10	0.60771	0.65431
11	0.89889	0.66820
12	0.73523	0.65970
13	0.33365	0.66487
14	0.55908	0.66171
15	1.14721	0.66364
16	0.81773	0.66246
17	−0.07551	0.66318
18	0.42423	0.66274
19	1.82088	0.66301
20	1.04035	0.66285
21	−1.19379	0.66295
22	0.05366	0.66289
23	3.69423	0.66292
24	1.66303	0.66290
25	−4.36075	0.66291
26	−1.00204	0.66291
27	9.09388	0.66291
28	3.46778	0.66291
29	−13.63928	0.66291
30	−4.11077	0.66291
31	25.15105	0.66291

For a long time the Newton–Cotes rules were preferred because of the simple formula for the nodes. Since the proliferation of computers the Gauss rules have become more popular.[6]

10.4 Theorems of Stieltjes and Erdős–Turán

In this section we assume that the weight function is defined on a *compact* interval.

[6] However, the composite Newton–Cotes rules (like in Sect. 10.9 below, p. 260) remain efficient.

In contrast to the Newton–Cotes rules, the efficiency of the Gauss rules increases with n. In the following theorem we express the dependence on n by writing A_k^n and x_k^n instead of A_k and x_k.

Theorem 10.4 (Stieltjes) *If $f : I \to \mathbb{R}$ is a continuous function defined on a compact interval, then*

$$\sum_{k=1}^{n} A_k^n f(x_k^n) \to \int_I f\, w\, dx$$

as $n \to \infty$.

We first prove a stronger result. We introduce the orthogonal polynomials p_n associated with w, and then we introduce for each n the Lagrange basis polynomials ℓ_k^n associated with the roots $x_1^n, \dots, x_n^n$ of p_n:

$$\ell_k^n \in \mathcal{P}_{n-1}, \quad \ell_k^n(x_j^n) = \delta_{kj}, \quad 1 \le k, j \le n.$$

We are investigating the sequence of Lagrange interpolation polynomials

$$L_n f := \sum_{k=1}^{n} f(x_k^n)\ell_k^n \in \mathcal{P}_{n-1}, \quad n = 1, 2, \dots.$$

Theorem 10.5 (Erdős–Turán) *If $f : I \to \mathbb{R}$ is a continuous function defined on a compact interval, then $\|f - L_n f\|_{2,w} \to 0$ as $n \to \infty$.*

Remark It is interesting to compare this result with Faber's theorem (p. 208): the norm $\|\cdot\|_{2,w}$ is weaker than $\|\cdot\|_\infty$.

Proof

(i) First we establish for each fixed $n \ge 1$ the following properties of the Lagrange basis polynomials:

$$\sum_{k=1}^{n} \ell_k^n = 1, \tag{10.10}$$

$$\int_I \ell_k^n \ell_j^n w\, dx = 0 \quad \text{if} \quad k \ne j, \tag{10.11}$$

$$\sum_{k=1}^{n} \int_I (\ell_k^n)^2 w\, dx = \|w\|_1 . \tag{10.12}$$

The equality (10.10) follows from the definition of interpolation because $L_n 1 = 1$. The fundamental orthogonality relation (10.11) is obtained by applying formula (10.5) of Theorem 10.3 (p. 236) to the polynomial $\ell_k^n \ell_j^n$ of degree $2n-2$:

$$\int_I \ell_k^n \ell_j^n w\,dx = \sum_{i=1}^n A_i^n (\ell_k^n \ell_j^n)(x_i^n) = \sum_{i=1}^n A_i^n \delta_{ki}\delta_{ji} = 0.$$

Finally, (10.12) follows from (10.10) and (10.11):

$$\sum_{k=1}^n \int_I (\ell_k^n)^2 w\,dx = \int_I \Big(\sum_{k=1}^n \ell_k^n\Big)^2 w\,dx = \int_I w\,dx = \|w\|_1 .$$

(ii) We simplify the notation by writing $\|f\|$ instead of $\|f\|_{2,w}$. Applying Weierstrass' Theorem 8.9 (p. 207), for any fixed $\varepsilon > 0$ we choose a polynomial p satisfying

$$\|f-p\|_\infty \le \varepsilon. \tag{10.13}$$

If $n \ge \deg p$, then $L_n p = p$, and therefore

$$\|f - L_n f\| \le \|f-p\| + \|L_n p - L_n f\|$$

by the triangle inequality. If we show that

$$\|f-p\| \le \varepsilon\,\|w\|_1^{1/2} \quad \text{and} \quad \|L_n p - L_n f\| \le \varepsilon\,\|w\|_1^{1/2}, \tag{10.14}$$

then we will deduce from the preceding inequality the estimate

$$\|f - L_n f\| \le 2\varepsilon\,\|w\|_1^{1/2}$$

for every $n \ge \deg p$. This will prove the relation $\|f - L_n f\| \to 0$.

The first inequality of (10.14) follows from (10.13):

$$\|f-p\|^2 = \int_I (f-p)^2 w\,dx \le \varepsilon^2 \int_I w\,dx = \varepsilon^2\,\|w\|_1 .$$

For the proof of the second one we start with the identity

$$L_n p - L_n f = L_n(p-f) = \sum_{k=1}^n (p-f)(x_k^n)\ell_k^n,$$

and we use (10.11), (10.13) and (10.12):

$$\begin{aligned}\|L_np - L_nf\|^2 &= \int_I \Big(\sum_{k=1}^n (p-f)(x_k^n)\ell_k^n \Big)^2 w\,dx \\ &= \sum_{k=1}^n \big|(p-f)(x_k^n)\big|^2 \int_I (\ell_k^n)^2 w\,dx \\ &\le \varepsilon^2 \, \|w\|_1 \,.\end{aligned}$$

□

Proof of Theorem 10.4 Using the Cauchy–Schwarz inequality we have

$$\begin{aligned}\|f - L_nf\|_{1,w} &= \int_I |f - L_nf| w\,dx \\ &\le \Big(\int_I w\,dx \Big)^{1/2} \Big(\int_I |f - L_nf|^2 w\,dx \Big)^{1/2} \\ &= \|w\|_1^{1/2} \cdot \|f - L_nf\|_{2,w} \,.\end{aligned}$$

The last expression tends to zero by the preceding theorem. Hence

$$\int_I (L_nf) w\,dx \to \int_I fw\,dx,$$

which is equivalent to the assertion. □

10.5 Euler's and Stirling's Formulas

Euler and Maclaurin developed an efficient way to numerically evaluate finite and infinite sums of the form

$$S^0 := \sum_{n=1}^N f(n),$$

like

$$1^\alpha + 2^\alpha + \cdots + N^\alpha \quad \text{or} \quad \ln 1 + \ln 2 + \cdots + \ln N = \ln(N!).$$

Let us introduce the more general sums

$$S^k := \sum_{n=1}^N f^{(k)}(n), \quad k = 0, 1, \ldots$$

where $f : (0, \infty) \to \mathbb{R}$ is a C^∞ function. Applying Taylor's formula to a primitive F of f, we obtain the *heuristic* estimate

$$\begin{aligned}\int_0^N f(x)\,dx &= F(N) - F(0)\\ &= \sum_{n=1}^N \big(F(n) - F(n-1)\big)\\ &\approx \sum_{n=1}^N \frac{F'(n)}{1!} - \frac{F''(n)}{2!} + \frac{F'''(n)}{3!} - \cdots ,\end{aligned}$$

i.e.,

$$\int_0^N f(x)\,dx \approx \frac{S^0}{1!} - \frac{S^1}{2!} + \frac{S^2}{3!} - \cdots . \tag{10.15}$$

Applying the same reasoning to the derivatives of f, we get the analogous formulas

$$\begin{aligned} f(N) - f(0) &\approx \frac{S^1}{1!} - \frac{S^2}{2!} + \frac{S^3}{3!} - \cdots ,\\ f'(N) - f'(0) &\approx \frac{S^2}{1!} - \frac{S^3}{2!} + \frac{S^4}{3!} - \cdots ,\\ &\vdots \end{aligned}$$

Using these relations we may eliminate the quantities $S^1, S^2, \ldots$ from (10.15) one by one, obtaining a relation of the form

$$\sum_{n=1}^N f(n) \approx \int_0^N f(x)\,dx + \sum_{k=1}^\infty a_k \left(f^{(k-1)}(N) - f^{(k-1)}(0)\right), \tag{10.16}$$

where the coefficients a_k converge rapidly to zero:

$$a_1 = \frac{1}{2},\ a_2 = \frac{1}{12},\ a_3 = 0,\ a_4 = -\frac{1}{720},\ a_5 = 0,\ a_6 = \frac{1}{30240}, \ldots .$$

Example Consider the sum

$$S := \sum_{n=1}^\infty \frac{1}{n^2} = \frac{\pi^2}{6} \approx 1.6449340644.$$

- Applying (10.16) with $f(x) = (x+1)^{-2}$ and $N = \infty$ we get

$$\sum_{n=1}^{\infty} \frac{1}{n^2} \approx \int_0^{\infty} (x+1)^{-2}\, dx + \frac{1}{2} + 2a_2 - 6a_3 + 24a_4 - \cdots ,$$

which yields the following approximations:

$$\begin{aligned} S &\approx 1 + 1 - a_1 = 1.5, \\ S &\approx 1 + 1 - a_1 + 2a_2 \approx 1.667, \\ S &\approx 1 + 1 - a_1 + 2a_2 - 6a_3 + 24a_4 \approx 1.633, \\ &\vdots \end{aligned}$$

- Applying (10.16) with $f(x) = (x+10)^{-2}$ and $N = \infty$ we get

$$\sum_{n=11}^{\infty} \frac{1}{n^2} \approx \int_0^{\infty} (x+10)^{-2}\, dx - \frac{a_1}{10^2} + \frac{2a_2}{10^3} - \frac{6a_3}{10^4} + \frac{24a_4}{10^5} - \cdots .$$

Computing the initial finite sum directly

$$\sum_{n=1}^{10} \frac{1}{n^2} \approx 1.5497677311,$$

we eventually obtain the following much more precise approximations:

$$\begin{aligned} S &\approx 1.5497677311 + \frac{1}{10} - \frac{a_1}{10^2} = 1.6447677311, \\ S &\approx 1.5497677311 + \frac{1}{10} - \frac{a_1}{10^2} + \frac{2a_2}{10^3} \approx 1.6449343978, \\ S &\approx 1.5497677311 + \frac{1}{10} - \frac{a_1}{10^2} + \frac{2a_2}{10^3} - \frac{6a_3}{10^4} + \frac{24a_4}{10^5} \approx 1.6449340644, \\ &\vdots \end{aligned}$$

In order to achieve the same precision by a direct summation, we should have to sum the first *billion* terms of the series!

Remark Continuing the previous computation, the next approximation is *less precise*:

$$\begin{aligned} S &\approx 1.5497677311 + 0.1 - \frac{a_1}{10^2} + \frac{2a_2}{10^3} - \frac{6a_3}{10^4} + \frac{24a_4}{10^5} - \frac{120a_5}{10^6} + \frac{720a_6}{10^7} \\ &\approx 1.6449344001. \end{aligned}$$

Thus we need a careful study of the error of this procedure.

A first result is the following, where we write f_j instead of $f(j)$ for brevity, and $[x]$ denotes the *integer part* of x, i.e., the largest integer p satisfying $p \le x$.

Proposition 10.6 *If* $f \in C^1([0,n])$, *then*

$$f_0 + \cdots + f_n = \int_0^n f(x)\,dx + \frac{f_0 + f_n}{2} + \int_0^n \Big(x - [x] - \frac{1}{2}\Big) f'(x)\,dx. \tag{10.17}$$

Proof Integrating by parts we obtain for each $k = 0, 1, \dots, n-1$ the equality

$$\int_k^{k+1} f(x)\,dx = \Big[\Big(x - k - \frac{1}{2}\Big) f(x)\Big]_k^{k+1} - \int_k^{k+1} \Big(x - k - \frac{1}{2}\Big) f'(x)\,dx,$$

which is equivalent to

$$\int_k^{k+1} f(x)\,dx = \frac{f_k + f_{k+1}}{2} - \int_k^{k+1} \Big(x - [x] - \frac{1}{2}\Big) f'(x)\,dx.$$

Summing them we get (10.17). □

Example Taking $f(x) = \frac{1}{1+x}$ and replacing n by $n-1$ we obtain the equality

$$1 + \frac{1}{2} + \cdots + \frac{1}{n} = \ln n + \frac{1}{2} + \frac{1}{2n} - \int_1^n \frac{x - [x] - \frac{1}{2}}{x^2}\,dx.$$

For $n \to \infty$ the last integral has a finite limit because

$$\int_1^\infty \left| \frac{x - [x] - \frac{1}{2}}{x^2} \right| dx \le \int_1^\infty \frac{1}{2x^2}\,dx < \infty,$$

and we obtain the existence of *Euler's constant*[7]

$$C := \lim_{n\to\infty} \Big(1 + \frac{1}{2} + \cdots + \frac{1}{n} - \ln n\Big) = \frac{1}{2} - \int_1^\infty \frac{x - [x] - \frac{1}{2}}{x^2}\,dx.$$

A more important application is the estimation of $n!$ for large n:

Proposition 10.7 (Stirling's Formula) *We have*

$$\lim_{n\to\infty} \frac{n!}{\left(\frac{n}{e}\right)^n \sqrt{2\pi n}} = 1.$$

[7] See also Exercise 10.2 (p. 263) for a geometric proof.

This result is widely used in the theory of probability, in combinatorics and in number theory.

Proof Applying Proposition 10.6 with $f(x) := \ln(1+x)$ and $n-1$ in place of n we get

$$\ln 1 + \cdots + \ln n = \int_1^n \ln x\,dx + \frac{\ln n}{2} + \int_1^n \frac{x-[x]-\frac{1}{2}}{x}\,dx.$$

Since

$$\int_1^n \ln x\,dx = [x\ln x - x]_1^n = n\ln n - n + 1,$$

this may be rewritten as

$$\ln n! - \left(n+\frac{1}{2}\right)\ln n + n = 1 + \int_1^n \frac{x-[x]-\frac{1}{2}}{x}\,dx.$$

Since the periodic function $\beta_1(x) := x-[x]-\frac{1}{2}$ has zero mean value, it has a periodic and hence bounded primitive function $\beta_2(x)$, and therefore

$$1 + \int_1^n \frac{x-[x]-\frac{1}{2}}{x}\,dx = 1 + \left[\frac{\beta_2(x)}{x}\right]_1^n + \int_1^n \frac{\beta_2(x)}{x^2}\,dx$$

is a Cauchy sequence of n. Denoting its finite limit by γ, we have

$$\ln n! - \left(n+\frac{1}{2}\right)\ln n + n \to \gamma$$

or equivalently

$$a_n := \frac{n!e^n}{n^{n+\frac{1}{2}}} \to e^\gamma.$$[8]

It remains to compute e^γ. We have

$$\begin{aligned}\frac{a_n^4}{a_{2n}^2} &= \frac{(n!)^4}{((2n)!)^2}\cdot\frac{(2n)^{4n+1}}{n^{4n+2}}\\ &= \frac{(2\cdot 4\cdots(2n))^4}{(1\cdot 2\cdots(2n))^2}\cdot\frac{2}{n}\end{aligned}$$

[8] See again Exercise 10.2 (p. 263) for a geometric proof.

$$
\begin{aligned}
&= \frac{(2\cdot 4\cdots(2n))^2}{(1\cdot 3\cdots(2n-1))^2}\cdot\frac{2}{n}\\
&= \frac{2\cdot 2}{1\cdot 3}\cdot\frac{4\cdot 4}{3\cdot 5}\cdots\frac{2n\cdot 2n}{(2n-1)\cdot(2n+1)}\cdot\frac{4n+2}{n}.
\end{aligned}
$$

Letting $n\to\infty$ and applying the Wallis formula[9]

$$\lim_{n\to\infty}\frac{2\cdot 2}{1\cdot 3}\cdot\frac{4\cdot 4}{3\cdot 5}\cdots\frac{2n\cdot 2n}{(2n-1)\cdot(2n+1)}=\frac{\pi}{2}$$

we conclude that $e^{2\gamma}=2\pi$, and hence $e^{\gamma}=\sqrt{2\pi}$. □

Proposition 10.6 will be greatly improved in Sect. 10.7.

10.6 Bernoulli Polynomials

Given a nonnegative continuous function $f_0:[a,b]\to\mathbb{R}$ on a compact interval, we define a sequence of functions $f_n:[a,b]\to\mathbb{R}$ by the recurrence relations

$$f_n'=f_{n-1}\quad\text{and}\quad\int_a^b f_n\,dx=0,\qquad n=1,2,\ldots.$$

Example Starting with $f_0(x)=\cos x$ in $[-\pi,\pi]$ we obtain a 4-periodic sequence of functions with

$$f_1(x)=\sin x,\quad f_2(x)=-\cos x\quad\text{and}\quad f_3(x)=-\sin x.$$

In the general case the sequence is not periodic, but certain properties of the trigonometric functions are preserved in a weaker form:

Lemma 10.8 *Assume that f_0 is nonnegative, and even with respect to the midpoint $c:=(a+b)/2$. Then we have the following properties:*

(a) *f_n is even (resp. odd) with respect to c when n is even (resp. odd);*
(b) *$f_n(a)=f_n(b)$ for $n=2,4,6,\ldots$;*
(c) *$f_1(c)=0$, and $f_n(a)=f_n(c)=f_n(b)=0$ for $n=3,5,7,\ldots$;*
(d) *$(-1)^k f_{2k-1}\ge 0$ in $[a,c]$ for $k=1,2,\ldots$.*
(e) *$(-1)^k f_{2k}(a)\le 0$ for $k=1,2,\ldots$.*[10]

[9] See Exercise 10.3, p. 264.

[10] The proof will show that the inequalities are strict except in the trivial case where $f_0(x)\equiv 0$.

Proof

(a) Setting $g_n(c+x) := (-1)^n f_n(c-x)$, a straightforward computation shows that

$$g_0 = f_0, \quad \text{furthermore} \quad g_n' = g_{n-1} \quad \text{and} \quad \int_a^b g_n\, dx = 0, \quad n = 1, 2, \ldots.$$

By the uniqueness of (f_n) we conclude that $g_n = f_n$ for all n.

(b) This follows from the evenness of f_n.

(c) It follows from the oddity of f_n that $f_n(c) = 0$ and $f_n(b) + f_n(a) = 0$. It remains to observe that $f_n(a) = f_n(b)$ for *all* $n \ge 2$ because

$$f_n(b) - f_n(a) = \int_a^b f_n'\, dx = \int_a^b f_{n-1}\, dx = 0$$

by definition.

(d) Since $f_1' = f_0 \ge 0$ in $[a, c]$ and $f_1(c) = 0$ by (c), $f_1 \le 0$ in $[a, c]$. We proceed by induction on k. If $(-1)^k f_{2k-1} \ge 0$ in $[a, c]$ for some $k \ge 1$, then the same property holds for $g := (-1)^{k+1} f_{2k+1}$ because $g'' = (-1)^{k+1} f_{2k-1} \le 0$ in $[a, c]$, so that g is concave in $[a, c]$, and $g(a) = g(c) = 0$ by (c).

(e) For $k = 2, 4, \ldots$ it follows from (a) and (d) that[11]

$$f_{2k}(x) - f_{2k}(a) = \int_a^x f_{2k-1}(t)\, dt \ge 0$$

for all $x \in [a, b]$. Hence

$$f_{2k}(a) \le \frac{1}{b-a} \int_a^b f_{2k}(x)\, dx = 0.$$

For $k = 1, 3, \ldots$ we may repeat the above proof by changing the sense of the inequalities, to conclude that $f_{2k}(a) \ge 0$. □

Definition Starting with $[a, b] = [0, 1]$ and $f_0 \equiv 1$ we obtain the sequence $b_0(x), b_1(x), \ldots$ of *Bernoulli polynomials.*[12]

See Figs. 10.2–10.7; the last four figures show the 4-periodic qualitative picture.

Example It follows directly from the definitions that

$$b_0(x) = 1, \quad b_1(x) = x - \frac{1}{2}, \quad b_2(x) = \frac{1}{2}x^2 - \frac{1}{2}x + \frac{1}{12}.$$

[11] The inequality follows from (d) if $a \le x \le c$. The case $c \le x \le b$ hence follows by observing that $\int_a^x = \int_a^{2c-x}$.

[12] In order to avoid any misunderstanding, henceforth we indicate the variable x of the Bernoulli polynomials. We observe that $\deg b_n(x) = n$.

Fig. 10.2 $b_0(x)$

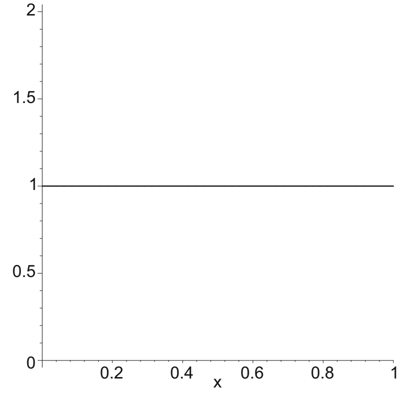


Fig. 10.3 $b_1(x)$

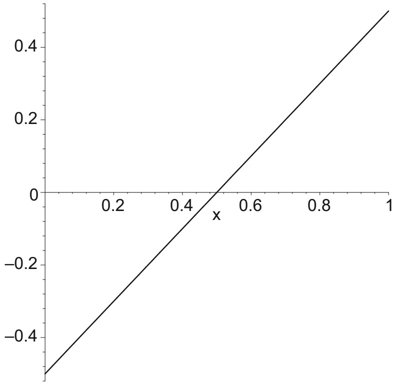


Fig. 10.4 $b_2(x)$

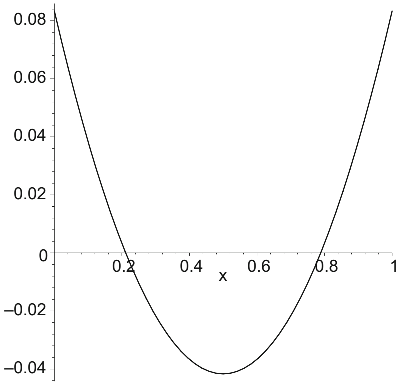


Fig. 10.5 $b_3(x)$

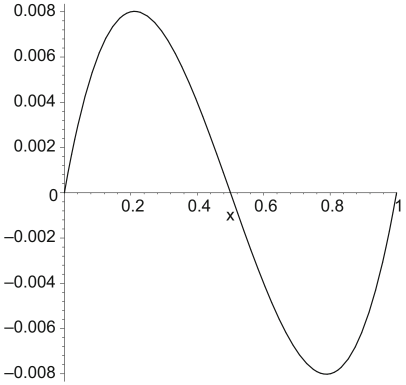


Fig. 10.6 $b_4(x)$

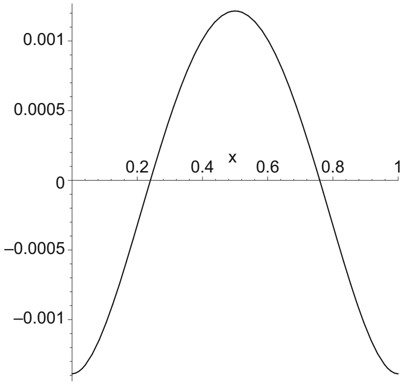
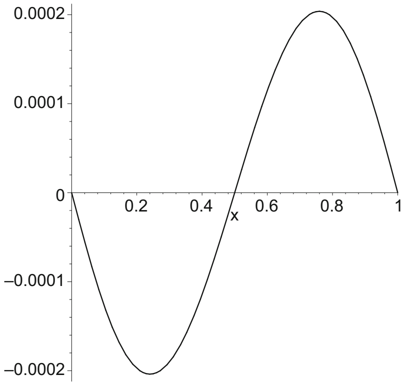


Fig. 10.7 $b_5(x)$



Proposition 10.9 *The Bernoulli polynomials are given by the formula*

$$b_n(x) = \sum_{k=0}^{n} b_{n-k} \frac{x^k}{k!}, \quad n = 0, 1, \ldots \tag{10.18}$$

for a suitable sequence $b_0, b_1, \ldots$ of real numbers.

Proof Since $b_0(x) \equiv 1$, the formula holds for $n = 0$ with $b_0 = 1$. If (10.18) holds for some $n \geq 0$, then

$$b_{n+1}(x) = \Big(\sum_{k=0}^{n} b_{n-k} \frac{x^{k+1}}{(k+1)!} \Big) + b_{n+1} = \sum_{k=0}^{n+1} b_{n+1-k} \frac{x^k}{k!}$$

by definition, with b_{n+1} chosen by the condition $\int_0^1 b_{n+1}(x)\, dx = 0$. □

Definition The coefficients $b_0, b_1, \ldots$ are called *Bernoulli numbers*.

Corollary 10.10 *The Bernoulli numbers have the following properties:*

(a) $b_n = b_n(0)$ *for all n;*
(b) $b_3 = b_5 = b_7 = \cdots = 0$*;*
(c) *the remaining numbers $b_0, b_1, b_2, b_4, b_6, \ldots$ have alternating signs.*

Proof (a) follows from Proposition 10.9; applying this to the above examples we see that $b_0 > 0$, $b_1 < 0$ and $b_2 > 0$. The remaining properties follow from Lemma 10.8 (c) and (e). □

Remarks

- Applying (10.18) for $n \geq 2$ with $x = 1$ and using the equality $b_n(1) = b_n(0) = b_n$ we obtain the equalities

$$b_{n-1} = -\sum_{k=2}^{n} \frac{b_{n-k}}{k!}, \quad n = 2, 3, \ldots. \tag{10.19}$$

  Starting with $b_0 = 1$ they allow us to recursively compute the Bernoulli numbers:

$$b_0 = 1, \quad b_1 = -\frac{1}{2}, \quad b_2 = \frac{1}{12}, \quad b_4 = -\frac{1}{720}, \quad b_6 = \frac{1}{30240}, \quad \ldots.$$

- Once the Bernoulli numbers are known, the Bernoulli polynomials may be computed by (10.18). Introducing the modified Bernoulli polynomials $B_n(x) := n!\, b_n(x)$ and numbers $B_k := k!\, b_k$, (10.18) and (10.19) may be written in the form

$$B_n(x) = \sum_{k=0}^{n} \binom{n}{k} B_{n-k} x^k \quad \text{for} \quad n \geq 0$$

and

$$B_n=\sum_{k=0}^{n}\binom{n}{k}B_{n-k}\quad\text{for}\quad n\ge 2,$$

reminding us of the binomial expansion of $(B+x)^n$ and $(B+1)^n$.

We have $B_3=B_5=B_7=\cdots=0$ and

$$B_0=1,\quad B_1=-\frac{1}{2},\quad B_2=\frac{1}{6},\quad B_4=-\frac{1}{30},\quad B_6=\frac{1}{42},\quad \ldots.$$

- It can be shown[13] that

$$b_{2n}\to 0\quad\text{and}\quad B_{2n}\to\infty.$$

10.7 Euler's General Formula

Now we are ready to generalize Proposition 10.6. For each $m=0,1,\ldots$ we denote by $\beta_m(x)$ the one-periodic function that coincides with the Bernoulli polynomial $b_m(x)$ in $[0,1)$.

Theorem 10.11 (Euler's Formula) *If $f\in C^m([0,n])$ for some integers $m,n\ge 1$, then*

$$f_0+\cdots+f_n=\int_0^n f(x)\,dx+\frac{f_0+f_n}{2}+\sum_{k=2}^{m}(-1)^k b_k\big[f^{(k-1)}\big]_0^n+R_m$$

with the remainder term

$$R_m=(-1)^{m-1}\int_0^n \beta_m(x)f^{(m)}(x)\,dx.$$

Remark Since $b_3=b_5=b_7=\cdots=0$ by Corollary 10.10 (b), we may also write the formula in the form

$$f_0+\cdots+f_n=\int_0^n f(x)\,dx+\frac{f_0+f_n}{2}+\sum_{k=1}^{[m/2]} b_{2k}\big[f^{(2k-1)}\big]_0^n+R_m$$

where $[x]$ denotes the integer part of x.

[13] See Exercise 10.7 below, p. 265.

Hence $R_{2j} = R_{2j+1}$ for all $j \ge 1$.

Proof [14] For $m = 1$ the equality holds by Proposition 10.6. If it holds for some $m \ge 1$, then it also holds for $m+1$ because, using the relations $\beta_k(0) = \beta_k(n) = b_k$ for $k \ge 2$,

$$\begin{aligned} R_m &= (-1)^{m-1} \int_0^n \beta_m(x) f^{(m)}(x)\, dx \\ &= (-1)^{m-1} [\beta_{m+1} f^{(m)}]_0^n + (-1)^m \int_0^n \beta_{m+1}(x) f^{(m+1)}(x)\, dx \\ &= (-1)^{m+1} b_{m+1} [f^{(m)}]_0^n + R_{m+1}. \end{aligned}$$

□

Example (Jacob Bernoulli) Applying the theorem for $f(x) := \dfrac{x^p}{p!}$ with a positive integer p and $m = p+1$ the remainder term vanishes because $f^{(p+1)} \equiv 0$, and we obtain the equality

$$\begin{aligned} \frac{1^p + \cdots + n^p}{p!} &= \int_0^n \frac{x^p}{p!}\, dx + \frac{n^p}{2(p!)} + \sum_{k=2}^{p+1} (-1)^k b_k \Big[\frac{x^{p+1-k}}{(p+1-k)!} \Big]_0^n \\ &= \frac{(n+1)^p}{(p+1)!} + \frac{n^p}{2(p!)} + \sum_{k=2}^{p} (-1)^k b_k \frac{n^{p+1-k}}{(p+1-k)!}. \end{aligned}$$

Since $b_0 = 1$, $b_1 = -1/2$ and $b_3 = b_5 = \cdots = 0$, this may be written in the form

$$\frac{1^p + \cdots + (n-1)^p}{p!} = \sum_{k=0}^{p} b_k \frac{n^{p+1-k}}{(p+1-k)!}.$$

Using the modified Bernoulli numbers $B_k := k! b_k$ this yields the formula

$$1^p + \cdots + (n-1)^p = \frac{1}{p+1} \sum_{k=0}^{p} \binom{p+1}{k} B_k n^{p+1-k}$$

or *symbolically*[15]

$$1^p + \cdots + (n-1)^p = \frac{(B+n)^{p+1} - B^{p+1}}{p+1}.$$

[14] Compare to that of Theorem 5.11, p. 129.

[15] We apply the binomial formula and then we replace each B^k by B_k.

For example,[16]

$$\begin{aligned}1^2+\cdots+(n-1)^2&=\frac{(B+n)^3-B^3}{3}=\frac{n^3+3B_1n^2+3B_2n}{3}\\&=\frac{2n^3-3n^2+n}{6}=\frac{n(n-1)(2n-1)}{6}\end{aligned}$$

and

$$\begin{aligned}1^3+\cdots+(n-1)^3&=\frac{(B+n)^4-B^4}{4}=\frac{n^4+4B_1n^3+6B_2n^2+4B_3n}{4}\\&=\frac{n^4-2n^3+n^2}{4}=\frac{n^2(n-1)^2}{4}.\end{aligned}$$

Let us investigate more closely the remainder term in Euler's formula.

Proposition 10.12 *If $f\in C^{2m+1}([0,n])$ for some integers $m,n\ge 1$ and both $f^{(2m-1)}$ and $f^{(2m+1)}$ are non-increasing, then*

$$R_{2m-1}=\theta b_{2m}\big[f^{(2m-1)}\big]_0^n$$

for some $\theta\in[0,1]$.

Remark Changing f to $-f$ we see that the proposition also holds if both $f^{(2m-1)}$ and $f^{(2m+1)}$ are non-decreasing.

Proof It suffices to show that R_{2m-1} and R_{2m+1} have opposite signs. Indeed, then we have either

$$0\le R_{2m-1}\le R_{2m-1}-R_{2m+1}=b_{2m}\big[f^{(2m-1)}\big]_0^n$$

or

$$0\ge R_{2m-1}\ge R_{2m-1}-R_{2m+1}=b_{2m}\big[f^{(2m-1)}\big]_0^n,$$

implying the assertion.

Using Lemma 10.8 (a) we may rewrite R_{2m-1} in the form

$$R_{2m-1}=\sum_{k=0}^{n-1}\int_0^{1/2}b_{2m-1}(t)h(k+t)\,dt$$

[16] Compare with Exercise 8.2, p. 214.

with

$$h(k+t) = f^{(2m-1)}(k+t) - f^{(2m-1)}(k+1-t).$$

Since $h \ge 0$ by our assumptions and $b_{2m-1}(t)$ has the constant sign $(-1)^m$ in $(0, \frac{1}{2})$ by Lemma 10.8 (d), R_{2m-1} has the same sign $(-1)^m$.

Similarly, R_{2m+1} has the opposite sign $(-1)^{m+1}$. □

If $f \in C^\infty([0,\infty))$, then we may study the limiting formulas as $n \to \infty$: Henceforth we write $f_n^{(j)}$ instead of $f^{(j)}(n)$ for brevity.

Corollary 10.13 *Let $f \in C^\infty([0,\infty))$, and assume that all odd-order derivatives of f converge monotonely to zero as $x \to \infty$. Then*

$$(f_0 + \cdots + f_n) - \frac{f_0 + f_n}{2} - \int_0^n f(x)\,dx$$

converges to a finite limit γ as $n \to \infty$.

Furthermore,

$$\gamma = -\sum_{k=1}^{m-1} b_{2k} f_0^{(2k-1)} - \theta_m b_{2m} f_0^{(2m-1)}$$

with some $\theta_m \in [0,1]$ for $m = 1, 2, \ldots$.

Remarks

- If f' is non-increasing,[17] then
  - f is non-decreasing;
  - *all* odd-order derivatives of f are non-increasing;
  - all even-order derivatives of f are non-decreasing, and converge to zero.

  See the following proof and Exercise 10.4, p. 265.
- Since $b_2, b_4, \ldots$ have alternating signs by Corollary 10.10 (c), it follows from the preceding remark that the partial sums of the series

$$-\sum_{k=1}^{\infty} b_{2k} f_0^{(2k-1)}$$

  are alternately $\le \gamma$ and $\ge \gamma$, and the errors do not exceed the first omitted terms.
- The series itself may be divergent: see the last remark in the next section and Exercise 10.8, p. 266.

[17] If f' is non-decreasing, then we may apply this remark to $-f$.

Proof Applying Proposition 10.6 we have

$$(f_0+\cdots+f_n)-\frac{f_0+f_n}{2}-\int_0^n f(x)\,dx=\int_0^n \big(x-[x]-\frac{1}{2}\big)f'(x)\,dx.$$

Since

$$\int_0^\infty \left|\big(x-[x]-\frac{1}{2}\big)f'(x)\right|\,dx\le\frac{1}{2}\int_0^\infty \left|f'(x)\right|\,dx=\frac{|f_0|}{2}<\infty$$

by our assumptions, the right-hand side of the above equality is a Cauchy sequence of n and hence it converges to some real number γ.

If $f^{(2m+1)}$ is non-decreasing for some $m\ge 1$, then $f^{(2m+1)}\le 0$ (because it tends to zero by assumption as $x\to\infty$), and therefore $f^{(2m-1)}$ is concave. Since $f^{(2m-1)}(x)$ is monotone, concave and tends to zero as $x\to\infty$, we conclude that $f^{(2m-1)}$ is also non-decreasing. Similarly, if $f^{(2m+1)}$ is non-increasing for some $m\ge 1$, then $f^{(2m-1)}$ is also non-increasing.

It follows that either all odd-order derivatives are non-increasing, or all odd-order derivatives are non-decreasing.[18] Changing f to $-f$ if necessary we may assume that all odd-order derivatives of f are non-increasing. Applying Theorem 10.11 and Proposition 10.12 we obtain for all integers $m,n\ge 1$ the relations

$$(f_0+\cdots+f_n)-\frac{f_0+f_n}{2}-\int_0^n f(x)\,dx=\sum_{k=1}^{m-1}b_{2k}\big[f^{(2k-1)}\big]_0^n+\theta_{m,n}b_{2m}\big[f^{(2m-1)}\big]_0^n$$

for some $\theta_{m,n}\in[0,1]$.

We conclude by letting $n\to\infty$ and using the hypothesis $\lim_\infty f^{(2k-1)}=0$ for all $k\ge 1$. □

10.8 Asymptotic Expansions: Stirling's Series

We give three applications of the results of the preceding section. We recall Euler's celebrated result[19]:

$$\sum_{j=1}^\infty \frac{1}{j^2}=\frac{\pi^2}{6}.$$

First we give a numerical evaluation of this constant:

[18] The two cases occur simultaneously only if f' is constant, and thus $f'\equiv 0$.

[19] A proof is given in Exercise 10.6, p. 265.

Proposition 10.14 *For any fixed integers $n, m \geq 1$ there exists a $\theta_{n,m} \in [0,1]$ such that*

$$\frac{\pi^2}{6} = \sum_{j=1}^{n} \frac{1}{j^2} + \sum_{k=0}^{2m-1} \frac{B_k}{n^{k+1}} + \theta_{n,m} \frac{B_{2m}}{n^{2m+1}}.$$

Proof Applying Corollary 10.13 for $f(x) := (x+n)^{-2}$ we have

$$f_0^{(j)} = (-1)^j \frac{(j+1)!}{n^{j+2}} \quad \text{for all} \quad j = 0, 1, \dots,$$

so that

$$\gamma_n = \sum_{k=1}^{m-1} \frac{b_{2k}(2k)!}{n^{2k+1}} + \theta_{n,m} \frac{b_{2m}(2m)!}{n^{2m+1}} = \sum_{k=1}^{m-1} \frac{B_{2k}}{n^{2k+1}} + \theta_{n,m} \frac{B_{2m}}{n^{2m+1}}$$

for some $\theta_{n,m} \in [0,1]$, where

$$\begin{aligned} \gamma_n &= \lim_{j \to \infty} \left(\frac{1}{n^2} + \cdots + \frac{1}{j^2} \right) - \left(\frac{1}{2n^2} + \frac{1}{2j^2} \right) - \int_n^j \frac{1}{x^2}\, dx \\ &= \lim_{j \to \infty} \left(\frac{1}{n^2} + \cdots + \frac{1}{j^2} \right) - \left(\frac{1}{2n^2} + \frac{1}{2j^2} \right) + \frac{1}{j} - \frac{1}{n} \\ &= \frac{\pi^2}{6} - \sum_{j=1}^{n} \frac{1}{j^2} + \frac{1}{2n^2} - \frac{1}{n}. \end{aligned}$$

Combining the two expressions of γ_n we obtain the equality

$$\frac{\pi^2}{6} = \sum_{j=1}^{n} \frac{1}{j^2} + \frac{1}{n} - \frac{1}{2n^2} + \sum_{k=1}^{m-1} \frac{B_{2k}}{n^{2k+1}} + \theta_{n,m} \frac{B_{2m}}{n^{2m+1}}.$$

The proposition follows by using the equalities $B_0 = 1$, $B_1 = -1/2$ and $B_3 = B_5 = \cdots = 0$. □

Examples

- For $m = 3$ the formula takes the form

$$\frac{\pi^2}{6} = \sum_{j=1}^{n} \frac{1}{j^2} + \frac{1}{n} - \frac{1}{2n^2} + \frac{1}{6n^3} - \frac{1}{30n^5} + \frac{\theta_{n,3}}{42n^7}.$$

For $n = 1$ this yields

$$\frac{\pi^2}{6} \approx 1.63$$

with an error < 0.03, while for $n = 10$ we get

$$\frac{\pi^2}{6} \approx 1.644\,934\,064$$

with an error $< 3 \cdot 10^{-9}$.

The proposition implies the *asymptotic expansion*

$$\frac{\pi^2}{6} \sim \sum_{j=1}^{n} \frac{1}{j^2} + \sum_{k=0}^{\infty} \frac{B_k}{n^{k+1}} \tag{10.20}$$

in the following sense:

Definition A series

$$a_0 + \frac{a_1}{n} + \frac{a_2}{n^2} + \cdots$$

is an *asymptotic expansion* of a sequence (f_n) if

$$f_n = \sum_{k=0}^{m} \frac{a_k}{n^k} + o\Big(\frac{1}{n^m}\Big) \quad \text{as} \quad n \to \infty$$

for each fixed $m = 0, 1, 2, \dots$.

We express this property by writing

$$f_n \sim a_0 + \frac{a_1}{n} + \frac{a_2}{n^2} + \cdots .$$

The more general relation

$$g_n \sim h_n + a_0 + \frac{a_1}{n} + \frac{a_2}{n^2} + \cdots$$

means that the preceding relation holds with $f_n = g_n - h_n$.

Our second application provides a numerical evaluation of Euler's constant C:[20]

Proposition 10.15 *For any fixed integers $n, m \ge 1$ there exists a $\theta_{n,m} \in [0, 1]$ such that*

$$C = \sum_{j=1}^{n} \frac{1}{j} - \ln n + \sum_{k=1}^{2m-1} \frac{B_k}{k \cdot n^k} + \theta_{n,m} \frac{B_{2m}}{2m \cdot n^{2m}}.$$

[20] See the example following Proposition 10.6, p. 244.

Proof Applying Corollary 10.13 for $f(x) := (x+n)^{-1}$ we have

$$f_0^{(j)} = (-1)^j \frac{j!}{n^{j+1}} \quad \text{for all} \quad j = 0, 1, \dots,$$

so that

$$\gamma_n = \sum_{k=1}^{m-1} \frac{b_{2k}(2k-1)!}{n^{2k}} + \theta_{n,m} \frac{b_{2m}(2m-1)!}{n^{2m}} = \sum_{k=1}^{m-1} \frac{B_{2k}}{2k \cdot n^{2k}} + \theta_{n,m} \frac{B_{2m}}{2m \cdot n^{2m}}$$

for some $\theta_{n,m} \in [0, 1]$, where

$$\begin{aligned}
\gamma_n &= \lim_{j\to\infty} \left(\frac{1}{n} + \cdots + \frac{1}{j}\right) - \left(\frac{1}{2n} + \frac{1}{2j}\right) - \int_n^j \frac{1}{x}\,dx \\
&= \lim_{j\to\infty} \left(\frac{1}{n} + \cdots + \frac{1}{j}\right) - \left(\frac{1}{2n} + \frac{1}{2j}\right) - \ln j + \ln n \\
&= \lim_{j\to\infty} \left(\frac{1}{1} + \cdots + \frac{1}{j} - \ln j\right) - \left(\frac{1}{1} + \cdots + \frac{1}{n-1}\right) - \frac{1}{2n} + \ln n \\
&= C - \left(\frac{1}{1} + \cdots + \frac{1}{n}\right) + \frac{1}{2n} + \ln n
\end{aligned}$$

Combining the two expressions of γ_n we obtain the equality

$$C = \left(\frac{1}{1} + \cdots + \frac{1}{n} - \ln n\right) - \frac{1}{2n} + \sum_{k=1}^{m-1} \frac{B_{2k}}{2k \cdot n^{2k}} + \theta_{n,m} \frac{B_{2m}}{2m \cdot n^{2m}}.$$

The proposition follows by using the equalities $B_1 = -1/2$ and $B_3 = B_5 = \cdots = 0$. □

The proposition implies the asymptotic expansion

$$C \sim \left(\frac{1}{1} + \cdots + \frac{1}{n} - \ln n\right) + \sum_{j=1}^{\infty} \frac{B_j}{j \cdot n^j}. \tag{10.21}$$

Example For $m = 3$ and $n = 10$ the proposition yields

$$C \approx 0.577\,215\,665$$

with an error $< 4 \cdot 10^{-9}$.

Finally improve Proposition 10.7 (p. 244):

Theorem 10.16 (Stirling's Series) *For any fixed integers $n, m \geq 1$ there exists a $\theta_{n,m} \in [0,1]$ such that*

$$\ln n! = \left(n + \frac{1}{2}\right) \ln n - n + \ln \sqrt{2\pi} + \sum_{k=1}^{m-1} \frac{B_{2k}}{2k(2k-1) \cdot n^{2k-1}} + \theta_{n,m} \frac{B_{2m}}{2m(2m-1) \cdot n^{2m-1}}.$$

Proof Fix two integers $m, n \geq 1$. Applying Corollary 10.13 for $f(x) := \ln(n + x)$ we get

$$\begin{aligned} \gamma_n &= -\sum_{k=1}^{m-1} \frac{b_{2k}(2k-2)!}{n^{2k-1}} - \theta_{n,m} \frac{b_{2m}(2m-2)!}{n^{2m-1}} \\ &= -\sum_{k=1}^{m-1} \frac{B_{2k}}{2k(2k-1) \cdot n^{2k-1}} - \theta_{n,m} \frac{B_{2m}}{2m(2m-1) \cdot n^{2m-1}} \end{aligned}$$

for some $\theta_{n,m} \in [0,1]$, where

$$\begin{aligned} \gamma_n &= \lim_{j \to \infty} (\ln n + \cdots + \ln j) - \frac{\ln n + \ln j}{2} - \int_n^j \ln x \, dx \\ &= \lim_{j \to \infty} (\ln n + \cdots + \ln j) - \frac{\ln n + \ln j}{2} - [x \ln x - x]_n^j \\ &= \lim_{j \to \infty} (\ln n + \cdots + \ln j) - \left(j + \frac{1}{2}\right) \ln j + j + \left(n - \frac{1}{2}\right) \ln n - n. \end{aligned}$$

Since

$$\ln n + \cdots + \ln j = \ln j! - \ln n! + \ln n,$$

hence (using Proposition 10.7 in the last step)

$$\begin{aligned} \gamma_n &= \lim_{j \to \infty} \ln j! - \left(j + \frac{1}{2}\right) \ln j + j - \ln n! + \left(n + \frac{1}{2}\right) \ln n - n \\ &= \lim_{j \to \infty} \ln \frac{j! e^j}{j^{j + \frac{1}{2}}} - \ln n! + \left(n + \frac{1}{2}\right) \ln n - n \\ &= \ln \sqrt{2\pi} - \ln n! + \left(n + \frac{1}{2}\right) \ln n - n. \end{aligned}$$

Combining the two expressions of γ_n we obtain

$$\ln n! = \left(n+\frac{1}{2}\right)\ln n - n + \ln\sqrt{2\pi} + \sum_{k=1}^{m-1}\frac{B_{2k}}{2k(2k-1)\cdot n^{2k-1}} + \theta_{n,m}\frac{B_{2m}}{2m(2m-1)\cdot n^{2m-1}}$$

for some $\theta_{n,m} \in [0,1]$. □

The preceding theorem yields the asymptotic expansion

$$\ln n! \sim \left(n+\frac{1}{2}\right)\ln n - n + \ln\sqrt{2\pi} + \sum_{k=1}^{\infty}\frac{B_{2k}}{2k(2k-1)\cdot n^{2k-1}}. \tag{10.22}$$

Remark The asymptotic expansions (10.20)–(10.22) of this section are divergent because $|B_{2n}| \to \infty$ very fast.[21] Nevertheless, even divergent asymptotic expansions are very useful in representing many important special functions.[22]

10.9 The Trapezoidal Rule: Romberg's Method

Let g be a continuous real function on a compact interval $I = [a,b]$. Subdividing I into n equal subintervals of length $h = (b-a)/n$, applying the trapezoidal rule on each of them, and adding them together, we obtain the *composite trapezoidal rule*:

$$\int_I g(y)\,dy \approx T_n(g) := \sum_{j=1}^{n}\frac{g(a+(j-1)h)+g(a+jh)}{2}h.$$

It was observed long ago[23] that for some functions the convergence of $T_n(g)$ to $\int_I g\,dx$ is unexpectedly fast as $n \to \infty$. This *superconvergence* is a consequence of the following result.

Proposition 10.17 *If $g \in C^m(I)$, then*

$$\int_I g(y)\,dy = T_n(g) + \sum_{k=1}^{[m/2]} c_{2k}n^{-2k} + O(n^{-m})$$

[21] See Exercises 10.7 and 10.8, p. 265.

[22] See, e.g., Burkill [70], Erdélyi [140], Watson [506], Whittaker–Watson [511].

[23] See Hairer–Wanner [217], pp. 129 and 164.

as $n \to \infty$, with the constants

$$c_{2k} := -b_{2k}\big[g^{(2k-1)}\big]_a^b (b-a)^{2k}.$$

Proof Applying Euler's formula for

$$f(x) := g(a+xh)h$$

we obtain the following relations:

$$\begin{aligned}
T_n(g) &= (f_0 + \cdots + f_n) - \frac{f_0 + f_n}{2}\\
&= \int_0^n f(x)\,dx + \sum_{k=1}^{[m/2]} b_{2k}\big[f^{(2k-1)}\big]_0^n + (-1)^{m-1}\int_0^n \beta_m(x) f^{(m)}(x)\,dx\\
&= \int_a^b g(y)\,dy + \sum_{k=1}^{[m/2]} b_{2k}\big[g^{(2k-1)}\big]_a^b h^{2k}\\
&\qquad + (-1)^{m-1}\int_a^b \beta_m\Big(\frac{y-a}{h}\Big) g^{(m)}(y)\,dy \cdot h^m\\
&= \int_a^b g(y)\,dy - \sum_{k=1}^{[m/2]} c_{2k} n^{-2k} + O(n^{-m})
\end{aligned}$$

as $n \to \infty$. In the last step we have used the equality $h = (b-a)/n$ and the boundedness of the functions β_m, $g^{(m)}$, implying that the last integral is bounded by a constant independent of h. □

Corollary 10.18 *If $g \in C^\infty(\mathbb{R})$ is a $(b-a)$-periodic function, then the constants c_{2k} all vanish, so that*

$$\int_I g\,dx = T_n(g) + O(n^{-m})$$

for each $m = 1, 2, \ldots$.

Remarks Romberg noticed that the above proposition allows us to construct efficient numerical integration formulas. We give two examples.

- If $g \in C^4(I)$, then applying the proposition with n and $2n$ we obtain the relations

$$\int_I g\,dx = T_n(g) + c_2 n^{-2} + O(n^{-4})$$

and

$$\int_I g\,dx = T_{2n}(g) + \frac{c_2}{4}n^{-2} + O(n^{-4}).$$

Combining them to eliminate c_2 we obtain the *composite Simpson's rule*

$$S_n(g) := \frac{4}{3}T_{2n}(g) - \frac{1}{3}T_n(g).$$

It is more efficient than the composite trapezoidal rule because

$$\int_I g\,dx = S_n(g) + O(n^{-4})$$

instead of

$$\int_I g\,dx = T_n(g) + O(n^{-2}).$$

- If $g \in C^6(I)$, then

$$\int_I g\,dx = T_n(g) + c_2 n^{-2} + c_4 n^{-4} + O(n^{-6}),$$

and repeating the preceding reasoning we now get

$$\int_I g\,dx = S_n(g) - \frac{c_4}{4}n^{-4} + O(n^{-6}).$$

We may eliminate c_4 by introducing the integration rule

$$R_n(g) := \frac{15}{16}S_{2n} - \frac{1}{16}S_n.$$

This is even more efficient than the previous one, because

$$\int_I g\,dx = R_n(g) + O(n^{-6}).$$

This is no longer a composite Newton–Cotes rule.

Continuing this procedure we may construct powerful integration rules for sufficiently smooth functions.

10.10 Exercises

Exercise 10.1 Let (p_n) be a sequence of orthogonal polynomials on some interval I and $x_{n,1}, \dots, x_{n,n}$ be the zeros of p_n, $n = 1, 2, \dots$. Prove that the set $\{x_{n,k}\}$ is dense in I.

Exercise 10.2 Let $f : [0, \infty) \to \mathbb{R}$ be a continuous, non-increasing function. As usual in this section, we write f_k instead of $f(k)$ for brevity. For each integer $k \geq 0$ we denote by t_k and T_k the open domains bounded by the vertical line $x = k$, the straight line joining the points (k, f_k) and $(k+1, f_{k+1})$, and the graph of f and its tangent line at $k+1$, respectively. See Fig. 10.8.

Prove the following:

(i) $t_k \subset T_k$ for each k;

(ii) if we translate the triangles T_k such that their right vertices move to $(1, f_1)$, then the translated triangles are pairwise disjoint, and they all belong to a right triangle of sides 1 and $f_0 - f_1$;

(iii) the sequence

$$(f_0 + \cdots + f_n) - \frac{f_0 + f_n}{2} - \int_0^n f(x)\, dx$$

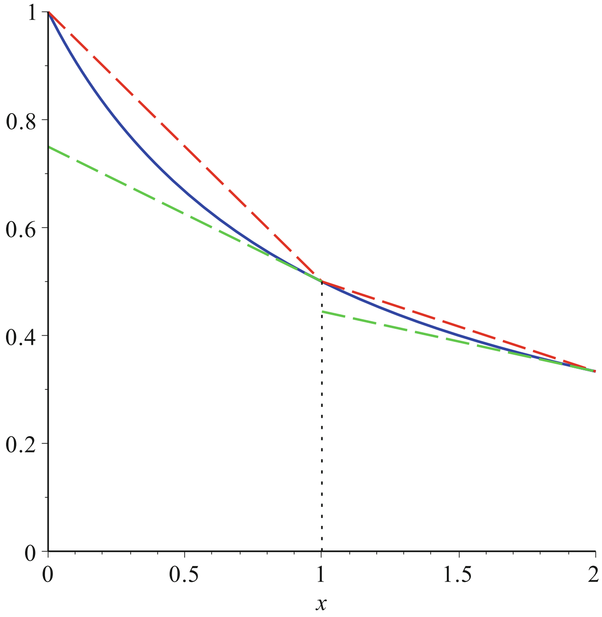


Fig. 10.8 Exercise 10.2 for $f(x) := 1/(1+x)$

has a finite limit γ as $n \to \infty$, and

$$0 \le \gamma \le \frac{f_0 - f_1}{2}.$$

Apply (iii) to the functions $f(x) := 1/(1+x)$ and $f(x) := -\ln(1+x)$ to prove the convergence of the sequences[24]

$$1 + \frac{1}{2} + \cdots + \frac{1}{n} - \ln n \quad \text{and} \quad \frac{n! e^n}{n^{n+\frac{1}{2}}}$$

as $n \to \infty$.

Exercise 10.3 (Wallis Formula) Set

$$J_m := \int_0^1 (1-x^2)^{\frac{m-1}{2}}\, dx, \quad m = 1, 2, \ldots.$$

(i) Show that $J_1 = 1$ and $J_2 = \frac{\pi}{4}$.

(ii) Verify the following recursive relation for $m = 3, 4, \ldots$:

$$J_m = (m-1)(J_{m-2} - J_m).$$

(iii) Introducing the *semifactorial* notation

$$(2n)!! := (2n)\cdot(2n-2)\cdots 4\cdot 2 \quad \text{and} \quad (2n+1)!! := (2n+1)\cdot(2n-1)\cdots 3\cdot 1,$$

show that

$$J_{2n} = \frac{(2n-1)!!}{(2n)!!}\cdot\frac{\pi}{2} \quad \text{and} \quad J_{2n+1} = \frac{(2n)!!}{(2n+1)!!}$$

for $n = 1, 2, \ldots$.

(iv) Show that $J_{2n+1} \le J_{2n} \le J_{2n-1}$ for all $n \ge 1$, and that they are equivalent to the inequalities

$$\frac{2n}{2n+1}\cdot\frac{\pi}{2} \le \frac{2\cdot 2}{1\cdot 3}\cdot\frac{4\cdot 4}{3\cdot 5}\cdots\frac{2n\cdot 2n}{(2n-1)\cdot(2n+1)} \le \frac{\pi}{2}.$$

(v) Infer from (iv) the *Wallis formula:*

$$\frac{\pi}{2} = \lim_{n\to\infty} \frac{2\cdot 2}{1\cdot 3}\cdot\frac{4\cdot 4}{3\cdot 5}\cdots\frac{2n\cdot 2n}{(2n-1)\cdot(2n+1)}.$$

[24] See Propositions 10.6 and 10.7, p. 244.

Exercise 10.4 Let f satisfy the hypotheses of Corollary 10.13, and assume that f' is non-increasing. Prove the following:

(i) all odd-order derivatives of f are non-increasing;
(ii) all even-order derivatives of f are non-decreasing, and converge to zero as $x \to \infty$;
(iii) f is non-decreasing.

Exercise 10.5 (Fourier Series of Bernoulli Polynomials) Starting from the Fourier series

$$\frac{\pi - x}{2} = \sum_{n=1}^{\infty} \frac{\sin nx}{n} \quad (0 < x < 2\pi)$$

prove that

$$b_1(x) = -\sum_{n=1}^{\infty} \frac{\sin 2n\pi x}{n\pi} \quad (0 < x < 1),$$

and then that

$$b_{2k}(x) = (-1)^{k-1} \sum_{n=1}^{\infty} \frac{2\cos(2n\pi x)}{(2n\pi)^{2k}}$$

and

$$b_{2k+1}(x) = (-1)^{k-1} \sum_{n=1}^{\infty} \frac{2\sin(2n\pi x)}{(2n\pi)^{2k+1}}$$

for all $k = 1, 2, \dots$ and $x \in [0, 1]$.

Exercise 10.6 (Sums of Hyperharmonic Series) Deduce from the preceding exercise the equalities

$$\sum_{n=1}^{\infty} \frac{1}{n^{2k}} = (2\pi)^{2k} \frac{|b_{2k}|}{2}$$

for $k = 1, 2, \dots$.

Exercise 10.7 (Size of the Bernoulli Numbers) Using the preceding exercise prove the following relations:

(i) $\dfrac{2}{(2\pi)^{2k}} < |b_{2k}| < \dfrac{4}{(2\pi)^{2k}}$ for $k = 1, 2, \dots$;

(ii) $(2\pi)^{2k}\,|b_{2k}| \to 2$;

(iii) $b_{2k} \to 0$ and $|B_{2k}| \to \infty$. Moreover, $p(k)b_{2k} \to 0$ and $\frac{p(k)}{B_{2k}} \to 0$ for any polynomial p.

Exercise 10.8 Prove that all three asymptotic expansions in Sect. 10.8 are divergent.

Exercise 10.9 (Generating Functions)

(i) Prove the identity

$$\left(1 + \frac{t}{2!} + \frac{t^2}{3!} + \cdots\right)\left(b_0 + b_1 t + b_2 t^2 + \cdots\right) = 1$$

for all t satisfying $|t| < 2\pi$, and infer from this the relation

$$\frac{t}{e^t - 1} = \sum_{n=0}^{\infty} b_n t^n \quad \text{if} \quad |t| < 2\pi. \tag{10.23}$$

(ii) Show that the series $\sum b_n t^n$ is divergent if $|t| \geq 2\pi$.

(iii) Take the Cauchy product of the series

$$\sum_{n=0}^{\infty} b_n t^n \quad \text{and} \quad e^{xt} = \sum_{n=0}^{\infty} \frac{x^n}{n!} t^n$$

to prove that

$$\frac{te^{xt}}{e^t - 1} = \sum_{n=0}^{\infty} b_n(x) t^n \quad \text{if} \quad x \in \mathbb{R} \quad \text{and} \quad |t| < 2\pi.$$

Exercise 10.10 Starting with (10.23), prove the following Taylor expansions:

$$x \coth x = 1 + \sum_{k=1}^{\infty} b_{2k} \cdot (2x)^{2k} \quad \text{if} \quad |x| < \pi;$$

$$x \cot x = 1 + \sum_{k=1}^{\infty} (-1)^k b_{2k} \cdot (2x)^{2k} \quad \text{if} \quad |x| < \pi;$$

$$\tan x = \sum_{k=1}^{\infty} (-1)^k b_{2k} \cdot (2^{2k} - 4^{2k}) x^{2k-1} \quad \text{if} \quad |x| < \pi/2.$$

Chapter 11
Finding Roots

Since Descartes's groundbreaking work the exact or numerical solution of polynomial equations has become a frequent task in analytical geometry. We also have to find the roots of orthogonal polynomials for the implementation of the Gauss rules of numerical integration. Based on his differential calculus, Newton invented a powerful method for localizing the roots of twice differentiable functions.

In this chapter we give an introduction to these questions.

11.1 * Descartes's Rule of Signs

We start with a theorem of Descartes.

Definition Given a polynomial $p(x) = a_n x^n + \cdots + a_0$, we denote by

- $n_+ = n_+(p)$ the number of its (strictly) positive roots (counted with multiplicity);
- $N_+ = N_+(p)$ the number of sign changes[1] in $a_n, a_{n-1}, \ldots, a_0$.

Theorem 11.1 (Descartes) *We have* $n_+ \le N_+$.

Proof More generally, we consider a function of the form

$$p(x) = a_n x^{b_n} + a_{n-1} x^{b_{n-1}} + \cdots + a_0 x^{b_0}$$

with non-zero real coefficients a_j and arbitrary *real* exponents $b_n > \cdots > b_0$. We prove the inequality $n_+(p) \le N_+(p)$ by induction on $N_+(p)$.

[1] After the removal of its zero elements. For example, the sequence $1, 3, -2, 0, 1, -4$ has 3 sign changes.

V. Komornik, *Topology, Calculus and Approximation*, Springer
Undergraduate Mathematics Series, DOI 10.1007/978-1-4471-7316-8_11

If $N_+(p) = 0$, then each term of $p(x)$ has the same sign for $x > 0$, so that $n_+(p) = 0$.

Proceeding by induction, let $N_+(p) > 0$, and consider an index i for which $a_{i+1}a_i < 0$. Since $n_+(p)$ and $N_+(p)$ do not change when we divide $p(x)$ by some real power of x, we may assume that $b_{i+1} > 0 > b_i$. If we differentiate p, then none of the coefficients $a_n, \dots, a_{i+1}$ changes sign, while all coefficients $a_i, \dots, a_0$ change sign, so that $N_+(p') = N_+(p) - 1$. Since $n_+(p') \geq n_+(p) - 1$ by the generalized Rolle theorem (Lemma 8.8, p. 205), using the induction hypothesis we obtain that

$$n_+(p) \leq n_+(p') + 1 \leq N_+(p') + 1 = N_+(p).$$

□

Remarks

- If a_{n_0} is the last non-zero coefficient, then the factorization of p shows that a_n and a_{n_0} have equal signs if $n_+(p)$ is even, and different signs if $n_+(p)$ is odd. Consequently, $N_+(p)$ and $n_+(p)$ always have the same parity.
- We can also estimate the number $n_- = n_-(p)$ of negative roots of p. Indeed, putting $q(x) := p(-x)$, we obviously have $n_-(p) = n_+(q)$, so that setting $N_- = N_-(p) := N_+(q)$ we deduce from the theorem that $n_- \leq N_-$.
- Since $N_-(p) - n_-(p) = N_+(q) - n_+(q)$, $N_-(p)$ and $n_-(p)$ also have the same parity.

Example For the polynomial $p(x) = x^3 + 3x^2 - 1$ we have $N_+(p) = 1$. Since $N_+(p)$ and $n_+(p)$ always have the same parity, we conclude that p has a unique positive root.

If the polynomial is known to have only real roots,[2] then the theorem may be sharpened:

Proposition 11.2

(a) *We have* $N_+(p) + N_-(p) \leq \deg p$ *for all non-zero polynomials.*
(b) *If* p *has only real roots, then*

$$n_+ = N_+ \quad \textit{and} \quad n_- = N_-.$$

Proof

(a) Let us write

$$p(x) = a_n x^{b_n} + a_{n-1}x^{b_{n-1}} + \cdots + a_0 x^{b_0}$$

[2] For example, when p is the characteristic polynomial of a symmetric matrix.

with non-zero real coefficients a_j and nonnegative *integers* $b_n > \cdots > b_0$. Then

$$p(-x) = (-1)^{b_n} a_n x^{b_n} + (-1)^{b_{n-1}} a_{n-1} x^{b_{n-1}} + \cdots + (-1)^{b_0} a_0 x^{b_0}.$$

Hence N_+ is the number of elements of the set of indices $1 \le i \le n$ satisfying $a_i a_{i-1} < 0$, and N_- is the number of elements of the set of indices $1 \le i \le n$ satisfying $(-1)^{b_i} a_i (-1)^{b_{i-1}} a_{i-1} < 0$.

If an index i belongs to both sets, then $b_i - b_{i-1}$ is a positive even integer, hence $b_i - b_{i-1} \ge 2$. Since the other differences $b_i - b_{i-1}$ are positive integers, this implies that

$$N_+(p) + N_-(p) \le \sum_{i=1}^{n} (b_i - b_{i-1}) = b_n - b_0 \le b_n = \deg p.$$

(b) Our assumption implies that p is not identically zero. If $p(0) = 0$, then dividing $p(x)$ by x the numbers n_+, N_+, n_-, N_- remain unchanged. We may therefore assume that $p(0) \ne 0$. Then p has only non-zero real roots and therefore $\deg p = n_+ + n_-$. Using the preceding theorem and part (a) above it follows that

$$\deg p = n_+ + n_- \le N_+ + N_- \le \deg p,$$

and hence $n_+ + n_- = N_+ + N_-$. This implies that none of the inequalities $n_+ \le N_+$ and $n_- \le N_-$ of the preceding theorem may be strict. □

11.2 * Sturm Sequences

Completing the unpublished work of Fourier, Sturm developed an efficient method of localizing the real roots of a polynomial.

Definition A finite sequence $p_0, p_1, \dots, p_n$ of C^1 functions $\mathbb{R} \to \mathbb{R}$ is a *Sturm sequence* if

(a) p_0 has no root;
(b) $p_k(\alpha) = 0 \Longrightarrow p_{k-1}(\alpha) p_{k+1}(\alpha) < 0, \quad k = 1, \dots, n-1;$
(c) $p_n(\alpha) = 0 \Longrightarrow (p_n p_{n-1})'(\alpha) > 0$.

Example If $(p_n)_{n \ge 0}$ is an orthogonal sequence of polynomials, then $p_0, p_1, \dots, p_n$ is a Sturm sequence for each n by Corollary 9.5 (p. 226).

Lemma 11.3 *If $p_0, p_1, \dots, p_n$ is a Sturm sequence, then*

(d) *p_n has no multiple roots;*
(e) *consecutive functions have no common roots.*

Proof If p_n had a multiple root α, then the equality $p_n(\alpha) = p'_n(\alpha) = 0$ would imply

$$(p_n p_{n-1})'(\alpha) = p_n(\alpha)p'_{n-1}(\alpha) + p'_n(\alpha)p_{n-1}(\alpha) = 0,$$

contradicting (c).

If $p_k(\alpha) = 0$ for some real α and integer $k < n$, then $k \geq 1$ by (a), and then $p_{k-1}(\alpha)p_{k+1}(\alpha) < 0$ by (b). Hence $p_{k+1}(\alpha) \neq 0$. □

Let $p_0, p_1, \dots, p_n$ be a Sturm sequence. For any real c we denote by $N(c)$ the number of sign changes in $p_0(c), p_1(c), \dots, p_n(c)$ (after the removal of the zero elements).

Remark If the sequence $p_0(c), p_1(c), \dots, p_n(c)$ contains three consecutive elements $p_{k-1}(c), p_k(c), p_{k+1}(c)$ such that $p_{k-1}(c)p_{k+1}(c) < 0$, then the removal of $p_k(c)$ does not alter the number of sign changes. Repeating this elimination procedure, we eventually arrive at a reduced sequence of length $1 + N(c)$.

Theorem 11.4 (Sturm) *If $p_0, p_1, \dots, p_n$ is a Sturm sequence, then p_n has exactly $N(\alpha) - N(\beta)$ roots in any interval $(\alpha, \beta]$.*

Proof Consider the function $c \mapsto N(c)$ defined on $\mathbb{R}$. If $p_0(c), \dots, p_n(c)$ are all nonzero for some c, then N keeps a constant value in a neighborhood of c because the continuous functions p_k do not change sign here.

This property remains valid under the weaker condition $p_n(c) \neq 0$. Indeed, if $p_k(c) = 0$ for some $k < n$, then $k \geq 1$ by (a), and $p_{k-1}(c)p_{k+1}(c) < 0$ by (b). Hence p_{k-1} and p_{k+1} keep their *different* signs in a neighborhood of c, so that the sign p_k in this neighborhood does not affect the number of sign changes by the remark preceding the statement of the theorem.

Since p_n has only simple roots by (d), it remains to show that $N(c)$ decreases by one when we cross a root c of p_n.

If $p_n(c) = 0$, then $(p_n p_{n-1})(c) = 0$, and $(p_n p_{n-1})'(c) > 0$ by property (c), so that $p_n p_{n-1} < 0$ just before c, and $p_n p_{n-1} > 0$ just after c. Hence $N(c') = N(c) + 1$ for c' just before c, and $N(c') = N(c)$ for c' just after c. □

Remarks

- If the sequence $p_0, p_1, \dots, p_n$ satisfies (a), (b) and (c') where (c') is obtained from (c) by changing the inequality $>$ to $<$, then p_n has exactly $N(\beta) - N(\alpha)$ roots in every interval $[\alpha, \beta)$: it suffices to apply the theorem for the functions $p_j(-x)$.
- If we consider a sequence of orthogonal polynomials p_k with positive leading coefficients in an interval (a, b), then $N(x) = 0$ for all $x \geq b$, because $p_k(x) > 0$ for all $k \geq 0$ and $x \geq b$. Hence p_n has exactly $N(c)$ roots in (c, ∞) for any c.

11.3 * Roots of Polynomials

We seek the real roots of a given polynomial p. Dividing p by the greatest common divisor of p and p',[3] we may assume that p has no multiple roots. We are going to construct a Sturm sequence allowing us to localize the roots of $p_n = p$.

We may assume that $\deg p \ge 1$. Set $q_0 = p$, $q_1 = p'$, and define the polynomials $q_2, q_3, \dots$ and $r_1, r_2, \dots$ by the Euclidean algorithm[4]:

$$q_{k-2} = q_{k-1}r_{k-1} - q_k, \quad \deg q_k < \deg q_{k-1}, \quad 2 \le k \le n;$$

we stop at the last non-zero polynomial q_n.

Since p has no multiple roots by our assumption, the greatest common divisor q_n of p and p' is a non-zero constant. This implies that consecutive polynomials q_k have no common roots, because such a root would also be a root of q_n by the algorithm.

Proposition 11.5 *The formula*

$$p_k := q_{n-k}$$

defines a Sturm sequence $p_0, p_1, \dots, p_n$ with $p_n = p$.

Proof Condition (a) is satisfied because $p_0 = q_n$ is a non-zero constant.

For the proof of (c) we remark that if $p_n(\alpha) = 0$ for some real α, then

$$(p_np_{n-1})'(\alpha) = (pp')'(\alpha) = p'(\alpha)^2 \ge 0;$$

moreover, the last inequality is strict because α is a *simple* root of p.

Finally, we establish (b) equivalently for the polynomials q_k instead of p_k. If $q_k(\alpha) = 0$ for some real α and for some $0 < k < n$, then multiplying the equality $q_{k-1} = q_kr_k - q_{k+1}$ by q_{k+1}, we obtain for $x = \alpha$ the equality

$$q_{k-1}(\alpha)q_{k+1}(\alpha) = -q_{k+1}(\alpha)^2 \le 0.$$

The last inequality is strict because consecutive polynomials q_k have no common roots, so that $q_{k+1}(\alpha) \ne 0$. □

Before starting Sturm's procedure it is useful to find an initial bounded interval containing all roots of the polynomial under investigation. We may apply the following simple result:

Proposition 11.6 *All roots α of*

$$p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$$

[3] It can be computed by the Euclidean algorithm.

[4] Notice the sign change for the remainders.

satisfy the inequality

$$|\alpha| < 1 + \max\{|a_{n-1}|, \ldots, |a_0|\}.$$

Proof Denote the above maximum by A. If $|x| \geq 1 + A$, then

$$\begin{aligned} |a_{n-1}x^{n-1} + \cdots + a_1 x + a_0| &\leq A(|x|^{n-1} + \cdots + |x| + 1) \\ &\leq (|x| - 1)(|x|^{n-1} + \cdots + |x| + 1) \\ &= |x|^n - 1 < |x|^n, \end{aligned}$$

so that $p(x) \neq 0$. □

Example Let us return to the polynomial $p(x) = x^3 + 3x^2 - 1$ of Sect. 11.1. By the preceding proposition its real roots belong to the interval $(-4, 4)$. A short computation shows that the roots of p are simple and its Sturm sequence is the following:

$$\begin{aligned} p_3(x) &= x^3 + 3x^2 - 1, \\ p_2(x) &= 3x^2 + 6x, \\ p_1(x) &= 2x + 1, \\ p_0(x) &= \frac{9}{4}. \end{aligned}$$

This yields the table

c	-3	-2	-1	0	1
$N(c)$	3	2	2	1	0

implying that p has exactly three real roots, one in each of the intervals $(-3, -2]$, $(-1, 0]$ and $(0, 1]$; see Fig. 11.1.

11.4 * The Method of Householder and Bauer

In this section we show that every symmetric matrix is similar to a tridiagonal symmetric matrix. This will be used in the next section for the evaluation of eigenvalues of symmetric matrices, a frequent task in geometry and physics.

We denote by B^* the transpose of the matrix B, and the elements of $\mathbb{R}^m$ are considered as column vectors. We recall that a matrix is orthogonal if and only if the linear map $v \mapsto Av$ is an isometry.[5]

[5] We consider the usual Euclidean norm on $\mathbb{R}^m$.

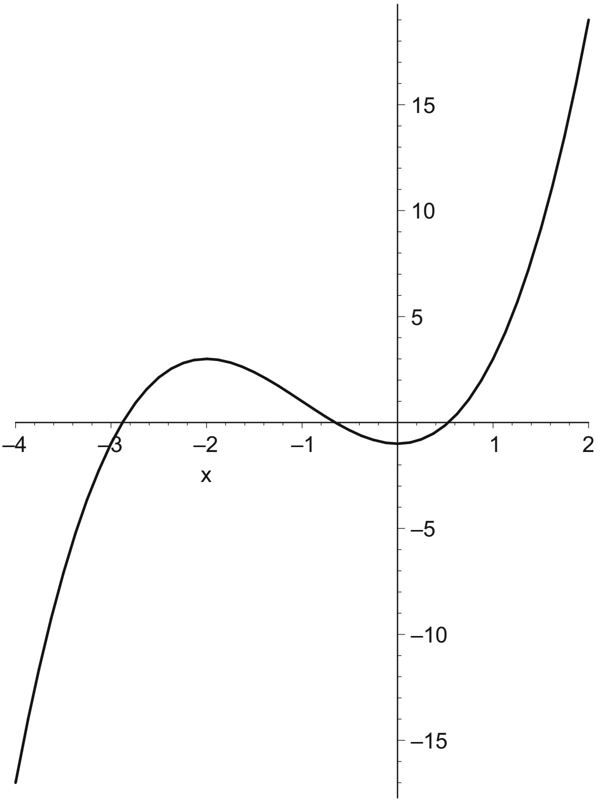


Fig. 11.1 Graph of $p(x) = x^3 + 3x^2 - 1$

Definition A square matrix $A = (a_{ij})$ is *tridiagonal* if

$$|i-j| \geq 2 \Longrightarrow a_{ij} = 0.$$

In other words, only the entries on the diagonal and those having a neighbor on the diagonal may be different from zero.

Proposition 11.7 (Householder–Bauer) *For every symmetric matrix A there exists an orthogonal matrix P such that P^*AP is tridiagonal.*

Remark The matrix P^*AP is also symmetric because

$$(P^*AP)^* = P^*A^*P^{**} = P^*AP.$$

For the proof we introduce the *Householder matrices*:

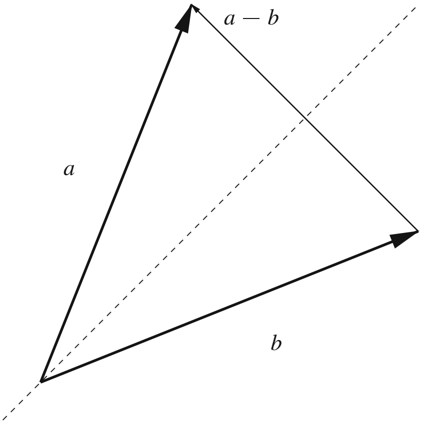


Fig. 11.2 $H(a-b)a=b$

Definition For any non-zero vector $v \in \mathbb{R}^m$ we denote by $H(v)$ the matrix of the reflection to the hyperplane that is orthogonal to v:

$$H(v)v=-v, \quad \text{and} \quad H(v)u=u \quad \text{for all} \quad u \perp v.$$

Furthermore, we set $H(0) := I$ (the identity matrix).[6]

Lemma 11.8

(a) *The Householder matrices are orthogonal.*
(b) *For each $a \in \mathbb{R}^m$ there exists a $v \in \mathbb{R}^m$ such that*

$$H(v)a=(|a|,0,\dots,0)^*.$$

Proof

(a) Neither the identity map nor the reflections change the norm of a vector.
(b) We claim that $H(a-b)a=b$ for $b:=(|a|,0,\dots,0)^*$.

The case $a=b$ is trivial. If $a \neq b$, then a and b are at the same distance from 0, and therefore they are symmetric to the hyperplane orthogonal to $a-b$ (see Fig. 11.2). Hence $H(a-b)a=b$ again. □

Proof of Proposition 11.7 There is nothing to prove if $n \le 2$. Given a square matrix $A=(a_{ij})$ of order $n \ge 3$, we construct by induction a sequence of orthogonal matrices $H_1,\dots,H_{n-2}$ such that setting

$$A_1 := A \quad \text{and} \quad A_{k+1} := H_k^* A_k H_k, \quad k=1,\dots,n-2,$$

the matrices $A_k=(a_{i,j}^k)$ satisfy the following conditions:

$$j \le k-1 \quad \text{and} \quad i \ge j+2 \Longrightarrow a_{ij}^k=0. \tag{11.1}$$

[6] The matrices $H(v)$ are symmetric, but we do not need this property here.

Since the matrices A_k are symmetric by construction,[7] this will imply that A_{n-1} is tridiagonal. Finally, the product matrix $P := H_1 \cdots H_{n-2}$ is orthogonal, and $A_{n-1} = P^*AP$.

The condition (11.1) is void for $k = 1$. Assume by induction that A_k has already been defined for some $1 \le k \le n-2$, and satisfies (11.1). Applying the lemma there exists a vector v_k such that

$$H(v_k)(a^k_{k+1,k}, \dots, a^k_{n,k})^* = (c, 0, \dots, 0)^*$$

with a suitable number $c \ge 0$. Then the formula

$$H_k := \begin{pmatrix} I_k & 0 \\ 0 & H(v_k) \end{pmatrix},$$

where I_k denotes the identity matrix of order k, defines an orthogonal matrix.

If we write

$$A_k := \begin{pmatrix} B_k & C_k^* \\ C_k & D_k \end{pmatrix}$$

where B_k is a square matrix of order k, then

$$A_{k+1} = \begin{pmatrix} B_k & C_k^* H(v_k) \\ H(v_k)C_k & H(v_k)D_kH(v_k) \end{pmatrix}$$

by a direct computation.

The first $k-1$ columns of C_k vanish by the induction hypothesis; they remain null vectors in $H(v_k)C_k$, too. Furthermore, the last column vector of C_k is replaced by $(c, 0, \dots, 0)^*$ in $H(v_k)C_k$ because of the choice of v_k. This shows that condition (11.1) is also satisfied for $k+1$ instead of k. □

11.5 * Givens' Method

We seek the eigenvalues of symmetric matrices. In view of the preceding section we consider only *tridiagonal* matrices:

$$A = \begin{pmatrix} b_1 & c_1 & & & \\ c_1 & b_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & b_{n-1} & c_{n-1} \\ & & & c_{n-1} & b_n \end{pmatrix}.$$

[7] See the remark following the statement of the proposition.

If $c_j = 0$ for some j, then A is the direct sum of two smaller symmetric tridiagonal matrices, and the *spectrum*[8] of A is the union of their spectra. Therefore it is sufficient to consider the case where none of the numbers c_i vanish.[9]

We introduce for $k = 1, \dots, n$ the submatrices

$$A_k := \begin{pmatrix} b_1 & c_1 & & & \\ c_1 & b_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & b_{k-1} & c_{k-1} \\ & & & c_{k-1} & b_k \end{pmatrix}$$

and their characteristic polynomials

$$p_k(\lambda) := \det(\lambda I_k - A_k).$$

We also set $p_0(\lambda) = 1$.

Proposition 11.9 (Givens) *If none of the numbers c_k vanish, then $p_0, p_1, \dots, p_n$ is a Sturm sequence.*

Proof We have to check the properties (a), (b), (c) on page 269. Property (a) is obvious because $p_0 \equiv 1$ has no root.

It follows from the definition of the determinant that p_k is a polynomial of degree k with leading coefficient 1.

Developing the determinant $p_{k+1}(\lambda)$ according its last column we obtain the recurrence relation

$$p_{k+1}(\lambda) = (\lambda - b_{k+1})p_k(\lambda) - c_k^2 p_{k-1}(\lambda)$$

for $k = 1, \dots, n-1$.

Since the numbers c_k are different from zero, henceforth we may repeat the proof of Proposition 9.4 (b), and then that of Corollary 9.5 (pp. 223–226).[10] □

The above proposition enables us to apply Sturm's theorem for the evaluation of the eigenvalues of A. The next result allows us to find an initial set of disks containing all eigenvalues:

[8] The spectrum of a matrix is by definition the set of its eigenvalues.

[9] We will find that A has n distinct (real) eigenvalues in this case.

[10] See also the second remark following the proof of Proposition 9.4.

Proposition 11.10 (Gerschgorin) *If λ is an eigenvalue of the matrix (a_{ij}), then there exists an index i such that*

$$|\lambda - a_{ii}| \le \sum_{j \ne i} |a_{ij}|.$$

Proof If $x = (x_1, \dots, x_n)^*$ is a non-zero eigenvector associated with λ, then

$$\sum_{j=1}^{n} a_{ij}x_j = \lambda x_i,$$

or equivalently

$$\sum_{j \ne i} a_{ij}x_j = (\lambda - a_{ii})x_i$$

for all $i = 1, \dots, n$. If $|x_i| = \max_{1 \le j \le n} |x_j|$, then

$$|\lambda - a_{ii}| = \Bigl|\sum_{j \ne i} a_{ij}\frac{x_j}{x_i}\Bigr| \le \sum_{j \ne i} |a_{ij}|. \qquad \square$$

11.6 Newton's Method

Newton invented a general method for the solution of equations of the form $f(x) = 0$ for sufficiently smooth functions f. Here we consider only a special case.[11]

Let $f : [a, b] \to \mathbb{R}$ be a C^2 function satisfying

$$f(a)f(b) < 0,$$

and[12]

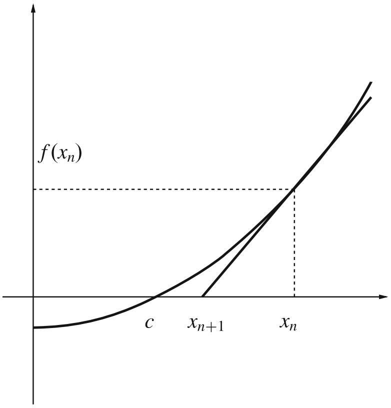
$$f'(x) > 0,\ f''(x) > 0 \quad \text{for all} \quad x \in [a, b].$$

Then the equation $f(c) = 0$ has a solution in (a, b) by Bolzano's theorem (p. 47), and this solution is unique because f is increasing by the condition $f' > 0$.

[11] See Exercise 11.3 for a more general result, p. 281.

[12] The more general condition "$f' \ne 0$ and $f'' \ne 0$ in $[a, b]$" may be reduced to this one by replacing $f(x)$ with $f(-x)$, $-f(x)$ or $-f(-x)$.

Fig. 11.3 Newton's method



Proposition 11.11 (Newton) *If $x_0 \in [a,b]$ and $f(x_0) > 0$, then the recursive formula*

$$x_{n+1} := x_n - \frac{f(x_n)}{f'(x_n)}, \quad n = 0, 1, \ldots \tag{11.2}$$

defines a (strictly) decreasing sequence, converging to c. (See Fig. 11.3.)

Proof Since f is strictly increasing and $f(x_0) > 0 = f(c)$, we have $c < x_0 \le b$. We show that if $c < x_n \le b$ for some n, then $c < x_{n+1} < x_n$. In particular, (x_n) is a well-defined, (strictly) decreasing sequence in $(c, b]$.

First we observe that

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} < x_n$$

because $f(x_n) > 0$ and $f'(x_n) > 0$. On the other hand, applying the Lagrange form of Taylor's formula (p. 127) we have

$$0 = f(c) = f(x_n) + f'(x_n)(c - x_n) + \frac{f''(d)}{2}(c - x_n)^2 \tag{11.3}$$

for a suitable point $d \in (c, x_n)$. Hence

$$f(x_n) + f'(x_n)(c - x_n) < 0$$

because $f''(d) > 0$, and this is equivalent to

$$c < x_n - \frac{f(x_n)}{f'(x_n)} = x_{n+1}$$

because $f'(x_n) > 0$.

The sequence (x_n) is monotone and bounded, hence it converges to some limit $x \in [c, b]$.

Letting $n \to \infty$ in (11.2) we get $f(x) = 0$ by continuity, and then $x = c$ by the injectivity of f. □

Remark Under additional assumptions we may estimate the convergence speed, too. For this we deduce from (11.3) the equality

$$x_{n+1} - c = x_n - c - \frac{f(x_n)}{f'(x_n)} = \frac{f''(d)}{2f'(x_n)}(c - x_n)^2.$$

Setting

$$A := \frac{\max f''}{2 \min f'}$$

we deduce from this equality the inequalities

$$|A(x_{n+1} - c)| \le |A(x_n - c)|^2, \quad n = 0, 1, \dots,$$

and then by induction the estimates

$$|A(x_n - c)| \le |A(x_0 - c)|^{2^n}, \quad n = 0, 1, \dots.$$

This shows that if $|x_0 - c| < 1/A$, then the convergence $x_n \to c$ is very fast.[13]

Example We apply Newton's method to the function $f(x) := x^2 - 2$ in an interval $[a, b]$ with $0 < a < \sqrt{2} < b$. Starting from an arbitrary point $\sqrt{2} < x_0 < b$, the sequence defined by the formula

$$x_{n+1} := x_n - \frac{x_n^2 - 2}{2x_n} = \frac{1}{2}\left(x_n + \frac{2}{x_n}\right)$$

converges rapidly to $\sqrt{2}$. It is interesting to compare the convergence speed estimate of the preceding remark with the one obtained after the fixed point theorem 1.10 (p. 17). The case $a < x_0 < \sqrt{2}$ is also worth investigating.

[13] A convergence of this kind is called *quadratic*.

11.7 Exercises

Exercise 11.1 Apply Newton's method to the function $f(x) := x^p - A$ where $p \geq 2$ is an arbitrary integer and A is an arbitrary positive number.[14]

What happens if we start Newton's method with a point $0 < x_0 < A^{1/p}$?

Exercise 11.2 (Secant Method) Let $f : [a,b] \to \mathbb{R}$ satisfy the conditions of Proposition 11.11 (p. 278). We change the formula (11.2) as follows. If x_n, x_{n+1} have already been defined for some $n \geq 0$, then consider the affine function g that coincides with f at x_n, x_{n+1}, and define x_{n+2} by the equation $g(x_{n+2}) = 0$. Given $x_0, x_1 \in [a,b]$, assume that this procedure allows us to define a sequence $(x_n) \subset [a,b]$.

(i) Show that

$$x_{n+2} = x_{n+1} - \frac{f(x_{n+1})(x_{n+1} - x_n)}{f(x_{n+1}) - f(x_n)}, \quad n = 0, 1, \dots.$$

(ii) Denoting by c the solution of the equation $f(c) = 0$, prove that

$$\frac{x_{n+2} - c}{(x_{n+1} - c)(x_n - c)} = \frac{\frac{f(x_{n+1})}{x_{n+1} - c} - \frac{f(x_n)}{x_n - c}}{f(x_{n+1}) - f(x_n)}. \tag{11.4}$$

(iii) Apply Cauchy's mean value theorem (Proposition 4.4 (c), p. 105) and Taylor's formula

$$0 = f(c) = f(\alpha) + f'(\alpha)(c - \alpha) + \frac{f''(\beta)}{2}(c - \alpha)^2$$

to write the right-hand side of (11.4) in the form

$$\frac{f''(\beta)}{2f'(\alpha)}$$

for a suitable β between x_n and x_{n+1}.

(iv) Assuming that

$$A := \frac{\max f''}{2 \min f'} < \infty,$$

infer from the above results that the quantities $d_n := A\,|x_n - c|$ satisfy the inequalities

$$d_{n+2} \leq d_{n+1} d_n, \quad n = 0, 1, \dots.$$

[14] Compare to the example on p. 280.

(v) Setting $d := \max\{d_0, d_1\}$, prove the estimates

$$A\,|x_n - c| \le d^{F_n}, \quad n = 0, 1, \ldots,$$

where the exponents are the Fibonacci numbers

$$F_0 = 1, \quad F_1 = 1, \quad \text{and} \quad F_{n+2} = F_{n+1} + F_n \quad \text{for} \quad n = 0, 1, \ldots.$$

We recall that

$$F_n = \frac{\varphi^{n+1} + (-\varphi)^{-n-1}}{\sqrt{5}} = \frac{\varphi^{n+1}}{\sqrt{5}} + o(1) \quad \text{as} \quad n \to \infty,$$

where $\varphi := (1 + \sqrt{5})/2 \approx 1.618$ denotes the Golden Ratio. Hence the convergence speed is slightly smaller than that of Newton's method.

Exercise 11.3 (Newton–Kantorovich Method) Let X be a Banach space, $x_0 \in X$, $R > 0$, and $F : B_R(x_0) \to X$ a function of class C^1. Furthermore, let $t_0 \in \mathbb{R}$, $0 < r < R$, and $\varphi : [t_0, t_0 + r]$ be a continuous function, of class C^1 in $(t_0, t_0 + r)$. Assume that the following hypotheses are satisfied:

$$\begin{aligned} &\|F(x_0) - x_0\| \le \varphi(t_0) - t_0; \\ &\left\|F'(x)\right\| \le \varphi'(t) \quad \text{if} \quad \|x - x_0\| \le t - t_0; \\ &\varphi \quad \text{has at least one fixed point.} \end{aligned}$$

(i) Prove by induction that, starting from t_0, the formula $t_{n+1} := \varphi(t_n)$ defines a non-decreasing sequence, converging to some $t^* \in [t_0, t_0 + r]$.
(ii) Show that t^* is the smallest fixed point of φ.
(iii) Prove by induction that, starting from x_0, the formula $x_{n+1} := F(x_n)$ defines a sequence satisfying the inequalities

$$\|x_{n+1} - x_n\| \le t_{n+1} - t_n, \quad n = 0, 1, \ldots.$$

(iv) Infer from the above results that (x_n) converges to a fixed point x^* of F, and that

$$\|x^* - x_n\| \le t^* - t_n, \quad n = 0, 1, \ldots.$$

Exercise 11.4 Given a linear operator A in a finite-dimensional normed space X and a vector $b \in X$, we seek the solution of the equation $x = Ax + b$ as the limit of the sequence (x_n) defined by the recursive formula

$$x_{n+1} := Ax_n + b, \quad n = 0, 1, \ldots, \tag{11.5}$$

where $x_0 \in X$ is an arbitrarily given vector.

Prove the following[15]:

(i) If (x_n) converges for some x_0, then $\lim x_n$ is a solution of the equation $x = Ax + b$.
(ii) If $\|A\| < 1$, then the equation $x = Ax + b$ has a unique solution x_b for any given b, and $x_n \to x_b$ for any given x_0.
(iii) If $\|A\| \geq 1$, and A is symmetric, then there exist b and x_0 for which (x_n) does not converge.

Exercise 11.5 We consider the same problem as in the preceding exercise, but in a finite-dimensional *complex* normed space X.[16] We define the *spectral radius* of A by the formula

$$\rho(A) := \max\{|\lambda_1|, \dots, |\lambda_r|\}, \tag{11.6}$$

where $\lambda_1, \dots, \lambda_r$ are the (complex) eigenvalues of A. Prove the following results:

(i) If (x_n) converges for some x_0, then $\lim x_n$ is a solution of the equation $x = Ax + b$.
(ii) We always have $\rho(A) \leq \|A\|$.
(iii) If $\rho(A) < 1$, then the equation $x = Ax + b$ has a unique solution x_b for any given b, and $x_n \to x_b$ for any given x_0.
(iv) If $\rho(A) \geq 1$, then there exist x_0 and b for which (x_n) does not converge.
(v) If X is a Euclidean space, then $\|A\| = \sqrt{\rho(A^*A)}$.
(vi) If X is a Euclidean space and A is selfadjoint, then $\|A\| = \rho(A)$.
(vii) The equality $\|A\| = \rho(A)$ may fail in general.

[15] Properties (ii) and (iii) will be improved in Exercise 12.3, p. 295.

[16] See Exercise 12.3 for the real case.

Chapter 12
Numerical Solution of Differential Equations

Integration is just a special case of the solution of differential equations. Similarly to numerical integration, the approximate solution of differential equations is of great importance in, among other subjects, physics, engineering and chemistry. We give a short introduction to this subject.

12.1 Approximation of Solutions

Differential equations of the form

$$x' = f(t,x), \quad x(\tau) = \xi \tag{12.1}$$

rarely may be solved explicitly. But we may find good approximations of the solutions. Euler proposed the following method. Assuming, for example,[1] that $\tau' > \tau$, let us divide the interval $[\tau, \tau']$ into n equal subintervals for some large integer n, and approximate f by a suitable constant in each of them. Introducing the notations

$$h := \frac{\tau' - \tau}{n} \quad \text{and} \quad t_k := \tau + kh, \quad k = 0, \dots, n, \tag{12.2}$$

this is equivalent to approximating $x(\tau')$ by $x_n := z_n$, where the numbers z_k are defined by the following recurrence formula (see Fig. 12.1):

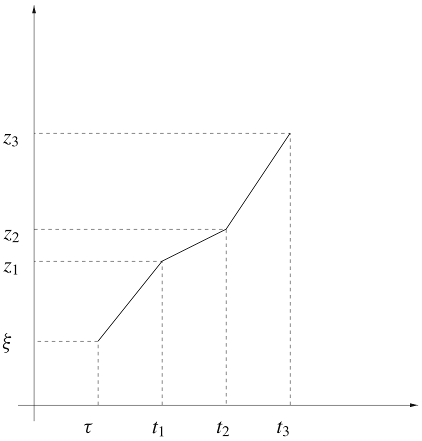
$$z_0 := \xi \quad \text{and} \quad z_{k+1} := z_k + f(t_k, z_k)h, \quad k = 0, \dots, n-1.$$

In order to justify this method we first establish an important inequality.

[1] The results of this chapter remain valid for $\tau' < \tau$ with obvious modifications.

V. Komornik, *Topology, Calculus and Approximation*, Springer
Undergraduate Mathematics Series, DOI 10.1007/978-1-4471-7316-8_12

Fig. 12.1 Euler's method



Definition A continuous function $x : I \to X$, defined on an interval I, is an ε-*approximate solution* of the differential equation $x' = f(t,x)$ if it is *piecewise continuously differentiable* with $\|x'(t) - f(t, x(t))\| \le \varepsilon$ for each $t \in I$, except at a finite number of points.[2]

In particular, every solution is also a 0-approximate solution.

Proposition 12.1 (Peano) *Let $x_i : I \to X$ be an ε_i-approximate solution of $x' = f(t,x)$ for $i = 1,2$. If $\|D_2 f\| \le L$, then*

$$\|x_1(t) - x_2(t)\| \le \|x_1(\tau) - x_2(\tau)\| \, e^{L|t-\tau|} + \frac{\varepsilon_1 + \varepsilon_2}{L}\left(e^{L|t-\tau|} - 1\right)$$

for all $\tau, t \in I$.

Proof Assuming by symmetry that $t \ge \tau$, we have

$$\begin{aligned}\|(x_1 - x_2)(t)\| &\le \|(x_1 - x_2)(\tau)\| + \int_\tau^t \left\|(x_1 - x_2)'(s)\right\| \, ds \\ &\le \|(x_1 - x_2)(\tau)\| + \int_\tau^t \varepsilon_1 + \varepsilon_2 + \|f(s, x_1(s)) - f(s, x_2(s))\| \, ds \\ &\le \|(x_1 - x_2)(\tau)\| + L\int_\tau^t \frac{\varepsilon_1 + \varepsilon_2}{L} + \|(x_1 - x_2)(s)\| \, ds.\end{aligned}$$

Hence the function

$$\varphi(t) := \frac{\varepsilon_1 + \varepsilon_2}{L} + \|(x_1 - x_2)(t)\|$$

[2] We assume that the one-sided derivatives exist and are continuous at the exceptional points.

satisfies the assumptions of Gronwall's lemma 6.8 (p. 155):

$$\varphi(t) \le \left(\frac{\varepsilon_1+\varepsilon_2}{L} + \|(x_1-x_2)(\tau)\|\right) + L\int_\tau^t \varphi(s)\,ds,$$

and we conclude that

$$\varphi(t) \le \left(\frac{\varepsilon_1+\varepsilon_2}{L} + \|(x_1-x_2)(\tau)\|\right) e^{L(t-\tau)}$$

for all $t \in I$, $t \ge \tau$. This is equivalent to the assertion of the proposition. □

Henceforth we assume that $f : D \to X$ satisfies the assumptions of the Cauchy–Lipschitz theorem. We consider the following generalization of Euler's recurrence formulas, where $\Phi : D \times (-r, r) \to X$ is some given function for some $r > 0$: given an integer $n > \frac{\tau'-\tau}{r}$, we define h and $t_0, \dots, t_n$ as in (12.2), and then we set

$$z_0 := \xi, \quad z_{k+1} := z_k + \Phi(t_k, z_k, h)h, \quad k = 0, \dots, n-1, \quad x_n := z_n.$$

Proposition 12.2 *Let the solution of* (12.1) *be defined in some interval* $[\tau, \tau']$, *and set*

$$K := \left\{(t, x(t), 0) \ : \ t \in [\tau, \tau']\right\}.$$

Assume that Φ *is uniformly continuous*[3] *in some neighborhood* U *of* K, *and*

$$\Phi(t, x, 0) = f(t, x)$$

for all $(t, x, 0) \in U$.

Fix $\varepsilon > 0$ *arbitrarily. If* n *is sufficiently large, then* $z_0, \dots, z_n$ *are well defined, and the formula*

$$z(t) := z_k + \Phi(t_k, z_k, h)(t - t_k), \quad t \in [t_k, t_{k+1}], \quad k = 0, \dots, n-1$$

defines an ε*-approximate solution of* (12.1).

Proof Fix a constant M satisfying $M > \|\Phi\|$ on the compact set K, and choose by continuity an open set V such that $K \subset V \subset U$ and $M > \|\Phi\|$ on V. Fix two positive numbers r_1, r_2 such that

$$(t, y, h) \in V \quad \text{whenever} \quad t \in [\tau, \tau'], \quad \|y - x(t)\| \le r_1 \quad \text{and} \quad |h| \le r_2.$$

[3] It is sufficient to assume the mere continuity if $\dim X < \infty$ because we may assume U to be compact and then Heine's theorem implies the uniform continuity.

We may assume that $\varepsilon > 0$ is sufficiently small such that

$$\frac{\varepsilon}{L}\left(e^{L(\tau'-\tau)} - 1\right) < r_1.$$

Choose n sufficiently large such that

$$0 < h < r_2, \quad \frac{\varepsilon}{L}\left(e^{L(\tau'-\tau)} - 1\right) + 2Mh < r_1,$$

and

$$\|\Phi(t_1, y_1, h) - \Phi(t_2, y_2, 0)\| < \varepsilon$$

whenever

$$(t_1, y_1, h), (t_2, y_2, 0) \in V, \quad |t_1 - t_2| \le h \quad \text{and} \quad \|y_1 - y_2\| \le Mh.$$

We prove by induction on k that $z(t)$ is well defined for $t \in [\tau, t_k]$, and z is an ε-approximate solution on $[\tau, t_k]$.

There is nothing to verify for $k = 0$. If our claim is true for some $0 \le k < n$, then applying Peano's inequality we have

$$\begin{aligned}\|z(t) - x(t)\| &\le \|z(t) - z(t_k)\| + \|z(t_k) - x(t_k)\| + \|x(t_k) - x(t)\|\\ &\le Mh + \frac{\varepsilon}{L}\left(e^{L(t_k-\tau)} - 1\right) + Mh\\ &< r_1\end{aligned}$$

for all $t \in [t_k, t_{k+1}]$. Hence $(t, z(t), 0), (t, z(t), h) \in V$ for all $t \in [t_k, t_{k+1}]$, and

$$\left\|z'(t) - f(t, z(t))\right\| = \|\Phi(t_k, z(t_k), h) - \Phi(t, z(t), 0)\| < \varepsilon$$

for all $t \in (t_k, t_{k+1})$. □

We are going to investigate the speed of convergence of $x_n = z_n$ to $x(\tau')$. If $f \in C^p$ for some $p \ge 1$, then we introduce for convenience the notation

$$f^{[0]} := f, \quad \text{and} \quad f^{[i]} := D_1 f^{[i-1]} + f D_2 f^{[i-1]} \quad \text{for} \quad i = 1, \dots, p.$$

It follows by induction on i that the maximal solution of (12.1) satisfies the equalities

$$x^{(i)}(t) = f^{[i-1]}(t, x(t)) \quad \text{for all} \quad t \in I \quad \text{and} \quad i = 1, \dots, p + 1. \tag{12.3}$$

Indeed, for $i = 0$ this is just the differential equation (12.1). If it holds for some $0 \le i \le p$, then it also holds for $i + 1$ because

$$\begin{aligned} x^{(i+1)}(t) &= \frac{d}{dt} x^{(i)}(t) \\ &= \frac{d}{dt} f^{[i-1]}(t, x(t)) \\ &= D_1 f^{[i-1]}(t, x(t)) + D_2 f^{[i-1]}(t, x(t)) x'(t) \\ &= D_1 f^{[i-1]}(t, x(t)) + f(t, x(t)) D_2 f^{[i-1]}(t, x(t)) \\ &= f^{[i]}(t, x(t)). \end{aligned}$$

Proposition 12.3 *Assume that the hypotheses of the preceding proposition are fulfilled. Assume furthermore that $\Phi \in C^p$ for some $p \ge 1$, and*

$$D_3^k \Phi(t, x, 0) = \frac{1}{k+1} f^{[k]}(t, x) \tag{12.4}$$

for all $(t, x) \in D$ and $k = 0, \dots, p-1$.

Then

$$x_n = x(\tau') + O(n^{-p}) \quad \textit{as} \quad n \to \infty. \tag{12.5}$$

Proof Fix two constants L, M satisfying

$$L > \|D_2 \Phi\|, \quad M > \|\Phi\| \quad \text{and} \quad M > \frac{2}{p!} \left\| D_3^p \Phi \right\| \tag{12.6}$$

on K, and fix an open set $K \subset V \subset U$ on which these inequalities are still satisfied.

If n is sufficiently large, then $z_0, \dots, z_n$ are well defined by the preceding proposition, and $(t, y, s) \in U$ whenever $t \in [\tau, \tau']$, $y \in [x(t), z(t)]$ and $s \in [0, h]$. Henceforth, we consider only such large values of n.

It suffices to establish the inequalities

$$\|x(t_{k+1}) - z_{k+1}\| \le (1 + Lh) \|x(t_k) - z_k\| + Mh^{p+1} \tag{12.7}$$

for $k = 0, \dots, n-1$. Indeed, since $x(t_0) - z_0 = \xi - \xi = 0$, they imply the estimate[4]

$$\begin{aligned} \|x(t_n) - x_n\| &= \|x(t_n) - z_n\| \\ &\le \sum_{k=0}^{n-1} (1 + Lh)^k M h^{p+1} \end{aligned}$$

[4] We recall that $nh = \tau' - \tau$ and that $(1 + x/n)^n < e^x$ for all $x > 0$.

$$\begin{aligned}
&\le n(1+Lh)^n Mh^{p+1}\\
&= (\tau'-\tau)\left(1+\frac{L(\tau'-\tau)}{n}\right)^n Mh^p\\
&< (\tau'-\tau)e^{L(\tau'-\tau)}Mh^p\\
&= M(\tau'-\tau)^{p+1}e^{L(\tau'-\tau)}n^{-p}.
\end{aligned}$$

For the proof of (12.7) first we apply Taylor's formula. We have

$$\left\| x(t_{k+1}) - x(t_k) - \sum_{i=0}^{p-1} \frac{x^{(i+1)}(t_k)}{(i+1)!}h^{i+1} \right\| \le \max_{t\in[t_k,t_{k+1}]} \left\| x^{(p+1)}(t) \right\| \frac{h^{p+1}}{(p+1)!}$$

and

$$\left\| \Phi(t_k,x(t_k),h) - \sum_{i=0}^{p-1} \frac{D_3^i\Phi(t_k,x(t_k),0)}{i!}h^i \right\| \le \max_{s\in[0,h]} \left\| D_3^p\Phi(t_k,x(t_k),s) \right\| \frac{h^p}{p!}.$$

Using (12.3), (12.4) and (12.6) we get

$$\|x(t_{k+1}) - x(t_k) - \Phi(t_k,x(t_k),h)h\| \le Mh^{p+1}. \tag{12.8}$$

Next we apply the Lagrange inequality to get

$$\begin{aligned}
\|z_{k+1} - z_k - \Phi(t_k,x(t_k),h)h\| &\le h\,\|\Phi(t_k,z_k,h) - \Phi(t_k,x(t_k),h)\|\\
&\le Lh\,\|z_k - x(t_k)\|\,.
\end{aligned}$$

Combining with (12.8) and using the triangle inequality, the estimate (12.7) follows. □

Definition A method is said to be of order p if it satisfies (12.5), but does not satisfy the similar estimate with $p+1$ instead of p.

Examples

- For *Euler's method* $\Phi(t,x,h) = f(t,x)$ the condition (12.4) of Proposition 12.3 is satisfied for $k=0$ by definition, so that it has at least order one.[5]
- The *modified Euler method* is defined by the formula

$$\Phi(t,x,h) = f\Big(t+\frac{h}{2}, x+\frac{h}{2}f(t,x)\Big).$$

[5] In fact, it is exactly one: see Exercise 12.1 below, p. 295.

Its order is at least two[6] because the condition (12.4) of Proposition 12.3 is satisfied for $k = 0$ and $k = 1$: $\Phi(t,x,0) = f(t,x)$ and

$$D_3\Phi(t,x,0) = \frac{1}{2}D_1 f(t,x) + D_2 f(t,x)\frac{f(t,x)}{2} = f^{[1]}(t,x).$$

12.2 The Runge–Kutta Method

According to an idea of Runge, developed by Heun and Kutta, we may construct methods of arbitrarily high order in the following way. We replace the integral in the equation

$$x(t+h) - x(t) = \int_t^{t+h} f(s,x(s))\,ds$$

with a suitable numerical integration formula

$$\int_t^{t+h} f(s,x(s))\,ds \approx h\sum_{i=1}^{m} b_i f(t+a_ih, x(t+a_ih)),$$

and we replace the integrals of the identities

$$x(t+a_ih) = x(t) + \int_t^{t+a_ih} f(s,x(s))\,ds$$

with suitable numerical integration formulas as well:

$$\int_t^{t+a_ih} f(s,x(s))\,ds \approx h\sum_{j=1}^{i-1} c_{ij} f(t+a_jh, x(t+a_jh)).$$

This leads to the investigation of functions of the form

$$\Phi(t,x,h) := \sum_{i=1}^{m} b_ik_i,$$

where the quantities $k_1,\dots,k_m$ are defined recursively by formulas of the type

$$k_i := f\Big(t+a_ih, x + h\sum_{j=1}^{i-1} c_{ij}k_j\Big), \quad i = 1,\dots,m.$$

[6] In fact, it is exactly two: see Exercise 12.2 below, p. 295.

Suitable choices of m, a_i, b_i, c_{ij} yield powerful numerical methods.

Examples

- For $m = 1, a_1 = 0$ and $b_1 = 1$ we recover *Euler's method*:

$$\Phi(t,x,h) = f(t,x).$$

- For $m = 2, b_1 = 0, b_2 = 1, a_1 = 0$ and $a_2 = c_{21} = 1/2$ we recover the *modified Euler's method*:

$$\Phi(t,x,h) = f\Big(t + \frac{h}{2}, x + \frac{h}{2} f(t,x)\Big).$$

- The *classical Runge–Kutta method* corresponds to the choice

$$\Phi(t,x,h) := \frac{k_1 + 2k_2 + 2k_3 + k_4}{6}$$

  with

$$\begin{aligned} k_1 &= f(t,x), \\ k_2 &= f(t + \frac{h}{2}, x + \frac{h}{2} k_1), \\ k_3 &= f(t + \frac{h}{2}, x + \frac{h}{2} k_2), \\ k_4 &= f(t + h, x + hk_3). \end{aligned}$$

  The last method is of order four.

12.3 The Dirichlet Problem

Important physical questions[7] require the solution of the *Dirichlet problem*

$$-\Delta u = 0 \quad \text{in} \quad \Omega, \quad u = f \quad \text{on} \quad \Gamma, \tag{12.9}$$

[7] See, e.g., Brezis–Browder [65], Evans [158], Feynman [169], Petrovsky [388], Sobolev [455], and Tikhonov–Samarskii [490].

where Ω is a given bounded domain in $\mathbb{R}^n$, $f : \Gamma \to \mathbb{R}$ is a given continuous function on its boundary Γ, and the *Laplacian* operator denotes the trace of the Hessian:

$$\Delta u := \sum_{i=1}^{n} D_i^2 u.$$

The solution of this problem being out of our scope here, we instead solve an approximate problem. If u solves (12.9), then applying Taylor's formula and using the canonical basis $e_1, \dots, e_n$ of $\mathbb{R}^n$, we obtain for each $x \in \Omega$ the relations

$$u(x \pm he_i) - u(x) = \pm hD_i u(x) + \frac{1}{2} D_i^2 u(x)h^2 + o(h^2), \quad h \to 0$$

for $i = 1, \dots, n$. Summing them we obtain the relation

$$\frac{1}{2n} \sum_{i=1}^{n} (u(x + he_i) + u(x - he_i)) = u(x) + \frac{h^2}{2n} \Delta u(x) + o(h^2), \quad h \to 0. \tag{12.10}$$

Fix a small positive real number h, and consider the set L_h of lattice points $x \in \mathbb{R}^n$ whose components are integer multiples of h. Each lattice point has $2n$ neighbors $y \in L_h$ satisfying

$$\sum_{i=1}^{n} |x_i - y_i| = 1.$$

Let us denote by $\Omega_h := \Omega \cap L_h$ the set of lattice points lying in Ω, and by Γ_h the set of lattice points outside Ω that have at least one neighbor in Ω. Pick for each $x \in \Gamma_h$ a point $y \in \Gamma \cap B_h(x)$ and set $f_h(x) := f(y)$.

If h is sufficiently small, then in view of (12.10) we may approximate the solution of (12.9) by the solution of the system of linear equations

$$A_h u = u \quad \text{in} \quad \Omega_h \quad \text{and} \quad u = f_h \quad \text{on} \quad \Gamma_h, \tag{12.11}$$

with

$$A_h u(x) := \frac{1}{2n} \sum_{i=1}^{2n} u(y_i) \quad \text{if} \quad x \in \Omega_h, \tag{12.12}$$

where $y_1, \dots, y_{2n}$ denote the neighbors of x.

For the solution of this problem we generalize the Banach–Cacciopoli fixed point theorem 1.10 (p. 16)[8]:

Proposition 12.4 *Let X be a complete metric space and $f : X \to X$. If there exists a positive integer m such that the m-fold composition $f^m := f \circ \cdots \circ f$ of f is a contraction, then f has a unique fixed point x.*

Moreover, $f^n(x_0) \to x$ for each fixed $x_0 \in X$.

Proof Since f^m has a unique fixed point by Theorem 1.10, for the first part it suffices to show that f and f^m have the same fixed points. Each fixed point of f is obviously also a fixed point of f^m. Conversely, if x is a fixed point of f^m, then $f^m(f(x)) = f(f^m(x)) = f(x)$, so that $f(x)$ is also a fixed point of f^m. Since f^m is a contraction, we conclude by uniqueness that $f(x) = x$.

For the second part it suffices to observe that each of the m subsequences $\big((f^m)^n f^r(x_0)\big)$, $r = 0, \ldots, m-1$, converges to this fixed point by Theorem 1.10 as $n \to \infty$. □

Proposition 12.5 *The problem (12.11) has a unique solution.*

Proof Let us denote by X the set of functions $u : \Omega_h \cup \Gamma_h \to \mathbb{R}$ satisfying $u = f_h$ on Γ_h. This is a complete metric space for the distance

$$d(u, v) := \|u - v\|_\infty = \max\left\{|u(x) - v(x)| \ : \ x \in \Omega_h\right\}.$$

Let us extend the definition of A_h for $\Omega_h \cup \Gamma_h$ by setting

$$A_h u(x) := f_h(x) \quad \text{if} \quad x \in \Gamma_h,$$

then $A_h : X \to X$.

If $u, v \in X$, then $A_h u(x) - A_h v(x) = 0$ for all $x \in \Gamma_h$, while for $x \in \Omega_h$ we have

$$|A_h u(x) - A_h v(x)| \le \frac{1}{2n} \sum_{i=1}^{2n} |u(y_i) - v(y_i)| \le d(u, v),$$

where $y_1, \ldots, y_{2n}$ denote the neighbors of x. Hence $d(A_h u, A_h v) \le d(u, v)$, and iterating this we obtain the inequalities

$$d(A_h^k u, A_h^k v) \le d(u, v) \tag{12.13}$$

for all $u, v \in X$ and $k = 1, 2, \ldots$.

We need a sharper estimate. Let us denote by $m(x)$ the distance of $x \in \Omega_h \cup \Gamma_h$ from the boundary Γ_h. More precisely, let $m(x) := 0$ if $x \in \Gamma_h$. Then, by induction,

[8] This proposition is closely related to the definition of the spectral radius in Exercise 12.3, p. 295.

let $m(x) := k + 1$ if $m(x)$ has not yet been defined, but x has a neighbor y satisfying $m(y) = k$.

By the definition of Ω_h and Γ_h, $m(x)$ is well defined for all $x \in \Omega_h \cup \Gamma_h$. Furthermore, since $\Omega_h \cup \Gamma_h$ is a finite set, there exists a positive integer m such that $m(x) \le m$ for all $x \in \Omega_h \cup \Gamma_h$.

We prove by induction that if $m(x) = k$ for some $x \in \Omega_h \cup \Gamma_h$, then

$$\left|A_h^k u(x) - A_h^k v(x)\right| \le (1 - (2n)^{-k}) d(u, v) \tag{12.14}$$

for all $u, v \in X$.

This is true for $k = 0$ because then $u(x) = v(x) = f_h(x)$. Let $k \ge 1$ and assume that the estimates hold for all $y \in \Omega_h \cup \Gamma_h$ with $m(y) = k - 1$.

If $m(x) = k$, then at least one neighbor y of x satisfies $m(y) = k-1$. Using (12.13) and applying the induction hypothesis, (12.14) follows:

$$\begin{aligned}\left|A_h^k u(x) - A_h^k v(x)\right| &\le \frac{1}{2n}(1 - (2n)^{1-k}) d(u, v) + \frac{2n-1}{2n} d(u, v)\\ &= (1 - (2n)^{-k}) d(u, v).\end{aligned}$$

It follows from (12.14) that

$$d(A_h^m u, A_h^m v) \le \left(1 - (2n)^{-m}\right) d(u, v)$$

for all $u, v \in X$. We conclude by applying Proposition 12.4 with $f = A_h$. □

Remark Since our problem was solved by applying the fixed point theorem of contractions, the proof also provides an algorithm for the numerical approximation of the solution.

12.4 The Monte Carlo Method

For the approximate solution of the Dirichlet problem another very efficient approach is provided by the so-called *Monte Carlo method* of von Neumann and Ulam. Let us consider the following *random walk* problem of Pólya. A person is walking on the lattice points of $\Omega_h \cup \Gamma_h$ as follows. If he is at some point x, then in the next step he moves to one of the $2n$ neighbors of x with equal probability $= \frac{1}{2n}$. When he arrives at a point y of the boundary, then he stops and receives the reward $f_h(y)$. Let us denote by $u(x)$ the expected value of his reward when he starts from x.

Proposition 12.6 *The function $u : \Omega_h \cup \Gamma_h \to \mathbb{R}$ solves* (12.11)*.*

Proof If $x \in \Gamma_h$, then the person does not move, so that $u(x) = f_h(x)$ by definition.

If $x \in \Omega_h$, then he moves to one of the $2n$ neighbors $y_1, \dots, y_{2n} \in \Omega_h \cup \Gamma_h$ with equal probability. Therefore

$$u(x) = \sum_{j=1}^{2n} \frac{1}{2n} u(y_j)$$

by elementary properties of the expected value, i.e., $u(x) = A_h u(x)$, as required. □

Remark This method provides a probabilistic method of solving (12.11) numerically, by repeating a large number of times the random walk, starting from a given point x, and taking the average of the corresponding rewards.

This is particularly useful if we only need to approximate the solution at some particular points x instead of all $x \in \Omega_h$: it requires much less computation than the method of the preceding section.

12.5 The Heat Equation

Given two positive numbers h and τ, we consider the following *diffusion* problem. If a particle is found in some position $x \in \mathbb{R}^n$ at time t, then, subject to independent random impacts, it will be at time $t + \tau$, with equal probability, in one of the neighboring positions $x \pm he_i$, $i = 1, \dots, n$.

We do not know the particle's position at time $t = 0$, but we know that it belongs to the lattice $L_h := h\mathbb{Z}^n$,[9] and we know the probability $g(x)$ that it is in position x. We are looking for the probability $u(t, x)$ that the particle will be in position x at time $t = \tau, 2\tau, \dots$.

We have

$$u(0, x) = g(x) \quad \text{for all} \quad x \in \Omega_h$$

by definition, and

$$u(t + \tau, x) = \frac{1}{2n} \sum_{i=1}^{2n} u(t, y_i) \quad \text{for all} \quad t = 0, \tau, 2\tau, \dots \quad \text{and} \quad x \in \Omega_h$$

by elementary probability considerations, where $y_1, \dots, y_{2n}$ denote the neighbors of x. In principle, they allow us to compute the probability distributions $u(k\tau, x)$, $x \in \Omega_h$, by induction on $k = 0, 1, \dots,$[10] but the results are not practical.

[9] See Sect. 12.3.

[10] Using the linear operator A_h of Sect. 12.3, now defined on L_h, we have $u(k\tau, x) = (A_h^k g)(x)$.

We are interested in the limit problem as h and τ tend to zero. Let us rewrite the last equation in the form

$$u(t+\tau,x)-u(t,x)=\frac{1}{2n}\sum_{i=1}^{2n}\bigl(u(t,y_i)-u(t,x)\bigr).$$

Applying Taylor's formula and using the relation (12.10) from Sect. 12.3, we obtain for $h,\tau\to 0$ the relation

$$\tau\frac{\partial u}{\partial t}(t,x)+o(\tau)=\frac{h^2}{2n}\Delta u(t,x)+o\bigl(h^2\bigr),$$

where we use the traditional notation $\frac{\partial u}{\partial t}$ for the first partial derivative $D_1 u$ of u.

Assuming that $\frac{h^2}{2n\tau}$ converges to some positive number a^2,[11] letting $h,\tau\to 0$ we obtain Fourier's heat equation

$$\frac{\partial u}{\partial t}=a^2\,\Delta u \quad \text{for} \quad t>0 \quad \text{and} \quad x\in\mathbb{R}^n.$$

12.6 Exercises

Exercise 12.1 Prove that the exact order of Euler's method is one.

Exercise 12.2 Prove that the exact order of the modified Euler method is two.

Exercise 12.3 Given a continuous linear operator A in a real or complex Banach space X, we define its *spectral radius* by the formula[12]

$$\rho(A):=\inf_{n\ge 1}\sqrt[n]{\|A^n\|}.$$

Prove the following results:

(i) In the finite-dimensional complex case this is equivalent to the definition (11.6) in Exercise 11.5, p. 282.

(ii) The following limit relation always holds:

$$\sqrt[n]{\|A^n\|}\to\rho(A) \quad \text{as} \quad n\to\infty.$$

(iii) If $\rho(A)<1$, then the equation $x=Ax+b$ has a unique solution $x_b\in X$ for each fixed $b\in X$. Furthermore, for any given $x_0\in X$ the sequence (x_n) defined

[11] This assumption is justified by physical motivations; a^2 is called the *diffusion coefficient*.

[12] Note that $\rho(A)\le\|A\|$ by definition.

by the recursive formula

$$x_{n+1} := Ax_n + b, \quad n = 0, 1, \dots$$

converges to x_b.

(iv) If $\rho(A) \geq 1$ and $\dim X < \infty$, then there exist x_0 and b for which (x_n) does not converge.

Exercise 12.4 Given a linear operator A in a finite-dimensional normed space X, prove that

$$\rho(A) = \inf \|A\|'$$

where $\|\cdot\|'$ runs over all norms on X, and $\|A\|'$ is the corresponding operator norm, i.e.,

$$\|A\|' := \sup_{\|x\|' \leq 1} \|Ax\|' .$$

Exercise 12.5 (Volterra operator) Let $a : T \to \mathbb{R}$ be a continuous function on the closed triangle

$$T := \left\{(t, s) \in \mathbb{R}^2 \ : \ 0 \leq s \leq t \leq 1\right\} .$$

Given a function $x \in C([0, 1])$, set

$$(Ax)(t) := \int_0^t a(t, s)x(s)\, ds, \quad t \in [0, 1].$$

Prove the following results:

(i) A is a continuous linear operator in $C([0, 1])$ with the supremum norm.
(ii) $\rho(A) = 0$.

Exercise 12.6 (Maximum Principle for Harmonic Functions) Let Ω be a bounded open domain with boundary Γ in $\mathbb{R}^n$, and $f : \overline{\Omega} \to \mathbb{R}$ a continuous function. Prove that if u is *harmonic* in Ω, i.e., if $\Delta u = 0$ in Ω, then

$$\min_{\Gamma} u \leq u(x) \leq \max_{\Gamma} u$$

for all $x \in \Omega$.

Exercise 12.7 (Maximum Principle for the Heat Equation) Let Ω and Γ be as in Exercise 12.6, and $T > 0$. Consider the cylinder $Q := (0, T) \times \Omega$, and denote by

Σ the union of its lower base and its lateral boundary:

$$\Sigma := (\{0\} \times \Omega) \cup [0, T] \times \Gamma.$$

Prove that if a continuous function $f : \overline{Q} \to \mathbb{R}$ satisfies in $(0, T] \times \Omega$ the heat equation

$$\frac{\partial u}{\partial t} - \Delta u = 0,$$

then

$$\min_{\Sigma} u \le u(t, x) \le \max_{\Sigma} u$$

for all $(t, x) \in Q$.[13]

[13] There is a natural physical interpretation of this property: if the boundary of a body is maintained at a constant temperature, then its temperature at any point always remains between the minimal and maximal temperatures of the body at the initial moment.

Hints and Solutions to Some Exercises

Exercise 1.1. Applying the second inequality with $x = y$ and using the first one we get the nonnegativity of d. Applying the second inequality with $z = y$ and using the first one we get $d(x, y) \le d(y, x) + d(y, y) = d(y, x)$. Exchanging the role of x and y this yields $d(x, y) = d(y, x)$.

Exercise 1.2.

(iv) If

$$d(x, y) \ge d(y, z) \ge d(z, x),$$

then

$$d(x, y) \le \max\{d(y, z), d(z, x)\} = d(y, z)$$

and hence $d(x, y) = d(y, z)$.

(v) If $y \in B_r(x)$ and $z \in B_r(y)$, then

$$d(x, z) \le \max\{d(x, y), d(z, y)\} < r,$$

so that $B_r(y) \subset B_r(x)$.

Since $y \in B_r(x) \iff x \in B_r(y)$, the converse inclusion also holds.

(vi) If $z \in B_r(x) \cap B_s(y)$ and say $r \le s$, then

$$B_r(x) = B_r(z) \subset B_s(z) = B_s(y)$$

by (v).

(vii) It suffices to show that the open sets

$$\{y \in X \ : \ d(x, y) > r\} \quad \text{and} \quad \{y \in X \ : \ d(x, y) < r\}$$

are also closed for any fixed $x \in X$ and $r > 0$.

V. Komornik, *Topology, Calculus and Approximation*, Springer
Undergraduate Mathematics Series, DOI 10.1007/978-1-4471-7316-8

First we prove the following: *if $y_n \to y \ne x$, then $d(x, y_n) = d(x, y)$ for all sufficiently large n.* Indeed, we have $d(y, y_n) \to 0$, and $d(x, y_n) \to d(x, y) > 0$ by the continuity of the metric, so that $d(x, y_n) > d(y, y_n)$ for all sufficiently large n, and then $d(x, y_n) = d(x, y)$ by (iv).

Now let $x \in X$ and $y_n \to y$.

If $d(x, y_n) > r$ for all n, then $y_n \not\to x$, so that $y \ne x$; applying the above claim we conclude that $d(x, y) > r$. This shows that the closed balls $\{y \in X \ : \ d(x, y) \le r\}$ are also open, because their complements are closed.

If $y \ne x$, and $d(x, y_n) < r$ for all n, then applying the above claim we conclude that $d(x, y) < r$. The result obviously holds if $y = x$, too. This shows that the open balls $B_r(x)$ are also closed.

Exercise 1.3.

(i) Consider a rotation of a circle.

(ii) Consider the function $f(x) := \sqrt{1 + x^2}$ in $[0, \infty)$ or $f(x) := x + \frac{1}{x}$ in $[1, \infty)$. We have $f(x) > x$ for all x, hence there is no fixed point. Furthermore, $0 < f' < 1$ in $(0, \infty)$ and $(1, \infty)$, respectively, so that $|f(x) - f(y)| < |x - y|$ whenever $x \ne y$ by the Lagrange mean value theorem.

(iii) Consider a minimal value of the function $x \mapsto \operatorname{dist}(x, f(x))$, and compare $\operatorname{dist}(x, f(x))$ with $\operatorname{dist}(f(x), f(f(x)))$.

(v) Consider the sign of $x - f(x)$.

Exercise 1.4.

(i) If a is not an accumulation point, then it has a neighborhood containing no accumulation points.

(ii) If F is the set of cluster points of a sequence (x_n) in X, then F belongs to the closure of the countable set $\{x_1, x_2, \dots\}$ so that F is separable. (Since X is a metric space, we can also find a countable dense subset of F.) If $a \notin F$ then a has an open neighborhood U containing at most a finite number of elements of the sequence (x_n). Since U is a neighborhood of each of its elements, it follows that no element of U belongs to F. This proves that $X \setminus F$ is open, and hence F is closed.

(iii) Choose a dense sequence (y_m) in F and then let (x_n) contain each element of (y_m) infinitely many times, and no other element, e.g.,

$$(x_n) = y_1, y_1, y_2, y_1, y_2, y_3, y_1, y_2, y_3, y_4, \dots .$$

The set F_0 of cluster points of (x_n) satisfies the inclusions

$$\{y_1, y_2, \dots\} \subset F_0 \subset \overline{\{y_1, y_2, \dots\}},$$

whence

$$\overline{F_0} = \overline{\{y_1, y_2, \dots\}} = F.$$

Since F_0 is closed by (ii), we conclude that $F_0 = F$.

(iv) If X is compact and A is an infinite set in X, then there exists a sequence (x_n) of distinct elements of A. By compactness this sequence has a cluster point, and this is an accumulation point of A.

If X is non-complete, then there exists a non-convergent Cauchy sequence (x_n) in X. Then (x_n) has no cluster point, so that (x_n) has no constant subsequences. It follows that $A := \{x_1, x_2, \ldots\}$ is an infinite set without accumulation points.

If X is not totally bounded, then there exists an $r > 0$ and a sequence (x_n) in X satisfying $d(x_k, x_n) \ge r$ for all $k \ne n$. Then $A := \{x_1, x_2, \ldots\}$ is an infinite set without accumulation points.

(v) We already know that the cluster points form a closed set. It remains to show that if $a < b < c$ and a, c are cluster points, then b is also a cluster point.

Since (x_n) has infinitely many elements satisfying $x_n < b$ (because a is a cluster point) and infinitely many elements satisfying $x_n > b$ (because c is a cluster point), there exists a subsequence (x_{n_k}) of (x_n) such that $x_{n_k} < b \le x_{n_k+1}$ for all k. Then

$$0 < b - x_{n_k} \le x_{n_k+1} - x_{n_k}.$$

Since the right-hand side tends to zero by assumption, we conclude that $x_{n_k} \to b$.

Exercise 1.5. Observe that $a_n \to 0 \iff f(a_n) \to 0$.

Exercise 1.6. If there is no such δ, then there exist two sequences of points $(x_n), (y_n)$ such that $d(x_n, y_n) \to 0$, and $x_n \in U_i \Longrightarrow y_n \notin U_i$ for all n, i. Since K is compact, there is a subsequence satisfying $x_{n_k} \to x \in K$, and then we have also $y_{n_k} \to x$.

There exists an i such that $x \in U_i$, and then $x_{n_k}, y_{n_k} \in U_i$ if k is sufficiently large. This contradicts the choice of the sequences (x_n) and (y_n).

Exercise 1.7. The functions

$$f_i(t) := \frac{t}{1 + 2^i t}$$

satisfy the conditions of Exercise 1.5, and $0 \le f_i \le 2^{-i}$. Hence

$$d(x, y) = \sum_{i=1}^{\infty} f_i(d_i(x_i, y_i))$$

is a metric satisfying $0 \le d \le 1$.

If $d(x^n, x) \to 0$, then $f_i(d_i(x_i^n, y_i)) \to 0$ for each i because

$$0 \le f_i(d_i(x_i^n, y_i)) \le d(x^n, x),$$

and hence $d_i(x_i^n, y_i) \to 0$ because $a_n \to 0 \iff f_i(a_n) \to 0$.

Conversely, if $d_i(x_i^n, y_i) \to 0$ for each i, then also $f_i(d_i(x_i^n, y_i)) \to 0$ for each i. For any fixed $\varepsilon > 0$, choosing $m = m(\varepsilon)$ such that $\sum_{i>m} 2^{-i} < \varepsilon$, we have

$$0 \le d(x^n, x) \le \varepsilon + \sum_{i=1}^{m} f_i(d_i(x_i^n, y_i)) \to \varepsilon.$$

Letting $\varepsilon \to 0$ we conclude that $d(x^n, x) \to 0$.

Exercise 1.8.

(iii) The function

$$2\Big(\frac{t_1}{3} + \frac{t_2}{3^2} + \cdots + \frac{t_n}{3^n} + \cdots\Big) \mapsto \frac{t_1}{2} + \frac{t_2}{2^2} + \cdots + \frac{t_n}{2^n} + \cdots$$

maps C onto $[0, 1]$.

Exercise 1.10.

(ii) Since $\mathbb{R}^n$ is the union of countably many compacts sets (for example, of closed balls), there exists a compact set K such that $A \cap K$ is uncountable. We claim that $A \cap K$ even has a condensation point. For otherwise each point $x \in K$ has an open neighborhood U_x containing at most countably many points of $A \cap K$. A finite number of these neighborhoods cover $A \cap K$, implying that $A \cap K$ itself is countable, contradicting the choice of K.

(iii) The complement of P is open.

Exercise 1.11.

(i) Set $F = X \setminus G$. For brevity we denote by $d(x, F)$ the distance of x to F. Since F is closed, $f(x) := 1/d(x, F)$ is well defined on G, and hence

$$D(x, y) := d(x, y) + |f(x) - f(y)|$$

is also well defined on G. The axioms of the metric are straightforward.

Since f is continuous, we have

$$d(y_n, y) \to 0 \iff D(y_n, y) \to 0$$

in G. Hence the two metrics define the same open, closed and compact sets in G.

If (y_n) is a Cauchy sequence in (G, D), then it is also a Cauchy sequence in (G, d), and the sequence $(f(y_n))$ is bounded. Therefore its limit in (X, d) belongs to G, so that $y_n \to y$ in (G, D). This proves the completeness of (G, D).

(ii) Set $F_i := X\setminus G_i$ and $f_i(x) := 1/d(x, F_i)$ on Y for each $i = 1, 2, \ldots$.

Generalizing the results of the first step we obtain that the formula

$$D(x,y) := d(x,y) + \sum_{i=1}^{\infty} \frac{\min\{1, |f_i(x) - f_i(y)|\}}{2^i}$$

defines a complete metric on Y which defines the same topology as the restriction of d onto Y.[1]

(iii) If F is a closed set, then the sets $G_n := \{x \in X \ : \ d(x,F) < 1/n\}$ are open for $n = 1, 2, \ldots$, and $F = \cap G_n$.

(iv) Now let D be a complete metric on some subset Y of X, equivalent to the restriction of d to Y. Then for each positive integer n and for each $y \in Y$ there exists $0 < \varepsilon(n,y) < 1/n$ such that

$$x \in Y \quad \text{and} \quad d(x,y) < \varepsilon(n,y) \Longrightarrow D(x,y) < \frac{1}{n}.$$

Then the sets

$$G_n := \cup_{y\in Y} \{x \in X \ : \ d(x,y) < \varepsilon(n,y)\}, \quad n = 1, 2, \ldots$$

are open in X, and $Y \subset \cap_{n=1}^{\infty} G_n$.

To prove the converse inclusion we fix $x \in \cap_{n=1}^{\infty} G_n$ arbitrarily, and we choose for each n a point $y_n \in Y$ satisfying $d(x, y_n) < \varepsilon(n, y_n)$. Since $\varepsilon(n,y) < 1/n$, we have $d(x, y_n) \to 0$.

Furthermore,

$$D(y_m, y_n) \le D(x, y_m) + D(x, y_n) < \frac{1}{m} + \frac{1}{n} \to 0 \quad \text{as} \quad m, n \to \infty,$$

so that (y_n) is a Cauchy sequence in (Y, D). By the completeness of D we have $D(y, y_n) \to 0$ for some $y \in Y$, and then also $d(y, y_n) \to 0$ by the equivalence of the metrics on Y.

Since $d(x, y_n) \to 0$ and $d(y, y_n) \to 0$, by uniqueness of the limit we conclude that $x = y \in Y$.

Exercise 1.13. If $f : \mathbb{R} \to \mathbb{R}$ is continuous and $f(x) \neq 0$ for some x, then $f(y) \neq 0$ for all y in a neighborhood of x, so that the complement of $f^{-1}(0)$ is open.

Conversely, if $F \subset \mathbb{R}$ is a non-empty closed set, then $F = f^{-1}(0)$ for the continuous function $f(x) := \operatorname{dist}(x, F)$. For $F = \varnothing$ we have $F = f^{-1}(0)$ for the constant function $f = 1$.

[1] We adapt the solution of Exercise 1.7.

Exercise 1.14.

(i) We prove the upper semi-continuity: if $\omega_f(x) < A$, then $\omega_f(y) < A$ for all y in a neighborhood of x. Choose $r > 0$ such that

$$\sup_{B_r(x)} f - \inf_{B_r(x)} f < A.$$

If $y \in B_{r/2}(x)$, then

$$\sup_{B_{r/2}(y)} f - \inf_{B_{r/2}(y)} f \le \sup_{B_r(x)} f - \inf_{B_r(x)} f < A.$$

(iii) Assume on the contrary that

$$\mathbb{Q} = \bigcap_{n=1}^{\infty} \left\{ y \in \mathbb{R} \ : \ \omega_f(y) < \frac{1}{n} \right\}$$

for some function $f : \mathbb{R} \to \mathbb{R}$. Then the closed sets $\left\{y \in \mathbb{R} \ : \ \omega_f(y) \ge \frac{1}{n}\right\}$ and the one-point closed sets $\{a\}$, $a \in \mathbb{Q}$, form a countable cover of $\mathbb{R}$. By Baire's Theorem 1.13 at least one of them contains a non-empty open interval I. This cannot be a one-point set, so that

$$I \subset \left\{ y \in \mathbb{R} \ : \ \omega_f(y) \ge \frac{1}{k} \right\}$$

for some k. Take a rational number $a \in I$, then

$$a \notin \left\{ y \in \mathbb{R} \ : \ \omega_f(y) < \frac{1}{k} \right\},$$

contradicting our hypothesis.

(v) Consider Thomae's function: let $f(0) = 1$, $f(x) = 0$ if x is irrational, and $f(p/q) = 1/q$ otherwise.

(vi) See Oxtoby [377], Theorem 7.1.

Exercise 1.15.

(ii) Assume by scaling that $M = 3$, and apply (i) with

$$A := \{x \in F \ : \ g(x) \le -1\} \quad \text{and} \quad B := \{x \in F \ : \ g(x) \ge 1\}.$$

(iii) Iterating (ii) we obtain a sequence of continuous functions $f_n : X \to \mathbb{R}$ satisfying

$$|f_n| \le \left(\frac{2}{3}\right)^n \frac{M}{2} \quad \text{on} \quad X, \quad \text{and} \quad \left|g - \sum_{k=1}^{n} f_k\right| \le \left(\frac{2}{3}\right)^n M \quad \text{on} \quad F$$

for $n = 1, 2, \dots$. Then $f := \sum_{k=1}^{\infty} f_k$ has the required properties.

(iv) Apply (iii) to $\arctan g(x)$.

Exercise 2.3.

(i) The maps $f(x) := (1-x)/x$ and $g(x) := \ln x$ show that $(0,1)$, $(0,\infty)$ and $(-\infty,\infty)$ are homeomorphic. The non-constant affine maps $h(x) := ax+b$ show that all other open intervals are homeomorphic to them.

(ii) It suffices to use the affine maps $h(x) := ax+b$ in (i).

(iii) Use the same maps as in (i).

(iv) The intervals $(0,1)$ and $[0,1)$ are not homeomorphic because every continuous and injective function $f : (0,1) \to [0,1)$ is strictly monotone, and hence it cannot take the value 0. The other cases are similar.

Alternatively, the removal of any point of $(0,1)$ yields a non-connected subspace topology, while removing 0 from $[0,1)$ or $[0,1]$ the resulting subspace topology is still connected. Furthermore, removing any two points from $[0,1)$ yields a non-connected subspace topology, while removing the two endpoints from $[0,1]$ we still obtain a connected subspace topology.

A third argument is that $[0,1]$ is compact, while the other two intervals are not, hence $[0,1]$ is not homeomorphic to any of the other two.

(v) If we remove an arbitrary point from a circle, then the remaining set is connected in the subspace topology. On the other hand, we may remove a point from any non-degenerate interval so that the remaining set is not connected in the subspace topology.

Exercise 2.4.

(i) $\Longrightarrow$ (ii) Every open set containing a is a neighborhood of a.

(ii) $\Longrightarrow$ (iii) Fix a point $x_V \in V \cap D$ for each neighborhood V of a. Then the net (x_V) converges to a for the usual order relation $V > W \Longleftrightarrow V \subset W$.

(iii) $\Longrightarrow$ (i) If a net (x_i) in D converges to a, and V is a neighborhood of a, then $x_i \in V$ for all sufficiently large i, and hence $D \cap V \neq \varnothing$.

For the definition of accumulation points, compare Lemma 1.3 (p. 9) with the above equivalences.

In a compact topological space every infinite set has an accumulation point: the solution of Exercise 1.4 (iv) remains valid. On the other hand, in the non-compact topological space of the second example on p. 58 every infinite set has an accumulation point.

Exercise 2.5.

(i) $\Longrightarrow$ (ii) If $\cap F_i = \varnothing$ for a family of closed sets F_i, then their complements form an open cover of K. By compactness there exists a finite subcover of K, but then the corresponding sets F_i have an empty intersection.

(ii) $\Longrightarrow$ (i) If a family of open sets U_i covers K, then the sets $K \setminus U_i$ have an empty intersection. By assumption (ii) there exists a finite number of these complements with an empty intersection, but then the corresponding finitely many open sets U_i already cover K.

Exercise 2.6.

(i) If i is continuous for some topology on Y, then $i^{-1}(U) = U \cap Y$ must be open in Y for each open set U of X by Proposition 2.10. Hence the topology of Y contains the open sets of the subspace topology.

(ii) If all projections $\pi_i : X \to X_i$ are continuous for some topology on X, then $\pi_i^{-1}(U_i)$ must be open in X for each index i and for each open set U_i of X_i by Proposition 2.10. Hence the topology of X contains all elements of the subbase defining the product topology.

(iii) If all functions f_i are continuous for some topology on X, then $f_i^{-1}(U_i)$ must be open in X whenever $i \in I$ and U_i is an open set in Y_i. On the other hand, they form a subbase for a topology on X.

Exercise 2.7.

(iv) If all functions f_i are continuous for some topology on Y, then $f_i^{-1}(U)$ must be open in X_i for each open set $U \subset Y$ and for each $i \in I$. On the other hand, the sets $U \subset Y$ having this property form a topology on Y by a direct verification.

Exercise 2.8.

(i) We have $\varnothing \in \mathcal{T}$ because $\varnothing$ is open in X, and $\overline{X} = \overline{X} \setminus \varnothing \in \mathcal{T}$ because $\varnothing$ is closed and compact in X.

It remains to show that $\mathcal{T}$ is stable under arbitrary unions and finite intersections. This is true separately for the subfamilies of the sets U (because this is a topology by assumption), and the sets $\overline{X} \setminus K$, because the intersections and finite unions of closed compact sets are again closed and compact. We conclude by observing that all sets $U \cap (\overline{X} \setminus K)$ and $U \cup (\overline{X} \setminus K)$ belong to $\mathcal{T}$. Indeed,

$$U \cap (\overline{X} \setminus K) = U \setminus K$$

is an open set in X, while using the closed set $F := X \setminus U$ of X we have

$$U \cup (\overline{X} \setminus K) = \overline{X} \setminus (F \cap K),$$

and $F \cap K$ is closed and compact in X.

(ii) Consider an arbitrary open cover of $\overline{X}$ by some sets U_i $(i \in I)$ and $\overline{X} \setminus K_j$ $(j \in J)$. Then $J \neq \varnothing$, and

$$\cap_{j \in J} K_j \subset \cup_{i \in I} U_i.$$

Since the intersection is compact in X, there exists a finite subcover

$$\cap_{j \in J} K_j \subset \cup_{k=1}^{m} U_{i_k} =: U,$$

whence

$$\cap_{j\in J}(K_j \setminus U) = \varnothing.$$

The sets $K_j \setminus U = K_j \cap (X \setminus U)$ being compact in X, by Exercise 2.5 (ii) there exist a finite number of K_js such that

$$\cap_{\ell=1}^{n}(K_{j_\ell} \setminus U) = \varnothing.$$

This is equivalent to

$$\cap_{\ell=1}^{n} K_{j_\ell} \subset U = \cup_{k=1}^{m} U_{i_k},$$

i.e., to

$$\left(\cup_{k=1}^{m} U_{i_k} \right) \cup \left(\cup_{\ell=1}^{n} (\overline{X} \setminus K_j) \right) = \overline{X}.$$

(iii) Since the sets $X \cap (\overline{X} \setminus K) = X \setminus K$ are open in X, X is a subspace of $\overline{X}$. The density follows by observing that each neighborhood of ∞ is of the form $\overline{X} \setminus K$ for some compact set K in X. Since X is not compact,

$$X \cap (\overline{X} \setminus K) = X \setminus K \neq \varnothing,$$

i.e., each neighborhood of ∞ meets X.

Exercise 2.9. Fix an arbitrary point $x = (x_i)_{i\in I} \in X$, where $X = \prod_{i\in I} X_i$ is a product of connected spaces, and consider the set D of points $y = (y_i)$ that coincide with x, except for at most finitely many components. Then D is the union of the sets

$$D_{I_0} := \{(y_i) \in X \ : \ y_i = x_i \quad \text{if} \quad i \in I \setminus I_0\},$$

where I_0 runs over the finite subsets of I. Each D_{I_0} is homeomorphic to $\prod_{i\in I_0} X_i$, and hence connected by Proposition 2.16 (b), p. 48. Furthermore, they have a common point (x_i), so that D is also connected by Proposition 2.16 (a).

In view of Proposition 2.16 (c) we complete the proof by showing that D is dense in X. Given an arbitrary non-empty open set of the form

$$U := \{(y_i) \in X \ : \ y_i \in U_i \quad \text{if} \quad i \in I_0\},$$

where I_0 is a finite subset of I, and U_i is an open set of X_i for each $i \in I_0$, it meets D_{I_0} and hence D.[2] Since they form a basis for the topology of X, we conclude that D is dense in X.

Exercise 2.10.

(i) If

$$t = 2\Big(\frac{t_1}{3} + \frac{t_2}{3^2} + \cdots + \frac{t_n}{3^n} + \cdots\Big)$$

and

$$t' = 2\Big(\frac{t_1'}{3} + \frac{t_2'}{3^2} + \cdots + \frac{t_n'}{3^n} + \cdots\Big)$$

are two points of C such that $t_n \neq t_n'$, then $|t-t'| \geq 1/3^n$. Indeed, if they first differ at the kth digit, then $k \leq n$, and therefore

$$\left|t-t'\right| = 2\left|\sum_{i\geq k} \frac{t_i - t_i'}{3^i}\right| \geq \frac{2}{3^k} - \sum_{i>k} \frac{2}{3^i} = \frac{1}{3^k} \geq \frac{1}{3^n}.$$

Hence, if $|t-t'| < 3^{-2n}$, then $t_k = t_k'$ for $k = 1, 2, \ldots, 2n$ and therefore the components $f_1, f_2 : C \to [0,1]$ of f satisfy the inequalities

$$\left|f_i(t) - f_i(t')\right| \leq 2^{-n}, \quad i = 1, 2.$$

This shows that the functions f_i are uniformly continuous.

(ii) Since $[0,1]\setminus C$ is a union of pairwise disjoint open intervals, and since f_1, f_2 are defined at the endpoints of these intervals, it suffices to extend each of them linearly to each open interval.

(iii) Define $f = (f_1, \ldots, f_N)$ with

$$f_i(t) := \sum_{j=1}^{\infty} \frac{t_{(j-1)N+i}}{2^j}, \quad i = 1, \ldots, N.$$

Exercise 2.11.

(iv) If $\beta := \sup f$ is not attained, then the sets

$$\{x \in X \;:\; f(x) < \alpha\}, \quad \alpha < \beta$$

[2] The intersection $U \cap D_{I_0}$ is formed by the points (z_i) satisfying $z_i \in U_i$ for $i \in I_0$ and $z_i = x_i$ for $i \in I \setminus I_0$.

form an open cover of X. Since X is compact, there exists a finite subcover, and then (by monotonicity) there exists $\alpha < \beta$ such that $X \subset \{x \in X \ : \ f(x) < \alpha\}$, contradicting the definition of β.

Exercise 3.1. Setting $A := \|x\|_\infty$ we have

$$A \le \|x\|_p \le (mA^p)^{1/p} = m^{1/p}A,$$

and $m^{1/p} \to 1$.

Exercise 3.2.

(iv) If

$$A = \left\{n + \frac{1}{100n} \ : \ n = 1, 2, \ldots\right\} \quad \text{and} \quad B = \{-n \ : \ n = 1, 2, \ldots\},$$

then

$$0 \in \overline{A + B} \setminus (A + B).$$

Exercise 3.3. The first two sets are not connected, because their images by the continuous determinant function are not connected. The symmetric matrices form a convex set.

Exercise 3.4. We may assume that the neighborhoods V_a are open. Since K is compact, it may be covered by a finite number of these neighborhoods, say

$$K \subset V_{a_1} \cup \cdots \cup V_{a_n}.$$

Set $L := \max\{L_{a_1}, \ldots, L_{a_n}\}$. Next, by Exercise 1.6 we may fix $\delta > 0$ such that if $x, y \in K$ and $d(x, y) < \delta$, then there exists $1 \le j \le n$ satisfying $x, y \in V_{a_j}$.

For any given $x, y \in K$, consider a subdivision of the segment $[x, y]$ into a finite number of subsegments $[x_0, x_1], \ldots, [x_{m-1}, x_m]$ of length $< \delta$ each, with $a = x$ and $b = y$. Then

$$\begin{aligned}\|f(x) - f(y)\| &= \left\|\sum_{j=1}^{m}(f(x_{j-1}) - f(x_j))\right\| \le \sum_{j=1}^{m} \left\|f(x_{j-1}) - f(x_j)\right\| \\ &\le \sum_{j=1}^{m} L \left\|x_{j-1} - x_j\right\| = L\|x - y\|.\end{aligned}$$

Exercise 3.5. Given any $\varepsilon > 0$, for each $x \in K$ there exists an index n_x such that $|(f_{n_x} - f)(x)| < \varepsilon$. Then by continuity there exists a neighborhood V_x of x such that $|(f_{n_x} - f)(y)| < \varepsilon$ for all $y \in V_x$. Furthermore, by the monotonicity of the sequence

(f_n) we even have

$$|(f_n - f)(y)| < \varepsilon \quad \text{for all} \quad y \in V_x \quad \text{and for all} \quad n \ge n_x.$$

Choosing a finite subcover $V_{x_1} \cup \cdots \cup V_{x_m}$ of the compact set K and setting $N(\varepsilon) :=$ max $\{n_{x_1}, \dots, x_{x_m}\}$, we conclude that

$$|(f_n - f)(y)| < \varepsilon \quad \text{for all} \quad y \in K \quad \text{and for all} \quad n \ge N(\varepsilon).$$

Exercise 3.6.

(i) If f and g are piecewise linear, then we may consider a common subdivision in their definition.
(ii) Use the uniform continuity of the functions in $C([a,b], X)$.
(iii) The following trivial estimate holds:

$$\left\| \int_a^b f(t)\, dt \right\| \le (b-a) \|f\|_\infty .$$

(iv) Apply Proposition 3.18.

Exercise 3.7. Assume for simplicity that $[a,b] = [0,2]$, and consider the functions

$$f_n(t) := \operatorname{med}\{0, n(t-1), 1\}, \quad 0 \le t \le 2, \quad n = 1, 2, \dots,$$

where med $\{x, y, z\}$ denotes the middle number among x, y and z. They form a Cauchy sequence for each norm $\|\cdot\|_p$, $1 \le p < \infty$.

If the sequence (f_n) converged in a norm $\|\cdot\|_p$ to some continuous function f, then we should have $f = 0$ in $[0,1]$ and $f = 1$ in $(1,2]$, contradicting its continuity.

Exercise 3.8. Let

$$x = c_1 x_1 + \cdots + c_m x_m$$

for some $c_1, \dots, c_m \in \mathbb{R}$. If $m > n+1$, then the number of the vectors $x_2 - x_1, x_3 - x_1, \dots, x_m - x_1$ exceeds the dimension of $\mathbb{R}^n$, hence there exist $d_2, \dots, d_m \in \mathbb{R}$, not all zero, such that

$$d_2(x_2 - x_1) + d_3(x_3 - x_1) + \cdots + d_m(x_m - x_1) = 0.$$

Setting $d_1 := -d_2 - \cdots - d_m$ this may be written in the form

$$0 = d_1 x_1 + \cdots + d_m x_m.$$

For each $t \in \mathbb{R}$ the above two equalities imply that

$$x = (c_1 + td_1)x_1 + \cdots + (c_m + td_m)x_m.$$

Choosing a non-zero coefficient d_j and then choosing $t = -c_j/d_j$ we obtain that x belongs to the convex hull of $x_1, \dots, x_{j-1}, x_{j-1}, \dots, x_m$.

The theorem follows by repeating this procedure until at most $n + 1$ points x_i remain.

Exercise 3.9. It is sufficient to consider a set of $n + 2$ points $x_1, \dots, x_{n+2} \in \mathbb{R}^n$. The homogeneous linear system

$$\sum_{i=1}^{n+2} c_i x_i = 0, \quad \sum_{i=1}^{n+2} c_i = 0$$

has $n + 2$ unknowns and $n + 1$ scalar equations, hence there is a non-trivial solution $c_1, \dots, c_{n+2} \in \mathbb{R}$. Then the convex hull A of $\{x_i \ : \ c_i > 0\}$ meets the convex hull B of $\{x_i \ : \ c_i < 0\}$, because

$$c := \sum_{c_i>0} c_i = -\sum_{c_i<0} c_i > 0,$$

and hence both contain the point

$$\frac{1}{c}\sum_{c_i>0} c_i x_i = \frac{1}{c}\sum_{c_i<0}(-c_i)x_i.$$

Exercise 3.10.

(i) Assume first that $k = n + 2$. By assumption we may choose for each $j = 1, \dots, n + 2$ a point x_j belonging to all C_i, except perhaps C_j. If two of these points coincide, then this point belongs to all C_i.

Otherwise, applying Radon's theorem (Exercise 3.9) there exists a partition

$$\{x_1, \dots, x_{n+2}\} = A_1 \cup A_2$$

such that the convex hull K_1 of A_1 meets the convex hull K_2 of A_2. We claim that each $x \in K_1 \cap K_2$ belongs to each C_j.

Indeed, if $x_j \in A_1$, then $x_j \notin A_2$ and therefore $x_i \in C_j$ for all $i \in A_2$. Since C_j is convex, the convex hull K_2 of A_2 is a subset of C_j, and therefore $x \in C_j$. The case $x_j \in A_2$ follows by exchanging the role of A_1 and A_2.

Proceeding by induction, let $k > n + 2$, and assume that the result holds for $k - 1$ sets. We already know that any $n + 2$ sets have a non-empty intersection. Hence, replacing C_k and C_{k+1} by their intersection, any $n + 1$ of the sets $C_1, \dots, C_{k-2}$ and $C_k \cap C_{k+1}$ have a non-empty intersection. Applying the induction hypothesis we conclude that

$$\cap_{i=1}^{k} C_i = \Big(\cap_{i=1}^{k-2} C_i \Big) \cap \Big(C_k \cap C_{k+1} \Big) \neq \varnothing .$$

(ii) The system (C_i) has the finite intersection property by (i). We conclude by applying Exercise 2.5, p. 62.

Exercise 3.13.

(i) Use Exercise 2.11.
(ii) Use the convexity-concavity conditions.
(iii) If x_0 belongs to all sets $K(y)$, then

$$\max_x \min_y f(x,y) \geq \min_y f(x_0,y) \geq \alpha = \min_y \max_x f(x,y).$$

(iv) We have

$$\max_x \min_y f(x,y) = \min_y f(x^*,y) \leq f(x^*,y^*) \leq \max_x f(x,y^*) = \min_y \max_x f(x,y)$$

by the definition of x^* and y^*. By (3.7) we have equality everywhere, and this implies (3.8).
(v) If (x^*,y^*) is a saddle point, then

$$\max_x \min_y f(x,y) \geq \min_y f(x^*,y) \geq f(x^*,y^*) \geq \max_x f(x,y^*) \geq \min_y \max_x f(x,y),$$

proving the non-trivial inequality of the minimax equality.

Exercise 4.1. All three partial derivatives of both components of f exist and are continuous on $\mathbb{R}^3$, so that $g \in C^1$, and (using the matrix notation)

$$f'(x,y,z) = \begin{pmatrix} \cos(x-e^z) & 0 & -e^z\cos(x-e^z) \\ 2x & 2y & 0 \end{pmatrix}$$

for all $(x,y,z) \in \mathbb{R}^3$.

Exercise 4.2. The partial derivatives of f exist and are continuous in $U := \mathbb{R}^2 \setminus \{(0,0)\}$, so that $f \in C^1$ in U. f is not continuous at $(0,0)$, because it takes all values between -1 and 1 in each neighborhood of $(0,0)$, so it is not differentiable at $(0,0)$ either.

Exercise 4.4. For any fixed $x \neq 0$ and $h \to 0$ the following relations hold:

$$\begin{aligned} f(x+h) - f(x) &= \frac{x+h}{\|x+h\|^2} - \frac{x}{\|x\|^2} \\ &= \frac{\|x\|^2 (x+h) - \|x+h\|^2 x}{\|x+h\|^2 \, \|x\|^2} \end{aligned}$$

$$= \|x\|^{-4} \frac{(x,x)h - 2(x,h)x + o(h)}{1 + 2\|x\|^{-2}(x,h) + o(h)}$$
$$= \|x\|^{-4}\left[(x,x)h - 2(x,h)x + o(h)\right] \cdot \left[1 - 2\|x\|^{-2}(x,h) + o(h)\right]$$
$$= \|x\|^{-4}\left[(x,x)h - 2(x,h)x\right] + o(h),$$

so that f is differentiable, and

$$f'(x)h = \|x\|^{-4}\left[(x,x)h - 2(x,h)x\right]$$

for all $x \in E \setminus \{0\}$ and $h \in E$.

Exercise 4.5. We have to show the relation

$$f(a+h) - f(a) - Ah = o(h) \quad \text{as} \quad h \to 0.$$

Equivalently, given $\varepsilon > 0$ arbitrarily, we have to find $r > 0$ such that

$$\|f(a+h) - f(a) - Ah\| \le \varepsilon \|h\|$$

for all $h \in X$ satisfying $0 < \|h\| < r$.

Since $\lim_a f'$ exists, we may fix $r > 0$ such that

$$\left\|f'(x) - f'(y)\right\| < \varepsilon \quad \text{for all} \quad x, y \in B_r(a) \setminus \{a\}.$$

Then, for any $h \in X$ satisfying $0 < \|h\| < r$ and for any $t \in (0,1)$, applying Theorem 4.7 there exists an $s \in (t,1)$ such that

$$\left\|f(a+h) - f(a+th) - f'(a+th)(1-t)h\right\|$$
$$\le \left\|(f'(a+sh) - f'(a+th))(1-t)h\right\| \le \varepsilon \|h\|.$$

Letting $t \to 0$ the required estimate follows.

Exercise 4.6.

(i) If f' is bounded by a constant L, then $\|f(x) - f(y)\| \le L\|x - y\|$ for all $x, y \in U$ by Theorem 4.7.

Conversely, if the last estimate holds, then

$$\left\|f'(a)h\right\| = \lim_{t \to 0}\left\|\frac{f(a+th) - f(a)}{t}\right\| \le L\|h\|$$

for all $a \in U$ and $h \in X$, so that $\|f'(a)\| \le L$ for all $a \in U$.

(ii) f' is continuous, hence locally bounded.

Exercise 4.7.

(i) Yes, of order 2 and 0.

(ii) Differentiating the composite function $t \mapsto f(g(t))$ where $g(t) = tx$, for any fixed $x \neq 0$ and using the identity $f(tx) = t^m f(x)$ we obtain for all $t > 0$ that

$$f'(tx)x = mt^{m-1}f(x) \quad \text{or equivalently} \quad x \cdot \nabla f(tx) = mt^{m-1}f(x).$$

For $t = 1$ this is the required result.

(iii) It suffices to show that the function $t^{-m}f(tx)$ is independent of $t > 0$ for any fixed $x \neq 0$. This is true because its derivative vanishes:

$$\begin{aligned}\frac{d}{dt}t^{-m}f(tx) &= -mt^{-m-1}f(tx) + t^{-m}x \cdot \nabla f(tx)\\ &= -mt^{-m-1}f(tx) + t^{-m-1}tx \cdot \nabla f(tx)\\ &= -mt^{-m-1}f(tx) + t^{-m-1}mf(tx) = 0.\end{aligned}$$

Exercise 4.8. Fix $x \in (-r, r)$ arbitrarily. Given any $\varepsilon > 0$, there exists a $\delta > 0$ such that $[x-\delta, x+\delta] \subset (-r, r)$, and

$$|D_1f(z,y) - D_1f(x,y)| < \varepsilon \quad \text{for all} \quad (z,y) \in [x-\delta, x+\delta] \times [a,b]$$

by the uniform continuity of D_1f on this compact rectangle.

If $|h| \leq \delta$, then

$$|f(x+h,y) - f(x,y) - D_1f(x,y)h| = |(D_1f(x+th,y) - D_1f(x,y))h| \leq \varepsilon h$$

with some $t = t(y,h) \in [0,1]$ by Theorem 4.7, and therefore

$$\left|\int_a^b f(x+h,y)\,dy - \int_a^b f(x,y)\,dy - \int_a^b D_1f(x,y)\,dyh\right| \leq (b-a)\varepsilon h.$$

This implies that

$$F(x+h) = F(x) + \Big(\int_a^b D_1f(x,y)\,dy\Big)h + o(h), \quad h \to 0.$$

Exercise 5.1. The case $m = 1$ has already been proved. Let $m \geq 2$, and assume that the property holds until $m-1$.

If $f : X_1 \times \cdots \times X_m \to Y$ is a continuous m-linear map, then it is differentiable, and

$$f'(a)h = \sum_{i=1}^m f(a_1, \ldots, a_{i-1}, h_i, a_{i+1}, \ldots, a_m)$$

for all $a, h \in X_1 \times \cdots \times X_m$. We may write

$$f' = \sum_{i=1}^{m} g_i \circ p_i$$

where

$$p_i : X_1 \times \cdots \times X_m \to X_1 \times \cdots \times X_{i-1} \times X_{i+1} \times \cdots \times X_m$$

is a continuous linear projection and

$$g_i : X_1 \times \cdots \times X_{i-1} \times X_{i+1} \times \cdots \times X_m \to L(X_i, Y)$$

is a continuous $(m-1)$-linear map.

This allows us to conclude by using the differentiability of composite functions.

Exercise 5.2. The function

$$f(x_1, \dots, x_n) = \sqrt{x_1^2 + \cdots + x_n^2}.$$

may be written in the form $f = g \circ h$ with the polynomial $h(x_1, \dots, x_n) := x_1^2 + \cdots + x_n^2$ and with $g(t) := \sqrt{t}$. Since h is of class C^∞ in $\mathbb{R}^n$, and g is of class C^∞ in $(0, \infty)$, the composite function f is of class C^∞ in $\mathbb{R}^n \setminus \{0\}$.

Let us compute the partial derivatives of f. We have

$$D_i f(x_1, \dots, x_n) = \frac{2x_i}{2\sqrt{x_1^2 + \cdots + x_n^2}} = \frac{x_i}{\|x\|}, \quad i = 1, \dots, n,$$

so that

$$f'(x) = \frac{x}{\|x\|}.$$

Furthermore, for $j \neq i$ we have

$$D_j D_i f(x_1, \dots, x_n) = -\frac{x_i}{\|x\|^2} \cdot \frac{x_j}{\|x\|} = -\frac{x_i x_j}{\|x\|^3},$$

while for $j = i$ we get

$$D_i^2 f(x_1, \dots, x_n) = \frac{1}{\|x\|} - \frac{x_i}{\|x\|^2} \cdot \frac{x_i}{\|x\|} = \frac{1}{\|x\|} - \frac{x_i^2}{\|x\|^3}.$$

Hence the Hessian matrix is equal to[3]

$$\frac{1}{\|x\|}I - \frac{1}{\|x\|^3}(x_i x_j)_{i,j=1}^n.$$

Exercise 5.3. Since $f'(x,y) = 0 \iff x = y = 0$, f may have an extremum only in $(0,0)$. It has a strict global minimum here by a direct inspection.

Similarly, g may have an extremum only in $(0,0)$. Since $g(0,0) = 0$, and g takes both positive and negative values in each neighborhood of $(0,0)$, g has no extremum here.

Exercise 5.4.

(i) Prove by induction on $k = 0, 1, \dots$ that

$$h^{(k)}(t) := \begin{cases} p_k(1/t)e^{-1/t} & \text{if } t > 0, \\ 0 & \text{if } t \le 0, \end{cases}$$

with suitable polynomials p_k, and then that

$$p_k(1/t)e^{-1/t} \to 0 \quad \text{as} \quad t \searrow 0.$$

(ii) Observe that $f = h \circ g$ with the polynomial

$$g(x) := 1 - \|x\|^2 = 1 - x_1^2 - \cdots - x_n^2.$$

Exercise 6.1. This is a separable equation. The solution is

$$x(t) = \frac{1}{1 + \ln(1 - t^2)}, \qquad t \in (-1, 1).$$

Exercise 6.2. Differentiating $x(t) = ty(t)$ we get $x'(t) = y(t) + ty'(t)$, and the equation is transformed into

$$y(t) + ty'(t) = f(y) \iff y' = \frac{f(y) - y}{t}.$$

The solutions of $x' = \frac{2tx - x^2}{t^2}$ are given by the formulas

$$x(t) = 0, \quad \text{and} \quad x(t) = \frac{t^2}{t - c} \quad \text{for} \quad c \in \mathbb{R}.$$

[3] We denote by I the identity matrix.

Exercise 6.3. If $x : I \to \mathbb{R}$ satisfies $F(t, x(t)) = c$ for all $t \in I$, then differentiating we get

$$0 = D_1F(t, x(t)) + D_2F(t, x(t))x'(t) = g(t, x(t)) + h(t, x(t))x'(t),$$

which is equivalent to (6.7).

The concrete differential equation may be rewritten in the form

$$(2t + 3t^2x)\ dt + (t^3 - 3x^2)\ dx = 0.$$

This is exact, because

$$\frac{d}{dx}(2t + 3t^2x) = 3t^2 = \frac{d}{dt}(t^3 - 3x^2),$$

hence there exists a primitive F satisfying

$$\frac{dF}{dt} = 2t + 3t^2x \quad \text{and} \quad \frac{dF}{dx} = t^3 - 3x^2.$$

The first condition implies that

$$F(t, x) = t^2 + t^3x + c_1(x)$$

with a function $c_1(x)$ not depending on t. Inserting this into the second condition we obtain $c_1'(x) = -3x^2$, whence $c_1(x) = -x^3$ and thus

$$F(t, x) = t^2 + t^3x - x^3$$

is a suitable primitive. We conclude that the solutions of the algebraic equations

$$t^2 + t^3x - x^3 + c = 0, \quad c \in \mathbb{R}$$

satisfy our differential equation.

Exercise 6.4.

(i) If

$$D_1F := \frac{dF}{dt} = g \quad \text{and} \quad D_2F := \frac{dF}{dx} = h$$

in D, then F is of class C^2 by Proposition 5.20 (p. 138), and then

$$D_2g = D_2D_1F = D_1D_2F = D_1h$$

by Theorem 5.6 (p. 123).

(ii) Fix $(t_0, x_0) \in D$ arbitrarily, and set

$$F_1(t,x) := \int_{t_0}^{t} g(t',x)\,dt'.$$

Then $D_1F_1 = g$ and

$$D_2F_1(t,x) = \int_{t_0}^{t} D_2g(t',x)\,dt' = \int_{t_0}^{t} D_1h(t',x)\,dt' = h(t,x) - h(t_0,x).$$

In order to eliminate the last term we set

$$F_2(x) := \int_{x_0}^{x} h(t_0,x')\,dx' \quad \text{and} \quad F(t,x) := F_1(t,x) + F_2(x).$$

Then

$$D_1F = g \quad \text{and} \quad D_2F = D_2F_1 + F_2' = h$$

as required.

Exercise 6.5. Dividing the equation by t^2 we obtain the exact equation

$$\frac{x}{t^2}\,dt + \left(-4x - \frac{1}{t}\right)\,dx = 0.$$

Indeed,

$$\frac{d}{dx}\frac{x}{t^2} = \frac{1}{t^2} = \frac{d}{dt}\left(-4x - \frac{1}{t}\right).$$

Therefore there exists a function F satisfying

$$\frac{dF}{dt} = \frac{x}{t^2} \quad \text{and} \quad \frac{dF}{dx} = -4x - \frac{1}{t}.$$

The first condition implies that

$$F(t,x) = -\frac{x}{t} + c_1(x)$$

with a function $c_1(x)$ independent of t. Inserting this expression into the second condition we obtain that $c_1'(x) = -4x$, so that

$$F(t,x) = -2x^2 - \frac{x}{t}$$

is a suitable primitive. We conclude that the solutions of the algebraic equations

$$2x^2 + \frac{x}{t} = c, \quad c \in \mathbb{R}$$

solve our differential equation.

Exercise 6.6. Let $n = 2$ to simplify the formulas. Using the equality

$$\begin{pmatrix} w'_{11} & w'_{12} \\ w'_{21} & w'_{22} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix},$$

we have

$$\begin{aligned} \begin{vmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{vmatrix}' &= \begin{vmatrix} w'_{11} & w'_{12} \\ w_{21} & w_{22} \end{vmatrix} + \begin{vmatrix} w_{11} & w_{12} \\ w'_{21} & w'_{22} \end{vmatrix} \\ &= \begin{vmatrix} a_{11}w_{11} + a_{12}w_{21} & a_{11}w_{12} + a_{12}w_{22} \\ w_{21} & w_{22} \end{vmatrix} \\ &\quad + \begin{vmatrix} w_{11} & w_{12} \\ a_{21}w_{11} + a_{22}w_{21} & a_{21}w_{12} + a_{22}w_{22} \end{vmatrix} \\ &= \begin{vmatrix} a_{11}w_{11} & a_{11}w_{12} \\ w_{21} & w_{22} \end{vmatrix} + \begin{vmatrix} w_{11} & w_{12} \\ a_{22}w_{21} & a_{22}w_{22} \end{vmatrix} \\ &= (a_{11} + a_{22}) \begin{vmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{vmatrix}. \end{aligned}$$

Exercise 6.7. The local Lipschitz condition is satisfied everywhere except at $(0,0)$. Determine all maximal solutions in the domain

$$\{(t,x) \in \mathbb{R}^2 \ : \ |x| \ne t^2\}.$$

Conclude that the initial value problem with $x(t_0) = x_0$ has a unique maximal solution if $|x_0| > t_0^2$, and it has infinitely many maximal solutions otherwise.

Exercise 6.8.

(i) The formula $x(t) = \frac{3-\sqrt{5}}{2}t^2$ defines a solution. Its uniqueness follows by a monotonicity argument.

(ii) Show that $x_{2n}(t) = 0$ and $x_{2n+1}(t) = t^2$, $n = 0, 1, \dots$.

Exercise 6.9. There are infinitely many solutions, for example $x(t) = c - \sqrt{t^4 + c^2}$ for $c \ge 0$, and $x(t) = c + \sqrt{t^4 + c^2}$ for $c \le 0$.

Exercise 6.10. Replace qu by $u'' + \lambda u$ under the integral sign and integrate by parts in $(x - t, x)$ and in $(x, x + t)$ to make u'' disappear.

Exercise 6.11. The function

$$f(x) := e^{\alpha x} \int_x^\infty E(s)\,ds, \quad x \ge 0$$

is non-increasing because

$$f'(x) = e^{\alpha x}\Big(\alpha \int_x^\infty E(s)\,ds - E(x)\Big) \le 0$$

almost everywhere. Hence

$$\alpha e^{\alpha x} \int_x^\infty E(s)\,ds = \alpha f(x) \le \alpha f(0) = \alpha \int_0^\infty E(s)\,ds \le E(0)$$

for all $x \ge 0$. Since E is nonnegative and non-increasing, we conclude that

$$E(0)e^{-\alpha x} \ge \alpha \int_x^\infty E(s)\,ds \ge \alpha \int_x^{x+\alpha^{-1}} E(s)\,ds \ge E(x+\alpha^{-1}).$$

Putting $t = x + \alpha^{-1}$ the inequality

$$E(x+\alpha^{-1}) \le E(0)e^{-\alpha x}, \quad x \ge 0$$

takes the form

$$E(t) \le E(0)e^{1-\alpha t}, \quad t \ge \alpha^{-1}.$$

The last inequality also holds for $0 \le t \le \alpha^{-1}$ because $E(t) \le E(0)$.

Exercise 7.2. The equations of the tangent lines at any given point (x_0, y_0) are obtained by *half substitutions*:

$$\frac{x_0 x}{a^2} + \frac{y_0 y}{b^2} = 1, \quad \frac{x_0 x}{a^2} - \frac{y_0 y}{b^2} = 1 \quad \text{and} \quad x_0 x = p(y_0 + y).$$

Exercise 7.3. Differentiating

$$x^3 + y^3 - 3xy = 0$$

we obtain that

$$3x^2 + 3y^2 y' - 3y - 3xy' = 0.$$

If $y'(x) = 0$, then $y(x) = x^2$, and hence

$$x^3 + x^6 - 3x^3 = 0.$$

This leaves two candidates: $(x, y) = (0, 0)$ and $(x, y) = (2^{1/3}, 2^{2/3})$.

Exercise 7.4.

(i) f' is not continuous.
(ii) f is still *locally invertible*.

Exercise 7.5.

(i) By homogeneity it suffices to show that if $x_1 \geq 0, \dots, x_n \geq 0$ and $x_1 + \dots + x_n = n$, then $x_1 \cdots x_n \leq 1$. By compactness the product $x_1 \cdots x_n$ has a maximum on this set. Since it is necessarily attained at a point satisfying $x_1 > 0, \dots, x_n > 0$ and $x_1 + \dots + x_n = n$, it is a maximum of $f_0(x) := x_1 \cdots x_n$ in the open set defined by the inequalities $x_1 > 0, \dots, x_n > 0$, under the condition

$$f_1(x) := x_1 + \dots + x_n - n = 0.$$

This leads to the condition $\lambda_0 f_0'(x) + \lambda_1 f_1'(x) = 0$, or equivalently

$$\lambda_0 x_1 \cdots x_n + \lambda_1 x_i = 0, \quad i = 1, \dots, n.$$

Hence $x_1 = \dots = x_n = 1$, and thus $f_0(x) = 1$.
(ii) Minimize $\frac{x^p}{p} + \frac{y^q}{q}$ under the condition $xy = 1$.
(iii) Maximize $x_1 y_1 + \dots + x_n y_n$ under the conditions $x_1^p + \dots + x_n^p = 1$ and $y_1^q + \dots + y_n^q = 1$.

Exercise 7.6.

(i) If $0 < x < y$ or $y < x < 0$, then apply Hölder's inequality for $a_1^x, \dots, a_n^x$ and $1, \dots, 1$ with $p = y/x$:

$$a_1^x + a_2^x + \dots + a_n^x \leq (a_1^y + a_2^y + \dots + a_n^y)^{x/y} n^{1-(x/y)}.$$

Hence $f(x) \leq f(y)$ in the first case, and $f(x) \geq f(y)$ in the second case.

It remains to deal with the case $x = 0$. Applying for $a_1^y, \dots, a_n^y$ the inequality between arithmetic and geometric means we obtain that

$$f(0)^y = \sqrt[n]{a_1^y \cdots a_n^y} \leq \frac{a_1^y + a_2^y + \dots + a_n^y}{n} = f(y)^y.$$

Hence $f(0) \leq f(y)$ if $y > 0$, and $f(0) \geq f(y)$ if $y < 0$.

If $a_1 = \dots = a_n$, then $f(x)$ does not depend on x. Otherwise f is (strictly) increasing.

(ii) Only the continuity at 0 is not obvious. Applying l'Hospital's rule for $\ln f(x)$ we obtain

$$\begin{aligned}\lim_{x\to 0}\ln f(x) &= \lim_{x\to 0}\frac{\ln\frac{a_1^x+\cdots+a_n^x}{n}}{x}\\ &= \lim_{x\to 0}\frac{n}{a_1^x+\cdots+a_n^x}\cdot\frac{a_1^x\ln a_1+\cdots+a_n^x\ln a_n}{n}\\ &= \frac{\ln a_1+\cdots+\ln a_n}{n};\end{aligned}$$

hence

$$\lim_{x\to 0} f(x) = \sqrt[n]{a_1\cdots + a_n}.$$

Exercise 7.7.

(i) Minimize the function $f_0(x,y,z) := 2xy + 2yz + 2zx$ under the condition $f_1(x,y,z) := xyz - V = 0$.

(ii) Recalling *Heron's formula*

$$4A = \sqrt{(x+y-z)(y+z-x)(z+x-y)(x+y+z)}$$

for the area of a triangle of sides x, y, z, maximize the product

$$f_0(x,y,z) : (x+y-z)(y+z-x)(z+x-y)(x+y+z)$$

under the condition $f_1(x,y,z) := x+y+z-P = 0$.

Exercise 8.2.

(ii) This is a special case of Newton's interpolation formula. A direct proof by induction is as follows. For $k = 0$ the formula reduces to the definition of $\Delta_0^{(0)}$. If it holds until some k, then

$$a_k = \sum_{j=0}^{k}\binom{k}{j}\Delta_0^{(j)} \quad\text{and}\quad \Delta_k^{(1)} = \sum_{j=0}^{k}\binom{k}{j}\Delta_0^{(j+1)}$$

because $(\Delta_k^{(1)})$ is also a generalized arithmetic sequence. It follows that

$$\begin{aligned}a_{k+1} &= a_k + \Delta_k^{(1)}\\ &= \sum_{j=0}^{k}\binom{k}{j}\left[\Delta_0^{(j)} + \Delta_0^{(j+1)}\right]\end{aligned}$$

$$
\begin{aligned}
&= \Delta_0^{(0)} + \left(\sum_{j=1}^{k} \left[\binom{k}{j} + \binom{k}{j-1} \right] \Delta_0^{(j)} \right) + \Delta_0^{(k+1)} \\
&= \Delta_0^{(0)} + \left(\sum_{j=1}^{k} \binom{k+1}{j} \Delta_0^{(j)} \right) + \Delta_0^{(k+1)} \\
&= \sum_{j=0}^{k} \binom{k}{j} \Delta_0^{(j)},
\end{aligned}
$$

so that the formula holds for $k+1$, too.

(iii) We have

$$1^2 + \cdots + k^2 = 0\binom{k}{0} + 1\binom{k}{1} + 3\binom{k}{2} + 2\binom{k}{3} = \frac{k(k+1)(2k+1)}{6}$$

and

$$1^3 + \cdots + k^3 = 0\binom{k}{0} + 1\binom{k}{1} + 7\binom{k}{2} + 12\binom{k}{3} + 6\binom{k}{4} = \left[\frac{k(k+1)}{2}\right]^2.$$

Exercise 8.3.

(i) For each $\ell \in \Sigma$, the expression

$$\prod_{j=0}^{n} \prod_{i=0}^{\ell_j - 1} \frac{x_j - i}{\ell_j - i}$$

defines a polynomial of total degree $\ell_1 + \cdots + \ell_n \leq k$, and it vanishes for all $x \in \Sigma$, except if $x_j \geq \ell_j$ for all $j = 0, \ldots, n$. Since

$$x_0 + \cdots + x_n = k = \ell_0 + \cdots + \ell_n,$$

this can only happen if $x = \ell$, and then the expression is equal to one.

Exercise 8.4. Apply Rolle's theorem $(n-1)$ times to $f - Lf$, and use Proposition 8.5 (a).

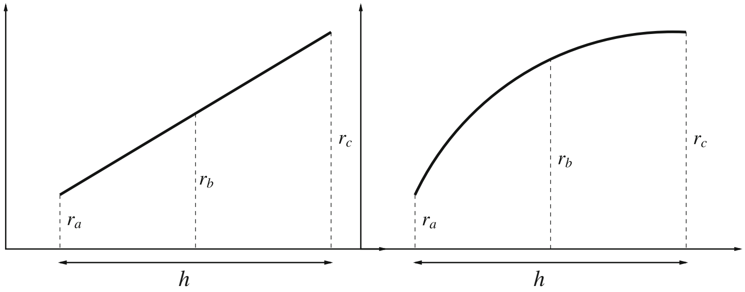


Fig. 1 Solution of Exercise 8.5

Exercise 8.5.

(i) and (ii) If we revolve the area between the graph of a function $f : [a,c] \to \mathbb{R}$ and the x-axis, then the volume of the body of revolution is given by the formula

$$V = \pi \int_a^c f(x)^2 \, dx.$$

Observe that $f(x) = \alpha x + \beta$ is an affine function in the first case, and $f(x) = \sqrt{R^2 - x^2}$ in the second case; see the figure where r_a, r_b, r_c denote the radii of the discs of area A, B, C. In particular, the fourth derivative of f^2 vanishes in both cases, and hence Simpson's formula is exact by formula (8.24), p. 207 (Fig. 1).

The proof of (8.36) is as follows (see Fig. 2):

$$V = \frac{\pi h}{6}\Big((R^2 - (R-h)^2) + 4\Big(R^2 - \Big(R - \frac{h}{2}\Big)^2\Big)\Big) = \pi h^2\Big(R - \frac{h}{3}\Big).$$

(iii) The proportion is $1 : 2 : 3$.

Exercise 8.6. Approximate $f^{(k)}$ uniformly with a sequence of polynomials q_n, and take suitable kth primitives.

Exercise 9.1. The functions $q_n(x) := (-1)^n p_n(x)$ have the same properties, hence $q_n = p_n$ by uniqueness.

Exercise 9.2. Orthogonal polynomials do not have multiple roots.

Exercise 9.3.

(i) Fix n arbitrarily and set $u(x) := (x^2-1)^n$. Since u is a polynomial of degree $2n$, $q_n = u^{(n)}$ is a polynomial of degree n. It remains to prove that $u^{(n)}$ is orthogonal to all polynomials v of degree $< n$. Integrating by parts n times we

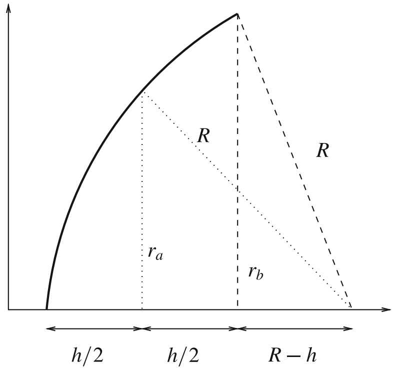


Fig. 2 Proof of (8.36)

obtain the equality

$$\int_{-1}^{1} u^{(n)} v\,dx = \left[u^{(n-1)}v - u^{(n-2)}v' + \cdots + (-1)^{n-1}uv^{(n-1)}\right]_{-1}^{1} + (-1)^n \int_{-1}^{1} uv^{(n)}\,dx.$$

We conclude by observing that all terms on the right-hand side vanish. Indeed, since -1 and 1 are zeros of multiplicity n of u, $u^{(j)}(\pm 1) = 0$ for $j = 0, \ldots, n-1$. Furthermore, the last integral also vanishes because v has degree $< n$, and hence $v^{(n)} \equiv 0$.

(ii) This follows from (i) and the uniqueness part of Proposition 9.1 (b), p. 220.

(iii) Fix n arbitrarily. Differentiating the function $u(x) := (x^2-1)^n$ we get

$$(x^2-1)u'(x) = 2nxu(x).$$

Differentiating $n+1$ more times and using the Leibniz formula for the derivation of products we obtain the equality

$$(x^2-1)u^{(n+2)}(x) + (n+1)2xu^{(n+1)}(x) + \frac{(n+1)n}{2}2u^{(n)}(x) = 2nxu^{(n+1)}(x) + (n+1)2nu^{(n)}(x).$$

Since $q_n(x) = u^{(n)}(x)$, this is equivalent to the differential equation

$$(1-x^2)q_n'' - 2xq_n' + n(n+1)q_n = 0.$$

Exercise 9.4. Adapt the solution of Exercise 9.3 for

$$u(x) := (1-x)^{n+\alpha}(1+x)^{n+\beta}.$$

Exercise 9.5.

(i) It may be proved by induction that

$$q_n(x) := 2^{1-n}\cos(n\arccos x)$$

is a polynomial of degree n with leading coefficient one. By uniqueness it remains to show that they are orthogonal. This follows from the orthogonality of the cosine system by the change of variables $x = \cos t$: if $n \ne k$, then

$$\begin{aligned}\int_{-1}^{1} q_n(x)q_k(x)(1-x^2)^{-1/2}\,dx \\ &= 2^{1-n}2^{1-k}\int_{\pi}^{0}\cos nt\cos kt\frac{-\sin t}{\sin t}\,dt \\ &= 2^{1-n}2^{1-k}\int_{0}^{\pi}\cos nt\cos kt\,dt = 0.\end{aligned}$$

(ii) It may be proved by induction that

$$q_n(x) := 2^{-n}\frac{\sin((n+1)\arccos x)}{\sin\arccos x}$$

is a polynomial of degree n with leading coefficient one. By uniqueness it remains to show that they are orthogonal. This follows by the same change of variables $x = \cos t$ as in (i): if $n \ne k$, then

$$\begin{aligned}\int_{-1}^{1} q_n(x)q_k(x)(1-x^2)^{1/2}\,dx \\ &= 2^{-n-k}\int_{\pi}^{0}\frac{\sin(n+1)t}{\sin t}\cdot\frac{\sin(k+1)t}{\sin t}\cdot\sin t\cdot(-\sin t)\,dt \\ &= 2^{-n-k}\int_{0}^{\pi}\sin(n+1)t\sin(k+1)t\,dt = 0.\end{aligned}$$

Exercise 9.6. Adapt the solution of Exercise 9.3 for $u(x) := x^{\alpha+n}e^{-x}$.
Exercise 9.7. Set $u(x) := e^{-x^2}$.

(i) We have $u^{(n)}(x) = q_n(x)e^{-x^2}$ for $n = 0, 1, \dots$ by induction with suitable polynomials q_n of degree n. Furthermore, the main coefficient of q_n is equal to $(-2)^n$. Finally, q_n is orthogonal to every polynomial v of degree $< n$, because integrating by parts n times we have

$$\begin{aligned}\int_{-\infty}^{\infty} q_n(x)v(x)e^{-x^2}\,dx &= \int_{-\infty}^{\infty} u^{(n)}v\,dx \\ &= \big[u^{(n-1)}v - u^{(n-2)}v' + \cdots + (-1)^{n-1}uv^{(n-1)}\big]_{-\infty}^{\infty} \\ &\quad + (-1)^n \int_{-\infty}^{\infty} uv^{(n)}\,dx.\end{aligned}$$

Since $v^{(n)} \equiv 0$, and all derivatives of u tend to zero in $\pm\infty$, the last expression vanishes. We conclude that $q_n(x) \equiv (-2)^n p_n(x)$.

(ii) We have $u'(x) = -2xu(x)$. Differentiating $n+1$ times this yields

$$u^{(n+2)}(x) + 2xu^{(n+1)}(x) + 2(n+1)u^{(n)}(x) = 0.$$

Since $u^{(n)}(x) = q_n(x)e^{-x^2}$, and hence

$$u^{(n+1)}(x) = \big[q_n'(x) - 2xq_n(x)\big]e^{-x^2}$$

and

$$u^{(n+2)}(x) = \big(q_n''(x) - 4xq_n'(x) - 2q_n(x) + 4x^2q_n(x)\big)e^{-x^2},$$

we conclude that

$$\begin{aligned}&\big[q_n''(x) - 4xq_n'(x) - 2q_n(x) + 4x^2q_n(x)\big]e^{-x^2} \\ &\qquad + 2x\big[q_n'(x) - 2xq_n(x)\big]e^{-x^2} + 2(n+1)q_n(x)e^{-x^2} = 0.\end{aligned}$$

This may be simplified to

$$q_n''(x) - 2xq_n'(x) + 2nq_n(x) = 0.$$

(iii) Applying Taylor's formula we have

$$e^{-(x+t)^2} = u(x+t) = \sum_{n=0}^{\infty} \frac{u^{(n)}(x)}{n!}t^n = \sum_{n=0}^{\infty} \frac{e^{-x^2}q_n(x)}{n!}t^n$$

for all $x, t \in \mathbb{R}$. Multiplying by e^{x^2} hence we obtain that

$$e^{-2xt-t^2} = \sum_{n=0}^{\infty} \frac{q_n(x)}{n!}t^n = \sum_{n=0}^{\infty} \frac{p_n(x)}{n!}(-2t)^n.$$

Exercise 9.8. See Jackson [254].

Exercise 9.9. We admit from electrostatics[4] that there is a unique equilibrium position $-1 < x_1 < \cdots < x_n < 1$, characterized by the system of equations[5]

$$\sum_{k\neq j} \frac{1}{x_j - x_k} + \frac{1}{2(x_j - 1)} + \frac{1}{2(x_j + 1)} = 0, \quad j = 1, \dots, n.$$

Setting

$$f(x) := (x - x_1) \cdots (x - x_n)$$

it may be rewritten in the form

$$\frac{f''(x_j)}{f'(x_j)} + \frac{1}{x_j - 1} + \frac{1}{x_j + 1} = 0, \quad j = 1, \dots, n.$$

Since $f(x_j) = 0$ for all j, this is equivalent to

$$(1 - x_j^2)f''(x_j) - 2x_jf'(x_j) + n(n + 1)f(x_j) = 0, \quad j = 1, \dots, n.$$

Since

$$(1 - x^2)f''(x) - 2xf'(x) + n(n + 1)f(x)$$

is a polynomial of degree $< n$ by a direct computation of the coefficient of x^n, it has to vanish identically, so that f satisfies Legendre's differential equation

$$(1 - x^2)f''(x) - 2xf'(x) + n(n + 1)f(x) = 0.$$

Exercise 10.1. Given an arbitrary non-empty open subinterval J of I, choose a continuous function $f : I \to \mathbb{R}$ vanishing outside J and positive in J. If none of the nodes x_k^n belonged to J, then the convergence in Theorem 10.4 (p. 239) could not hold.

Exercise 10.2.

(i) By convexity the graph of f is between the straight line joining the points (k, f_k) and $(k + 1, f_{k+1})$, and the tangent line at $k + 1$.

(ii) By convexity the slope of the tangent line at $k + 1$ is between the slopes of the straight lines joining the point $(k + 1, f_{k+1})$ to (k, f_k) and $(k + 2, f_{k+2})$, respectively. Furthermore, all slopes are ≤ 0 because f is non-increasing.

[4] See Szegő [476], Section 6.7 for more details and to a generalization to all Jacobi polynomials.

[5] For each fixed j the sum is taken over the $n - 1$ indices k different from j.

(iii) We infer from (i) and (ii) that

$$0 \le \frac{f_k + f_{k+1}}{2} - \int_k^{k+1} f(x)\, dx = t_k \le T_k$$

for each k, and

$$\sum_{k=0}^{\infty} T_k \le \frac{f_0 - f_1}{2}.$$

Summing for $k = 0, \dots, n-1$ this yields

$$0 \le (f_0 + \cdots + f_n) - \frac{f_0 + f_n}{2} - \int_0^n f(x)\, dx \le \frac{f_0 - f_1}{2}$$

for each n. Since the expression in the middle is non-decreasing, it has a limit γ as $n \to \infty$, and

$$0 \le \gamma \le \frac{f_0 - f_1}{2}.$$

Applying this result with $f(x) := (1 + x)^{-1}$ and $f(x) := -\ln(1 + x)$ we obtain respectively

$$\frac{1}{2} \le \left(\frac{1}{1} + \cdots + \frac{1}{n} - \ln n\right) - \frac{1}{2n} \le \frac{3}{4}$$

and

$$1 \ge \ln n! - \left(n + \frac{1}{2}\right) \ln n + n \ge 1 - \frac{\ln 2}{2}$$

for every n, and hence

$$0.5 \le \lim_{n\to\infty} \left(\frac{1}{1} + \cdots + \frac{1}{n} - \ln n\right) - \frac{1}{2n} \le 0.75$$

and

$$1.922 \approx \frac{e}{\sqrt{2}} \le \lim_{n\to\infty} \frac{n! e^n}{n^{n+\frac{1}{2}}} \le e \approx 2.718.$$

We recall that the limits are equal to $C \approx 0.522$ and $\sqrt{2\pi} \approx 2.507$, respectively.

Exercise 10.3.

(i) J_2 represents the area of a quarter of the unit disk.
(ii) Integrate by parts.

Exercise 10.4.

(i) Since $f'(x) \searrow 0$ as $x \to \infty$, $f' \geq 0$ and hence f is non-decreasing.
(ii) This was established during the proof of the corollary.
(iii) If $m \geq 1$, then $f^{(2m+1)}$ is non-increasing by (ii). Since $f^{(2m+1)} \to 0$ by hypothesis, we conclude that $f^{(2m+1)} \geq 0$ and hence $f^{(2m)}$ is non-decreasing.

It remains to show that $\alpha := \lim_\infty f^{(2m)} = 0$. In case $\alpha > 0$ there would exist an $x_0 \geq 0$ such that $f^{(2m)}(x) > \alpha/2$ for all $x \geq x_0$, and this would lead to a contradiction:

$$0 = \lim_\infty f^{(2m-1)}(x) \geq \lim_\infty f^{(2m-1)}(x_0) + \frac{\alpha}{2}(x - x_0) = \infty.$$

In case $\alpha < 0$ we would obtain similarly the contradiction

$$0 = \lim_\infty f^{(2m-1)}(x) \leq \lim_\infty f^{(2m-1)}(x_0) + \frac{\alpha}{2}(x - x_0) = -\infty.$$

Exercise 10.5. The function

$$f(x) := \frac{\pi - x}{2}$$

is differentiable on $(0, 2\pi)$, hence its Fourier series converges to $f(x)$ at each $x \in (0, 2\pi)$.[6] This proves the first equality, which is equivalent to the series of $b_1(x)$. The others follow by induction, by taking successive primitive functions.

Exercise 10.8. Apply Exercise 10.7.

Exercise 10.9.

(i) The sum of the first series is $(e^t - 1)/t$, and the second series converges by Exercise 10.7. Apply Proposition 10.9 (p. 250).
(ii) Use Exercise 10.7.

[6] We recall that a piecewise continuous, 2π-periodic function is the sum of its Fourier series at each point where it is differentiable. This is a special case of a theorem of Lipschitz and Dini; see, e.g., Chernoff [101] (or [285, Exercise 8.5]) for a simple proof.

Exercise 10.10. If $|x| < \pi$, then applying 10.9 (iii) we get

$$\begin{aligned} x\coth x &= x\frac{e^x + e^{-x}}{e^x - e^{-x}} = x\frac{e^{2x}+1}{e^{2x}-1} \\ &= x + \frac{2x}{e^{2x}-1} = x + \sum_{n=0}^{\infty} b_n \cdot (2x)^n \\ &= x + 1 - x + \sum_{n=2}^{\infty} b_n \cdot (2x)^n = 1 + \sum_{k=1}^{\infty} b_{2k} \cdot (2x)^{2k}. \end{aligned}$$

Using Euler's relation $e^{ix} = \cos x + i \sin x$ it follows that[7]

$$x\cot x = (ix)\coth(ix) = 1 + \sum_{k=1}^{\infty} (-1)^k b_{2k} \cdot (2x)^{2k}.$$

Finally, using the identity

$$\tan x = \cot x - 2\cot(2x),$$

for $|x| < \pi/2$ we obtain

$$x\tan x = x\cot x - 2x\cot(2x) = \sum_{k=1}^{\infty} (-1)^k b_{2k} \cdot \left[(2x)^{2k} - (4x)^{2k}\right]$$

and hence

$$\tan x = \sum_{k=1}^{\infty} (-1)^k b_{2k} \cdot (2^{2k} - 4^{2k}) x^{2k-1}.$$

Exercise 11.1. We obtain the iteration

$$x_{n+1} := x_n - \frac{f(x_n)}{f'(x_n)} = \frac{(p-1)x^p + A}{px^{p-1}}.$$

[7] The use of complex numbers may be avoided, but the computations are longer; see, e.g., [170, Section 449].

Exercise 11.2.

(i) and (ii) Direct computation.

(iii) Applying Cauchy's mean value theorem (Proposition 4.4 (c), p. 105), the right-hand side of (11.4) is equal to

$$\frac{f'(\alpha)(\alpha-c)-f(\alpha)}{(\alpha-c)^2 f'(\alpha)}$$

for a suitable α between x_n and x_{n+1}.

Applying Taylor's formula

$$0=f(c)=f(\alpha)+f'(\alpha)(c-\alpha)+\frac{f''(\beta)}{2}(c-\alpha)^2$$

this expression is equal to

$$\frac{f''(\beta)}{2f'(\alpha)}$$

for a suitable β between x_n and x_{n+1}.

(iv) The last equality is equivalent to

$$x_{n+2}-c=\frac{f''(\beta)}{2f'(\alpha)}(x_{n+1}-c)(x_n-c).$$

Hence

$$|A(x_{n+2}-c)|\le|A(x_{n+1}-c)|\cdot|A(x_n-c)|,$$

i.e.,

$$d_{n+2}\le d_{n+1}d_n,\quad n=0,1,\ldots.$$

(v) Easy proof by induction.

Exercise 11.4.

(i) Use the continuity of A.

(ii) The map $F(x):=Ax+b$ is a contraction.

(iii) There exists a (non-zero) eigenvector x_0 with an eigenvalue λ of modulus ≥ 1. If $\lambda=1$, then $I-A$ is not onto, and the equation $x=Ax+b$ has no solution for any b outside the range of $I-A$. Otherwise for $b=0$ the sequence $(x_n)=(\lambda^n x_0)$ does not converge.

Exercise 11.5.

(i) Use the continuity of A again.

(ii) If $Ax = \lambda x$ for some $x \neq 0$, then

$$|\lambda| \cdot \|x\| = \|Ax\| \le \|A\| \cdot \|x\|,$$

whence $|\lambda| \le \|A\|$.

(iii) If X has a basis formed by eigenvectors $e_1, \dots, e_r$, then the map $F(x) := Ax + b$ is a contraction for the equivalent Euclidean norm

$$\|c_1 e_1 + \cdots + c_r e_r\| := \sqrt{|c_1|^2 + \cdots + |c_r|^2}.$$

The general case may be treated similarly, by considering the Jordan decomposition of A; see the details, e.g., in Varga [498].

(iv) Repeat the proof of Exercise 11.4 (iii).

(v) Observe that A^*A is selfadjoint, and hence[8]

$$\sup_{\|x\| \le 1} (A^*Ax, x) = \rho(A^*A).$$

Since

$$\|Ax\|^2 = (Ax, Ax) = (A^*Ax, x)$$

for all x, this is equivalent to

$$\sup_{\|x\| \le 1} \|Ax\| = \sqrt{\rho(A^*A)}.$$

(vi) The definition of the spectral radius implies that $\rho(A^2) = \rho(A)^2$. Since $A^* = A$, it follows that

$$\|A\| = \sqrt{\rho(A^*A)} = \sqrt{\rho(A^2)} = \rho(A).$$

(vii) Consider the matrix

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

[8] See the *proof* of Lemma 7.5, p. 174.

Exercise 12.1. Consider the differential equation $x' = x$ with the initial condition $x(0) = 1$. Show that for $\tau' = t > 0$ and $h = t/n$ we have $x_n = (1 + t/n)^n$ and

$$\inf_n n(e^t - x_n) > 0.$$

Exercise 12.2. Consider the differential equation $x' = 4t^3$ with the initial condition $x(0) = 0$. Its solution is $x(t) = t^4$. Applying the modified Euler method with $\tau' = 1$ and $h = 1/n$ we have $z_0 = 0$ and

$$z_{k+1} = z_k + 4\left(\frac{k}{n} + \frac{1}{2n}\right)^3 \frac{1}{n} = z_k + \frac{(2k+1)^3}{2n^4}$$

for $k = 0, 1, \dots, n-1$. Hence

$$\begin{aligned} x_n &= \frac{1}{2n^4}\sum_{k=0}^{n-1}(2k+1)^3 = \frac{1}{2n^4}\left(\sum_{k=1}^{2n} j^3 - \sum_{k=1}^{n}(2j)^3\right) \\ &= \frac{1}{8n^4}\left((2n)^2(2n+1)^2 - 8n^2(n+1)^2\right) = \frac{(2n+1)^2 - 2(n+1)^2}{2n^2} \\ &= 1 - \frac{1}{2n^2}, \end{aligned}$$

and therefore

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

Exercise 12.3.

(i) Since both definitions remain unchanged if we replace the norm by an equivalent norm, we may consider a norm as in the proof of Exercise 11.5 (iii).

(ii) It suffices to show that

$$\limsup_{n\to\infty} \sqrt[n]{\|A^n\|} < \alpha$$

for every fixed $\alpha > \rho(A)$. Fixing a positive integer m such that $\sqrt[m]{\|A^m\|} < \alpha$, it suffices to show that

$$\limsup_{k\to\infty} \sqrt[km+r]{\|A^{km+r}\|} \le \alpha \quad \text{for} \quad r = 1, \dots, m.$$

This follows from the estimate

$$\left\|A^{km+r}\right\| \le \|A^r\| \cdot \|A^m\|^k \le \|A^r\|\,\alpha^{mk} = \|A^r\|\,\alpha^{-r}\alpha^{mk+r},$$

implying that

$$\sqrt[km+r]{\|A^{km+r}\|} \le \left(\|A^r\| \, \alpha^{-r} \right)^{1/(km+r)} \alpha \to \alpha.$$

(iii) If $\rho(A) < 1$, then there exists a positive integer m such that $\|A^m\| < 1$. The the mth iteration of the function $f(x) := Ax + b$ is a contraction in X, and we may apply Proposition 12.4.

(iv) The complex case was solved in Exercise 11.5; henceforth we assume that X is a finite-dimensional real normed space.

If $I - A$ is not onto, or if A has a real eigenvalue of modulus ≥ 1, then we may repeat the proof of Exercise 11.5 (iv). Otherwise identify X with $\mathbb{R}^d$, A with a real square matrix, and consider the operator defined by this matrix in $\mathbb{C}^d$. It has a non-real eigenvalue $\lambda = re^{i\varphi}$ of modulus $r \ge 1$ and a corresponding non-zero complex eigenvector $v \in \mathbb{C}^d$. Then $x_0 := v + \overline{v} \in \mathbb{R}^d$, and for $b = 0$ the sequence

$$x_n = A^n x_0 = r^n \cos(n\varphi) x_0$$

does not converge as $n \to \infty$.

Exercise 12.4. Given $\alpha > \rho(A)$ arbitrarily, choose an n such that $\sqrt[n]{\|A^n\|} \le \alpha$, and define

$$\|x\|' := \sum_{k=0}^{n-1} \frac{\|A^k x\|}{\alpha^k}.$$

This is a norm on X, and

$$\frac{\|Ax\|'}{\alpha} = \sum_{k=1}^{n} \frac{\|A^k x\|}{\alpha^k} \le \left(\sum_{k=1}^{n-1} \frac{\|A^k x\|}{\alpha^k} \right) + \|x\| = \|x\|'$$

for all $x \in X$.

Exercise 12.5. Show by induction on n that

$$|(A^n x)(t)| \le \frac{M^n t^n}{n!} \|x\|_\infty$$

for all $t \in [0, 1]$ and $n = 1, 2, \dots$, where M denotes the maximum of $|a(t, s)|$ on T.

Exercise 12.6. We prove the maximum case: the other case is obtained by considering $-u$ instead of u. If[9] $\max_{\overline{\Omega}}$ is not attained on Γ, then the modified function

$$v(x) := u(x) + \varepsilon \left(x_1^2 + \cdots + x_n^2 \right)$$

[9] Observe that $\overline{\Omega}$ and Γ are compact, so that $\max_{\overline{\Omega}} u$ and $\max_\Gamma u$ exist.

has the same property if $\varepsilon > 0$ is chosen sufficiently small. Indeed, denoting by M the maximum of $x_1^2 + \cdots + x_n^2$ on $\overline{\Omega}$, it suffices to take $\varepsilon > 0$ satisfying

$$\max_{\Gamma} u + \varepsilon M < \max_{\overline{\Omega}} u.$$

It follows that $\max_{\overline{\Omega}} v$ is attained at some interior point $x \in \Omega$. Then $D_j^2 v(x) \le 0$ for all $j = 1, \ldots, n$, and hence $\Delta v(x) \le 0$. This, however contradicts the relation

$$\Delta v(x) = \Delta u(x) + 2n\varepsilon = 2n\varepsilon > 0.$$

Exercise 12.7. By linearity it suffices to consider again the maximum case. Assume on the contrary that u takes at some point of $\overline{Q}$ a larger value than $\max_{\Sigma} u$, and choose, similarly to the solution of the preceding exercise, a small $\varepsilon > 0$ such that the modified function

$$v(t,x) := u(t,x) + \varepsilon\bigl(x_1^2 + \cdots + x_n^2\bigr), \quad (t,x) \in \overline{Q}$$

has a maximum at some point

$$(t,x) \in \overline{Q} \setminus \Sigma = Q \cup (\{T\} \times \Omega).$$

If $(t,x) \in Q$, then $D_1 v(t,x) = 0$ and $D_j^2 v(t,x) \le 0$ for all $j = 2, \ldots, n+1$, and hence $D_1 v(t,x) - \Delta v(t,x) \ge 0$. If $(t,x) \in \{T\} \times \Omega$, then $D_1 v(t,x) \ge 0$ and $D_j^2 v(t,x) \le 0$ for all $j = 2, \ldots, n+1$, and hence $D_1 v(t,x) - \Delta v(t,x) \ge 0$ again.

This, however, is impossible, because

$$D_1 v(t,x) - \Delta v(t,x) = (D_1 u(t,x) - \Delta u(t,x)) - 2n\varepsilon < 0.$$

Comments and Historical References

The most cited mathematicians in this book are the following:

- Descartes, René, 1596–1650
- Newton, Isaac, 1642–1715
- Bernoulli, Jacob, 1654–1705
- Euler, Leonard, 1707–1783
- Lagrange, Joseph Louis, 1736–1813
- Gauss, Carl Friedrich, 1777–1855
- Bolzano, Bernard, 1781–1848
- Cauchy, Augustin-Louis, 1789–1857
- Weierstrass, 1815–1897
- Chebyshev, Pafnuty Lvovich, 1821–1894
- Schwarz, Hermann Amandus, 1843–1921
- Cantor, Georg, 1845–1918
- Dini, Ulisse, 1845–1918
- Peano, Giuseppe, 1858–1932
- Hausdorff, Felix, 1868–1942
- Banach, Stefan, 1892–1945

Metric Spaces

The distance $d(x, y) := |x_1 - y_1| + \cdots + |x_m - y_m|$ was first used in Jordan [263], p. 18.

Metric spaces were introduced by Fréchet [174] in his Ph.D. thesis, under the supervision of Hadamard.

Open sets: Cantor [76], p. 135; see also Lebesgue [319], p. 242.

Product sets: Grassmann [202], p. 5.

Closed sets: Cantor [79], p. 217.

V. Komornik, *Topology, Calculus and Approximation*, Springer
Undergraduate Mathematics Series, DOI 10.1007/978-1-4471-7316-8

Convergent sequences: Bolzano [53], Cauchy [87], Fréchet [174].

Uniform convergence: Weierstrass [507], Seidel [448], Stokes [469], pp. 279–281.

Subsequence: Cantor [74], p. 89.

Proposition 1.4: Cantor [79], p. 226.

Limit: d'Alembert [6], p. 542, Cauchy [87].

Cluster points and isolated points: Cantor [75].

Continuous functions: Bolzano [52, 53], Cauchy [87], Weierstrass [507].

Uniform continuity: Heine [228], p. 361. See the remark on Theorem 1.27 below.

Lipschitz continuity: Lipschitz [336].

Cauchy sequence: Bolzano [53]; Cauchy [87], pp. 115–116.

Completeness of $\mathbb{R}$: Cauchy [87], pp. 115–116.

Completeness in metric spaces: Fréchet [174].

Fixed point theorem 1.10: Banach [26], Cacciopoli [71].

Brouwer's fixed point theorem: Brouwer [67]; see, e.g., Lax [317], Milnor [349], Pontryagin [397], Rogers [425], Aigner–Ziegler [4] for proofs.

Example on the application of the fixed point theorem: W. Bolyai [51] I, 442, 447–449.

Cantor's intersection theorem (Propositions 1.12, 1.23): Cantor [79], p. 217.

Density: Cantor [76], p. 140.

Baire's lemma (Propositions 1.13): Osgood [376], pp. 163–164, Baire [24], p. 65, Kuratowski [294], Banach [28].

Continuous extension of functions: See the comments on Exercise 1.15 (p. 34) below.

Completion of metric spaces (Proposition 1.16): Hausdorff [225]. Generalizing a method of Cantor [75] and Méray [344], he considered as the elements of (X', d') certain equivalence classes of Cauchy sequences. The short proof given here, based on an idea of Fréchet [175], p. 161, is due to Kuratowski [296].

Proposition 1.18: Bolzano [53], Weierstrass [509].

Proposition 1.19: See Kürschák [300], p. 60. We do not know of an earlier publication of this result. See the *Rising Sun lemma* of Riesz [418, 419] (see also in Riesz–Sz.-Nagy [421], p. 6) for an important application of this notion to integration theory.

Compactness: Fréchet [174].

Theorem 1.24: the first version on continuous functions on bounded closed intervals was published by Weierstrass [508] and Cantor [73], p. 82. However, the result had already appeared in Bolzano's book [54], written between 1833 and 1841, but published only in 1930.

Theorem 1.26: Hausdorff [225].

Theorem 1.27: Heine [229], p. 188. It was discovered recently that the notion of uniform continuity and Heine's theorem was already known to Bolzano; see [55].

Precompact sets: Hausdorff [225].

Theorem 1.28: The characterization by finite open subcovers goes back to Heine [229], p. 188, Cousin [115], p. 22, Borel [58], p. 51, Lindelöf [333], Lebesgue [320],

p. 105. See also the historical account of Hildebrandt [241]. The equivalence with complete precompact sets is due to Hausdorff [225].

Exercises 1.9–1.10: Cantor [78], p. 575, Bendixson [34], p. 415, Lindelöf [334]. See, e.g., Alexandroff [11] (pp. 135–147), Kuratowski [295] (pp. 140–141), Sz.-Nagy [479] (pp. 45–50) for more details and further results.

Exercise 1.11: Alexandroff [9]. A simple proof was given by Hausdorff [227]. See also Oxtoby [377].

Exercise 1.12: Hausdorff [225].

Exercises 1.13–1.14: See Oxtoby [377] for a nice exposition of these and related results. The set of differentiability of a continuous function $f : \mathbb{R} \to \mathbb{R}$ was also characterized by Zahorski [518]. Thomae's function was defined in [483, p. 14].

Using his theorem, Baire also proved that if a sequence of continuous functions $f_n : \mathbb{R} \to \mathbb{R}$ converges *pointwise* to $f : \mathbb{R} \to \mathbb{R}$, then f is continuous on a dense set of $\mathbb{R}$. We recall[1] that if the convergence is uniform, then the limit function is (everywhere) continuous.

In fact, Baire proved sharper results: see Oxtoby [377] again.

Exercise 1.15: Tietze's theorem [484] was proved earlier in $\mathbb{R}^2$ by Lebesgue [321], pp. 99–100.

Hausdorff [226] gave an explicit formula for a Tietze type extension. Assuming without loss of generality that $g : F \to [0, 1]$, we may take

$$f(x) := \inf_{y \in F} \left(f(y) + \frac{|x-y|}{\operatorname{dist}(x, F)} - 1 \right), \quad x \in X \setminus F.$$

See also Kuratowski [295], pp. 212–213.

Urysohn [495] extended Tietze's theorem to normal topological spaces; see, e.g., Császár [118], Engelking [139], Kelley [273]. See also [285, Proposition 8.6] for a related result.

Topological Spaces

Topological space: Riesz [412], Hausdorff [225].

Homeomorphism: Poincaré [392], p. 9.

Interior point: Cantor [76], Riesz [412].

Closure: Riesz [412].

Boundary point: Cantor [76], p. 135.

Exterior point: Hausdorff [225]. *Neighborhood:* Cantor [75].

Continuity in topological spaces and Proposition 2.10: Hausdorff [225].

Counterexamples following Corollary 2.11: they are taken from Steen–Seebach [459], Examples 12, 14.

[1] See Proposition 2.13, p. 46.

The space $C_b(K)$: Fréchet [174].

Connected set: Riesz [412].

Theorem 2.15: Lagrange 1769, pp. 541–542, Bolzano [53], Hausdorff [225].

A simple proof of von Neumann's *minimax theorem* [265], a fundamental theorem of game theory, is based on the notion of connectedness and mathematical economy.

Cantor's intersection theorem (Proposition 2.20): Cantor [79], p. 217.

Theorem 2.21: Hausdorff [225].

Theorem 2.22: Weierstrass [508], Cantor [73] p. 82.

Proposition 2.23: Hausdorff [225].

Tychonoff's theorem 2.24: Tychonoff [488, 489], Čech [99].

Zorn's lemma 2.25: Kuratowski [293], Zorn [519].

Proposition 2.26: Alexander [7].

Nets: Vietoris [499], Moore and Smith [355], Picone [391].

Filter: Vietoris 1921, H. Cartan [84].

Counterexamples following Proposition 2.27: they are inspired by Steen–Seebach [459].

Exercise 2.8 (Alexandroff's one-point compactification): Alexandroff [10].

Exercise 2.10 (Peano curve): Peano [383]. The construction given here is due to Lebesgue [323], pp. 44–45. This is a modification of *Cantor's function* $f : [0,1] \to [0,1]$, a continuous function defined by the formula

$$f\Big(\frac{2t_1}{3} + \frac{2t_2}{3^2} + \cdots + \frac{2t_n}{3^n} + \cdots\Big) := \frac{t_1}{2} + \frac{t_3}{2^2} + \cdots + \frac{t_{2n-1}}{2^n} + \cdots$$

on Cantor's ternary set C, and constant on the connected components of $[0,1] \setminus C$. See [242] for a survey of various other interesting features of Cantor's function.

Other interesting constructions of Peano curves were given by Hilbert [237] (see also Hilbert–Cohn-Vossen [240]) and Schoenberg [438].

Normed Spaces

Norm: Riesz [416].

Proposition 3.1 (Cauchy–Schwarz inequality): Lagrange [305], pp. 662–663 in $\mathbb{R}^3$, Cauchy [87], pp. 375–377 in $\mathbb{R}^n$, Bouniakowsky [59], p. 4 and Schwarz [442], p. 251 for integrals of functions of one or several variables.

Optimality of Proposition 3.1: Jordan–von Neumann [265].

The norms $\|x\|_1$, $\|x\|_2$, $\|x\|_\infty$ of $\mathbb{R}^m$ were introduced respectively by Jordan [263], p. 18, Cantor [78], p. 197 and Peano [380], p. 450, [382], p. 186.

Proposition 3.2: Young [516], Rogers [424] and Hölder [248], Minkowski [350], pp. 115–117.

Proposition 3.3: Riesz [415, 417].

Theorem 3.9: Tychonoff [489].

Proposition 3.11: Kirchberger [276], Borel [58], p. 82, Riesz [416].

Proposition 3.12: Legendre [326], Gauss [186]. See the historical account of Goldstine [197].

Lemma 3.13: Banach [26].

Dual space: Hahn [215].

Theorem 3.20: Helly [230], Hahn [215], Banach [27]. Helly's name is seldom mentioned in this theorem, although the crucial Lemma 3.21 below is due to him. See Hochstadt [243].

Uniqueness of norm-preserving extensions: Taylor [480], Foguel [171].

Lemma 3.21: Helly [230].

Norm in a complex vector space: Wiener [512].

Hilbert space: Hilbert [238], von Neumann [360], Löwig [338], Rellich [407].

The complexification formula $\varphi(x) := \psi(x) - i\psi(ix)$ was discovered by Murray [357], Bohnenblust–Sobczyk [49] and Soukhomlinov [457].

Exercise 3.1: observation of Fischer; see Riesz [414], in [411] I, p. 404.

Exercise 3.5: Dini [133].

Exercise 3.8: Carathéodory [82].

Exercise 3.9: Radon [402]. See Bajmóczy–Bárány [25] for a generalization.

Exercise 3.10: Helly [231].

Exercise 3.11: Jordan–von Neumann [265]. We follow Yosida [514], p. 39.

Exercise 3.12: Joó [261]; see also Stachó [458].

Exercise 3.13: von Neumann [361]; this is the founding theorem of game theory. See also von Neumann–Morgenstern [363] for many applications, including economy. See also [496] for an elementary introduction.

The present proof is due to Joó [261]; see also Stachó [458] and Komornik [280] for some generalizations.

The Derivative

Lemma 4.1 (a): Weierstrass [508], Stolz [470], Fréchet [176].

Lemma 4.1 (b): Hadamard [213], Carathéodory [83].

Directional derivative: Gâteaux [185]. The last example on p. 104 is due to Peano; see Genocchi and Peano [191], Section 123.

Proposition 4.3: Oresme [375], Kepler [274], Fermat [167], Newton [367].

Proposition 4.4 (a): Rolle [426]. He considered only polynomials. Cauchy [88], pp. 45–46 proved the theorem for functions of class C^1, including the endpoints. The present proof is due to Bonnet, published by Serret [449] I, pp. 17–19.

Proposition 4.4 (b): Lagrange [310], p. 154.

Proposition 4.4 (c): Cauchy [87].

Corollary 4.5: Johann Bernoulli [39], L'Hospital [245]. See Truesdell [492] for the interesting story of its discovery.

Proposition 4.6: Lagrange [310], p. 154.

The example preceding Proposition 4.9 is taken from Gelbaum–Olmsted [189].

Exercise 4.2: The example is due to Peano; see Genocchi and Peano [191], Section 123.

Higher-Order Derivatives

Definition of higher-order derivatives: Newton [367], Leibniz [327], Johann Bernoulli [40], Fréchet [178].

Remarks preceding Theorem 5.6: Euler [147, p. 177], [155, §226], Schwarz [441], Peano in Genocchi–Peano [191, §123].

Theorem 5.6: Young [515], p. 22.

Early versions of Taylor's formula: Gregory [206], Johann Bernoulli [41], Taylor [481].

Theorem 5.8: first stated in Genocchi and Peano [191]; the first proofs were published by J. König [290], pp. 532–538 and Peano [381].

Proposition 5.9: Lagrange [310], p. 154. The proof given here is due to Cauchy [88], p. 152. Previously Ampére [12] applied Rolle's theorem only once as follows.[2]

For any fixed number A the function

$$g(x) := \Big(\sum_{k=0}^{n-1} \frac{f^{(k)}(x)}{k!}(b-x)^k\Big) + A(b-x)^n$$

is continuous on $[a, b]$, differentiable on (a, b), and

$$g'(x) = \Big(\frac{f^{(n)}(x)}{(n-1)!} - nA\Big)(b-x)^{n-1}.$$

Choose A such that $g(a) = g(b)$, i.e., set

$$A = (b-a)^{-n}\Big(f(b) - \sum_{k=0}^{n-1} \frac{f^{(k)}(a)}{k!}(b-a)^k\Big).$$

Then Rolle's theorem yields $c \in (a, b)$ such that $g'(c) = 0$, and then

$$\frac{f^{(n)}(c)}{(n-1)!} - nA = 0.$$

[2] See Valiron [497]. This proof works under the weaker assumptions that f is continuous on $[a, b]$, n times differentiable on (a, b), and $n-1$ times continuously differentiable at a and b.

Taking the value of A into account, the proposition follows.

Theorem 5.11: Johann Bernoulli [41], Cauchy [87], Graves [204].

Theorem 5.12: Lagrange [302], Hesse [234], p. 251.

The example at the end of Sect. 5.4 is due to Peano in Genocchi and Peano [191], Section 123.

Definition of convex functions: Hölder [248] (for C^2 functions), Jensen [260].

Proposition 5.13: Hölder [248], Jensen [260], p. 180 in the French translation.

Proposition 5.15: Jensen [260], Minty [351].

Proposition 5.17: Hölder [248].

Proposition 5.18: Stolz [471], Jensen [260], pp. 189–190 in the French translation, Blumberg [46]. The present proof follows Roberts–Varberg [422].

Proposition 5.21: Hesse [234], p. 251.

Ordinary Differential Equations

Section 6.1: Riemann introduced his integral in [408]. The simpler integral of this section was studied for scalar functions in Dieudonné [130].

Section 6.2: Peano's example was published in Peano [382].

Theorem 6.2: Cauchy [89], Lipschitz [336], Elconin–Michal [138]. Cauchy could not publish this result; see, e.g., Hairer–Wanner [217] for possible reasons. It was rediscovered and published only in 1981 (!): see Cauchy [89], p. XIX. See also Choquet [103, p. 22] on a weaker geometric condition ensuring the uniqueness of the solutions.

Remarks following Theorem 6.2: Peano [379, 382]. See also Coddington–Levinson [112] and Walter [504] for proofs of Peano's theorem.

Generalizing an example of Dieudonné [129], Godunov [196] proved that Peano's theorem fails in *every* infinite-dimensional Banach space.

Piecewise linear approximation: Euler [156], p. 424. See, e.g., Coddington–Levinson [112]. We will also use Euler's method in Chap. 12.

Successive approximation: Cauchy [89] (?, part of his notes is still missing), Liouville [335], p. 19, Peano [380], Picard [389], Bendixson [35], Lindelöf [332].

Equivalent integral equation: Laplace [315], p. 236.

Proof of Theorem 6.2: The idea to use the equivalent norm $\|y\|_L := \sup_{t\in J} \|y(t)\|\, e^{-L|t-\tau|}$ is due to Bielecki [45].

Proposition 6.4: Volterra [500, 501].

Remark following Proposition 6.6: The result mentioned on infinite-dimensional spaces was proved in Dieudonné [129] and Deimling [123] in special cases, and in Komornik–Martinez–Pierre–Vancostenoble [286] for all infinite-dimensional Banach spaces.

Proposition 6.7: Niccoletti [374].

Lemma 6.8: Peano [379], Gronwall [212].

Section 6.5, variation of constants: Lagrange [306, 307].

Exercise 6.4: The second result remains valid for all simply connected domain D instead of rectangles: see, e.g., Walter [505], p. 39.

Exercise 6.8: This exercise is taken from Walter [505], pp. 79–80. Concerning the convergence of the successive approximations, see also Rosenblatt [428] and Hartman [224].

Exercise 6.10: This formula is very useful in the theory of non selfadjoint differential operators; see, e.g., Il'in–Joó [251], Joó–Komornik [262], and [281, 284].

Exercise 6.11: Haraux [222], Lagnese [301], Komornik [282], p. 103. See also the nonlinear generalization of this result in [282], p. 124.

Implicit Functions and Their Applications

Proposition 7.2: Descartes [125], Dini [134], pp. 153–164.

Folium of Descartes: Descartes [126]. The shape of the curve was determined by Roberval [423] for $x, y > 0$ and by Huygens [249] in the general case.

Proposition 7.3: Euler [151], Lagrange [308], pp. 78–79.

Lemma 7.5 and Theorem 7.4: Cauchy [90]. We present his original proof.

Proof of Theorem 7.7: The use of the fixed point theorem in this proof is due to Goursat [199]. The alternative proof in finite dimensions, indicated in the last remark of Sect. 7.4, is due to Kowalewski [288].

Theorem 7.8: Dini [134], pp. 153–164, Graves–Hildebrandt [205]. Originally Dini proceeded by induction on the dimension; see also Fichtenholz [170].

Theorem 7.9: Lagrange [308].

Remarks following Theorem 7.8: Lyusternik [339] (see also Alekseev–Tikhomirov–Fomin [5]), Kuhn–Tucker [292] (see also [285], Theorem 1.7 and Corollary 1.8).

Proposition 7.10: Bendixson [36], Picard [390], Peano [384].

Exercise 7.1 (Cayley–Hamilton theorem): Frobenius [182]. We follow Bernhardt [37].

Exercise 7.5.

(i) Euclid [145] II. 5, V. 25 for $n = 2$, Cauchy [87], pp. 373–374 (in his complete works). There is also an elementary proof as follows.[3] The case $n = 1$ is obvious. If the inequality holds for some n, and $a_1 a_2 \cdots a_{n+1} = 1$, then we may choose two elements, say a_1 and a_2, such that $a_1 \le 1$ and $a_2 \ge 1$. Then $(a_1 - 1)(a_2 - 1) \le 0$, so that $a_1 + a_2 \ge 1 + a_1 a_2$. Since

$$a_1 a_2 + a_3 + \cdots + a_{n+1} \ge n$$

[3] See, e.g., Beckenbach–Bellman [32], Korovkin [291].

by the induction hypothesis, we infer that

$$a_1 + a_2 + a_3 + \cdots + a_{n+1} \ge 1 + a_1a_2 + a_3 + \cdots + a_{n+1} \ge 1 + n.$$

The case $a_1a_2 \cdots a_{n+1} \ne 1$ follows by homogeneity.

(ii) Young [516]. The inequality follows at once from the concavity of the logarithmic function: we have

$$\ln\left(\frac{1}{p}a^p + \frac{1}{q}b^q\right) \ge \frac{1}{p}\ln a^p + \frac{1}{q}\ln b^q,$$

and we conclude by taking exponentials.

There is also an elementary proof. We may assume by continuity that $p = n/m$ and $q = n/(n-m)$ with suitable positive integers $m < n$. Applying (i) with

$$a_1 = \cdots = a_m = a^p \quad \text{and} \quad a_{m+1} = \cdots = a_n = b^q$$

we obtain the equivalent inequality

$$\sqrt[n]{(a^p)^m(b^q)^{n-m}} \le \frac{ma^p + (n-m)b^q}{n}.$$

There is also a transparent geometric proof; see, e.g., [285, Proposition 2.14].

(iii) Rogers [424], Hölder [248], Cauchy [87], pp. 457–459 (see also [93] (2) III, pp. 375–377). Let us recall an elementary proof. We may assume by homogeneity that $a_1^p + a_2^p + \cdots + a_n^p = b_1^q + b_2^q + \cdots + b_n^q = 1$. Applying (ii) for each product a_ib_i we obtain

$$\begin{aligned} a_1b_1 + a_2b_2 + \cdots + a_nb_n &\le \frac{a_1^p + a_2^p + \cdots + a_n^p}{p} + \frac{b_1^q + b_2^q + \cdots + b_n^q}{q} \\ &= \frac{1}{p} + \frac{1}{q} = 1, \end{aligned}$$

which is equivalent to our assertion.

Exercise 7.6: Duhamel–Reynaud [136], p. 155, Schlömilch [436].

Exercise 7.7: Heron's formula was probably known by Archimedes two centuries before.

Interpolation

Proposition 8.1: Gregory [208], Newton [369], Lagrange [309], pp. 284–287.

Theorem 8.2: Cauchy [92].

Theorem 8.3: Chebyshev [95], p. 301. He also proved that equality holds only in the above mentioned case: see, e.g., Achiezer [3], Natanson [359] or Cheney [100].

Proposition 8.5: Gregory [208], Newton [369].

Proposition 8.6: Hermite [233].

Theorem 8.7: Stieltjes [462], pp. 142–144.

Simpson's formula: Cavalieri [94], p. 446, Gregory [207], Simpson [452], pp. 109–111.

Theorem 8.9: Weierstrass [510], p. 5. Among the many proofs, those of Lebesgue [318], Landau [314] and Bernstein [43] are reproduced in [285], Exercises 8.1–8.3, Theorem 8.1 and Proposition 8.16, respectively. See also Borel [58].

Theorem 8.10: Faber [159]. His proof was simplified by Fejér [162]. See, e.g., Natanson [359], or [285].

Erdős–Vértesi theorem: Erdős–Vértesi [144].

Theorem 8.11: Fejér [161].

Definition of splines: Schoenberg [439].

Proposition 8.12: Schoenberg–Whitney [440], p. 258.

Proposition 8.14: Holladay [244].

Exercise 8.2. This is an application of Newton's interpolation formula. We follow Erdős [141].

Exercise 8.4: Cauchy [92].

Exercise 8.5: Archimedes [13] computed the volume of a spherical segment. In the same work he attributes the discovery of the volume of cones to Eudoxos. However, the formula (8.35) for truncated pyramids appears already in the Moscow papyrus [356] around 1850 BC; see van der Waerden [502].

See Gould [198] for the method of discovery of Archimedes, and Heath in [14] for more on Archimedes' life and work.

Exercise 8.6: Painlevé [378].

Orthogonal Polynomials

Proposition 9.1: Gram [200], Schmidt 1907.

Proposition 9.2: Stieltjes [463], Schmidt [437].

Legendre polynomials: Legendre [325].

Chebyshev polynomials: Chebyshev [95], p. 301, [96], p. 418.

Jacobi polynomials: Jacobi [257].

Laguerre polynomials: Lagrange [303], pp. 534–539, Abel [1], p. 284, Chebyshev [97], Laguerre [313], pp. 78–81, Sonine [456], p. 41.

Hermite polynomials: Sturm [474], pp. 424–426, Chebyshev [97], Hermite [232].

Proposition 9.3: Stieltjes [463], Schmidt [437].

Proposition 9.4: Stieltjes [463].

Exercise 9.9: Stieltjes [464–466].

Numerical Integration

Proposition 10.2: Korkin–Zolotarev [287], Stieltjes [461].

Newton–Cotes rules: Newton [369], Cotes [113].

Méray's example: Méray [345, 346].

Runge's example: Runge [434]. See Montel [354] or Steffensen [460] for a complete analysis. The Newton–Cotes formulas are unstable for $n = 8$ and for all $n \geq 10$ because some of the coefficients A_k^n become negative: see Bernstein [44].

Theorem 10.3: Gauss [187], Jacobi [255], Christoffel [104]. The last estimate is due to Markov [342].

Theorem 10.4: Stieltjes [463].

Theorem 10.5: Erdős–Turán [143]. The special cases of Legendre polynomials and of the Chebyshev polynomials of the first kind were studied earlier by Fejér [164] and Erdős–Feldheim [142], using a different method.

Proposition 10.7: De Moivre [353] without the exact constant $\sqrt{2\pi}$, Stirling [468]. See Pearson [386], Hald [219].

Bernoulli polynomials: Jacobi [256].

Bernoulli numbers: Jacob Bernoulli [38]; see p. 99 of the German translation.

Theorem 10.11: Euler [146, 148, 155], p. 310, Maclaurin [340], Book II, Chap. IV, p. 663. It is also called the Euler–Maclaurin formula. The remainder term first appeared in the work of Poisson [394]. The first rigorous proof was given by Jacobi [256]. The proof given here is due to Wirtinger [513] and Jordan [264].

Example to compute $1^p + \cdots + n^p$: Jacob Bernoulli [38].

Theorem 10.16: De Moivre [353], Stirling [468].

Romberg's method: Romberg [427].

Exercise 10.3. We follow the original reasoning of Wallis.

Finding Roots

Theorem 11.1: Descartes [125]. The first proofs were published by Segner [446, 447]; his first paper was believed to have been lost for more than two centuries: see Szénássy [477]. The proof given here is taken from Komornik [283].

Theorem 11.4: Sturm [473]. See its introduction in Fourier's contribution.

Proposition 11.7: Householder–Bauer [247].

Proposition 11.9: Givens [193, 194].

Proposition 11.10: Gerschgorin [192].

Proposition 11.11: Newton [368].

Exercise 11.2: We follow Szidarovszky [478].

Exercise 11.3: Kantorovich [270, 271]. We follow Kantorovich–Akilov [272].

Exercises 11.4–11.5: We follow Varga [498].

Exercise 12.3: (ii) is closely related to a classical theorem of Kakeya [267, 268] on subadditive sequences; see also [396], Part 1, Exercise 131.

Exercise 12.4: We follow Walter [505], pp. 352–353, where the following corollary is also proved:

$$\rho(A+B) \le \rho(A) + \rho(B) \quad \text{and} \quad \rho(AB) \le \rho(A)\rho(B)$$

whenever $AB = BA$.

The results and the proofs remain valid in infinite-dimensional normed spaces if we define the spectral radius as in Exercise 12.3, and the infimum is taken over all norms on X that are *equivalent* to $\|\cdot\|$.

Numerical Solution of Differential Equations

Euler's method: Euler [156], p. 424.

Proposition 12.1: Peano [379].

Modified Euler method: Runge [433], p. 168.

Runge–Kutta method: Runge [433], Heun [235], Kutta [299].

Monte-Carlo method: Ulam–von Neumann, see [347, 494], pp. 196–200, and also [454].

Random walk: Pearson [386], Pólya [395], see also Alexanderson [8].

Heat equation: Fourier [172], and also Bochner [48] and Kahane [266] for historical comments. A slightly more general model leads to the *Fokker–Planck equation*: see, e.g., Gnedenko [195], pp. 288–290.

Exercise 12.6: The proof given here is due to Privalov [400].

Bibliography

1. N.H. Abel, *Sur une espèce particulière de fonctions entières nées du développement de la fonction* $\frac{1}{1-v}e^{-\frac{xv}{1-v}}$ *suivant les puissances de* v, *Mémoires de mathématiques par N.H. Abel*, Paris, 1826; [2] II, 248.
2. *Oeuvres complètes de Niels Henrik Abel I-II*, Christiania, Imprimerie de Grondahl & Son, 1881.
3. N.I. Achieser, *Theory of Approximation*, Dover, New York, 1992.
4. M. Aigner, G.M. Ziegler, *Proofs from the Book*, Springer, New York, 4th ed. 2010, VIII, 274 p. 250 illus.
5. V.M. Alekseev, V.M. Tikhomirov, S.V. Fomin, *Optimal Control*, Contemporary Soviet Mathematics, Springer, New York, 1987.
6. J. le Rond d'Alembert, *Limite*, Encyclopédie IX, Neufchastel, 1765.
7. J.W. Alexander, *Ordered sets, complexes and the problem of compactifications*, Proc. Nat. Acad. Sci. U.S.A. 25 (1939), 296–298.
8. G.L. Alexanderson, *The Random Walks of George Pólya*, The Math. Association of America, 2000.
9. P. Alexandroff, *Sur les ensembles de la première classe et les ensembles abstraits*, C.R. Acad. Paris 178 (1924), 185–187.
10. P. Alexandroff, *Über die Metrisation der im Kleinen kompakten topologischen Räume*, Math. Ann. 92 (1924), 3–4, 294–301.
11. P.S. Alexandroff, *Introduction to Set Theory and General Topology [in Russian]*, Nauka, Moscow, 1977
12. A.-M. Ampère, *Recherches sur quelques points de la théorie des fonctions dérivées qui conduisent à une nouvelle démonstration de la série de Taylor, et à l'expression finie des termes qu'on néglige lorsqu'on arrête cette série à un terme quelconque*, J. de l'École Polytechnique, treizième cahier, tome VI, avril 1806, 148–181.
13. Archimedes, *On the sphere and cylinder I-II*; see [14], 1–90.
14. Archimedes, *The Works of Archimedes*, (Ed. T. Heath), Dover, 2002.
15. V.I. Arnold, *Mathematical Methods of Classical Mechanics*, Springer, New York, 1978.
16. V.I. Arnold, *Ordinary Differential Equations*, The MIT Press, New York, 1978.
17. V.I. Arnold, *Geometrical Methods in the Theory of Ordinary Differential Equations*, Springer, New York, 1988.
18. C. Arzelà, *Funzioni di linee*, Atti della R. Accad. dei Lincei Rend. Cl. Sci. Fis. Mat. Nat. (4) 5I (1889), 342–348.
19. C. Arzelà, *Sulle funzioni di linee*, Mem. Accad. Sci. Ist. Bologna Cl. Sci. Fis. Mat. (5) 5 (1895), 55–74.

V. Komornik, *Topology, Calculus and Approximation*, Springer
Undergraduate Mathematics Series, DOI 10.1007/978-1-4471-7316-8

20. G. Ascoli, *Le curve limiti di una varietà data di curve*, Mem. Accad. dei Lincei (3) 18 (1883), 521–586.
21. C.E. Aull, R. Lowen (ed.), *Handbook of the History of General Topology I*, Kluwer, Dordrecht, 1997.
22. V. Avanissian, *Initiation à l'analyse fonctionnelle*, Presses Universitaires de France, Paris, 1996.
23. R. Baire, *Sur les fonctions discontinues qui se rattachent aux fonctions continues*, C.R. Acad. Sci. Paris 126 (1898), 1621–1623.
24. R. Baire, *Sur les fonctions à variables réelles*, Ann. di Mat. (3) 3 (1899), 1–123.
25. E.G. Bajmóczy, I. Bárány, *On a common generalization of Borsuk's and Radon's theorem*, Acta Math. Acad. Sci. Hungar. (3–4) 34 (1979), 347–350.
26. S. Banach, *Sur les opérations dans les ensembles abstraits et leurs applications aux équations intégrales*, Fund. Math. 3 (1922), 133–181; [30] II, 305–348.
27. S. Banach, *Sur les fonctionnelles linéaires*, Studia Math. 1 (1929), 211–216, 223–239; [30] II, 375–395.
28. S. Banach, *Théorèmes sur les ensembles de première catégorie*, Fund. Math. 16 (1930), 395–398; [30] I, 204–206.
29. S. Banach, *Théorie des opérations linéaires*, Monografie Matematyczne 1, Warszawa, 1932; [30] II, 13–219. English translation: *Theory of Linear Operations*, Dover, New York, 1987.
30. S. Banach, *Oeuvres I-II*, Pánstwowe Wydawnictvo Naukowe, Warszawa, 1967–1979.
31. M.E. Baron, *The Origins of the Infinitesimal Calculus*, Pergamon Press, Elmsford, New York, 1969.
32. E.F. Beckenbach, R. Bellman, *Inequalities*, Springer Verlag, Berlin–Göttingen–Heidelberg, 1961.
33. M. Beke, *Bevezetés a differenciál- és integrálszámításba [Introduction to Differential and Integral Calculus]*, Gondolat, Budapest, 1967.
34. I. Bendixson, *Quelques théorèmes de la théorie des ensembles*, Acta Math. 2 (1883), 415–429.
35. I. Bendixson, *Sur le calcul des intégrales d'un système d'équations différentielles par des approximations successives*, Stock. Akad. Forh. 51 (1893), 599–612.
36. I. Bendixson, *Démonstration de l'existence de l'intégrale d'une équation aux dérivées partielles linéaires*, Bull. Soc. Math. France 24 (1896), 220–225.
37. C. Bernhardt, *A proof of the Cayley–Hamilton theorem*, Amer. Math. Monthly 116 (2009), 5, 456–457.
38. Jacob Bernoulli, *Ars conjectandi*, Basel, 1713. German translation: *Warscheinlichkeitsrechnung*, Ostwald's Klassiker Nr. 107–108, Engelmann, Leipzig, 1899.
39. Johann Bernoulli, *Lectiones mathematicae de methodo integralium, aliisque conscriptae in usum Ill. Marchionis Hospitalii cum auctor Parisiis ageret annis 1691 et 1692*; [42] III, 385–558. Partial English translation: *Mathematical Lectures on the Method of Integrals, and on Other Subjects Written for the Use of the Marquis de l'Hôpital, as the Author Gave Them in Paris during 1691 and 1692*, [472], 324–328.
40. Johann Bernoulli, *Die Differentialrechnung von Johann Bernoulli*, 1691/92, Nach der in der Basler Universitätsbibliothek befindlichen Handschrift übersetzt von Paul Schafheitlin, Akademische Verlagsgesellschaft Leipzig, 1924.
41. Johann Bernoulli, *Effectionis omnium quadraturam & rectificationum curvarum per seriem quandam generalissimam*, Acta Erud. Lips. 13 (1694), "Additamentum" to *Modus generalis construendi aequationes differentiales primi gradus*, Acta Erud. Lips. 13 (1694), 435. o.; [42] I, 123–125.
42. Johann Bernoulli, *Opera Omnia I-IV*, Lausannae & Genavae, 1742.
43. S.N. Bernstein, *Démonstration du théorème de Weierstrass fondée sur le calcul de probabilités*, Comm. Kharkov Math. Soc. 13 (1912), 1–2.
44. S.N. Bernstein, *Sur les formules de quadrature de Cotes et de Tchebycheff*, C.R. de l'Académie des Sciences de l'URSS 14 (1937), 323–326.

45. A. Bielecki, *Une remarque sur la méthode de Banach-Caccioppoli- Tikhonov*, Bull. Acad. Polon. Sci. 4 (1956), 261–268.
46. H. Blumberg, *On convex functions*, Trans. Amer. Math. Soc. 20 (1919), 40–44.
47. S. Bochner, *Integration von Funktionen, deren Werte die Elemente eines Vektorräumes sind*, Fund. Math. 20 (1933), 262–276.
48. S. Bochner, *Fourier series came first*, Amer. Math. Monthly 86 (1979), no. 3, 197–199.
49. H.F. Bohnenblust, A. Sobczyk 1938, *Extensions of functionals on complex linear spaces*, Bull. Amer. Math. Soc. 44 (1938), 91–93.
50. I. Bolyai, *Appendix prima scientia spatii, a veritate aut falsitate axiomatis XI-mi Euclidei (a priori haud unquam decidenda) independens: atque ad casum falsitatis qudratura circuli geometrica]*, Marosvásárhely, April 1831. English translation: *The science absolute of space: independent of the truth or falsity of Euclid's axiom XI (which can never be decided a priori)*, The Neomon, Austin, 1896.
51. W. Bolyai, *Tentamen juventutem studiosam in elementa matheseos purae, elementaris ac sublimioris methodo intuitiva, evidentiaque huic propria, introducendi*, Marosvásárhely, 1832–33.
52. B. Bolzano, *De binomische Lehrsatz und als Folgerung aus ihm der polynomische und die Reihen, die zur Berechnung der Logarithmen und Exponentialgrössen dienen, genauer als bisher erwiesen*, Prag, Enders, 1816. English translation: *The binomial theorem, and as a consequence from it the polynomial theorem, and the series, which serve for the calculation of loqarithmic and exponential quantities, proved more strictly than before*, [55], 155–248.
53. B. Bolzano, *Rein analytischer Beweis des Lehrsatzes, dass zwischen je zwei Werthen, die ein entgegengesetztes Resultat gewähren, wenigstens eine reelle Wurzel der Gleichung liege*, Prag, Haase, 1817. New edition: Ostwald's Klassiker der exakten Wissenschaften, No. 153, Leipzig, 1905. English translation: *Purely analytic proof of the theorem the between any two values which give results of opposite sign, there lies at least one real root of the equation*, Historia Math. 7 (1980), 156–185 and [55], 251–277.
54. B. Bolzano, *Functionenlehre*, written between 1833 and 1841, but published only in 1930; see [259]. English translation: *Theory of Functions* and *Improvements and Additions to the Theory of Functions*, [55], 429–572, 573–589.
55. B. Bolzano, *The Mathematical Works of Bernard Bolzano*, edited and translated by S. Russ, Oxford Univ. Press, 2004.
56. R. Bonola, *Non-Euclidean Geometry*, Chicago, 1912. Dover reprint 1955.
57. E. Borel, *Sur quelques points de la théorie des fonctions*, Ann. École Norm. Sup. (3) 12 (1895), 9–55.
58. E. Borel, *Leçons sur les fonctions de variables réelles et les développements en séries de polynômes*, Gauthier-Villars, Paris, 1905.
59. V. Bouniakowsky, *Sur quelques inégalités concernant les intégrales ordinaires et les intégrales aux différences finies*, Mémoires de l'Acad. de St-Pétersburg (vii) 1 (1859), No. 9, 1–18.
60. N. Bourbaki, *Elements of the History of Mathematics*, Springer, Berlin, 1994.
61. C.B. Boyer, *The History of the Calculus and its Conceptual Development*, Dover, New York, 1959.
62. C.B. Boyer, *A History of Mathematics*, John Wiley & Sons, New York, 1968.
63. D. Bressoud, *A Radical Approach to Real Analysis*, The Mathematical Association of America, Washington, 1994.
64. H. Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Springer, New York, Universitext, 2010.
65. H. Brezis, F. Browder, *Partial differential equations in the 20th century*, Advances in Math. 135, (1998), 76–144.
66. H. Briggs, *Arithmetica Logarithmica ...*, Londres, 1624.
67. L.E.J. Brouwer, *Über Abbildung von Mannigfaltigkeiten*, Math. Ann. 71 (1912), 97–115.
68. J. Bürgi, *Aritmetische und geometrische Progress Tabulen ...*, Prague, 1620.
69. J.C. Burkill, *The Lebesgue Integral*, Cambridge Univ. Press, 1951.

70. J.C. Burkill, *The Theory of Ordinary Differential Equations*, Oliver & Boyd, Edinburgh, 1956.
71. R. Caccioppoli, *Un teorema generale sull'esistenza di elementi uniti in una trasformazione funzionale*, Rend. Acc. Naz. Lincei 11 (1930), 794–799.
72. F. Cajori, *A History of Mathematics*, AMS-Chelsea, New York, 1991.
73. G. Cantor, *Beweis, dass für jeden Werth von x durch eine trigonometrische Riehe gegebene Function $f(x)$ sich nur eine einzige Weise in dieser Form darstellen lässt*, J. reine angew. Math. 72 (1870), 139–142; [80], 80–83.
74. G. Cantor, *Ueber trigonometrische Reihen*, Math. Ann. 4 (1871), 139–143; [80], 87–91.
75. G. Cantor, *Über die Ausdehnung eines Satzes aus der Theorie des trigonometrischen Reihen*, Math. Ann. 5 (1872), 123–132; [80], 92–101.
76. G. Cantor, *Über einen Satz aus der Theorie des stetigen Mannigfaltigkeiten*, Göttinger Nachr. 1879, 127–135; [80], 134–138.
77. G. Cantor, *Über unendliche lineare Punktmannigfaltigkeiten I*, Math. Ann. 15 (1879), 1–7; [80], 139–145.
78. G. Cantor, *Über unendliche lineare Punktmannigfaltigkeiten V*, Math. Ann. 21 (1883), 545–586; [80], 165–208.
79. G. Cantor, *Über unendliche lineare Punktmannigfaltigkeiten VI*, Math. Ann. 23 (1884), 453–488; [80], 210–244.
80. G. Cantor, *Gesammelte Abhandlungen*, Springer, Berlin, 1932.
81. M. Cantor, *Vorlesungen ueber Geschichte der Mathematik I-IV*, Teubner, Leipzig, 1880–1908.
82. C. Carathéodory, *Über den Variabilitätsbereich der Fourierschen Konstanten von positiven harmonischen Funktionen*, Rendiconti del Circolo Matematico di Palermo 32 (1911), 193–217.
83. C. Carathéodory, *Funktionentheorie*, Erster Band, Birkhäuser Verlag, Basel, 1950.
84. H. Cartan, *Théorie des filtres*, C.R. Acad. Sci. Paris 205 (1937), 595–598.
85. H. Cartan, *Filtres et ultrafiltres*, C.R. Acad. Sci. Paris 205 (1937), 777–779.
86. H. Cartan, *Differential calculus*, Hermann, Paris, 1971.
87. A.L. Cauchy, *Cours d'analyse algébrique*, Paris, 1821; [93] (2) III, 1–476.
88. A.L. Cauchy, *Résumé des leçons sur le calcul infinitésimal*, École Royale Polytechnique, Paris, 1823; [93] (2) IV, 1–261.
89. A.L. Cauchy, *Résumé des leçons données à l'École Royale Polytechnique. Suite du calcul infinitésimal*, 1824, published as *Équations différentielles ordinaires*, ed. Chr. Gilain, Johnson 1981.
90. A.L. Cauchy, *Leçons sur le calcul différentiel*, Paris, 1829; [93] (2) IV, 267–615.
91. A.L. Cauchy, *Sur l'équation à l'aide de laquelle on détermine les inégalités séculaires des mouvements des planètes*, Exercices de mathématiques (anciens exercices), année 1829; [93] (2) IX, 174–195.
92. A.L. Cauchy, *Sur les fonctions interpolaires*, C.R. Acad. Sci. Paris 11 (1840), 775–789; [93] (1) V, 409–424.
93. A.L. Cauchy, *Oeuvres*, 2 series, 22 volumes, Gauthier-Villars, Paris, 1882–1905.
94. B.F. Cavalieri *Centuria di varii problemi*, Bologna, 1639.
95. P.L. Chebyshev [Tchebychef], *Sur les questions de minima qui se rattachent à la représentation approximative des fonctions*, Mémoires de l'Acad. Imp. des Sciences de Saint-Pétersbourg (6) Sciences math. et phys. 7 (1859), 199–291; [98] I, 273–378.
96. P.L. Chebyshev [Tchebychef], *Sur l'interpolation dans le cas d'un grand nombre de données fournies par les observations*, Mémoires de l'Acad. Imp. des Sciences de Saint-Pétersbourg (7) 1 (1859), 5, 1–81; [98] I, 387–469.
97. P.L. Chebyshev [Tchebychef], *Sur le développement des fonctions à une seule variable*, Bulletin de l'Acad. Imp. des Sciences de Saint-Pétersbourg 1 (1859), 193–200; [98] I, 499–508.
98. P.L. Chebyshev [Tchebychef], *Oeuvres I-II*, Chelsea, New York, 1962.
99. E. Čech, *On bicompact spaces*, Ann. of Math. (2) 38 (1937), 823–844.
100. E.W. Cheney, *Introduction to Approximation Theory*, McGraw-Hill, New York, 1966.

101. P.R. Chernoff, *Pointwise convergence of Fourier series*, Amer. Math. Monthly 87 (1980), 399–400.
102. G. Choquet, *Cours d'analyse, tome II. Topologie*, Masson, Paris, 1964.
103. G. Choquet, *Équations différentielles*, Les cours de Sorbonne, Centre de Documentation Universitaire, 1963.
104. E.B. Christoffel, *Über die Gaussische Quadratur une eine Verallgemeinerung derselben*, J. reine angew. Math. 55 (1858), 61–82.
105. G. Christol, A. Cot, C.M. Marle, *Calcul différentiel*, Ellipses, Paris, 1997.
106. G. Christol, A. Cot, C.M. Marle, *Topologie*, Ellipses, Paris, 1997.
107. P.G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
108. P.G. Ciarlet, *Introduction to Numerical Linear Algebra and Optimisation*, Cambridge Univ. Press, 1989.
109. P.G. Ciarlet, *Linear and Nonlinear Functional Analysis with Applications*, SIAM, Philadelphia, 2013.
110. P.-G. Ciarlet, B. Miara, J.-M. Thomas, *Exercices d'analyse numérique matricielle et d'optimisation*, Masson, Paris, 1991.
111. E.A. Coddington, *An Introduction to Ordinary Differential Equations*, Dover, New York, 1989.
112. E.A. Coddington, N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.
113. R. Cotes, *De methodo differentiali Newtonia*, appendice de *Harmonia mensurarum, Cantabrigiae*, 1722. German translation: *Über die Newtonsche Differentialmethode*, [289], 12–25.
114. R. Courant, D. Hilbert, *Methods of Mathematical Physics I*, Wiley, New York, 1953.
115. P. Cousin, *Sur les fonctions de n variables complexes*, Acta Math. 19 (1895), 1–61.
116. H.S.M. Coxeter, *Introduction to Geometry*, John Wiley & Sons, New York, 1961.
117. H.S.M. Coxeter, S.L. Greitzer, *Geometry Revisited*, The Mathematical Association of America, Washington, 1967.
118. Á. Császár, *Foundations of General Topology*, International series of monographs on pure and applied mathematics v. 35, Pergamon Press, New York, 1963.
119. Á. Császár, *Valós analízis I-II [Real Analysis I-II]*, Tankönyvkiadó, Budapest, 1983.
120. H.T. Davis, *Introduction to Nonlinear Differential and Integral Equations*, Dover, New York, 1962.
121. P.J. Davis, *Interpolation and Approximation*, Dover, New York, 1975.
122. C. De Boor, *Best approximation properties of spline functions of odd degree*, J. Math. Mech. 12 (1963), 5, 747–749.
123. K. Deimling, *Ordinary Differential Equations in Banach Spaces*, Lecture Notes in Mathematics, Vol. 596, Springer, Berlin-New York, 1977.
124. B.P. Demidovich, *Problems in Mathematical Analysis*, Gordon & Breach, 1971.
125. R. Descartes, *La géométrie*, Leyden, 1637.
126. R. Descartes, *Letter to Mersenne on January 18, 1638*; see [128].
127. R. Descartes, *Unpublished manuscript*, 1639; see [132] and [277], Chapter L, Section 3.
128. R. Descartes, *Oeuvres complètes I-XI*, Vrin, Paris, 1996.
129. J. Dieudonné, *Deux exemples singuliers d'équations différentielles*, Acta Sci. Math. (Szeged) 12 (1950), 38–40.
130. J. Dieudonné, *Foundations of Modern Analysis*, Academic Press, 1960.
131. J. Dieudonné, *History of Functional Analysis*, North-Holland, Amsterdam, 1981.
132. J. Dieudonné, *Une brève histoire de la topologie*: Pier, Jean-Paul (ed.) *Development of mathematics 1900–1950 (Luxembourg, 1992)*, Birkhäuser, Basel, 1994, 35–155.
133. U. Dini, *Serie di Fourier e altre rappresentazioni analitiche delle funzioni di una variabile reale*, Nistri, Pisa, 1880.
134. U. Dini, *Analisi infinitesimale I-IV*, I. Calcolo differenziale, Autografia Bertini, Pisa, 1878.
135. J. Dugundji, *Topology*, Allyn and Bacon, Boston, 1970.

136. A.A.L. Reynaud, J.M.C. Duhamel, *Problèmes et développements sur diverses parties des mathématiques*, Paris, Bachelier, 1823.
137. C.H. Edwards, *The Historical Development of the Calculus*, Springer, New York, 1979.
138. V. Elconin, A.D. Michal, *Completely integrable differential equations in abstract spaces*, Acta Math. 68 (1937), 71–107.
139. R. Engelking, *General Topology*, Pánstwowe Wydawnictvo Naukowe, Warszawa, 1977.
140. A. Erdélyi, *Asymptotic Expansions*, Dover, New York, 1956.
141. P. Erdős, *A magasabb rendű számtani sorokról [On arithmetic sequnces of higher order]*, Középiskolai matematikai lapok, 1929. február, 187–189.
142. P. Erdős, E. Feldheim, *Sur le mode de convergence pour l'interpolation de Lagrange*, C.R. Acad. Sci. Paris Sér. I. Math. 203 (1936), 913–915.
143. P. Erdős, P. Turán, *On interpolation I. Quadrature- and mean-convergence in the Lagrange interpolation*, Ann. of Math. 38 (1937), 142–155.
144. P. Erdős, P. Vértesi, *On the almost everywhere divergence of Lagrange interpolatory polynomials for arbitrary system of nodes*, Acta Math. Acad. Sci. Hungar. 36 (1980), 71–89, 38 (1981), 263.
145. Euclid, *The Thirteen Books of Euclid's Elements I-III*, translation and commentaries by Heath, Thomas L. Dover Publications, New York, 1956
146. L. Euler, *Methodus generalis summandi progressiones*, Comm. Acad. Sci. Petrop. 6 (1732/3), published in 1738, 68–97; [157] (1) XIV, 42–72.
147. L. Euler, *De infinitis curvis eiusdem generis seu methodus inveniendi aequationes pro infinitis curvis eiusdem generis*, Comm. Acad. Sci. Petrop. 7 (1734/5), published in 1740, 174–189, 180–183; [157] (1) XXII, 36–56.
148. L. Euler, *Methodus universalis serierum convergentium summas quam proxime inveniendi*, Comm. Acad. Sci. Petrop. 8 (1736), published in 1741, 3–9; [157] (1) XIV, 101–107.
149. L. Euler, *Solutio problematis ad geometriam situs pertinentis*, Comm. Acad. Sci. Petrop. 8 (1736), published in 1741, 128–140; [157] (1) VII, 1–10. English translation: *The seven bridges of Königsberg*, [366] I, 573–580.
150. L. Euler, *Methodus universalis series summandi ulterius promota*, Comm. Acad. Sci. Petrop. 8 (1736), published in 1741, 147–158; [157] (1) XIV, 124–137.
151. L. Euler, *Methodus inviniendi lineas curvas maximi minimive proprietate gaudentes sive solutio problematis isoperimetrici latissimo sensu accepti*, Lausanne et Genève, 1744; [157] (1) XXIV, 1–308.
152. L. Euler, *Recherches sur la question des inégalités du mouvement de Saturne et de Jupiter*, Pièce qui a remporté le prix de l'académie royale des sciences 1748, published in 1749, 1–123; [157] (2) XXV, 45–157.
153. L. Euler, *Elementa doctrinae solidorum*, Novi Comm. Acad. Sci. Petrop. 4 (1752–53), published in 1758, 109–140; [157] (1) XXVI, 71–93.
154. L. Euler, *Demonstratio nonnullarum insignium proprietatum, quibus solida hedris planis inclusa sunt praedita*, Novi Comm. Acad. Sci. Petrop. 4 (1752–53), published in 1758, 140–160; [157] (1) XXVI, 94–108.
155. L. Euler, *Institutiones calculi differentialis cul eius vsu in analysi finitorum ac doctrina serierum*, Imp. Acad. Imper. Scient. Petrop., Petropoli, 1755; [157] (1) X, 1–676.
156. L. Euler, *Institutionum calculi integralis I*, Petropoli, 1768; [157] (1) XI, 1–462.
157. L. Euler, *Opera Omnia*, 4 series, 73 volumes, Teubner, Leipzig and Berlin, followed by Füssli, Zürich, 1911–.
158. L.C. Evans, *Partial Differential Equations*, Amer. Math. Soc., Providence, RI, 1999.
159. G. Faber, *Über die interpolatorische Darstellung stetiger Funtionen*, Jahresbericht der deutschen Mathematiker-Vereinigung 23 (1914), 192–210.
160. G.F. Fagnano, *Problemata quaedam ad methodum maximorum et minimorum spectantia*, Nova Acta Eruditorum (1775), 281–303.
161. L. Fejér, *Über Interpolation*, Götting. Nachr., 1916, 66–91; [165] II, 25–48.
162. L. Fejér, *Die Abschätzung eines Polynoms in einem Intervalle, wenn Schranken für seine Werte und ersten Ableitungswerte in einzelnen Punkten des Intervalles gegeben sind, und ihre*

Anwendung auf die Konvergenzfrage Hermitescher Interpolationsreihen, Math. Z. 32 (1930), 426–457; [165] II, 285–317.

163. L. Fejér, *Bestimmung derjenigen Abszissen eines Intervalles, für welche die Quadratsumme der Grundfunktionen der Lagrangeschen Interpolation im Intervalle ein möglichst kleines Maximum besitzt*, Ann. Scuola Norm. Sup. Pisa (2) 1 (1932), 263–276; [165] II, 432–447.
164. L. Fejér, *On the characterization of some remarkable systems of points of interpolation by means of conjugate points*, Amer. Math. Monthly 41 (1934), p. 1–14; [165] II, 527–539.
165. L. Fejér, *Gesammelte Arbeiten I-II*, Akadémiai Kiadó, Budapest, 1970.
166. P. Fermat, *Ad locos planos et solidos. Isagoge [Introduction aux lieus plans et solides]*, 1629, published in 1679; [168] I, 91–110 (latinul), III, 85–101 (in French). English translation: *Introduction to Plane and Solid Loci*, [453], 389–396.
167. P. Fermat, *Methodus ad disquirendam maximam et minimam [Méthode pour la recherche du maximum et du minimum]*, 1638; [168] I, 133–136 (latinul), III, 121–123 (in French). English translation: *On a method for the evaluation of maxima and minima*, [472], 222–224.
168. P. Fermat, *Oeuvres I-III*, Gauthier-Villars, Paris, 1891–1896.
169. R.P. Feynman, R.B. Leighton, M. Sands, *The Feynman Lectures on Physics*, Addison Wesley Longman, 1970.
170. G.M. Fichtenholz, *Differential- und Integralrechnung I-III*, Deutscher Verlag der Wissenschaften, Berlin, 1975–1987.
171. S.R. Foguel, *On a theorem by A.E. Taylor*, Proc. Amer. Math. Soc. 9 (1958), 325.
172. J.B.J. Fourier, *Théorie analytique de la chaleur*, Didot, Paris 1822; [173] I.
173. J.B.J. Fourier, *Oeuvres I-II*, Gauthier-Villars, Paris, 1888–1890.
174. M. Fréchet, *Sur quelques points du calcul fonctionnel*, Rend. Circ. Mat. Palermo 22 (1906), 1–74.
175. M. Fréchet, *Les dimensions d'un ensemble abstrait*, Math. Ann. 68 (1910), 145–168.
176. M. Fréchet, *Sur les fonctionnelles continues*, Ann. École Norm. Sup. 27 (1910), 193–216.
177. M. Fréchet, *Sur la notion de la différentielle totale*, Nouv. Ann. Math. (4) 12 (1912), 385–403, 433–449.
178. M. Fréchet, *Sur les fonctionnelles bilinéaires*, Trans. Amer. Math. Soc. 16 (1915), 215–234.
179. I. Fredholm, *Sur une nouvelle méthode pour la résolution du problème de Dirichlet*, Kong. Vetenskaps-Akademiens Förh. Stockholm (1900), 39–46; [181], 61–68.
180. I. Fredholm, *Sur une classe d'équations fonctionnelles*, Acta Math. 27 (1903), 365–390; [181], 81–106.
181. I. Fredholm, *Oeuvres complètes de Ivar Fredholm*, Litos Reprotryck, Malmo, 1955.
182. G. Frobenius, *Über lineare Substitutionen und bilineare Formen*, J. reine angew. Math. 84 (1878), 1–63; [183] I, 343–405.
183. G. Frobenius, *Gesammelte Abhandlungen I-III*, Springer, Berlin, 1968.
184. I.S. Gál, *On sequences of operations in complete vector spaces*, Amer. Math. Monthly 60 (1953), 527–538.
185. R. Gâteaux, *Sur les fonctionnelles continues et les fonctionnelles analytiques*, C.R. Acad. Sci. Paris 157 (1913), 325–327.
186. C.F. Gauss, *Theoria Motus coelestium in sectionibus conicis solem ambientium*, Hamburg, 1809.
187. C.F. Gauss, *Methodus nova integralium valores per approximationem inveniendi*, Göttingen, 1814; [188] III, 163–196. German translation: *Neue Methode zur näherungsweisen Auffindung von Integralwerten*, [289].
188. C.F. Gauss, *Werke I-XII*, Königl. Ges. Wiss. Göttingen, 1863–1929.
189. B.R. Gelbaum, J.M.H. Olmsted, *Counterexamples in Analysis*, Holden-Day, Inc., San Francisco, 1964.
190. B.R. Gelbaum, J.M.H. Olmsted, *Theorem and Counterexamples in Mathematics*, Springer-Verlag, New York, 1990.
191. A. Genocchi, G. Peano, *Calcolo differenziale e principii di calcolo integrale*, Torino, 1884.
192. S. Gerschgorin, *Über die Abgrenzung der Eigenwerte einer Matrix*, Izv. Akad. Nauk SSSR Otd. Mat. Estest. 1931, 749–754.

193. W. Givens, *A method of computing eigenvalues and eigenvectors suggested by classical results on symmetric matrices*, Nat. Bur. Standards Appl. Math. 29 (1953), 117–122.
194. W. Givens, *Numerical computation of the characteristic values of a real symmetric matrix*, Oak Ridge Nat. Lab. Rep. ORNL 1574, 1954.
195. B.V. Gnedenko, *The Theory of Probability*, Mir, Moscow, 1978.
196. A.N. Godunov, *Peano's theorem in Banach spaces*, Funkcional. Anal. i Priložen. 9 (1974), no. 1, 59–60 (in Russian), Functional Anal. Appl. 9 (1975), no. 1, 53–55 (in English).
197. H.H. Goldstine, *A History of Numerical Analysis from the 16th Through the 19th Century*, Springer, New York, 1977.
198. S.H. Gould, *The Method of Archimedes*, Amer. Math. Monthly 62 (1955), no. 7, 473–476.
199. É. Goursat, *Sur la théorie des fonctions implicites*, Bull. Soc. Math. France 31 (1903), 184–192.
200. J.P. Gram, *Om Rackkendvilklinger bestemte ved Hjaelp af de mindste Kvadraters Methode*, Copenhagen, 1879. German translation: *Ueber die Entwicklung reeller Funktionen in Reihen mittelst der Methode der kleinsten Quadrate*, J. reine angew. Math. 94 (1883), 41–73.
201. H. Grassmann, *Die lineale Ausdehnungslehre*, Verlag von Otto Wigand, Leipzig, 1844; [203], I_1, 1–319.
202. H. Grassmann, *Die Ausdehnungslehre*, Verlag von Th. Chr. Fr. Enslin, 1862; [203] I_2, 1–383.
203. H. Grassmann, *Gesammelte mathematische und physikalische Werke I-III*, Teubner, Leipzig, 1894–1911.
204. L.M. Graves, *Riemann integration and Taylor's formula in general analysis*, Trans. Amer. Math. Soc. 29 (1927), 163–177.
205. L.M. Graves, T.H. Hildebrandt, *Implicit functions and their differentials in general analysis*, Trans. Amer. Math. Soc. 29 (1927), 127–153.
206. J. Gregory, *Vera circuli et hyperbolae quadratura*, 1667.
207. J. Gregory, *Exercitationes geometricae...*, London, 1668.
208. J. Gregory, *Letter to John Collins, November 23, 1670*, see [210] or [373] I, 46. o.
209. J. Gregory, *Letter to John Collins, February 15, 1671.*
210. J. Gregory, *James Gregory Tercentanary Memorial Volume*, ed. H.W. Turnbull, London, 1939.
211. P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Michigan, 1985.
212. T.H. Gronwall, *Note on the derivatives with respect to a parameter of the solutions of a system of differential equations*, Ann. of Math. 20 (1919), 292–296.
213. J. Hadamard, *Leçons sur la propagation des ondes et les équations de l'hydrodynamique*, Hermann, Paris, 1903.
214. H. Hahn, *Ueber Folgen linearer Operationen*, Monatsh. Math. Phys. 32 (1922), 3–88.
215. H. Hahn, *Ueber lineare Gleichungssysteme in linearen Räumen*, J. reine angew. Math. 157 (1927), 214–229.
216. E. Hairer, S.P. Norsett, G. Wanner, *Solving Ordinary Differential Equations I-II*, Springer, Berlin, 1987–1991.
217. E. Hairer, G. Wanner, *Analysis by Its History*, Springer, New York, 1996.
218. A. Hajnal, P. Hamburger, *Set Theory*, Cambridge Univ. Press, 1999.
219. A. Hald, *A History of Probability and Statistics and Their Applications Before 1750*, Wiley, 2003.
220. P.R. Halmos, *Measure Theory*, D. Van Nostrand Co., Inc., Princeton, N.J., 1950.
221. R.S. Hamilton, *Three-manifolds with positive Ricci curvature*, J. Differential Geom. 17 (1982), 2, 255–306.
222. A. Haraux, *Semi-groupes linéaires et équations d'évolution linéaires périodiques*, Publication du Laboratoire d'Analyse Numérique No. 78011, Université Pierre et Marie Curie, Paris, 1978.
223. T. Harriot, *De Numeris Triangularibus et inde de Progressionibus Arithmeticis Magisteria magna*, 1611.
224. P. Hartman, *Ordinary Differential Equations*, John Wiley & Sons, New York, 1964.
225. F. Hausdorff, *Grundzüge der Mengenlehre*, Verlag von Veit, Leipzig, 1914.

226. F. Hausdorff, *Über halbstetigen Funktionen und deren Verallgemeinerung*, Math. Z. 5 (1919), 292–309.
227. F. Hausdorff, *Die Mengen G_δ in vollständigen Räumen*, Fund. Math. 6 (1924), 2, 146–148.
228. E. Heine, *Über trigonometrischen Reihen*, J. reine angew. Math. 71 (1870), 353–365.
229. E. Heine, *Die Elemente der Funktionenlehre*, J. reine angew. Math. 74 (1872), 172–188.
230. E. Helly, *Über lineare Funktionaloperationen*, Sitzber. Akad. Wiss. Wien 121 (1912), 265–297.
231. E. Helly, *Über Mengen konvexer Körper mit gemeinschaftlichen Punkten*, Jahresbericht der Deutschen Mathematiker-Vereinigung 32 (1923), 175–176.
232. Ch. Hermite, *Sur un nouveau développement en série des fonctions*, C.R. Acad. Sci. Paris 58 (1864), 93–100, 266–273.
233. Ch. Hermite, *Sur la formule d'interpolation de Lagrange*, J. reine angew. Math. 84 (1878), 70–79.
234. O. Hesse, *Ueber die Criterien des Maximums und Minimums der einfachen Integrale*, J. reine angew. Math. 54 (1857), 227–273.
235. K. Heun, *Neue Methode zur approximativen Integration der Differentialgleichungen einer unabhängigen Veränderlichen*, Z. Math. Phys. 45 (1900), 23–38.
236. E. Hewitt, K. Stromberg, *Real and Abstract Analysis*, Springer, Berlin, 1965.
237. D. Hilbert, *Ueber die stetige Abbildung einer Line auf ein Flächenstück*, Mathematische Annalen, 38 (1891), 3, 459–460.
238. D. Hilbert, *Grundzüge einer allgemeinen Theorie der Integralgleichungen I*, Nachr. Akad. Wiss. Göttingen. Math.-Phys. Kl. (1904), 49–91.
239. D. Hilbert, *Grundzüge einer allgemeinen Theorie der linearen Integralgleichungen IV*, Nachr. Akad. Wiss. Göttingen. Math.-Phys. Kl. 8 (1906), 157–227.
240. D. Hilbert, S. Cohn-Vossen, *Geometry and the Imagination*, AMS Chelsea Publishing, 1999.
241. T.H. Hildebrandt, *The Borel theorem and its generalizations*, Bull. Amer. Math. Soc. 32 (1926), 423–474.
242. E. Hille, J.D. Tamarkin, *Remarks on a known example of a monotone continuous function*, Amer. Math. Monthly 36 (1929), 5, 255–264.
243. H. Hochstadt, *Eduard Helly, father of the Hahn-Banach theorem*, Math. Intelligencer 2 (1979), 3, 123–125.
244. J.C. Holladay, *Smoothest curve approximation*, Math. Tables Aids Comp. 11 (1957), 233–243.
245. G. de L'Hospital, *Analyse des infiniment petits, pour l'intelligence des lignes courbes*, A Paris, de l'Impremerie Royale. M.DC.XCVI. Partial English translation in [472], 312–316.
246. A.S. Householder, *The Theory of Matrices in Numerical Analysis*, Dover, New York, 1975.
247. A.S. Householder, F.L. Bauer, *On certain methods of expanding the characteristic polynomial*, Numer. Math. 1 (1959), 29–37.
248. O. Hölder, *Ueber einen Mittelwerthsatz*, Götting. Nachr., 1889, 38–47.
249. Ch. Huygens, *Letter to l'Hospital on December 29, 1692*; [250] X, 348–355.
250. Ch. Huygens, *Oeuvres complètes I-XXII*, Nijhoff, La Haye, 1888–1950.
251. V.A. Il'in, I. Joó, *Uniform estimate of eigenfunctions and upper estimate of eigenvalues of Sturm–Liouville operators with L^p potential*, Diff. Equations (Russian) 15 (1979), 1164–1174.
252. E.L. Ince, *Ordinary Differential Equations*, Dover, New York, 1956.
253. E.L. Ince, *Integration of Ordinary Differential Equations*, Oliver and Boyd, Edinburgh and London, 1952.
254. D. Jackson, *Fourier Series and Orthogonal Polynomials*, Menasha, Wisconsin, 1941.
255. C.G.J. Jacobi, *Über Gauss' neue Methode die Werthe der Integrale näherungsweise zu finden*, J. reine angew. Math. 1 (1826), 301–308; [258] VI, 3–11. See also: [289].
256. C.G.J. Jacobi, *De usu legitimo formulae summatoriae Maclaurinianae*, J. reine angew. Math. 12 (1834), 263–272; [258] VI, 64–75.
257. C.G.J. Jacobi, *Untersuchungen über die Differentialgleichung der hypergeometrischen Reihe*, J. reine angew. Math. 56 (1859), 149–165; [258] VI, 184–202.
258. C.G.J. Jacobi, *Gesammelte Werke I-VIII*, Berlin, 1881–1891.

259. V. Jarník, *Bolzano and the Foundations of Mathematical Analysis*, Society of Czechoslovak Mathematicians and Physicists, Prague, 1981.
260. J.L.W.V. Jensen, *Om konvexe Funktioner og Uligheder mellem Middelvaerdier*, Nyt Tidsskr. Math. 16B (1905), 49–69. French translation: *Sur les fonctions convexes et les inégalités entre les valeurs moyennes*, Acta Math. 30 (1906), 175–193.
261. I. Joó, *A simple proof for von Neumann's minimax theorem*, Acta Sci. Math. (Szeged) 42 (1980), 91–94.
262. I. Joó, V. Komornik, *On the equiconvergence of expansions by Riesz bases formed by eigenfunctions of the Schrödinger operator*, Acta Sci. Math. (Szeged) 46 (1983), 357–375.
263. C. Jordan, *Cours d'analyse*, École Polytechnique, Paris, 1882.
264. Ch. Jordan, *On a new demonstration of Maclaurin's or Euler's summation formula*, Tohoku Math. J. 21 (1922), 244–246.
265. P. Jordan, J. von Neumann, *On inner products in linear, metric spaces*, Ann. of Math. (2) 36 (1935), no. 3, 719–723.
266. J.-P. Kahane, *Fourier Series*, see J.-P. Kahane and P.-G. Lemarié-Rieusset, *Fourier Series and Wavelets*, Gordon and Breach, 1995.
267. S. Kakeya, *On the set of partial sums of an infinite series*, Proc. Tokyo Math.-Phys. Soc (2) 7 (1914), 250–251.
268. S. Kakeya, *On the partial sums of an infinite series*, Tôhoku Sc. Rep. 3 (1915), 159–163.
269. E. Kamke, *Differentialgleichungen, Lösungsmethoden un Lösungen*, Leipzig, 1943.
270. L.V. Kantorovich, *On Newton's method* (in Russian), Trudy Mat. Inst. Steklov. 28 (1949), 104–144.
271. L.V. Kantorovich, *The majorant principle and Newton's method* (in Russian), Doklady Akad. Nauk SSSR (N.S.) 76 (1951), 17–20.
272. L.V. Kantorovich, G.P. Akilov, *Functional analysis*, Second edition. Pergamon Press, Oxford-Elmsford, N.Y., 1982.
273. J. Kelley, *General Topology*, Van Nostrand, New York, 1954.
274. J. Kepler, *Stereometria Doliorum Vinariorum [Erklärung und Bestättigung der Oesterreichischen Weinvisier-Ruthen]*, Linz, 1615; [275] IV, 545–665 (in Latin), V, 495–610 (in German). See also the comments in [5] (20. o.), [487], (47–53. o.).
275. J. Kepler, *Opera Omnia I-VIII*, Heyder & Zimmer, Frankfurt-Erlangen, 1858–1871.
276. P. Kirchberger, *Über Tchebychefsche Annäherungsmethoden*, Math. Ann. 57 (1903), 509–540.
277. M. Kline, *Mathematical Thought from Ancient to Modern Times*, Oxford Univ. Press, New York, 1972.
278. A.N. Kolmogorov, S.V. Fomin, *Elements of the Theory of Functions and Functional Analysis, Volume 1*, Dover, New York, 1999.
279. K. Knopp, *Theory and Application of Infinite Series*, Dover, 1990.
280. V. Komornik, *Minimax theorems for upper semicontinuous functions*, Acta Math. Hungar. 40 (1982), 159–163.
281. V. Komornik, *On the equiconvergence of eigenfunction expansions associated with ordinary linear differential operators*, Acta Math. Hungar. 47 (1986), 261–280.
282. V. Komornik, *Exact Controllability and Stabilization. The Multiplier Method*, Masson, Paris, and John Wiley & Sons, Chicester, 1994.
283. V. Komornik, *Another short proof of Descartes's rule of signs*, Amer. Math. Monthly 113 (2006), 9, 829–831.
284. V. Komornik, *Uniformly bounded Riesz bases and equiconvergence theorems*, Bol. Soc. Paran. Mat. (3s.) 25 (2007), 1–2, 139–146.
285. V. Komornik, *Lectures on Functional Analysis and the Lebesgue integral*, Springer, Universitext, 2016.
286. Komornik V., P. Martinez, M. Pierre, J. Vancostenoble, *"Blow-up" of bounded solutions of differential equations*, Acta Sci. Math. (Szeged) 69 (2003), 3–4, 651–657.
287. A.N. Korkin, E.I. Zolotarev, *Sur un certain minimum*, Oeuvres de E.I. Zolotarev, 138–153.

288. G. Kowalewski, *Einführung in die Determinantentheorie einschliesslich der unendlichen und der fredholmschen Determinanten*, Leipzig, Verlag von Veit, 1909.
289. A. Kowalewski, *Newton, Cotes, Gauss, Jacobi. Vier grundlegende Abhandlungen über Interpolation und genäherte Quadratur*, Leipzig, 1917.
290. Gy. König [Julius König], *Analízis [Analysis]*, Budapest, 1887.
291. P.P. Korovkin, *Inequalities*, Little Mathematics Library, Mir, Moscow, 1975.
292. H.W. Kuhn, A.W. Tucker, *Nonlinear programming*, Proc. of Second Berkeley Symp., Univ. of California Press, Berkeley, 1951, 481–492.
293. C. Kuratowski, *Une méthode d'élimination des nombres transfinis des raisonnements mathématiques*, Fund. Math. 3 (1922), 76–108.
294. C. Kuratowski, *La propriété de Baire dans les espaces métriques*, Fund. Math. 16 (1930), 390–394.
295. C. Kuratowski, *Topologie I*, 4th ed., Monografje Matematyczne, Warszawa, 1958.
296. C. Kuratowski, *Quelques problèmes concernant les espaces métriques non-séparables*, Fund. Math. 25 (1935), 534–545.
297. C. Kuratowski, *Topologie I-II*, Pánstwowe Wydawnictvo Naukowe, Warszawa, 1948–1950.
298. A.G. Kuros, *Higher Algebra*, Mir, Moscow, 1980.
299. W. Kutta, *Beitrag zur näherungsweisen Integration totaler Differentialgleichungen*, Z. Math. Phys. 46 (1901), 435–453.
300. J. Kürschák, *Analízis és analitikus geometria [Analysis and Analytic Geometry]*, Budapest, 1920.
301. J.E. Lagnese, *Boundary Stabilization of Thin Plates*, SIAM Studies in Appl. Math., Philadelphia, 1989.
302. J.L. Lagrange, *Recherches sur la méthode de maximis et minimis*, Misc. Taurinensia, Torino 1 (1759); [312] I, 3–20.
303. J.L. Lagrange, *Solution de différents problèmes de calcul intégral*, Miscellanea Taurinensia III (1762–1766); [312] I, 471–668.
304. J.L. Lagrange, *Sur la résolution des équations numériques*, Mém. Acad. royale des Sciences et Belles-Lettres de Berlin 23 (1769); [312] II, 539–578.
305. J.L. Lagrange, *Solutions analytiques de quelques problèmes sur les pyramides triangulaires*, Nouv. Mém. Acad. Berlin 1773; [312] III.
306. J.L. Lagrange, *Recherches sur les suites récurrentes*, Nouveaux Mémoires de l'Académie royale des Sciences et Belles Lettres de Berlin, année 1775; [312] IV, 151–251.
307. J.L. Lagrange, *Sur différentes questions d'analyse relatives à la théorie des intégrales particulières*, Nouveaux Mémoires de l'Académie royale des Sciences et Belles Lettres de Berlin, année 1779; [312] IV, 585–634.
308. J.L. Lagrange, *Méchanique analitique [sic]*, Paris, 1788; [312] XI.
309. J.L. Lagrange, *Leçons élémentaires sur les mathématiques*, Cours á l'École Normale, Paris, 1795; [312] VII, 183–288.
310. J.L. Lagrange, *Théorie des fonctions analytiques*, Paris, 1797, new edition 1813; [312] IX.
311. J.L. Lagrange, *Leçons sur le calcul des fonctions analytiques*, Cours de 1799, École Polytechnique, Paris, 1801, new edition in 1806; [312] X.
312. J.L. Lagrange, *Oeuvres I-XIV*, Gauthier-Villars, Paris, 1867–1882.
313. E. Laguerre, *Sur l'intégrale* $\int_x^\infty \frac{e^{-x}}{x}\,dx$, Bull. Soc. Math. France 7 (1879), 72–81.
314. E. Landau, *Über die Approximation einer stetigen Funktion durch eine ganze rationale Funktion*, Rend. Circolo Mat. Palermo 25 (1908), 337–345.
315. P.S. Laplace, *Mémoires sur les approximations des formules qui sont fonctions de très grands nombres*, Mém. de l'Acad. royale des Sci. de Paris (1782), 1–88, published in 1785; *Oeuvres X*, 209–291.
316. P.-J. Laurent, *Approximation et optimisation*, Hermann, Paris, 1972.
317. P. Lax, *Change of variables in multiple integrals*, Amer. Math. Monthly 106 (1999), 497–501.
318. H. Lebesgue, *Sur l'approximation des fonctions*, Bull. Sci. Math. 22 (1898); [324] III, 11–20.
319. H. Lebesgue, *Intégrale, longueur, aire*, Ann. Mat. Pura Appl. (3) 7 (1902), 231–359; [324] I, 201–331.

320. H. Lebesgue, *Leçons sur l'intégration et la recherche des fonctions primitives*, Paris, 1904; [324] II, 11–154.
321. H. Lebesgue, *Sur le problème de Dirichlet*, Rend. Circ. Mat. Palermo 24 (1907), 371–402; [324] IV, 91–122.
322. H. Lebesgue, *Notice sur les travaux scientifiques de M. Henri Lebesgue*, Toulouse, 1922; [324] I, 97–175.
323. H. Lebesgue, *Leçons sur l'intégration*, Paris, 1928.
324. H. Lebesgue, *Oeuvres scientifiques I-V*, Université de Genève, 1972–73.
325. A.M. Legendre, *Recherches sur l'attraction des sphéroïdes homogènes*, Mémoires math.-phys. présentés á l'Acad. Sci. 10 (1785), 411–434.
326. A.M. Legendre, *Nouvelles méthodes pour la détermination des orbites des comètes; Appendice sur la méthode des moindres carrées*, Paris, 1805.
327. G.W. Leibniz, *Nova methodus pro maximis et minimis, itemque tangentibus, quoe nec fractas, nec irrationales quantitates moratur, & singolare pro illis calculi genus*, Acta Eruditorum 3 (1684), 467–473. o.; [330] V, 220–226. English translation: *A new method for maxima and minima as well as tangents, which is neither impeded by fractional nor irrational quantities, and a remarkable type of calculus for them*, [472], 272–280.
328. G.W. Leibniz, *De geometria recondita et analysi indivisibilium atque infinitorum*, Acta Eruditorum 5 (1686), 292–300; [330] V, 226–233. French translation: *Sur la géométrie profonde et l'analyse des indivisibles et des infinis*, [331], 126–143.
329. G.W. Leibniz, *Symbolismus memorabilis calculi algebraici et infinitesimalis in comparatione potentiarum et differentiarum, et de lege homogeneorum transcendantali*, Miscellanea Berolinensia ad incrementum scientiarum, 1710; [330] V, 377–382. French translation: *Présentation d'une notation algébrique remarquable pour comparer les puissances et les différences, et de la loi d'homogénéité transcendentale*, [331], 409–421.
330. G.W. Leibniz, *Leibnizens mathematische Schriften I-IX*, Asher, Berlin, 1849–1850, Schmidt, Berlin 1855–1863.
331. G.W. Leibniz, *La naissance du calcul différentiel*, 26 articles des Acta Eruditorum, introduction, traduction et notes par Marc Parmentier, Vrin, Paris, 1989.
332. E. Lindelöf, *Sur l'application des méthodes d'approximations successives à l'étude des intégrales réelles des équations différentielles ordinaires*, J. de Math. (4) 10 (1894), 117–128.
333. E. Lindelöf, *Sur quelques points de la théorie des ensembles*, C.R. Acad. Sci. Paris Sér. I Math. 137 (1903), 697–700.
334. E. Lindelöf, *Remarques sur un théorème fondamental de la théorie des ensembles*, Acta Math. 29 (1905), 183–190.
335. J. Liouville, *Sur le développement des fonctions ou parties de fonctions en séries dont les divers termes sont assujettis à satisfaire à une même équation différentielle du second ordre contenant un paramètre variable*, Liouville J. Math. (1) 2 (1837), 16–35.
336. R. Lipschitz, *Disamina della possibilità d'integrare completamente un dato sistema di equazioni differenziali ordinarie*, Ann. Mat. Pura Appl. (2) 2 (1868–69), 288–302. French translation: *Sur la possibilité d'intégrer complètement un système donné d'équations différentielles*, Bull. Sci. Math. Astr. 10 (1876), 149–159.
337. N.I. Lobachevsky, *On the Foundations of Geometry* (in Russian), Kasan Messenger 25 (1829), February–March, 178–187, 28 (1830), March–April, 251–283, 28 (1830), July–August, 571–636. German translation: *Ueber die Anfangsgründe der Geometrie*, see *Nikolaj Iwanowitsch Lobatschefskij. Zwei geometrische Abhandlungen*, Teubner, Leipzig, 1898, 1–66. English translation: *Geometrical Investigations on the Theory of Parallel Lines*, 1891. Reprinted in [56].
338. H. Löwig, *Komplexe euklidische Räume von beliebiger endlicher oder unendlicher Dimensionszahl*, Acta Sci. Math. Szeged 7 (1934), 1–33.
339. L.A. Lyusternik, *Conditional extrema of functionals* (in Russian), Mat. Sb. 41 (1934), 3, 390–401.
340. C. Maclaurin, *A Treatise on Fluxions I-II*, Edinburgh, 1742.
341. J.H. Manheim, *The Genesis of Point Set Topology*, Oxford, 1964.

342. A.A. Markov [Markoff], *Some applications of algebraic continuous fractions* (in Russian), Bull. St-Petersburg, 1884.
343. A.A. Markov [Markoff], *Sur la méthode de Gauss pour le calcul approché des intégrales*, Math. Ann. 25 (1885), 427–432.
344. C. Méray, *Nouveau précis d'analyse infinitésimal*, Paris, 1872.
345. C. Méray, *Observations sur la légitimité de l'interpolation*, Ann. Sci. École Norm. Sup. 1 (1884), 165–176.
346. C. Méray, *Nouveaux exemples d'interpolations illusoires*, Bull. Sci. Math. 20 (1896), 266–270.
347. N. Metropolis, S. Ulam *The Monte Carlo Method*, J. Amer. Stat. Association 44 (1949), 247, 335–341.
348. J. Milnor, *Topology from the differentiable viewpoint*, The University Press of Virginia, Charlottesville, 1965.
349. J. Milnor, *Analytic proofs of the "hairy ball theorem" and the Brouwer fixed-point theorem*, Amer. Math. Monthly 85 (1978), 521–524.
350. H. Minkowski, *Geometrie der Zahlen I*, Leipzig, 1896.
351. G.J. Minty, *On the monotonicity of the gradient of a convex function*, Pacific J. Math. 14 (1964), 243–247.
352. F. Moigno, *Leçons de calcul différentiel et de calcul intégral, rédigées d'après les méthodes et les ouvrages publiés ou inédits de M. A.L. Cauchy I-IV*, Paris, 1840–1861.
353. A. de Moivre, *Miscellania analytica de seriebus et quadraturis, Londini, 1730*, *Miscellaneis Analyticis Supplementum*, 22 pages.
354. P. Montel, *Leçons sur les séries de polynômes à une variable complexe*, Gauthier-Villars, Paris, 1910.
355. E.H. Moore, H.L. Smith, *A general theory of limits*, Amer. J. Math. 44 (1922), 102–121.
356. *Moscow Mathematical Papyrus,* written around 1850 BC.
357. F.J. Murray, *On complementary manifolds and projections in spaces L_p and l_p*, Trans. Amer. Math. Soc. 41 (1937), 138–152.
358. J. Napier, *Mirifici Logarithmorum Canonis descriptio. . .* , Edinburgh, 1614.
359. I.P. Natanson, *Constructive Function Theory I-III*, Frederick Ungar, New York, 1961–1965.
360. J. von Neumann, *Mathematische Begründung der Quantenmechanik*, Nachr. Gesell. Wiss. Göttingen. Math.-Phys. Kl. (1927), 1–57. [364] I, 151–207.
361. J. von Neumann, *Zur Theorie der Gesellschaftsspiele*, Math. Ann. 100 (1928), 295–320; [364] VI, 1–26. English translation: *Contributions to the Theory of Games*, Volume IV, ed. by A.W. Tucker and R.D. Luce, Annals of Mathematics Studies 4, Princeton Univ. Press, New York, 1959, 13–42.
362. J. von Neumann, *Mathematische Grundlagen der Quantenmechanik*, Springer, Berlin, 1932. English translation: *Mathematical Foundations of Quantum Mechanics*, Princeton Univ. Press, New York, 1955.
363. J. von Neumann, O. Morgenstern, *Theory of Games and Economic Behavior*, Princeton Univ. Press, New York, 1944.
364. J. von Neumann, *Collected works I-VI*, Pergamon Press, Oxford, New York, 1972–1979.
365. M.A. Neumark, *Normierte Algebren*, VEB Deutscher Verlag der Wissenschaften, Berlin, 1959.
366. J. Newman (ed.), *The World of Mathematics I-IV*, Dover, New York, 2000.
367. I. Newton, *Annotations from Wallis*, manuscript of 1665, see [372] I.
368. I. Newton, *Methodus Fluxonum et Serierum Infinitarum*, manuscript of 1671, published in: Opuscula mathematica I, London 1736. French translation: *La méthode des fluxions et des suites infinies*, Paris, 1790.
369. I. Newton, *Methodus Differentialis*, manuscript of 1676, published in: *Analysis per quantitatum series, fluxiones, ac differentias*, W. Jones, London, 1711. English translation: [371] II, 165–173. German translation: *Die Differentialmethode*, [289].

370. I. Newton, *Philosophiae naturalis principia mathematica*, London, 1687. English translation: *Newton's Mathematicalk Principles of Natural Philosophy*, A. Motte's translation revised, ed. F. Cajori, Univ. of California Press, 1934.
371. I. Newton, *The Mathematical Works of Isaac Newton I-II*, ed. D.T. Whiteside, New York, Johnson Reprint, 1964.
372. I. Newton, *The Mathematical Papers of Isaac Newton I-VII*, ed. D.T. Whiteside, Cambridge Univ. Press, 1967–1976.
373. I. Newton, *The Correspondence of Isaac Newton I-VII*, ed. H.W. Turnball et al., Cambridge Univ. Press, 1959–1978.
374. O. Niccoletti, *Sugli integrali delle equazioni differenziali considerati come funzioni dei loro valori iniziali*, Atti R. Accad. Rend. Lincei. Cl. Sci. Fis. Mat. Nat. (5) 4 (1895), 316–324.
375. N. Oresme, *Tractatus de latitudinibus formarum*, manuscript around 1361, see [61].
376. W. Osgood, *Non uniform convergence and the integration of series term by term*, Amer. J. Math. 19 (1897), 155–190.
377. J.C. Oxtoby, *Measure and Category*, Springer, New York, Heidelberg, Berlin, 1971.
378. P. Painlevé, *Sur le développement des fonctions analytiques pour les valeurs réelles des variables*, C.R. Acad. Sci. Paris 126 (1898), 385–388.
379. G. Peano, *Sull'integrabilitá delle equazioni differenziali di primo ordine*, Atti delle Reale Accad. delle Scienze di Torino 21 A (1886), 677–685; [385] I, 74–81.
380. G. Peano, *Intégration par séries des équations différentielles linéaires*, Math. Ann. 32 (1888), 450–456; [385] I, 83–90.
381. G. Peano, *Une nouvelle formule du reste dans la formule de Taylor*, Mathesis 9 (1889), p. 182–183; [385] I, 95–96.
382. G. Peano, *Démonstration de l'intégrabilité des équations différentielles linéaires*, Math. Ann. 37 (1890), 182–228; [385] I, 119–170.
383. G. Peano, *Sur une courbe, qui remplit toute une aire plane*, Math. Annalen 36 (1890), 1, 157–160.
384. G. Peano, *Generalità sulle equazioni differenziali ordinarie*, Atti R. Accad. Sci. Torino 33 (1897), 9–18; [385] I, 285–293.
385. G. Peano, *Opere scelte I-III*, Edizioni cremonese, Roma, 1957–1959.
386. K. Pearson, *The problem of the random walk*, Nature, 72 (1905), 294 (July 27).
387. K. Pearson, *Historical note on the origin of the normal curve of error*, Biometrika, 16 (1924), 402–404.
388. I.G. Petrovsky, *Lectures on Partial Differential Equations*, Dover, New York, 1991.
389. É. Picard, *Mémoire sur la théorie des équations aux dérivées partielles et la méthode des approximations successives*, J. Math. Pures Appl. (4) 6 (1890), 145–210, 231.
390. É. Picard, *Sur les méthodes d'approximations successives dans la théorie des équations différentielles*, see G. Darboux, *Leçons sur la théorie générale des surfaces IV*, Gauthier-Villars, Paris, 1896, 353–367.
391. M. Picone, *Lezioni di analisi infinitesimale*, Circolo matematico di Catania, 1923.
392. H. Poincaré, *Analysis situs*, J. de l'École Polytechnique (2) 1 (1895), 1–122.
393. H. Poincaré, *Cinquième complément à l'analysis situs*, Rendiconti del Circolo Matematico di Palermo 18 (1904), 45–110.
394. S.-D. Poisson, *Sur le calcul numérique des intégrales définies*, Mémoires Acad. scienc. Inst. France 6 (1823), 571–602.
395. G. Pólya, *Über eine Aufgabe der Wahrscheinlichkeitsrechnung betreffend die Irrfahrt im Straßennetz*, Math. Ann. 83 (1921), 149–160.
396. G. Pólya, G. Szegő, *Problems and Exercises in Analysis*, Vol. I, Springer-Verlag, Berlin, New York, 1972.
397. L.S. Pontryagin [Pontriagin], *Foundations of Combinatorial Topology*, Graylock, Rochester, 1952.
398. L.S. Pontryagin, *Ordinary Differential Equations*, Addison-Wesley; Pergamon, 1962.

399. K.A. Posse, *Sur les quadratures*, Nouvelles Annales de Math. (2) 14 (1875), 49–62.
400. I.I. Privalov, *Sur les fonctions harmoniques*, Mat. Sbornik 32 (1925), 3, 464–471.
401. H. Rademacher, O. Toeplitz, *The Enjoyment of Mathematics*, Dover, New York, 1990.
402. J. Radon, *Mengen konvexer Körper, die einen gemeinsamen Punkt enthalten*, Math. Annalen 83 (1921) (1–2), 113–115.
403. J. Raphson, *Analysis Aequationim Universalis...*, London, 1690.
404. A. Ralston, P. Rabinowitz, *A First Course in Numerical Analysis*, McGraw-Hill, New York, 1978.
405. P.A. Raviart, J.M. Thomas, *Introduction à l'analyse numérique des équations aux dérivées partielles*, Masson, Paris, 1983.
406. M. Reed, B. Simon, *Methods of Modern Mathematical Physics I-IV*, Academic Press, New York, 1972–1979.
407. F. Rellich, *Spektraltheorie in nichtseparabeln Räumen*, Math. Ann. 110 (1935), 342–356.
408. B. Riemann, *Ueber die Darstellberkeit einer Function durch eine trigonometrische Reihe*, Abhandlungen der Königlichen Gesellschaft der Wissenschaften zu Göttingen 13 (1867); [411], 213–251. French translation: *Sur la possibilité de représenter une fonction par une série trigonométrique*, [411], 225–272.
409. B. Riemann, *Ueber die Hypothesen, welche der Geometrie zu Grunde liegen*, Habilitationsschrift, 1854, Abhandlungen der Königlichen Gesellschaft der Wissenschaften zu Göttingen 13 (1867); [411], 254–269. French translation: *Sur les hypothèses qui servent de base à la Géométrie*, [411], 280–299.
410. B. Riemann, *Theorie der Abel'schen Functionen*, J. reine angew. Math. 54 (1857); [411], 88–142. French translation: *Théorie des fonctions abéliennes*, [411], 89–162.
411. B. Riemann, *Werke*, Teubner, Leipzig, 1876. French translation: *Oeuvres mathématiques de Riemann*, Gauthier-Villars, Paris, 1898.
412. F. Riesz, *A térfogalom genezise I [Genesis of the notion of space I]*, Math. és Phys. Lapok 15 (1906), 97–122, 16 (1907), 145–161; [420] I, 67–109. German translation: *Die Genesis des Raumbegriffes*, Math. u. Naturwiss. Berichte aus Ungarn 24 (1907), 309–353; [420] I, 110–154.
413. F. Riesz, *Stetigkeitsbegriff und abstrakte Mengenlehre*, Atti del IV. Congr. Internaz. dei Mat. Roma 2 (1908), 18–24; [420] I, 155–161.
414. F. Riesz, *Sur certains systèmes d'équations fonctionnelles et l'approximation de fonctions continues*, C.R. Acad. Sci. Paris 150 (1910), 674–677.
415. F. Riesz, *Untersuchungen über Systeme integrierbar Funktionen*, Math. Ann. 69 (1910), 449–497.
416. F. Riesz, *Lineáris függvényegyenletekről [On linear functional equations]*, Math. és Természettudományi Ért. 35 (1917), 544–579; [420] II, 1017–1052. German translation: *Über lineare Funktionalgleichungen*, Acta Math. 41 (1918), 71–98; [420] II, 1053–1080.
417. F. Riesz, *Su alcune disuguglianze [On some inequalities]*, Boll. dell'Unione Mat. Ital. 7 (1928), 77–79; [420] I, 519–521.
418. F. Riesz, *A monoton függvények differenciálhatóságáról [On the differentiability of monotone functions]*, Mat. és Fiz. Lapok 38 (1931), 125–131; [420] I, 243–249.
419. F. Riesz, *Sur l'existence de la dérivée des fonctions monotones et sur quelques problèmes qui s'y rattachent [On the differentiability of monotone functions and related problems]*, Acta Sci. Math. (Szeged) 5 (1930–32), 208–221; [420] I, 250–263.
420. F. Riesz, *Collected works I-II]*, Akadémiai Kiadó, Budapest, 1960.
421. F. Riesz, B. Sz.-Nagy, *Functional Analysis*, Dover, 1990.
422. A.W. Roberts, D.E. Varberg, *Another proof that convex functions are locally Lipschitz*, Amer. Math. Monthly 81 (1974), 1014–1016.
423. G.P. de Roberval, *Letter to Fermat on June 1st, 1638*; Fermat II, 147–151.
424. L.J. Rogers, *An extension of a certain theorem in inequalities*, Messenger of Math. 17 (1888), 145–150.

425. C.A. Rogers, *A less strange version of Milnor's proof of Brouwer's fixed-point theorem*, Amer. Math. Monthly 87 (1980), 525–527.
426. M. Rolle, *Démonstration d'une Methode pour résoudre les Egalitez de tous les degres...*, Paris, Chez Jean Cusson, ruë Saint Jacques, 1691.
427. W. Romberg, *Vereinfachte numerische Integration*, Norske Vid. Selsk. Forhdl. 28 (1955), 30–36.
428. A. Rosenblatt, *Über the Existenz von Integralen gewöhnlicher Differentialgleichungen*, Arkiv Mat. Astron. Fys. 5 (1909), No. 2.
429. A. Rosental, L. Zoretti, *Die Punktmengen*, Encyklopädie der Mathematischen Wissenschaften, II C 9a, Teubner, Leipzig, 1924.
430. W. Rudin, *Principles of Mathematical Analysis*, Third edition, McGraw Hill, New York, 1976.
431. W. Rudin, *Real and Complex Analysis*, Third edition, McGraw Hill, New York, 1986.
432. W. Rudin, *Functional Analysis*, McGraw-Hill, New York, 1991.
433. C. Runge, *Ueber die numerische Auflösung von Differentialgleichungen*, Math. Ann. 46 (1895), 167–178.
434. C. Runge, *Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten*, Z. für Math. Phys. 46 (1901), 224–243.
435. S. Russ, *Bolzano's analytic programme*, Math. Intelligencer 14 (1992), 3, 45–53.
436. O. Schlömilch, *Über Mittelgrössen verschiedener Ordnungen*, Z. für Math. Phys. 3 (1858), 308–310.
437. E. Schmidt, *Entwickelung willkürlicher Functionen nach Systemen vorgeschriebener*, Inaugural-Dissertation, Göttingen, 1905, 4–6 and Math. Ann. 63 (1907), 433–476.
438. I.J. Schoenberg, *On the Peano curve of Lebesgue*, Bull. Amer. Math. Soc. 44 (1938), 8, 519.
439. I.J. Schoenberg, *Contributions to the problem of approximation of equidistant data by analytic functions*, Quart. Appl. Math. 4 (1946), 45–99, 112–141.
440. I.J. Schoenberg, A. Whitney, *On Pólya frequency functions, III. The positivity of translation determinants with an application to the interpolation problem by spline curves*, Trans. Amer. Math. Soc. 74 (1953), 246–259.
441. H.A. Schwarz, *Ueber ein vollständiges System von einander unabhängiger Voraussetzungen zum Beweise des Satzes* $\frac{\partial}{\partial y}\left(\frac{\partial f(x,y)}{\partial x}\right) = \frac{\partial}{\partial x}\left(\frac{\partial f(x,y)}{\partial y}\right)$, Verhandlungen der Schweizerischen Naturf. Ges. (1873), 259–270; [443] II, 275–284.
442. H.A. Schwarz, *Ueber ein die Flächen kleinsten Flächeninhalts betreffendes Problem der Variationsrechnung*, Acta Soc. Scient. Fenn. 15 (1885), 315–362; [443] I, 223–269.
443. H.A. Schwarz, *Gesammelte Mathematische Abhandlungen I-II*, Springer, Berlin, 1890.
444. J. Sebestik, *Bernard Bolzano et son mémoire sur le théorème fondamental de l'analyse*, Revue d'Histoire des Sciences et de Leurs Applications 17 (1964), 136–164.
445. J. Sebestik, *Logique et mathématique chez Bernard Bolzano*, Vrin, Paris, 1992.
446. J.A. Segner, *Dissertatio epistolica, qua regulam Harriotti...*, Jena, 1728.
447. J.A. Segner, *Démonstration de la règle de Descartes...*, Hist. de l'Acad. de Berlin, 1756, 292–299.
448. Ph.L. Seidel, *Note über eine Eigenschaft der Reihen, welche discontinuierliche Functionen darstellen*, München, 1847; see: Ostwald's Klassiker der exakten Wissenschaften, No. 116, Leipzig, 35–45.
449. J.A. Serret, *Cours de calcul différentiel et intégral I-II*, Gauthier-Villars, Paris, 1868.
450. G.E. Shilov, *Elementary Real and Complex Analysis*, Dover, New York, 1996.
451. G.E. Shilov, *Elementary Functional Analysis*, Dover, New York, 1996.
452. T. Simpson, *Mathematical Dissertations etc.*, London, 1743, 109–111.
453. D.E. Smith, *A Source Book in Mathematics*, Dover, New York, 1959.
454. I.M. Sobol, *The Monte Carlo Method*, The Univ. of Chicago Press, 1974.
455. S.L. Sobolev, *Partial Differential Equations of Mathematical Physics*, Dover, New York, 1989.
456. N.Y. Sonine, *Recherches sur les fonctions cylindriques et le développement des fonctions continues en séries*, Math. Ann. 16 (1880), 1–80.

457. G.A. Soukhomlinov, *Über Fortsetzung von linearen Funktionalen in linearen komplexen Räumen und linearen Quaternionräumen* (in Russian, with a German abstract), Mat. Sbornik N. S. (3) 4 (1938), 355–358.
458. L.L. Stachó, *Minimax theorems beyond topological vector spaces*, Acta Sci. Math. (Szeged) 42 (1980), 157–164.
459. L.A. Steen, J.A. Seebach, Jr., *Counterexamples in Topology*, Dover, New York, 1995.
460. J.F. Steffensen, *Interpolation*, Chelsea Publ. Co., New York, 1927.
461. T.J. Stieltjes, *Iets over de benaderde voorstelling van eene functie door eene andere [De la représentation approximative d'une fonction par une autre]*, Delft, 1876; [467] I, 89–98 (in Dutch), 99–108 (in French).
462. T.J. Stieltjes, *Over Lagrange's Interpolatieformule [A propos de la formule d'interpolation de Lagrange]*, Versl. K. Akad. Wet. Amsterdam (2) 17 (1882), 239–254; [467] I, 121–134 (in Dutch), 135–148 (in French).
463. T.J. Stieltjes, *Quelques recherches sur les quadratures dites mécaniques*, Ann. Sci. École Norm. Sup. (3) I (1884), 406–426; [467] I, 377–394.
464. T.J. Stieltjes, *Sur quelques théorèmes d'algèbre*, C.R. Acad. Sci. Paris 100 (1885), 439–440; [467] I, 440–441.
465. T.J. Stieltjes, *Sur les polynômes de Jacobi*, C.R. Acad. Sci. Paris 100 (1885), 620–622; [467] I, 442–444.
466. T.J. Stieltjes, *Sur les racines de l'équation* $X_n = 0$, Acta Math. 9 (1886), 385–400; [467] II, 73–88.
467. T.J. Stieltjes, *Oeuvres complètes I-II*, Springer, Berlin, 1993.
468. J. Stirling, *Methodus differentialis* ..., London, 1730.
469. G.G. Stokes, *On the critical values of the sums of periodic series*, Cambridge, 1847, see: *Mathematical and Physical Papers I*, 236–285.
470. O. Stolz, *Bemerkungen zur Theorie der Functionen von mehreren unabhängigen Veränderlichen*, Innsbrucker Berichte, 1887; see also [471].
471. O. Stolz, *Grundzüge der Differential- und Integralrechnung*, Teubner, Leipzig, 1893.
472. D.J. Struik, *A Source Book in Mathematics 1200–1800*, Harvard Univ. Press, Cambridge, 1969.
473. Ch. Sturm, *Analyse d'un mémoire sur la résolution des équations numériques*, Bulletin de Férussac 11 (1829), 419–422.
474. Ch. Sturm, *Sur une classe d'équations à différences partielles*, J. de Math. (1) 1 (1836), 373–444.
475. P. Szász, *A differenciál- és integrálszámítás elemei I-II [Elements of Differential and Integral Calculus I-II]*, Közoktatásügyi Kiadóvállalat, Budapest, 1951.
476. G. Szegő, *Orthogonal Polynomials*, American Math. Soc., Providence, Rhode Island, 1975.
477. B. Szénássy, *History of Mathematics in Hungary until the 20th Century*, Springer, Berlin, 1992.
478. F. Szidarovszky, *Bevezetés a numerikus módszerekbe [Introduction to Numerical Methods]*, Közgazdasági és Jogi Könyvkiadó, Budapest, 1974.
479. B. Sz.-Nagy, *Introduction to Real Functions and Orthogonal Expansions*, Oxford University Press, New York, 1965.
480. A.E. Taylor, *The extension of linear functionals*, Duke Math. J. 5 (1939), 538–547.
481. B. Taylor, *Methodus incrementorum directa & inversa*, LL.D. & Regiae Societatis Secretario, Londini, 1715.
482. P.L. Tchebychef, see P.L. Chebyshev.
483. K.J. Thomae, *Einleitung in die Theorie der bestimmten Integrale*, Halle, 1875.
484. H. Tietze, *Über Funktionen, die auf einer abgeschlossenen Menge stetig sind*, J. reine angew. Math. 145 (1910), 9–14.
485. H. Tietze, L. Vietoris, *Beziehungen zwischen den verschiedenen Zweigen der Topologie*, Encyklopädie der Mathematischen Wissenschaften, III AB 13, Leipzig, 1930.
486. V.M. Tikhomirov, *Fundamental Principles of the Theory of Extremal Problems*, John Wiley & Sons, Chichester, 1986.

487. V.M. Tikhomirov, *Stories about Maxima and Minima*, Amer. Math. Soc., Providence, Rhode Island, 1990.
488. A. Tychonoff [A.N. Tikhonov], *Über die topologische Erweiterung von Räumen*, Math. Ann. 102 (1930), 544–561.
489. A. Tychonoff [A.N. Tikhonov], *Ein Fixpunktsatz*, Math. Ann. 111 (1935), 767–776.
490. A.N. Tikhonov, A.A. Samarskii, *Equations of Mathematical Physics*, Dover, New York, 1990.
491. A.F. Timan, *Theory of Approximation of Functions of a Real Variable*, Pergamon Press Ltd., Oxford, 1963.
492. C. Truesdell, *The New Bernoulli Edition*, Isis, 49 (1958), No. 1, 54–62.
493. P. Turán, *On some open problems of approximation theory*, J. Approx. Theory 29 (1980), 23–85.
494. S. Ulam, *Adventures of a Mathematician*, Charles Scribner's Sons, New York, 1976.
495. P. Urysohn, *Über die Mächtigkeit der zusammenhängenden Mengen*, Math. Ann. 94 (1925), 262–295.
496. S. Vajda, *Theory of Games and Linear Programming*, London: Methuen & Co. Ltd, New York: John Wiley & Sons, Inc., 1956.
497. G. Valiron, *Théorie des fonctions*, Masson, Paris, 1942.
498. R.S. Varga, *Matrix Iterative Analysis*, Prentice-Hall, Inc., Englewood Cliffs, 1962; Second Revised and Expanded Edition, Springer, Heidelberg, 2000.
499. L. Vietoris, *Stetige Mengen*, Monatsh. Math. 31 (1921), 173–204.
500. V. Volterra, *Sulla inversione degli integrali definiti*, Rend. Accad. Lincei 5 (1896), 177–185; 289–300.
501. V. Volterra, *Sopra alcune questioni di inversione di integrali definiti*, Ann. di Math. (2) 25 (1897), 139–187.
502. B.L. van der Waerden, *Science Awakening*, Oxford University. Press, New York, 1961.
503. J.L. Walsh, J.H. Ahlberg, E.N. Nilson, *Best approximation properties of the spline function fit*, J. Math. Mech. 1 (1962), 2, 225–234.
504. W. Walter, *There is an elementary proof of Peano's existence theorem*, Amer. Math. Monthly 78 (1971), 170–173.
505. W. Walter, *Ordinary Differential Equations*, Springer, 1998.
506. G.N. Watson, *A Treatise on the Theory of Bessel Functions*, Cambridge Univ. Press, 1995.
507. K. Weierstrass, *Zur Theorie der Potenzreihen*, manuscript of 1841, published in 1894: *Mathematische Werke I*, Mayer & Müller, Berlin, 1894, 67–74.
508. K. Weierstrass, *Differential Rechnung*, Vorlesung an dem Königlichen Gewerbeinstitute, manuscript of 1861, Math. Bibl., Humboldt Universität, Berlin.
509. K. Weierstrass, *Theorie der analytischen Funktionen*, Vorlesung an der Univ. Berlin, manuscript of 1874, Math. Bibl., Humboldt Universität, Berlin.
510. K. Weierstrsass, *Über die analytische Darstellbarkeit sogenannter willkürlicher Funktionen reeller Argumente, Erste Mitteilung*, Sitzungsberichte Akad. Berlin, 1885, 633–639; *Mathematische Werke III*, Mayer & Müller, Berlin, 1903, 1–37. French translation: *Sur la possibilité d'une représentation analytique des fonctions dites arbitraires d'une variable réelle*, J. Math. Pures Appl. 2 (1886), 105–138.
511. E.T. Whittaker, G.N. Watson, *A Course of Modern Analysis*, Cambridge Univ. Press, 1996.
512. N. Wiener, *Limit in terms of continuous transformation*, Bull. Soc. Math. France 50 (1922), 119–134.
513. W. Wirtinger, *Einige Anwendungen der Euler–Maclaurin'schen Summenformel, insbesondere auf eine Aufgabe von Abel*, Acta Math. 26 (1902), 255–271.
514. K. Yosida, *Functional Analysis*, Springer, Berlin, 1980.
515. W.H. Young, *The fundamental theorems of differential calculus*, Cambridge Tracts No. 11, Cambridge, 1910.
516. W.H. Young, *On classes of summable functions and their Fourier series*, Proc. Royal Soc. (A) 87 (1912), 225–229.

517. W.H. Young, *The progress of mathematical analysis in the 20th century*, Proc. London Math. Soc. (2) 24 (1926), 421–434.

518. Z. Zahorski, *Sur l'ensemble des points de non-dérivabilité d'une fonction continue*, Bulletin de la S. M. F. 74 (1946), 147–178.

519. M. Zorn, *A remark on a method in transfinite algebra*, Bull. Amer. Math. Soc. 41 (1935), 667–670.

Teaching Suggestions

We mention some main differences from conventional textbooks.

Topology

- It would be more logical to start with topological spaces, and then to treat successively the more special metric and normed spaces. Our teaching experience shows that the present treatment is easier for the majority of students. Incidentally, this reflects the historical evolution of the subject.
- The very short and elegant proof of the completion of metric spaces (p. 21) does not seem to be well-known.
- The simple proof of the Cauchy–Schwarz inequality (p. 67) is not well-known either.
- Baire's theorem (p. 17) is not used in this book (except in Exercise 1.14, p. 34), but is very important in Functional Analysis; see, e.g., [285].
- General topological spaces are rarely used in this book, but they are important, for example, when describing weak convergence in Functional Analysis; see, e.g., [285] again.

Differential Calculus

- The general framework of arbitrary normed spaces *simplifies* the statement and proof of most theorems. The readers are encouraged, however, to consider systematically the special cases of $\mathbb{R}$ and $\mathbb{R}^n$.
- Carathéodory's equivalent definition of the derivative (formula (4.3), p. 98) simplifies the proof of Propositions 4.2 and 4.10 on composite functions, on the relationship between total and partial derivatives, and of the inverse function theorem 7.7 (pp. 101, 111 and 179).
- Applying Proposition 3.19 (p. 86) we obtain short and transparent proofs for vectorial versions of the mean value theorems and of Taylor's formula. Moreover, we obtain optimal results which are sharper than usual. When dealing with $\mathbb{R}^n$ or with Hilbert spaces, Theorem 3.20 may be avoided by using the simple Euclidean case of Proposition 3.19.

V. Komornik, *Topology, Calculus and Approximation*, Springer
Undergraduate Mathematics Series, DOI 10.1007/978-1-4471-7316-8

- We give a simple alternative proof of an important special case of the Schwarz–Young theorem (p. 123).
- The proof of the local Lipschitz continuity of convex functions (p. 136) is not well-known either.
- Before treating the general case, we give short and transparent proofs of the scalar case of the implicit function theorem and of the Lagrange multiplier theorem (Sects. 7.1 and 7.2, pp. 165 and 171). This special case suffices for many applications: we illustrate this by Cauchy's proof of the diagonalizability of symmetric matrices.

Approximation Methods

- We present numerical analysis by starting from simple problems and arriving at important general theorems in a natural way. This approach enables us to present several beautiful classical results, which had almost been forgotten.
- As a curiosity, Sturm sequences appear unexpectedly in various places.
- A short and transparent geometric proof is given for Stirling's formula in Exercise 10.2, p. 263.
- We stress some fundamental and rarely taught results of Hungarian mathematicians (Segner, Kürschák, Fejér, Riesz, Erdős, Turán).

Subject Index

V. Komornik, *Topology, Calculus and Approximation*, Springer
Undergraduate Mathematics Series, DOI 10.1007/978-1-4471-7316-8

Name Index

V. Komornik, *Topology, Calculus and Approximation*, Springer Undergraduate Mathematics Series, DOI 10.1007/978-1-4471-7316-8