Christian Pötzsche

# Introduction to Functional Analysis

WS 2010/11
TU München

February 16, 2011

Christian Pötzsche
Centre for Mathematical Sciences
Munich University of Technology
Boltzmannstraße 3
D-85748 Garching
Germany

poetzsch@ma.tum.de
http://www-m12.ma.tum.de/poetzsche

# Preface

The modern theory of functional analysis can be traced back to the first half of the 20th century. According to [Lax02] the earliest text book on functional analysis was the one by Banach in 1932 (cf. [Ban32]), although prior contributions are also due to Hans Hahn, David Hilbert, Frigyes Riesz, Erhard Schmidt, Vito Volterra and in particular the polish school including Stefan Banach himself, Stanislav Mazur, Julius Schauder, Hugo Steinhaus, Stanislaw Ulam. It can be seen as a synthesis of two more classical areas, namely linear algebra and analysis. Since then it turned into a vital field of contemporary mathematics with various applications.

From linear algebra it draws the concepts of a vector space (linear space) and of a linear mapping (operator), which is typically given in matrix form

$$Tu := \begin{pmatrix} T_{11} & \dots & T_{d1} \\ \vdots & & \vdots \\ T_{d1} & \dots & T_{dd} \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_d \end{pmatrix}.$$

However, the linear spaces of interest in functional analysis are spaces of functions or sequences, and therefore infinite dimensional. The linear mappings of relevance act on such spaces and are typically differential operators, like for instance Sturm-Liouville operators

$$(Tu)(x) := -\frac{d}{dx}\left(p(x)\frac{du}{dx}(x)\right) + q(x)u(x)$$

or integral operators, like e.g. Fredholm operators

$$(Tu)(x) := \int_a^b k(x,y)u(y)\,dy.$$

Consequently, also linear differential and integral equations can be written in the abstract form

$$Tu = f,$$

as known for systems of linear equations from linear algebra.

Yet, differing from linear algebra also topological properties of the function spaces, as well as continuity properties of the linear operators under consideration are of crucial importance. This is where analysis comes into play with concepts like completeness, convergence or continuity. Such analytical tools become even more important when dealing with nonlinear operators.

Referring to the high level of generality and abstraction, applications of functional analysis are widespread and include mathematical areas like probability theory, and in particular integral or partial differential equations. A further important playground for functional analytical tools is numerical or computational mathematics. This is due to the fact that many problems can be posed as (linear) equations in function spaces. Yet, this is of little use if one is interested in an actual solution and numerical methods come into play. Taking rounding errors and their finite floating point arithmetic aside, computers can solve only linear equations in finite-dimensional equations. For this reason one has to find appropriate "approximations" of function spaces and for operators acting on them. In this sense, finite differences replace derivatives and differential operators, and finite elements or finite volumes approximate infinite-dimensional function spaces. We refer to [Col66] for related applications of functional analysis in numerical mathematics.

This course forms the basis of a two hours per week class for students in *Computational Mechanics* and *Computational Science and Engineering*. We develop some basic functional analytical tools and skills needed for a variational formulation of boundary value problems and the finite element method. Being addressed to students in engineering, our focus is directed towards understanding and insight, and not only mathematical rigor and abstraction. Such an emphasis also explains that we neglect various pillars of functional analysis, like the open mapping theorem or the principle of uniform boundedness.

We close this preface with a hint to the related literature. As excellent introduction to the field of functional analysis we recommend [NS82] — mathematically inclined readers might consult [Con90, Yos80]. Finally, the monograph [LV03] has a focus on applications in mechanics.

München, February 16, 2011 Christian Pötzsche

# Contents

# Chapter 1
# Basic structures

This first chapter introduces some fundamental concepts for mathematics as a whole and in particular for functional analysis. A much more comprehensive approach can be found in [NS82, pp. 2ff, Chapter 1] regarding proof methods, and in [NS82, pp. 11ff, Chapter 2] concerning the preliminaries on sets and functions. In the first place, on an abstract level we generalize essential properties of the well-known spaces $\mathbb{R}^d$ or $\mathbb{C}^d$ in order to make them work in much more general settings, like function spaces.

Let $X, Y$ be sets. A relation $f : X \to Y$, $x \mapsto f(x)$ which assigns to every $x \in X$ exactly one $y = f(x) \in Y$ is called a *function* or synonymously a *map* or a *mapping*. In this context, we denote the set $X$ as *domain* and $Y$ as *codomain*, while

$$f(X) := \{y \in Y \mid \text{there exists a } x \in X \text{ with } y = f(x)\} \subseteq Y$$

as the *image* or the *range* of $f$ (cf. Fig. 1.1). For a given subset $X_0 \subseteq X$, the function

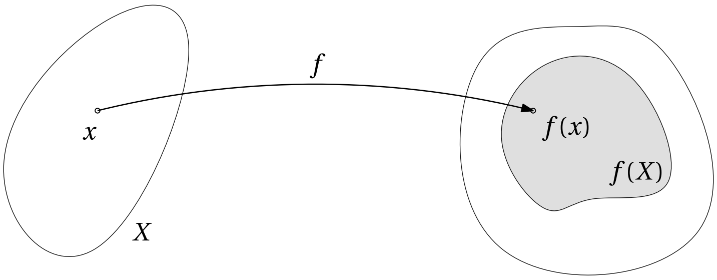

**Fig. 1.1** Domain $X$, codomain $Y$ and image $f(X)$ of a mapping $f : X \to Y$

$g : X_0 \to Y$, $g(x) := f(x)$ is called the *restriction* of $f$ to $X_0$ and will be denoted by $f|_{X_0}$. Given a superset $X_1 \supseteq X$, a function $F : X_1 \to Y$ is called *extension* of $f$ to $X_1$, provided one has $F|_X = f$.

In classical analysis the sets $X, Y$ are typically subsets of $\mathbb{R}^d$ or $\mathbb{C}^d$. In functional analysis, however, $X, Y$ consist of functions itself.

## 1.1 Metric spaces

In mathematics a "space" is typically a set equipped with an additional structure. For instance, the notion of a metric is an abstraction of the naive concept of distance known from the familiar spaces $\mathbb{R}^2$ (the plane) or $\mathbb{R}^3$. A set on which we can measure distances, is called a metric space.

**Definition 1.1.1** (metric space)**.** Let $X$ be a nonempty set. If $d : X \times X \to \mathbb{R}$ is a mapping satisfying the properties

(i) $d(x, y) = 0 \Leftrightarrow x = y$,
(ii) $d(x, y) \leq d(x, z) + d(y, z)$ (*triangle inequality*)

for all $x, y, z \in X$, then $d$ is called a *metric* and the pair $(X, d)$ a *metric space*.

*Remark* 1.1.2 (product metric). If $(X, d_X)$ and $(Y, d_Y)$ are metric spaces, then also the *cartesian product* $X \times Y := \big\{(x, y) : x \in X,\ y \in Y\big\}$ is a metric space by means of the *product metric* $D\big((x_1, y_1), (x_2, y_2)\big) := d_X(x_1, x_2) + d_Y(y_1, y_2)$ for all $x_1, x_2 \in X$ and $y_1, y_2 \in Y$.

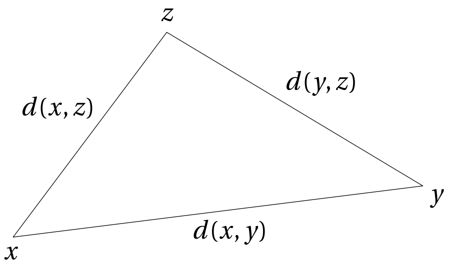

**Fig. 1.2** Triangle inequality from Def. 1.1.1(ii)

*Example* 1.1.3 (discrete metric). Every nonempty set $X$ trivially becomes a metric space by virtue of the mapping $d : X \times X \to \{0, 1\}$,

$$d(x, y) := \begin{cases} 1, & x \neq y, \\ 0, & x = y; \end{cases}$$

one speaks of the *discrete metric*.

The above basic conditions on a metric yield further natural properties:

**Corollary 1.1.4** (properties of a metric)**.** *A metric $d : X \times X \to \mathbb{R}$ fulfills*

*(a)* $d(x, y) \geq 0$,

*(b)* $d(x, y) = d(y, x)$ *(*symmetry*),*
*(c)* $\big|d(x,z) - d(y,w)\big| \leq d(x, y) + d(z, w)$ *(*quadrangle inequality*)*

*for all* $w, x, y, z \in X$.

The quadrangle inequality allows geometrical interpretations relating the lengths of sides and diagonals for arbitrary quadrangles.

*Proof.* Let $w, x, y, z \in X$.

(a) From the properties (i) and (ii) of Def. 1.1.1 we obtain

$$0 = d(x, x) \leq d(x, y) + d(x, y) = 2d(x, y)$$

and consequently $d(x, y) \geq 0$.

(b) The two relations

$$d(x,y) \overset{(ii)}{\leq} d(x,x) + d(y,x) \overset{(i)}{=} d(y,x), \qquad d(y,x) \overset{(ii)}{\leq} d(y,y) + d(x,y) \overset{(i)}{=} d(x,y)$$

immediately imply $d(x, y) = d(y, x)$.

(c) Using the triangle inequality (ii) and the symmetry shown above, one has

$$\begin{aligned} d(x,z) - d(y,w) &\leq d(x,y) + d(z,y) - d(y,w) \\ &\leq d(x,y) + d(y,w) + d(z,w) - d(y,w) = d(x,y) + d(z,w), \\ d(y,w) - d(x,z) &\leq d(y,x) + d(x,w) - d(x,z) \\ &\leq d(y,x) + d(x,z) + d(z,w) - d(x,z) = d(x,y) + d(z,w), \end{aligned}$$

which yields the quadrangle inequality. □

*Example* 1.1.5. Each one of the mappings $d_1, d_2, d_\infty : S \times S \to \mathbb{R}$,

$$d_1(x,y) := \sum_{j=1}^{d} \big|x_j - y_j\big|, \quad d_2(x,y) := \sqrt{\sum_{j=1}^{d} \big|x_j - y_j\big|^2}, \quad d_\infty(x,y) := \max_{j=1}^{d} \big|x_j - y_j\big|$$

defines a metric on an arbitrary nonempty subset $S \subseteq \mathbb{R}^d$. The same statement holds for $\mathbb{R}^d$ replaced by $\mathbb{C}^d$.

Further results on and examples of metric spaces can be found in [NS82, pp. 43ff, Chapter 3].

*Exercises* 1.1.6. Given a nonempty set $S$ solve the following problems:

(1) Show that the discrete metric defined in Ex. 1.1.3 fulfills the conditions of Def. 1.1.1.

(2) The *modulus*[1] of a complex number $z \in \mathbb{C}$ is given by $|z| := \sqrt{z\overline{z}}$.[2] Verify that the function $d(z, w) := |z - w|$ defines a metric on every subset $S \subseteq \mathbb{C}$.

(3) Given the metrics from Ex. 1.1.5, can you find real constants $c_i, C_i > 0$ such that the following relations hold for all $x, y \in S \subseteq \mathbb{R}^n$ or $\mathbb{C}^n$,

$$c_1 d_1(x,y) \le d_2(x,y) \le C_1 d_1(x,y),$$
$$c_2 d_1(x,y) \le d_\infty(x,y) \le C_2 d_1(x,y),$$
$$c_3 d_2(x,y) \le d_\infty(x,y) \le C_3 d_2(x,y)?$$

## 1.2 Normed spaces

We continue to mimic structures well-known from the spaces $\mathbb{R}^2$ or $\mathbb{R}^3$. Having a metric available means that we have equipped a set with a *metric structure*.

Now we are concerned with a so-called *algebraic structure*. In doing so, one can define algebraic operations like an addition or a (scalar) multiplication on more general sets, yielding the concept of a linear space. Therefore, in this section we introduce some notions which originally stem from linear algebra and refer to [NS82, pp. 159ff, Chapter 4] for a general survey. Later we will define a metric on linear spaces, which is compatible with the algebraic operations we are about to define. This means we have combined both metric and algebraic structure.

Let $\mathbb{K}$ denote one of the fields $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$.

**Definition 1.2.1** (linear space)**.** A nonempty set $X$ is called a *linear space* or a *vector space* over the *field* $\mathbb{K}$, provided there exists

(i) an *addition* $+ : X \times X \to X$, $(x, y) \mapsto x + y$ with the properties

$$\begin{aligned} \exists 0 \in X : x + 0 &= x && \text{(existence of zero vector)},\\ \exists -x \in X : x + (-x) &= 0 && \text{(existence of inverse vector)},\\ x + y &= y + x && \text{(commutativity)},\\ x + (y + z) &= (x + y) + z && \text{(associativity)}, \end{aligned}$$

(ii) a *scalar multiplication* $\cdot : \mathbb{K} \times X \to X$, $(\alpha, x) \mapsto \alpha \cdot x$ with the properties

$$\begin{aligned} 1 \cdot x &= x, \\ (\alpha + \beta) \cdot x &= \alpha \cdot x + \beta \cdot x && \text{(distributitvity)},\\ \alpha \cdot (x + y) &= \alpha \cdot x + \alpha \cdot y && \text{(distributitvity)},\\ \alpha \cdot (\beta \cdot x) &= (\alpha\beta) \cdot x \end{aligned}$$

[1] the modulus of a real number is called its *absolute value*

[2] the *complex conjugate* of a complex number $z = x + iy$ is given by $\overline{z} = x - iy$

for all $x, y, z \in X$ and $\alpha, \beta \in \mathbb{K}$. In this context, the elements of $\mathbb{K}$ are called *scalars*, and the elements of $X$ are called *vectors*.

Depending on the field $\mathbb{K}$, one speaks of a *real* ($\mathbb{K} = \mathbb{R}$) or a *complex* linear space ($\mathbb{K} = \mathbb{C}$).

*Remark* 1.2.2. (1) One usually abbreviates $\alpha x := \alpha \cdot x$.

(2) From Def. 1.2.1 one easily deduces the relations[3]

$$0 \cdot x = 0, \qquad\qquad (-1) \cdot x = -x \quad \text{for all } x \in X.$$

*Example* 1.2.3 (trivial linear spaces). The smallest linear space is $X = \{0\}$ — we speak of the *trivial space*. Moreover, also $\mathbb{K}$ is a linear space over itself, where the scalar multiplication is given by the usual product on $\mathbb{K}$.

*Example* 1.2.4 ($d$-tuples). Let $d \in \mathbb{N}$. The set of (real or complex) $d$-tuples

$$\mathbb{K}^d := \{(x_1, \dots, x_d) : x_1, \dots, x_d \in \mathbb{K}\}$$

becomes a linear space by virtue of the addition resp. scalar multiplication

$$x + y = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} + \begin{pmatrix} y_1 \\ \vdots \\ y_d \end{pmatrix} := \begin{pmatrix} x_1 + y_1 \\ \vdots \\ x_d + y_d \end{pmatrix}, \qquad \alpha \cdot x = \alpha \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} := \begin{pmatrix} \alpha x_1 \\ \vdots \\ \alpha x_d \end{pmatrix}$$

for all $x, y \in \mathbb{K}^d$ and $\alpha \in \mathbb{K}$. Here, the zero vector reads as $0 = (0, \dots, 0)$.

*Example* 1.2.5 ($m \times n$-matrices). Let $m, n \in \mathbb{N}$ and $\mathbb{K}^{m \times n}$ denote the set of matrices with $m$ rows, $n$ columns and entries from a field $\mathbb{K}$. Using a component wise definition of the addition and scalar multiplication makes $\mathbb{K}^{m \times n}$ a linear space.

*Example* 1.2.6 (polynomials). Let $\Omega \subseteq \mathbb{K}$ be nonempty. A function $p : \Omega \to \mathbb{K}$ is called a *polynomial* over $\Omega$, if it allows the representation

$$p(x) = \sum_{j=0}^{n} p_j x^j \quad \text{for all } x \in \Omega$$

with given $n \in \mathbb{N}_0$ and so-called *coefficients* $p_0, \dots, p_n \in \mathbb{K}$. Depending on $\mathbb{K}$ one speaks of a *real* or a *complex polynomial*. The *degree* $\deg p$ of a polynomial $p$ is the largest $j \in \mathbb{N}_0$ such that $p_j \neq 0$. The set of all polynomials $P(\Omega)$ is a linear space over $\mathbb{K}$. Here, the sum of two polynomials $p, q \in P(\Omega)$ is given by

$$(p + q)(x) := \sum_{j=0}^{n} p_j x^j + \sum_{j=0}^{n} q_j x^j = \sum_{j=0}^{n} (p_j + q_j) x^j \quad \text{for all } x \in \Omega$$

[3] the first relation follows from $x = 1 \cdot x = (1+0) \cdot x = 1 \cdot x + 0 \cdot x = x + 0 \cdot x$ by subtraction of $x$, while the second relation is a consequence of $(-1) \cdot x + x = (-1) \cdot x + 1 \cdot x = (-1+1) \cdot x = 0 \cdot x = 0$.

and the scalar multiplication of $\alpha \in \mathbb{K}$ and $p \in P(\Omega)$ reads as

$$(\alpha p)(x) := \alpha \sum_{j=0}^{n} p_j x^j = \sum_{j=0}^{n} \alpha p_j x^j \quad \text{for all } x \in \Omega.$$

The zero vector in $P(\Omega)$ is the polynomial whose coefficients identically vanish, i.e. the null function $0 : x \mapsto 0$.

A subspace of a linear space is a subset, which itself is a linear space again.

**Definition 1.2.7** (subspace)**.** Let $X$ be a linear space over $\mathbb{K}$. A subset $Y \subseteq X$ is called a *subspace* of $X$, if one has

$$\alpha x + \beta y \in Y \quad \text{for all } x, y \in Y,\ \alpha, \beta \in \mathbb{K}.$$

In this case one also says $Y$ is *algebraically closed*.

*Remark* 1.2.8. (1) Every linear space $X$ has the two trivial subspaces $\{0\}$ and $X$. Hence, a subspace must contain the element 0.

(2) A subspace is also a linear space. The intersection $Y_1 \cap Y_2$ of two subspaces $Y_1, Y_2 \subseteq X$ is also a subspace, as well as the sum

$$Y_1 + Y_2 := \{y_1 + y_2 \in X : y_1 \in Y_1,\ y_2 \in Y_2\},$$

while the union $Y_1 \cup Y_2$ of subspaces needs not to be a subspace anymore.

*Example* 1.2.9 (polynomials). For $n \in \mathbb{N}_0$, $\Omega \subseteq \mathbb{K}$, the polynomials of maximal degree $n$ over $\Omega$ given by $P_n(\Omega) := \{p \in P(\Omega) : \deg p \le n\}$ are a subspace of $P(\Omega)$. However, the set $\{p \in P(\Omega) : \deg p = n\}$ is not a subspace for $n > 0$, since it does not contain the zero polynomial 0.

**Definition 1.2.10** (basis)**.** Let $X$ be a nontrivial linear space over $\mathbb{K}$. A subset $B \subseteq X$ is called a *basis* of $X$, if

(i) $B$ is a *generating system* of $X$, i.e. for every $x \in X$ there exist an $n \in \mathbb{N}$, scalars $\alpha_1, \ldots, \alpha_n \in \mathbb{K}$ and vectors $x_1, \ldots, x_n \in B$ such that

$$x = \sum_{j=1}^{n} \alpha_j x_j, \tag{1.2a}$$

(ii) $B$ is *minimal*, i.e. for every $x \in B$ the set $B \setminus \{x\}$ is not a generating system.

The *dimension* $\dim X$ of $X$ is the cardinality of $B$.

An expression of the form (1.2a) is called a *linear combination* of $x_1$ to $x_n$.

*Remark* 1.2.11. (1) A basis cannot contain the zero vector.

(2) Note that the basis of a linear space is not uniquely determined. Indeed, for many numerical schemes it becomes important to choose an appropriate basis of the solution space in order to make a problem as simple as possible. However, the representation (1.2a) of an element $x \in X$ w.r.t. a given basis is unique.

(3) Every basis of a linear space has the same cardinality.[4] For this reason one says the notion of a dimension $\dim X$ of $X$ is *well-defined*, i.e. independent of the particular basis. We will see in Ex. 1.2.14 that a basis can be infinite.

*Example* 1.2.12. The unit vectors $e_1 = (1,0,\dots,0), \dots, e_d := (0,\dots,0,1)$ form a basis of $\mathbb{K}^d$ — one speaks of the *canonical basis*. Another basis of $\mathbb{K}^d$ consists of the vectors $(1,0,0,\dots,0), (1,1,0\dots,0), \dots, (1,1,1\dots,1)$. Similarly, the matrices $E_{jk} \in \mathbb{K}^{m\times n}$ given by

$$E_{11} := \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}, E_{12} := \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}, \dots, E_{mn} := \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}$$

form a basis of $\mathbb{K}^{m\times n}$. We consequently obtain

$$\dim \mathbb{K}^d = d, \qquad \dim \mathbb{K}^{m\times n} = mn.$$

*Example* 1.2.13 (subsets of $\mathbb{K}^{n\times n}$). (1) A matrix $A \in \mathbb{K}^{n\times n}$ is called *symmetric*, if $A = A^T$ holds.[5] The set of symmetric matrices

$$S_n(\mathbb{K}) := \left\{ A \in \mathbb{K}^{n\times n} : A = A^T \right\}$$

is a subspace of $\mathbb{K}^{n\times n}$. With the matrices $E_{jk}$ introduced in Ex. 1.2.12, a basis of $S_n(\mathbb{K})$ is given by the set $\left\{ E_{jk} + E_{jk}^T \right\}_{1\le j\le k\le n}$ and consequently $\dim S_n = \frac{n(n+1)}{2}$.

(2) The *Hermitian matrices*[6] $H_n := \left\{ A \in \mathbb{C}^{n\times n} : A = A^* \right\}$ form a subset, but not a subspace of $\mathbb{C}^{n\times n}$. Also the *invertible matrices* $GL_n(\mathbb{K}) := \{A \in \mathbb{K}^{n\times n} : \det A \neq 0\}$ are a subset, but not a subspace of $\mathbb{K}^{n\times n}$, since $0 \notin GL_n(\mathbb{K})$.

*Example* 1.2.14 (polynomials). Let $\Omega \subseteq \mathbb{R}$ be infinite. A basis of $P_n(\Omega)$ is given by the *monomials*, i.e. functions $e_k : \Omega \to \mathbb{K}$, $e_k(x) := x^k$ with $k \in \{0,\dots,n\}$. From this we observe that $\dim P_n(\Omega) = n+1$. However, the relation

$$P(\Omega) = \bigcup_{n\in\mathbb{N}} P_n(\Omega)$$

shows us $\dim P(\Omega) = \infty$ and $P(\Omega)$ is an infinite-dimensional space.

Our next goal is to measure distances on general linear spaces, as well as on their subsets, which leads us to the concept of a norm.

[4] naively, the *cardinality* of a set is the number of its elements

[5] the *transpose* of a matrix $A = (a_{jk})$ is given by $A^T = (a_{kj})$

[6] for $A = (a_{jk}) \in \mathbb{C}^{n\times n}$ one defines $A^* := (\bar{a}_{kj})$

**Definition 1.2.15** (normed space)**.** Let $X$ be a linear space over $\mathbb{K}$. Provided $\|\cdot\| : X \to \mathbb{R}$ is a mapping satisfying the properties

(i) $\|x\| = 0 \Leftrightarrow x = 0$,
(ii) $\|\alpha x\| = |\alpha| \, \|x\|$ (*positive homogeneity*),
(iii) $\left\|x + y\right\| \le \|x\| + \left\|y\right\|$ (*triangle inequality*)

for all $x, y \in X$, $\alpha \in \mathbb{K}$, then $\|\cdot\|$ is called a *norm* on $X$ and the pair $(X, \|\cdot\|)$ is called a *normed space.*

*Remark* 1.2.16. (1) Sometimes it is appropriate to write $\|\cdot\|_X$ for the norm on $X$. In accordance with Rem. 1.1.2 the product $X \times Y$ of two normed spaces $X, Y$ is also a normed space, whose norm (the so-called *product norm*) is given by

$$\left\|(x, y)\right\|_{X \times Y} := \|x\|_X + \left\|y\right\|_Y \quad \text{for all } x \in X,\ y \in Y.$$

(2) Every subspace of a normed space $(X, \|\cdot\|)$ inherits its norm from $X$. Every subset $S \subseteq X$ of a normed space $(X, \|\cdot\|)$ becomes a metric space $(S, d)$ by virtue of the metric $d(x, y) := \left\|x - y\right\|$. Thus, norms measure distances in linear spaces. Nevertheless, not every metric on a linear space is induced by a norm as above.

*Example* 1.2.17. (1) The modulus $|\cdot| : \mathbb{K} \to \mathbb{R}$ is a norm on $\mathbb{K}$.

(2) Each one of the mappings $\|\cdot\|_1, \|\cdot\|_2, \|\cdot\|_\infty : \mathbb{K}^d \to \mathbb{R}$,

$$\|x\|_1 := \sum_{j=1}^{d} \left|x_j\right|, \qquad \|x\|_2 := \sqrt{\sum_{j=1}^{d} \left|x_j\right|^2}, \qquad \|x\|_\infty := \max_{j=1}^{d} \left|x_j\right|$$

defines a norm on $\mathbb{K}^d$. Here, $\|\cdot\|_1$ is called the 1-norm, $\|\cdot\|_2$ the *Euclidean norm* and $\|\cdot\|_\infty$ the *maximum norm.* These norms reduce to the modulus $|\cdot|$ in case $d = 1$ and give rise to the metrics defined in Ex. 1.1.5. In Fig. 1.3 we have illustrated the corresponding *closed unit balls*

$$\bar{B}_j := \left\{x \in \mathbb{K}^d : \|x\|_j \le 1\right\} \quad \text{for } j \in \{1, 2, \infty\}. \tag{1.2b}$$

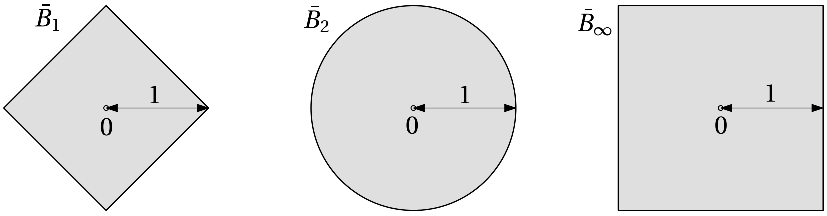

**Fig. 1.3** Closed unit balls $\bar{B}_j$ in $\mathbb{R}^2$ with respect to the norms $\|\cdot\|_1$ (left), $\|\cdot\|_2$ (middle) and $\|\cdot\|_\infty$ (right) from Ex. 1.2.17(2)

**Corollary 1.2.18** (properties of a norm)**.** *A norm $\|\cdot\| : X \to \mathbb{R}$ fulfills*

*(a)* $\|x\| \geq 0$*,*
*(b)* $\big|\|x\| - \|y\|\big| \leq \|x \pm y\| \leq \|x\| + \|y\|$ *(*general triangle inequality*)*

*for all* $x, y \in X$*.*

We refer to Fig. 1.4 for an illustration of the (generalized) triangle inequality.

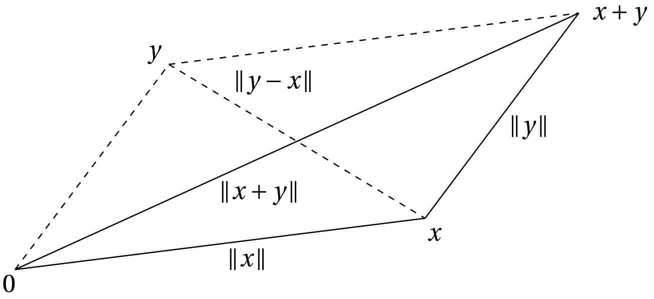

**Fig. 1.4** Triangle inequality from Cor. 1.2.18(ii) in the form $\|x \pm y\| \leq \|x\| + \|y\|$

*Proof.* Let $x, y \in X$.

(a) From Def. 1.2.15 we obtain for $y = -x$ that

$$0 \overset{(i)}{=} \|x + (-x)\| \overset{(iii)}{\leq} \|x\| + \|-x\| \overset{(ii)}{=} 2\,\|x\|$$

and thus $\|x\| \geq 0$.

(b) Above all, we get $\|x - y\| = \|x + (-y)\| \overset{(iii)}{\leq} \|x\| + \|-y\| \overset{(ii)}{=} \|x\| + \|y\|$. Since every norm induces a metric via $d(x,y) = \|x - y\|$ we have the quadrangle inequality from Cor. 1.1.4(c) at our disposal. Setting $z = w = 0$ yields

$$\big|\|x\| - \|y\|\big| = \big|d(x,0) - d(y,0)\big| \leq d(x,y) + d(0,0) = \|x - y\|$$

and if we replace $y$ by $-y$ one finally obtains the remaining estimate

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

which finishes our proof. □

*Exercises* 1.2.19. Solve the following problems:

(1) Make a sketch of the sets

$$Y_1 := \left\{(x,y) \in \mathbb{R}^2 : y = 0\right\}, \qquad Y_2 := \left\{(x,y) \in \mathbb{R}^2 : y = x\right\},$$

$$Y_3 := \{(x,y) \in \mathbb{R}^2 : |y| = |x|\},$$

decide whether they are subspaces of $\mathbb{R}^2$, and if so determine a basis! Moreover, sketch $Y_1 \cap Y_2$, $Y_1 \cup Y_2$ and $Y_1 + Y_2$.

(2) The *Legendre polynomials* $p_n : \mathbb{R} \to \mathbb{R}$ are defined by

$$p_n(x) := \frac{1}{2^n n!} \frac{d^n}{dx^n} (x^2 - 1)^n \quad \text{for all } n \in \mathbb{N}_0.$$

Show that $\{p_0, p_1, p_2\}$ is a basis of $P_2(\mathbb{R})$.

(3) Let $(X, \|\cdot\|)$ be a normed space and suppose $\varphi : [0,\infty) \to [0,\infty)$ is a continuously differentiable function satisfying $\varphi(0) = 0$ such that $\varphi$ is strictly increasing and $\varphi'$ is nondecreasing.[7] Show that the mapping $d : X \times X \to \mathbb{R}$, $d(x,y) := \varphi(\|x - y\|)$ is a metric on $X$ (cf. Rem. 1.2.16(2)) and give examples for functions $\varphi$ fulfilling the above properties.

(4) Which of the mappings $\|\cdot\|, \|\cdot\|' : \mathbb{R}^{n\times n} \to \mathbb{R}$ given by

$$\|T\| := \sum_{j=1}^{n} \sum_{k=1}^{n} |T_{jk}|, \qquad \|T\|' := |\det T|$$

defines a norm on $\mathbb{R}^{n\times n}$? Here, $\det T$ denotes the *determinant* of $T \in \mathbb{R}^{n\times n}$.

## 1.3 Inner product spaces

So far we have considered general normed spaces $X$, on which we had algebraic operations $+, \cdot$ available, plus the feature that we can measure distances using norms $\|\cdot\|$. However, in order to obtain a geometric intuition as in the familiar 2- or 3-dimensional Euclidean geometry, also the additional concept of orthogonality or perpendicularity is desirable.

This can be motivated as follows: Given a subspace $Y \subseteq X$ and a point $x \in X$, is there a (unique) point $y_0 \in Y$ such that $\|x - y_0\| = \min_{y \in Y} \|x - y\|$ (see Figure 1.5)? Such a $y_0$ is called *orthogonal projection* of $x$ onto $Y$. The first ingredient to tackle this problem are inner products, which will also give rise to a natural norm (cf. Prop. 1.3.4). Yet, a complete solution has to be postponed until Sect. 2.4.

**Definition 1.3.1** (inner product space)**.** Let $X$ be a linear space $X$ over $\mathbb{K}$. An *inner product* on $X$ is a mapping $\langle\cdot,\cdot\rangle : X \times X \to \mathbb{K}$ with the properties

(i) $\langle \alpha x + \beta y, z\rangle = \alpha \langle x, z\rangle + \beta \langle y, z\rangle$ (*linearity in the first argument*),

(ii) $\langle x, y\rangle = \overline{\langle y, x\rangle}$ (*conjugate symmetry*),

(iii) $\langle x, x\rangle \geq 0$ with equality only for $x = 0$ (*positive definiteness*)

[7] this means for any $s \leq t$ one has $\varphi'(s) \geq \varphi'(t)$

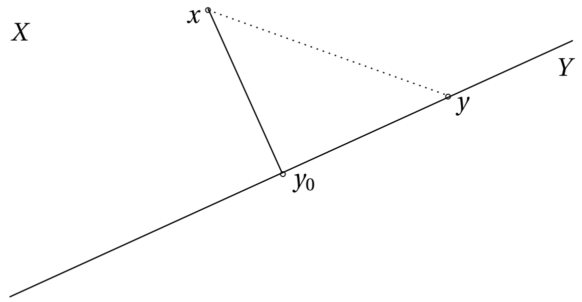

**Fig. 1.5** Triangle inequality from Cor. 1.2.18(ii) in the form $\|x \pm y\| \le \|x\| + \|y\|$

for all $x, y, z \in X$ and $\alpha, \beta \in \mathbb{K}$. A linear space with an inner product is called *inner product space* $(X, \langle\cdot,\cdot\rangle)$ or a *pre-Hilbert space*.

One also uses the notion *scalar product* instead of inner product. However, do not confuse the scalar product $\langle x, y\rangle \in \mathbb{K}$ of two vectors $x, y \in X$ with the scalar multiplication $\alpha \cdot x \in X$ of a scalar $\alpha \in \mathbb{K}$ with a vector $x \in X$ from Def. 1.2.1.

*Remark* 1.3.2. (1) Due to its conjugate symmetry an inner product always fulfills $\langle x, x\rangle \in \mathbb{R}$, while the positive definiteness guarantees $\langle x, x\rangle > 0$ for all $x \neq 0$. Moreover, one has $\langle x, 0\rangle = \langle 0, x\rangle = 0$ for all $x \in X$.

(2) An inner product is *semilinear in the second argument*, i.e.

$$\langle x, \alpha y + \beta z\rangle = \overline{\alpha}\langle x, y\rangle + \overline{\beta}\langle x, z\rangle^{8} \quad \text{for all } x, y, z \in X,\ \alpha, \beta \in \mathbb{K}. \tag{1.3a}$$

(3) Two elements $x, y \in X$ are called *orthogonal*, if $\langle x, y\rangle = 0$ holds. A basis $B$ of an inner product space $X$ is called *orthogonal basis*, if $\langle x, y\rangle = 0$ for all $x, y \in B$, $x \neq y$ and *orthonormal basis*, if

$$\langle x, y\rangle = \begin{cases} 1, & x = y, \\ 0, & x \neq y \end{cases} \quad \text{for all } x, y \in B.$$

Hence, every orthonormal basis is an orthogonal basis. For an orthonormal basis $B$, the coefficients $\alpha_1, \ldots, \alpha_n \in \mathbb{K}$ in a linear combination (1.2a) of $x \in X$ can be obtained from

$$\alpha_k = \sum_{j=1}^{n} \alpha_j \langle x_j, x_k\rangle = \left\langle \sum_{j=1}^{n} \alpha_j x_j, x_k \right\rangle = \langle x, x_k\rangle \quad \text{for all } 1 \le k \le n.$$

*Example* 1.3.3. (1) Let $y^* \in \mathbb{K}^{1\times d}$ denote the conjugate transpose of $y \in \mathbb{K}^d$. With this, on the linear space $\mathbb{K}^d$ we can define an inner product by

[8] one has $\langle x, \alpha y + \beta z\rangle \overset{(ii)}{=} \overline{\langle \alpha y + \beta z, x\rangle} \overset{(i)}{=} \overline{\alpha\langle y, x\rangle + \beta\langle z, x\rangle} = \overline{\alpha}\overline{\langle y, x\rangle} + \overline{\beta}\overline{\langle z, x\rangle} \overset{(ii)}{=} \overline{\alpha}\langle x, y\rangle + \overline{\beta}\langle x, z\rangle$ for all $x, y, z \in X$ and $\alpha, \beta \in \mathbb{K}$ from the properties given in Def. 1.3.1

$$\langle x,y\rangle := y^* x = \sum_{j=1}^{d} \overline{y_j} x_j \quad \text{for all } x,y \in \mathbb{K}^d \tag{1.3b}$$

and an *Euclidean space* is $\mathbb{R}^d$ equipped with the above inner product. Given this inner product, the unit vectors $e_1 = (1,0,\dots,0)^*$, ..., $e_d := (0,\dots,0,1)^*$ defined in Ex. 1.2.12 form an orthonormal basis of $\mathbb{K}^d$. Also the eigenvectors of a symmetric matrix in $\mathbb{R}^{n\times n}$ can be arranged to form an orthogonal basis of $\mathbb{R}^n$.

(2) For a symmetric positive-definite matrix $A \in \mathbb{R}^{d\times d}$, also $\langle x,y\rangle := y^* Ax$ is an inner product on $\mathbb{K}^d$.

Our next result ensures that every inner product space is a normed space, and in turn a metric space.

**Proposition 1.3.4** (natural norm)**.** *If* $(X,\langle\cdot,\cdot\rangle)$ *is an inner product space, then*

$$\|x\| := \sqrt{\langle x,x\rangle} \quad \textit{for all } x \in X$$

*defines a norm on* $X$.

*Remark* 1.3.5. (1) If there exist different norms on an inner product space, then the norm induced by the inner product is denoted as *natural norm.*

(2) Every nonempty subset $S \subseteq X$ of an inner product space $(X,\langle\cdot,\cdot\rangle)$ becomes a metric space $(S,d)$ with

$$d(x,y) := \sqrt{\langle x-y, x-y\rangle} \quad \text{for all } x,y \in S.$$

(3) The elements of an orthonormal basis $B$ have norm 1, since one has the relation $\langle x,x\rangle = \|x\|^2 = 1$ for all $x \in B$.

*Proof.* See Exercise 1.3.9(2). □

*Example* 1.3.6. The inner product $\langle x,y\rangle := \sum_{j=1}^d x_j y_j$ on the Euclidean space $\mathbb{R}^d$ induces the 2-norm as natural norm $\|x\|_2 = \sqrt{\sum_{j=1}^d |x_j|^2}$. Yet, we have seen in Ex. 1.2.17 that there exist other norms in $\mathbb{R}^d$ as well, which are not natural, though.

**Proposition 1.3.7** (Cauchy-Schwarz inequality)**.** *If* $(X,\langle\cdot,\cdot\rangle)$ *is an inner product space, then*

$$|\langle x,y\rangle| \le \|x\|\,\|y\| \quad \textit{for all } x,y \in X.$$

*Proof.* Let $x,y \in X$. Obviously, the claim holds for $y=0$ and we can assume $y \neq 0$ and consequently $\|y\| \neq 0$. Thus, we can define $\lambda := \frac{\langle x,y\rangle}{\|y\|^2}$ and obtain

$$\begin{aligned}
0 &\le \left\langle x-\lambda y, x-\lambda y\right\rangle = \left\langle x, x-\lambda y\right\rangle - \lambda\left\langle y, x-\lambda y\right\rangle \\
&\overset{(1.3a)}{=} \langle x,x\rangle - \overline{\lambda}\left\langle x,y\right\rangle - \lambda\left\langle y,x\right\rangle + \lambda\overline{\lambda}\left\langle y,y\right\rangle \\
&= \langle x,x\rangle - \frac{\overline{\left\langle x,y\right\rangle}}{\left\|y\right\|^2}\left\langle x,y\right\rangle - \frac{\left\langle x,y\right\rangle}{\left\|y\right\|^2}\left\langle y,x\right\rangle + \frac{\left\langle x,y\right\rangle}{\left\|y\right\|^2}\frac{\overline{\left\langle x,y\right\rangle}}{\left\|y\right\|^2}\left\langle y,y\right\rangle \\
&= \langle x,x\rangle - \frac{\left\langle x,y\right\rangle\overline{\left\langle x,y\right\rangle}}{\left\|y\right\|^2}.
\end{aligned}$$

This immediately implies $\left\langle x,y\right\rangle\overline{\left\langle x,y\right\rangle} \le \|x\|^2\left\|y\right\|^2$ and the claimed relation. □

A well-known result in Euclidean geometry is the parallelogram law (for this, see Fig. 1.6). It states that in a parallelogram the squares of the diagonals are equal twice the squares of the sides. Our next proposition states that such a property holds in general inner product spaces.

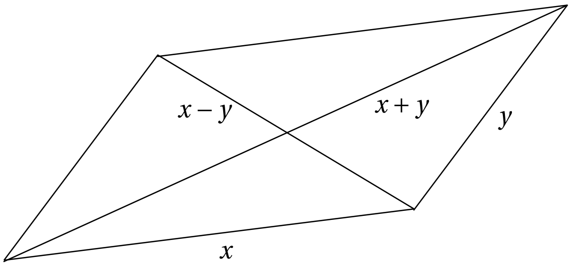

**Fig. 1.6** Parallelogram identity from Thm. 1.3.8(a)

**Proposition 1.3.8** (parallelogram identity). *(a) If* $(X,\langle\cdot,\cdot\rangle)$ *is an inner product space, then the* parallelogram identity *holds:*

$$\left\|x+y\right\|^2 + \left\|x-y\right\|^2 = 2\|x\|^2 + 2\left\|y\right\|^2 \quad \textit{for all } x,y\in X. \tag{1.3c}$$

*(b) Conversely, if* $(X,\|\cdot\|)$ *is a normed space over* $\mathbb{K}$ *on which* (1.3c) *holds, then* $X$ *is an inner product space by virtue of*

$$\left\langle x,y\right\rangle = \tfrac{1}{4}\left(\left\|x+y\right\|^2 - \left\|x-y\right\|^2\right), \quad \textit{if } \mathbb{K}=\mathbb{R},$$
$$\left\langle x,y\right\rangle = \tfrac{1}{4}\left(\left\|x+y\right\|^2 - \left\|x-y\right\|^2\right) + \tfrac{i}{4}\left(\left\|ix-y\right\|^2 - \left\|ix+y\right\|^2\right), \quad \textit{if } \mathbb{K}=\mathbb{C}.$$

*Proof.* (a) The proof is a straight forward computation using $\|x\|^2 = \langle x,x\rangle$.
(b) See [NS82, p. 276, Thm. 5.12.8]. □

*Exercises* 1.3.9. Solve the following problems:

(1) Let $(X, \langle\cdot,\cdot\rangle)$ be an inner product space. Show the *law of cosines*

$$\left\|x+y\right\|^2 = \|x\|^2 + \left\|y\right\|^2 + 2\Re\left\langle x,y\right\rangle \quad \text{for all } x,y \in X.$$

Can you conclude a *Pythagorean theorem*?

(2) Prove that the mapping $\|\cdot\|$ defined Prop. 1.3.4 is a norm. You will need the Cauchy-Schwarz inequality.

(3) Show that the *Legendre polynomials* $p_0, p_1, p_2$ defined in Exercise 1.2.19(2) are an orthogonal basis of $P_2[-1,1]$ with respect to the inner product

$$\left\langle p,q\right\rangle := \int_{-1}^{1} p(x)q(x)\,dx;$$

are they even an orthonormal basis? Are the monomials $e_0, e_1, e_2$ an orthogonal basis of $P_2[-1,1]$?

## 1.4 Classical function spaces

Let $\Omega$ be a (nonempty) set and $\mathbb{K}$ stand for one of the fields $\mathbb{R}$ or $\mathbb{C}$. We denote the set of all functions $u : \Omega \to \mathbb{K}$ by $F(\Omega)$. It is left to the interested reader to show that $F(\Omega)$ is a linear space over $\mathbb{K}$, w.r.t. the algebraic operators

$$(u+v)(x) := u(x)+v(x), \qquad (\alpha u)(x) := \alpha u(x) \quad \text{for all } u,v \in F(\Omega),\ \alpha \in \mathbb{K},$$

where the zero vector is given by the null function $0 : x \mapsto 0$ and the inverse vector to $u \in F(\Omega)$ reads as $(-u)(x) := -u(x)$ for $x \in \Omega$.

In this spirit, linear spaces whose elements are functions are called *function spaces*. Sequences are functions defined on the integers $\mathbb{Z}$ and so, a special case of general function spaces are *sequence spaces*, where $\Omega = \mathbb{Z}$ or $\Omega = \mathbb{N}$.

Since $F(\Omega)$ is a very large space, it is hard to introduce a meaningful metric or norm. Thus, in the following, we investigate certain relevant subspaces of $F(\Omega)$.

**Important remark**: Dealing with function spaces it is important to distinguish between *functions* and their *arguments*. An element of $F(\Omega)$ is a function $u : \Omega \to \mathbb{K}$ while the argument $u(x)$ (at the value $x \in \Omega$) is a real or complex number. Hence, speaking about a "function $u(x)$" makes no sense!

### *1.4.1 Bounded functions*

A function $u : \Omega \to \mathbb{K}$ is called *bounded*, if there exists a real $C \geq 0$ satisfying

$$|u(x)| \leq C \quad \text{for all } x \in \Omega \tag{1.4a}$$

and the smallest such $C \geq 0$ is denoted as *supremum* $\sup_{x\in\Omega}|u(x)| < \infty$.[9] For the subset of all bounded functions we write $B(\Omega)$.

*Example* 1.4.1. Every polynomial $p : \Omega \to \mathbb{K}$ over a bounded set $\Omega \subseteq \mathbb{K}$ is bounded. In particular, the function $u_1 : [a,b] \to \mathbb{R}$, $u_1(x) := x^2 - 1$ with $a < 0 < b$ is bounded with $\sup_{x\in[a,b]}|u_1(x)| = \max\{|a^2-1|, |b^2-1|, 1\}$. Also the function $u_2 : \mathbb{R} \to \mathbb{R}$, $u_2(x) := \arctan x$ is bounded with $\sup_{x\in\mathbb{R}}|u_2(x)| = \frac{\pi}{2}$. However, for example the functions $u_3 : \mathbb{R} \to \mathbb{R}$, $u_3(x) := x^2 - 1$ or $u_4 : (0,\infty) \to \mathbb{R}$, $u_4(x) := \frac{1}{x}$ are not bounded since $u_3(\mathbb{R}) = [-1,\infty)$ and $u_4((0,\infty)) = (0,\infty)$.

We refer to Fig. 1.7 for an illustration of the following

**Proposition 1.4.2.** *The set $B(\Omega)$ is a normed space with the norm*

$$\|u\| := \sup_{x\in\Omega}|u(x)|.$$

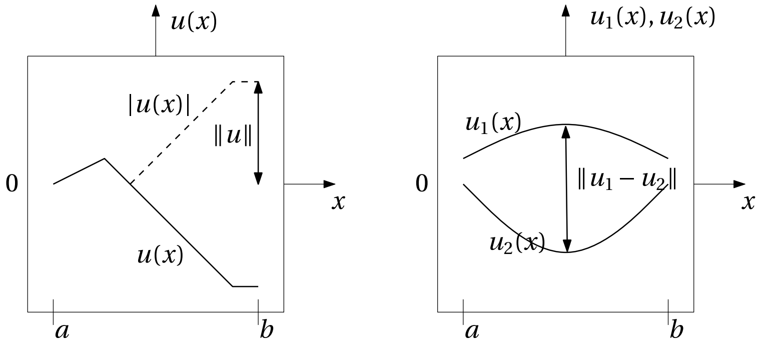

**Fig. 1.7** Supremum norm $\|u\|$ of a function $u \in B[a,b]$ (left) and distance $\|u_1 - u_2\|$ of two functions $u_1, u_2 \in B[a,b]$ (right)

*Proof.* Let $\alpha, \beta \in \mathbb{K}$ and $u_1, u_2 : \Omega \to \mathbb{K}$ be bounded functions, i.e. $u_1, u_2 \in B(\Omega)$. Then there exist reals $C_1, C_2 \geq 0$ with

$$|u_1(x)| \leq C_1, \qquad |u_2(x)| \leq C_2 \quad \text{for all } x \in \Omega.$$

(I) We first show that $B(\Omega)$ is a linear space. Thereto, it is sufficient to show that $B(\Omega)$ is a subspace of $F(\Omega)$, i.e. it is algebraically closed w.r.t. the addition and the scalar multiplication inherited from $F(\Omega)$. This means for all scalars $\alpha, \beta \in \mathbb{K}$ also the function $\alpha u_1 + \beta u_2 : \Omega \to \mathbb{K}$ is bounded, which follows from

$$|\alpha u_1(x) + \beta u_2(x)| \leq |\alpha||u_1(x)| + |\beta||u_2(x)| \leq |\alpha| C_1 + |\beta| C_2 \quad \text{for all } x \in \Omega$$

[9] one also denotes $C$ in (1.4a) as an *upper bound* of the function $|u(\cdot)| : \Omega \to [0,\infty)$ and obtains that the supremum is the smallest upper bound

Hence, we concluded $\alpha u_1 + \beta u_2 \in B(\Omega)$ and $B(\Omega)$ is a linear space.

(II) It remains to prove that $\|u\| := \sup_{x\in\Omega} |u(x)|$ defines a norm on $B(\Omega)$. It is clear that the null function $0 : x \mapsto 0$ has norm 0 and conversely. Due to

$$\|\alpha u_1\| = \sup_{x\in\Omega} |\alpha u_1(x)| = |\alpha| \sup_{x\in\Omega} |u_1(x)| = |\alpha| \, \|u_1\|$$

one has positive homogeneity. The triangle inequality in $B(\Omega)$ follows from

$$|u_1(x) + u_2(x)| \le |u_1(x)| + |u_2(x)| \le \|u_1\| + \|u_2\| \quad \text{for all } x \in \Omega,$$

when we pass over to the supremum over all $x \in \Omega$ in this inequality, yielding

$$\|u_1 + u_2\| = \sup_{x\in\Omega} |u_1(x) + u_2(x)| \le \|u_1\| + \|u_2\| .$$

Thus, $B(\Omega)$ is also a normed space. □

### 1.4.2 Continuous functions

From now on we prescribe a subset $\Omega \subseteq \mathbb{K}^d$. The space of all continuous functions $u : \Omega \to \mathbb{K}$ is denoted by $C(\Omega)$. Since continuous functions on compact[10] sets are bounded and achieve their maximum and minimum (see [NS82, p. 148, Thm. 3.19.21]), one has the inclusion

$$C(\Omega) \subseteq B(\Omega) \tag{1.4b}$$

for compact subsets $\Omega \subseteq \mathbb{K}^d$. Without the compactness assumption on $\Omega$ the inclusion (1.4b) is wrong, as demonstrated by the continuous, but unbounded function $u : (0,1] \to \mathbb{R}$, $u(x) := \frac{1}{x}$. For additional examples, note that all the functions from Ex. 1.4.1 have been continuous.

**Proposition 1.4.3.** *If $\Omega \subseteq \mathbb{K}^d$ is compact, then $C(\Omega)$ is a normed space with the norm*

$$\|u\| := \max_{x\in\Omega} |u(x)| .$$

*Remark* 1.4.4. (1) For reals $a < b$ one has the inclusions $P[a,b] \subseteq C[a,b] \subseteq B[a,b]$ and therefore both $C[a,b]$ and $B[a,b]$[11] are infinite-dimensional linear spaces.

[10] a subset $\Omega \subseteq \mathbb{K}^d$ is called *compact*, if it is bounded and if every convergent sequence in $\Omega$ has its limit in $\Omega$ (see [NS82, p. 147, Thm. 3.17.20]). For instance, finite sets, the cartesian products $[a_1,b_1] \times \ldots \times [a_d,b_d] \subseteq \mathbb{R}^d$ or $\bar{B}_j \subseteq \mathbb{K}^d$ (see (1.2b)) are compact, while $\mathbb{R}$, $\mathbb{C}$, $(a,\infty)$ or $(a,b)$, $[a,b)$ and $(a,b]$ are not compact subsets of $\mathbb{R}$ resp. $\mathbb{C}$

[11] we prefer the brief notation $B[a,b]$ to $B([a,b])$ and proceed similarly with other function spaces to be defined

(2) The norm in $B(\Omega)$ or $C(\Omega)$ is called *supremum norm* and denoted by $\|\cdot\|_\infty$.

(3) For later use we also introduce the space $C_0(\Omega)$ of *continuous functions vanishing on the boundary*. Thereto, let $\partial\Omega \subseteq \mathbb{K}^d$ denote the *boundary*[12] of the domain $\Omega$. Then the set (see Fig. 1.8(right))

$$C_0(\Omega) := \{u \in C(\Omega) : u(x) = 0 \text{ for all } x \in \partial\Omega\}$$

is a subspace of $C(\Omega)$ and itself a normed space w.r.t. the norm $\|\cdot\|_\infty$.

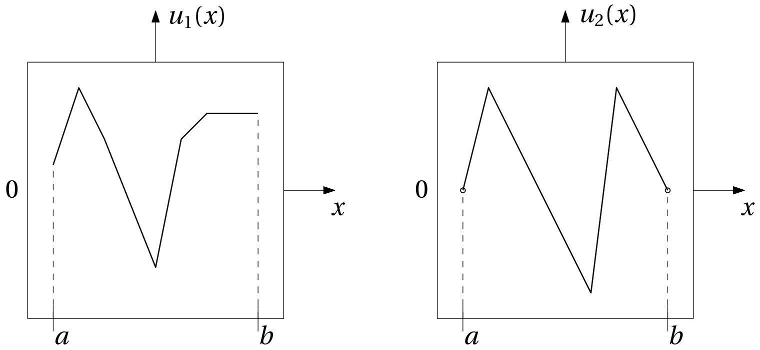

**Fig. 1.8** A function $u_1 \in C[a,b]$ and a function $u_2 \in C_0[a,b]$

*Proof.* First of all, from the above inclusion (1.4b) we know that $C(\Omega)$ is a subset of $B(\Omega)$. In order to show that $B(\Omega)$ is a subspace, it remains to verify that also the linear combination $\alpha u_1 + \beta u_2$, $\alpha, \beta \in \mathbb{K}$ of continuous functions $u_1, u_2 : \Omega \to \mathbb{K}$ is continuous again. However, this is clear from elementary calculus, and we conclude that $C(\Omega)$ is a linear space. On the other hand, every continuous function $u : \Omega \to \mathbb{K}$ on a compact set $\Omega$ attains its maximum and minimum (see [NS82, p. 148, Thm. 3.19.21]), which guarantees

$$\|u\| = \sup_{x\in\Omega} |u(x)| = \max_{x\in\Omega} |u(x)|$$

and as in the proof of Prop. 1.4.2 one shows that $\|\cdot\|$ is a norm on $C(\Omega)$. □

We finally illustrate that the norm on $C(\Omega)$ is not induced by an inner product.

*Example* 1.4.5. Let $\mathbb{K} = \mathbb{R}$ and $\Omega := [\frac{1}{2}, 1]$ be given. For the continuous functions $u, v : [\frac{1}{2}, 1] \to \mathbb{R}$, $u(x) := 1$ and $v(x) := x$ we obtain

$$\|u+v\| = \max_{x\in[\frac{1}{2},1]} |u(x)+v(x)| = \max_{x\in[\frac{1}{2},1]} |1+x| = 2, \qquad \|u\| = 1,$$

$$\|u-v\| = \max_{x\in[\frac{1}{2},1]} |u(x)-v(x)| = \max_{x\in[\frac{1}{2},1]} |1-x| = \tfrac{1}{2}, \qquad \|v\| = 1$$

[12] for instance, $\partial[a,b] = \partial(a,b) = \{a,b\}$ or $\partial\{(x,y) \in \mathbb{R}^2 : x^2+y^2 \le 1\} = \{(x,y) \in \mathbb{R}^2 : x^2+y^2 = 1\}$, but see [NS82, p. 110] for a general definition

and therefore $\|u+v\|^2 + \|u-v\|^2 = \frac{17}{4} \neq 4 = 2\|u\|^2 + 2\|v\|^2$. Consequently, the parallelogram identity (1.3c) does not hold and Prop. 1.3.8(a) shows that the norm on $C[\frac{1}{2}, 1]$ is not given by an inner product.

### 1.4.3 Continuously differentiable functions

Classical formulations of differential equations involve differentiable functions as their solutions. We are about to introduce the corresponding function spaces.

Thereto, let $m \in \mathbb{N}_0$ and assume that $\Omega \subseteq \mathbb{R}^d$ is open.[13] Then $C^m(\Omega)$ is the set of all $m$-times differentiable functions $u : \Omega \to \mathbb{K}$, whose $n$th derivatives $D^n u$ are continuous on $\Omega$ for $n \in \{0, \ldots, m\}$. In addition, we define the set

$$C^\infty(\Omega) := \bigcap_{m \in \mathbb{N}_0} C^m(\Omega)$$

of infinitely often continuously differentiable functions, which is also a linear space. We obtain the inclusions $C^\infty(\Omega) \subseteq C^m(\Omega) \subseteq C^0(\Omega) = C(\Omega)$.

*Example* 1.4.6. One clearly has $P(\mathbb{R}) \subset C^\infty(\mathbb{R})$, but also the exponential, trigonometric or hyperbolic functions are in $C^\infty(\mathbb{R})$. Another example of a $C^\infty$-function is the following $u : \mathbb{R} \to \mathbb{R}$ given in Fig. 1.9.

$$u(x) := \begin{cases} 0, & x \leq 0, \\ e^{-\frac{1}{x}}, & x > 0. \end{cases}$$

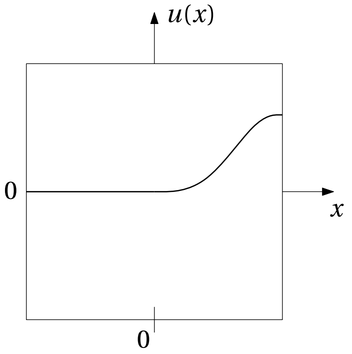

**Fig. 1.9** The $C^\infty$-function $u$ from Ex. 1.4.6

For general open sets $\Omega \subseteq \mathbb{R}^d$ a function in $C^m(\Omega)$ might not be bounded with bounded derivatives. We therefore define the subspace

$$C^m_b(\Omega) := \left\{u \in C^m(\Omega) : D^n u \in B(\Omega) \text{ for } n \in \{0, \ldots, m\}\right\}$$

of both $C^m(\Omega)$ and $B(\Omega)$.

[13] a set $\Omega \subseteq \mathbb{R}^d$ is called *open*, if for every $x \in \Omega$ there exists an $r > 0$ such that the $r$-ball with center $x$ is contained in $\Omega$, i.e. $\left\{y \in \mathbb{R}^d : \left\|y - x\right\| < r\right\} \subseteq \Omega$. For instance, the sets $\mathbb{R}$, intervals $(a, b)$, boxes $(a_1, b_1) \times \ldots \times (a_d, b_d)$ or balls $\left\{x \in \mathbb{R}^d : \|x\| < 1\right\}$ are open, while $[a, b)$ or $\left\{x \in \mathbb{R}^d : \|x\| \leq 1\right\}$ are not open

**Proposition 1.4.7.** *If $\Omega \subseteq \mathbb{R}^d$ is open, then the set $C_b^m(\Omega)$ is a normed space with the norm*

$$\|u\| := \sum_{k=0}^{m} \sup_{x\in\Omega} \left|D^k u(x)\right| = \sum_{k=0}^{m} \left\|D^k u\right\|_\infty .$$

*Remark* 1.4.8. (1) For $0 \le m \le n$ the $n$-times continuously differentiable functions $C_b^n(\Omega)$ are a subspace of $C_b^m(\Omega)$ with $\|u\|_{C_b^m(\Omega)} \le \|u\|_{C_b^n(\Omega)}$ for all $u \in C_b^n(\Omega)$.

(2) Let $\Omega \subseteq \mathbb{R}^d$ be bounded and open. An important subspace of $C_b^m(\Omega)$ are the *continuously differentiable functions vanishing on the boundary.* It is given by

$$C_0^m(\Omega) := \left\{ u \in C^m(\Omega) : \begin{array}{l} u \text{ has a continuous extension } U \text{ to } \Omega \cup \partial\Omega \\ \text{satisfying } U(x) = 0 \text{ for all } x \in \partial\Omega \end{array} \right\}.$$

(3) For compact domains $\Omega \subseteq \mathbb{R}^d$ a function $u : \Omega \to \mathbb{K}$ is said to be $m$-times continuously differentiable, if there exists an extension $U \in C^m(\Omega_1)$ of $u$ to an open set $\Omega_1 \supseteq \Omega$. Then one has $C_b^m(\Omega) = C^m(\Omega)$ and $C^m(\Omega)$ is a normed space.

*Proof.* The proof is similar to that of Prop. 1.4.3. □

### 1.4.4 Integrable functions

Let $\mathbb{K}$ denote $\mathbb{R}$ or $\mathbb{C}$ and $\Omega \subseteq \mathbb{R}^d$ be a *measurable set*. We do not give a precise definition of "measurability" and refer to the reader's naive intuition: For a measurable set $\Omega \subseteq \mathbb{R}^d$ it makes sense to define its length ($d = 1$), its area ($d = 2$) or its volume ($d = 3$) and to denote it by $\mu(\Omega) \ge 0$. For example, intervals $(a, b), [a, b]$, boxes $[a_1, b_1] \times \ldots \times [a_d, b_d]$, balls $\{x \in \mathbb{R}^d : \|x\| \le r\}$ or compact subsets of $\mathbb{R}^d$, as well as their finite union and intersection, fall into the category of measurable sets. The inclined reader is referred to the special literature (cf. [Coh80]). In this sense, a set of *measure zero* has measure $\mu(\Omega)$ (length, area, or volume) zero.

The well-known *Riemann integral* (cf., e.g., [NS82, pp. 559ff]) allowed a simple construction using upper and lower sums, and turned out to be sufficient for many applications. However, it has some serious drawbacks making it inappropriate in functional analysis. Among them is the fact that there exist sequences $(u_n)_{n\in\mathbb{N}}$ of Riemann integrable functions $u_n : \Omega \to \mathbb{K}$ with the property that their pointwise limit function

$$u : \Omega \to \mathbb{K}, \qquad u(x) := \lim_{n\to\infty} u_n(x) \quad \text{for all } x \in \Omega$$

is not Riemann integrable (see [NS82, p. 564]). A further disadvantage will become apparent in Section 2.3.

To avoid such problems, one needed a more flexible integral notion — the *Lebesgue integral.* An introduction to the required mathematical preliminaries can be found in, e.g., [NS82, pp. 589ff]. In particular, every Lebesgue integrable function is also Riemann integrable. Therefore, if we speak of an integrable function from now on, we always mean Lebesgue integrable.

In this class we identify two functions $u_1, u_2 : \Omega \to \mathbb{K}$, if the set $\Omega_0 \subseteq \mathbb{R}^d$ of points $x \in \Omega$ with $u_1(x) \neq u_2(x)$ is negligible, i.e. of measure zero. For example, every *denumerable set*[14] is negligible in this sense.

After these preliminary remarks, given $p \in [1,\infty)$ we can define the so-called *Lebesgue spaces* of *p-integrable functions*

$$L^p(\Omega) := \left\{ u : \Omega \to \mathbb{K} | \int_\Omega |u(x)|^p \, dx < \infty \right\}.$$

*Example* 1.4.9. For every parameter $\alpha > 0$ one can easily show the relations

$$\int_0^1 \frac{dx}{x^\alpha} = \begin{cases} \frac{1}{1-\alpha}, & \alpha \in (0,1), \\ \infty, & \alpha \geq 1, \end{cases} \qquad \int_1^\infty \frac{dx}{x^\alpha} = \begin{cases} \infty, & \alpha \in (0,1], \\ \frac{1}{\alpha-1}, & \alpha > 1. \end{cases}$$

For parameters $\beta > 0$ and functions $u : (0,1) \to \mathbb{R}$, $v : [1,\infty) \to \mathbb{R}$ given by $u(x) := v(x) := \frac{1}{x^\beta}$ one has $\int_0^1 |u(x)|^p \, dx = \int_0^1 \frac{dx}{x^{p\beta}}$ yielding the inclusion $u \in L^p(0,1)$ for $p \in [1, \beta^{-1})$, but $u \notin L^p(0,1)$ for $p \geq \beta^{-1}$. This means functions very singular at 0 (i.e. $\beta$ is large) are not contained in any of the spaces $L^p(0,1)$. On the other hand, $v \in L^p[1,\infty)$ for $p > \beta^{-1}$ and $v \notin L^p[1,\infty)$ for $p \in [1, \beta^{-1}]$. As an interpretation, note that functions decaying slowly ($\beta$ is small) are in $L^p$-spaces with large $p$.

To handle $p$-integrable functions, some preparations are due:

**Lemma 1.4.10** (Young inequality)**.** *If reals $p, q \geq 1$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$, then*

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q} \quad \text{for all } a, b \in [0,\infty).$$

*Proof.* Since the claimed estimate obviously holds for $a = 0$ or $b = 0$, we restrict to the case $a, b > 0$ and set $t := 1/p$, $1 - t = 1/q$. Using the strict concavity of the logarithm function $\ln : (0,\infty) \to \mathbb{R}$ we deduce

$$\ln(ta^p + (1-t)b^q) \geq t \ln a^p + (1-t) \ln b^q = \ln a + \ln b = \ln(ab)$$

and exponentiating yields the assertion. □

**Lemma 1.4.11** (Hölder inequality)**.** *Let $\Omega \subseteq \mathbb{R}^d$ be measurable and suppose the reals $p, q \geq 1$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$. If $u \in L^p(\Omega)$, $v \in L^q(\Omega)$, then $uv \in L^1(\Omega)$ and*

$$\int_\Omega |u(x)v(x)| \, dx \leq \left( \int_\Omega |u(x)|^p \, dx \right)^{1/p} \left( \int_\Omega |v(x)|^q \, dx \right)^{1/q}. \tag{1.4c}$$

[14] a set is called *denumerable*, if it is finite or has the same cardinality as $\mathbb{N}$

*Proof.* Let $u \in L^p(\Omega)$ and $v \in L^q(\Omega)$.

(I) Using Young's inequality from Lemma 1.4.10 we have

$$\alpha\beta = (\alpha t)\left(\frac{\beta}{t}\right) \le \frac{t^p}{p}\alpha^p + \frac{1}{qt^q}\beta^q \quad \text{for all } \alpha, \beta \ge 0,\ t > 0 \tag{1.4d}$$

and we abbreviate

$$I := \int_\Omega |u(x)v(x)|\, dx, \qquad P := \int_\Omega |u(x)|^p\, dx, \qquad Q := \int_\Omega |v(x)|^q\, dx.$$

Thanks to (1.4d) it is $|u(x)v(x)| \le \frac{t^p}{p}|u(x)|^p + \frac{1}{qt^q}|v(x)|^q$ and by integration

$$I \le \frac{t^p}{p}P + \frac{1}{qt^q}Q \quad \text{for all } t > 0. \tag{1.4e}$$

If $P = 0$ or $Q = 0$, then $I = 0$, since otherwise (1.4e) cannot hold for all $t > 0$. In these cases we have established (1.4c).

(II) In the following we suppose $P, Q > 0$. Setting $t = 1$, $\alpha = |u(x)|P^{-1/p}$ and $\beta = |v(x)|^{-1/q}$ in (1.4d) implies $\frac{|u(x)v(x)|}{P^{-1/p}Q^{-1/q}} \le \frac{|u(x)|^p}{pP} + \frac{|v(x)|^q}{qQ}$ and integration yields

$$\frac{I}{P^{-1/p}Q^{-1/q}} \le \frac{P}{pP} + \frac{Q}{qQ} = \frac{1}{p} + \frac{1}{q} = 1,$$

which implies the assertion. □

These preparations allow us to introduce a norm on the general spaces $L^p(\Omega)$.

**Proposition 1.4.12.** *If $\Omega \subseteq \mathbb{R}^d$ is measurable and $p \ge 1$, then the set $L^p(\Omega)$ is a normed space with the norm*

$$\|u\|_p := \left(\int_\Omega |u(x)|^p\, dx\right)^{1/p}.$$

*Remark* 1.4.13. (1) The triangle inequality in $L^p(\Omega)$ is also denoted as *Minkowski inequality* (cf. [NS82, p. 548]).

(2) The $L^1$-norm can be interpreted as the area between two functions, while the supremum norm $\|\cdot\|_\infty$ measures the maximal distance between corresponding function values (see Fig. 1.10).

(3) For a set $\Omega \subseteq \mathbb{R}^d$ with finite measure $\mu(\Omega) < \infty$ and $1 \le p < q$ one can show $L^q(\Omega) \subseteq L^p(\Omega)$ with $\|u\|_p \le \mu(\Omega)^{\frac{1}{p}-\frac{1}{q}}\|u\|_q$ for all functions $u \in L^q(\Omega)$.

*Proof.* The simple case $p = 1$ will be treated in the exercises. So we can assume $p > 1$ and choose $q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$, i.e. $(p-1)q = p$. We only establish

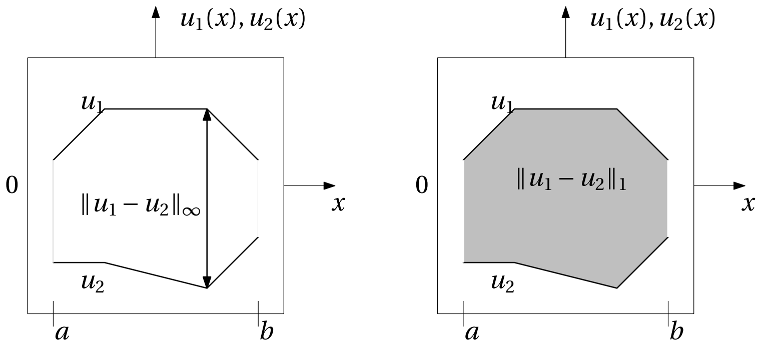

**Fig. 1.10** The distance between the functions $u_1, u_2 : [a,b] \to \mathbb{R}$ in the supremum norm (left) and in the $L^1$-norm (right)

the triangle inequality and leave a proof of the remaining linear space and norm properties to the interested reader. Let $u, v \in L^p(\Omega)$. We define

$$I := \int_\Omega |u(x) + v(x)|^p \, dx = \|u + v\|_p^p$$

and observe that the claim is trivial for $I = 0$. Hence, we can assume $I > 0$ and obtain from the Hölder inequality from Lemma 1.4.11 that

$$\begin{aligned} I &\le \int_\Omega |u(x)|\,|u(x) + v(x)|^{p-1} \, dx + \int_\Omega |v(x)|\,|u(x) + v(x)|^{p-1} \, dx \\ &\overset{(1.4c)}{\le} \left(\int_\Omega |u(x)|^p \, dx\right)^{1/p} \left(\int_\Omega |u(x) + v(x)|^{(p-1)q} \, dx\right)^{1/q} \\ &\quad + \left(\int_\Omega |v(x)|^p \, dx\right)^{1/p} \left(\int_\Omega |u(x) + v(x)|^{(p-1)q} \, dx\right)^{1/q} \\ &= I^{1/q} \left[ \left(\int_\Omega |u(x)|^p \, dx\right)^{1/p} + \left(\int_\Omega |v(x)|^p \, dx\right)^{1/p} \right]. \end{aligned}$$

Division by $I^{1/p}$ yields $\|u + v\|_p \le \|u\|_p + \|v\|_p$, since we have $\frac{1}{p} = 1 - \frac{1}{q}$. □

A particular role plays the case $p = 2$ of square-integrable functions $L^2(\Omega)$. Note that its inner product yields a natural norm, which is consistent with the norms on general spaces $L^p(\Omega)$.

**Proposition 1.4.14.** *If $\Omega \subseteq \mathbb{R}^d$ is measurable, then the set $L^2(\Omega)$ is an inner product space with inner product*

$$\langle u, v \rangle := \int_\Omega u(x) \overline{v(x)} \, dx.$$

*Proof.* Let $u_1, u_2, v \in L^2(\Omega)$ and $\alpha, \beta \in \mathbb{K}$. First, thanks to Prop. 1.4.12 with $p = 2$ the space $L^2(\Omega)$ is linear. It remains to verify the inner product conditions from Def. 1.3.1. We remark that Hölder's inequality with $p = q = 2$ in Lemma 1.4.11 guarantees $u(\cdot)\overline{v(\cdot)} \in L^1(\Omega)$ for functions $u, v \in L^2(\Omega)$ and the inner product on $L^2(\Omega)$ is well-defined. Using the linearity of the integral we have

$$\begin{aligned}\langle \alpha u_1 + \beta u_2, v\rangle &= \int_\Omega (\alpha u_1(x) + \beta u_2(x))\overline{v(x)}\,dx \\ &= \alpha \int_\Omega u_1(x)\overline{v(x)}\,dx + \beta \int_\Omega u_2(x)\overline{v(x)}\,dx = \alpha\langle u_1, v\rangle + \beta\langle u_2, v\rangle\end{aligned}$$

yielding linearity in the first argument (i). Moreover, it is

$$\langle u_1, v\rangle = \int_\Omega u_1(x)\overline{v(x)}\,dx = \int_\Omega \overline{v(x)\overline{u_1(x)}}\,dx = \overline{\int_\Omega v(x)\overline{u_1(x)}\,dx} = \overline{\langle v, u_1\rangle}$$

proving the conjugate symmetry (ii). Finally, it is

$$\langle v, v\rangle = \int_\Omega v(x)\overline{v(x)}\,dx = \int_\Omega \underbrace{|v(x)|^2}_{\geq 0}\,dx \geq 0$$

and $0 = \langle v, v\rangle$ also implies $v(x) = 0$ for all $x \in \Omega$ (except from a negligible set); we have shown property (iii). □

*Exercises* 1.4.15. Solve the following problems:

(1) Show that $L^1(\Omega)$ is a normed space.

(2) Let $\Omega = [-1, 1]$ and consider the functions $u_j : \Omega \to \mathbb{R}$ given by

$$u_1(x) := x^2 - \tfrac{1}{2}, \qquad u_2(x) := \sin(\pi x),$$
$$u_3(x) := |x|, \qquad u_4(x) := \begin{cases} 1, & x \geq 0, \\ -1, & x < 0. \end{cases}$$

For every index $j \in \{1, 2, 3, 4\}$ determine the function spaces discussed in the present Section 1.4 which contain the function $u_j$. Moreover, compute their corresponding norm. In case of $L^p(\Omega)$ you can restrict to cases $p \in \{1, 2\}$.

(3) Let $n \in \mathbb{N}$. Consider the function $u_n : [-1, 1] \to \mathbb{R}$ given by

$$u_n(x) := \begin{cases} 0, & x \leq 0, \\ x^n, & x > 0 \end{cases}$$

and determine a maximal $m \in \mathbb{N}_0$ such that $u_n \in C^m[-1, 1]$ holds. Furthermore, compute the corresponding $C^m$-norm of $u_n$!

## 1.5 Linear mappings

With given linear spaces $X, Y$ we consider functions $T : X \to Y$. In case $X$ and $Y$ are function spaces, such mappings are also called *operators*. A particular important class of such mappings are those which preserve the linear structure of the domain $X$. For these so-called linear mappings we shortly write $Tx := T(x)$.

**Definition 1.5.1** (linear mapping)**.** Let $X, Y$ be linear spaces over $\mathbb{K}$. A mapping $T : X \to Y$ is called *linear*, if $T(\alpha x + \beta y) = \alpha Tx + \beta Ty$ for all $\alpha, \beta \in \mathbb{K}$, $x, y \in X$. In case $Y = \mathbb{K}$ one speaks of a *linear functional*.

*Remark* 1.5.2. In accordance with Rem. 1.3.2 a mapping $T : X \to Y$ is called *semilinear*, if $T(\alpha x + \beta y) = \overline{\alpha} Tx + \overline{\beta} Ty$ for all $\alpha, \beta \in \mathbb{K}$, $x, y \in X$.

(1) Every linear mapping fulfills $T0 = 0$.

(2) The set of all linear mappings $T : X \to Y$ forms a linear space.

(3) If $B = \{x_1, x_2, \dots\}$ is a basis of $X$, then a linear mapping $T$ is completely determined by its values on $B$. This means given $x = \sum_{k=1}^n \xi_k x_k$, $n \in \mathbb{N}$, $\xi_k \in \mathbb{K}$, one has

$$Tx = T \sum_{k=1}^n \xi_k x_k = \sum_{k=1}^n \xi_k T x_k$$

and the knowledge of the coefficients $\xi_1, \dots, \xi_n \in \mathbb{K}$ and of $Tx_1, \dots, Tx_n \in Y$ enables us to compute $Tx$.

*Example* 1.5.3 (linear mappings). Let $a < b$ and $A \in \mathbb{K}^{m \times n}$.

(1) The *zero mapping* $T : X \to Y$, $Tx := 0 \in Y$ is linear with the image $TX = \{0\}$. Also the *identity mapping* $\mathrm{id} : X \to X$, $\mathrm{id}\, x := x$ is linear with image $\mathrm{id}\, X = X$.

(2) Clearly, $Tx := Ax$ defines a linear mapping $T : \mathbb{K}^n \to \mathbb{K}^m$.

(3) The mapping $D : C^1[a, b] \to C[a, b]$ given by $Du := u'$ is linear.

(4) The mapping $T : C[a, b] \to C^1[a, b]$, $(Tu)(x) := \int_a^x u(t)\, dt$ is linear.

*Example* 1.5.4 (differential operators). Suppose that $\Omega \subseteq \mathbb{R}^d$ is an open set. Then the *gradient*

$$\nabla : C^1(\Omega) \to C(\Omega)^d, \qquad \nabla u := \left( \frac{\partial u}{\partial x_1}, \dots, \frac{\partial u}{\partial x_d} \right)$$

and the *Laplace operator*

$$\Delta : C^2(\Omega) \to C(\Omega), \qquad \Delta u := \sum_{j=1}^d \frac{\partial^2 u}{\partial x_j^2}$$

are linear mappings. One can verify that both the $C^2$-functions $u_1 : \mathbb{R}^2 \setminus \{0\} \to \mathbb{R}$, $u_1(x) := \ln(x_1^2 + x_2^2)$ and $u_2 : \mathbb{R}^2 \to \mathbb{R}$, $u_2(x) := e^{x_1} \sin x_2$ fulfill $\Delta u_i = 0$ for $i = 1, 2$.

*Example* 1.5.5 (linear functionals). With $a < b$ the mappings $E_t : F[a, b] \to \mathbb{K}$, $E_t u := u(t)$ (evaluation at a point $t \in [a, b]$), $\partial_t : C^1[a, b] \to \mathbb{K}$, $\partial_t u := u'(t)$ (differentiation in $t \in [a,b]$) and furthermore $I_a^b : L^1[a,b] \to \mathbb{K}$, $I_a^b u := \int_a^b u(x)\,dx$ (integration) are linear functionals.

Usually it is desirable that (numerical) schemes preserve properties of the mathematical objects that are approximated. In particular, this is true for the differentiation $D$ and the integral operator $I_a^b$ defined above:

*Example* 1.5.6 (numerical differentiation). Given $a,b \in \mathbb{R}$, $a < b$ and $h > 0$ one defines the forward difference operator $D_h u(x) := \frac{u(x+h)-u(x)}{h}$ for all $x \in [a,b-h]$. We observe that $D_h : C[a,b-h] \to C[a,b-h]$ is linear. Similarly, an approximation of the Laplace operator from Ex. 1.5.4 is given by $\Delta_h u(x) := \frac{u(x+h)-2u(x)+u(x-h)}{h^2}$ for $d = 1$ and in $d = 2$ dimension one has the well-known 5-*point stencil*

$$\Delta_h u(x) := \frac{u(x_1+h,x_2)+u(x_1-h,x_2)-4u(x)+u(x_1,x_2+h)+u(x_1,x_2-h)}{h^2}.$$

*Example* 1.5.7 (numerical quadrature). Given $a,b \in \mathbb{R}$, $a \le b$, the following mappings $R_a^b, T_a^b, S_a^b : C[a,b] \to \mathbb{R}$ are linear functionals with applications in numerical quadrature:

- $R_a^b u := (b-a)\,u\left(\frac{b+a}{2}\right)$ (*rectangle rule*)
- $T_a^b u := (b-a)\,\frac{u(b)+u(a)}{2}$ (*trapezoidal rule*)
- $S_a^b u := \frac{b-a}{6}\left[u(a) + 4u\left(\frac{a+b}{2}\right) + u(b)\right]$ (*Simpson's rule*)

We close this section by introducing mappings with values in $\mathbb{K}$ depending on two arguments. These so-called sesquilinear[15] mappings are linear in the first and semilinear in the second argument — a property they share with inner products. Yet, we do not require them to be positive definite or symmetric.

As we will see later, sesquilinear mappings arise in variational formulations of boundary value problems and in particular in a mathematical theory of the finite element method.

**Definition 1.5.8** (sesquilinear and bilinear form)**.** Let $X, Y$ be linear spaces over $\mathbb{K}$. A mapping $a : X \times Y \to \mathbb{K}$ is called *sesquilinear form*, if

$$a(\alpha_1 x_1 + \alpha_2 x_2, y_1) = \alpha_1 a(x_1,y_1) + \alpha_2 a(x_2,y_1),$$
$$a(x_1, \alpha_1 y_1 + \alpha_2 y_2) = \overline{\alpha_1} a(x_1,y_1) + \overline{\alpha_2} a(x_1,y_2)$$

holds for all $x_1, x_2 \in X$, $y_1, y_2 \in Y$ and $\alpha_1, \alpha_2 \in \mathbb{K}$. In case $\mathbb{K} = \mathbb{R}$ one speaks of a *bilinear form*.

*Remark* 1.5.9. With a sesquilinear form $a : X \times Y \to \mathbb{K}$ and $y \in X$, the mapping $a(\cdot, y) : X \to \mathbb{K}$ is a linear functional.

[15] the prefix *sesqui* (lat.) means "one and a half"

*Example* 1.5.10. (1) Every inner product $\langle\cdot,\cdot\rangle : X \times X \to \mathbb{K}$ is a sesquilinear form.

(2) Given a matrix $A \in \mathbb{K}^{m\times n}$, the mapping $a : \mathbb{K}^n \times \mathbb{K}^m \to \mathbb{K}$, $a(x,y) := y^* Ax$ is a sesquilinear form.

(3) If $\Omega \subseteq \mathbb{R}^d$ is a compact set, then the mappings $a_1 : C(\Omega) \times C(\Omega) \to \mathbb{K}$ and also $a_2 : L^2(\Omega) \times L^2(\Omega) \to \mathbb{K}$,

$$a_1(u,v) := a_2(u,v) := \int_\Omega u(x)\overline{v(x)}\,dx$$

are inner products on $C(\Omega)$ resp. $L^2(\Omega)$. On the other hand, for real numbers $a < b$ the mappings

$$\begin{aligned} &b : C^1[a,b] \times C^1[a,b] \to \mathbb{K}, && b(u,v) := \int_a^b u'(x)\overline{v'(x)}\,dx,\\ &c : C^1[a,b] \times C[a,b] \to \mathbb{K}, && c(u,v) := \int_a^b u'(x)\overline{v(x)}\,dx \end{aligned}$$

are sesquilinear forms, but not inner products. Indeed, for each constant function $u : [a,b] \to \mathbb{R}$ one has $b(u,u) = c(u,u) = 0$ and thus $b, c$ are not positive definite.

*Example* 1.5.11 (energy norm). Suppose that $\Omega \subseteq \mathbb{R}^d$ is a compact set. Let us introduce the sesquilinear form $c : C^1(\Omega) \times C^1(\Omega) \to \mathbb{K}$,

$$c(u,v) := \int_\Omega \nabla u(x) \cdot \overline{\nabla v(x)}\,dx = \int_\Omega \sum_{j=1}^d \frac{\partial u(x)}{\partial x_j}\overline{\frac{\partial v(x)}{\partial x_j}}\,dx.$$

We observe that $\langle u,v\rangle := c(u,v)$ does not define an inner product on $C^1(\Omega)$, since $\langle u,u\rangle = 0$ holds for all constant functions $u \in C^1(\Omega)$ and not only for the null function. However, $c$ defines an inner product on the subspace $C_0^1(\Omega) \subseteq C^1(\Omega)$ of functions vanishing on the boundary and consequently

$$\|u\| := \sqrt{c(u,u)} = \sqrt{\int_\Omega \sum_{j=1}^d \frac{\partial u(x)}{\partial x_j}\overline{\frac{\partial u(x)}{\partial x_j}}\,dx}$$

defines a norm on $C_0^1(\Omega)$ — the so-called *energy norm.*[16]

We close this first chapter with a review to the concepts we have introduced so far — however, compared to the text, in reverse order from a very specific case to the most general situation:

- The set with the most structure needed here are the real numbers $\mathbb{R}$: We can add and multiply real numbers ($\mathbb{R}$ is a field), and we additionally can compare reals, i.e. inequalities make sense.
- The various properties of the reals $\mathbb{R}$ are shared by the field of complex numbers $\mathbb{C}$, minus the fact that inequalities do not make sense in $\mathbb{C}$.

[16] this terminology comes from physics: If $u(x)$ denotes the deflection of a string at a point $x$ under the influence of a form, then its elastic energy reads as $\frac{1}{2}\int_a^b u'(x)^2\,dx$

- In inner product spaces we have an inner product (a scalar product) at hand, which enables us to define orthogonality. Examples include the Euclidean space $\mathbb{R}^d$ equipped with the inner product (1.3b) or the space of square-integrable functions $L^2(\Omega)$. On inner product spaces one uses the natural norm induced by the inner product as in Prop. 1.3.4.
- On general normed spaces there is a norm available, which needs not to be induced by an inner product. Examples include $\mathbb{K}^d$ equipped with the 1- or the maximum norm, $B(\Omega)$ and $C(\Omega)$, $C_0(\Omega)$ with the supremum norm, as well as the $m$-times continuously differentiable functions $C^m(\Omega)$, or the $p$-integrable functions $L^p(\Omega)$ for $p \neq 2$.
- In linear spaces we can add elements, and multiply them with scalars from $\mathbb{K}$, but there needs not to exist a norm (or a metric). Hence, every normed space is a linear space, but not conversely. For instance, the polynomials $P(\Omega)$ or all $\mathbb{K}$-valued functions $F(\Omega)$ form linear spaces, but we have not defined a norm on these spaces.
- On metric spaces there exists a distance function (a metric), which enables us to measure distances, but there is not necessarily an addition or a scalar multiplication available. On normed spaces the metric is given by the norm (see Rem. 1.2.16(2)), on inner product spaces by their natural norm, and on $\mathbb{K}$ simply by the modulus (absolute value). Thus, every nonempty subset of a normed space is a metric space. Without indicating a norm or metric, a linear space needs not to be a metric space.
- Finally, on arbitrary sets we have no algebraic or metric structures available. This limits our possibilities to do interesting mathematics on general sets.

We have illustrated the above set inclusions in Fig. 1.11.

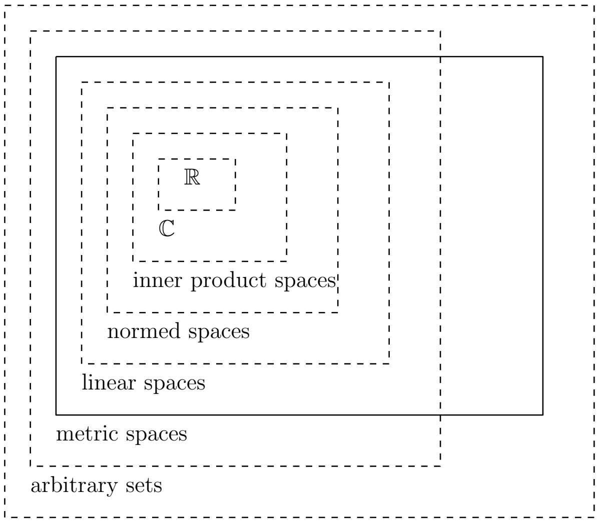

**Fig. 1.11** Hierarchy of spaces

*Exercises* 1.5.12. Solve the following problems:

(1) Let $T : X \to Y$ denote a linear mapping between linear spaces $X, Y$ over $\mathbb{K}$. Show that the *kernel* $N(T) := \{x \in X : Tx = 0\}$ and the *range* $R(T) := TX$ of $T$ are subspaces of $X$ resp. $Y$.

(2) Determine kernel and range of the linear mapping $D : P(\mathbb{R}) \to P(\mathbb{R})$, $Dp := p'$ (the derivative). Can you describe the set $DP_n(\mathbb{R})$ (the range of $D$)?

(3) Determine the kernel of the linear mapping $T : C^2(\mathbb{R}) \to C(\mathbb{R})$ given by

$$(Tu)(x) := u''(x) + 2au'(x) + u(x)$$

with a coefficient $a \in \mathbb{R}$. What is $\dim N(T)$?

(4) Given the linear functionals $I_a^b$ from Ex. 1.5.5 and $T_a^b$ from Ex. 1.5.7 restricted to the polynomials $P_2[a, b]$ as domain. Determine the kernel $N(I_a^b - T_a^b)$ and interpret your result.

# Chapter 2
# Topological structures

The notion of convergence is a central pillar in calculus, since it is fundamental to understand concepts like continuity, differentiability and integrability of functions $f : \Omega \subseteq \mathbb{K}^d \to \mathbb{K}^n$.

In this chapter we aim to generalize the notion of continuity to mappings between general metric or normed spaces. Surprisingly, this can be done on the basis of convergence for real sequences alone. In this context, we remind the reader that a real or complex sequence $(\xi_n)_{n\in\mathbb{N}}$ is said to *converge* to a point $\xi \in \mathbb{K}$, if for every $\varepsilon > 0$ there exists a $N > 0$ such that

$$|\xi_n - \xi| < \varepsilon \quad \text{for all } n \geq N;$$

in this case we write $\xi = \lim_{n\to\infty} \xi_n$.

## 2.1 Convergence

On finite-dimensional spaces one typically used the Euclidean norm $\|\cdot\|_2$ in order to study convergence. Nevertheless, on more general sets we have seen that it is possible to introduce different metrics.

**Definition 2.1.1** (convergent sequence)**.** Let $X$ be a nonempty set and suppose $d : X \times X \to \mathbb{R}$ is a metric on $X$. A sequence $(x_n)_{n\in\mathbb{N}}$ is said to *converge* to the *limit* $x \in X$ w.r.t. the metric $d$, if

$$\lim_{n\to\infty} d(x_n, x) = 0.$$

In this case we write $\lim_{n\to\infty} x_n = x$ or $x_n \to x$ for $n \to \infty$ w.r.t. $d$.

*Remark* 2.1.2. (1) In normed spaces $(X, \|\cdot\|)$ convergence is always understood w.r.t. the metric $d(x,y) := \|x-y\|$. Without proof, we remark for convergent sequences $(\alpha_n)_{n\in\mathbb{N}}, (\beta)_{n\in\mathbb{N}}$ in $\mathbb{K}$ and $(x)_{n\in\mathbb{N}}, (y)_{n\in\mathbb{N}}$ in $X$ with respective limits $\alpha, \beta \in \mathbb{K}$ and $x, y \in X$ also the sequence $(\alpha_n x_n + \beta_n y_n)_{n\in\mathbb{N}}$ converges with limit

$$\lim_{n\to\infty} \left(\alpha_n x_n + \beta_n y_n\right) = \alpha x + \beta y.$$

This means the limit operator is linear.

(2) Every convergent sequence is bounded.

On subsets $S$ of finite-dimensional spaces $\mathbb{K}^d$ we had introduced the metrics $d_1, d_2, d_\infty$ in Ex. 1.1.5. From Exercise 1.1.6(3) one can see that convergence of a sequence in $\mathbb{K}^d$ w.r.t. the metric $d_2$ implies convergence in $d_1$ or $d_\infty$ as well. Since these metrics on $\mathbb{K}^d$ are induced by the norms from Ex. 1.2.17(2) this means that all norms on finite-dimensional spaces are *equivalent.*[1]

On infinite-dimensional spaces, and in particular on function spaces, the situation is different and one has to carefully distinguish between various forms of convergence. First of all, a sequence $(u_n)_{n\in\mathbb{N}}$ in $F(\Omega)$ is said to *converge pointwise*, if the sequences $(u_n(x))_{x\in\Omega}$ converge in $\mathbb{K}$ for all $x \in \Omega$. One can define the *limit function* $u : \Omega \to \mathbb{K}$ by

$$u(x) := \lim_{n\to\infty} u_n(x) \quad \text{for all } x \in \Omega.$$

While there is no norm on $F(\Omega)$, we now retreat to certain normed subspaces. For instance, convergence in $B(\Omega)$ equipped with the norm $\|\cdot\|_\infty$ is called *uniform convergence*. Pointwise and uniform convergence are related as follows:

**Corollary 2.1.3.** *Let $\Omega$ be a nonempty set, $u \in B(\Omega)$ and $(u_n)_{n\in\mathbb{N}}$ be a sequence of functions $u_n : \Omega \to \mathbb{K}$ in $B(\Omega)$. If* $\lim_{n\to\infty} u_n = u$ *in $B(\Omega)$, then $(u_n)_{n\in\mathbb{N}}$ converges pointwise to $u$.*

*Proof.* One obviously has the limit relation

$$|u_n(x) - u(x)| \le \sup_{x\in\Omega} |u_n(x) - u(x)| = \|u_n - u\|_\infty \xrightarrow[n\to\infty]{} 0$$

and therefore $\lim_{n\to\infty} u_n(x) = u(x)$ for all $x \in \Omega$. □

The next example shows that pointwise convergence needs not to be uniform:

*Example* 2.1.4. Consider the functions $u_n : [0,1] \to \mathbb{R}$, $n \in \mathbb{N}$, given by the monomials $u_n(x) := x^n$ and we obviously have $u_n \in C[0,1]$ for all $n \in \mathbb{N}$. Hence, $(u_n)_{n\in\mathbb{N}}$ is a sequence in $C[0,1]$ (cf. Fig. 2.1). Using the pointwise limit we obtain the limit function $u : [0,1] \to \mathbb{R}$,

[1] more precisely, two norms $\|\cdot\|$ and $\|\cdot\|'$ on a linear space are called *equivalent*, if convergence w.r.t. $\|\cdot\|$ implies convergence in $\|\cdot\|'$ and conversely

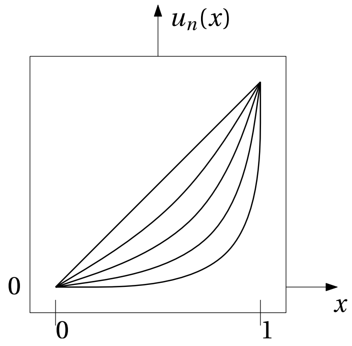

**Fig. 2.1** Graphs of the monomials $u_n : [0,1] \to \mathbb{R}$, $u_n(x) := x^n$

$$u(x) := \lim_{n\to\infty} u_n(x) = \lim_{n\to\infty} x^n = \begin{cases} 0, & x \in [0,1), \\ 1, & x = 1, \end{cases}$$

which is bounded, but discontinuous. Now we equip $C[0,1]$ with the $L^p$-norm $\|\cdot\|_p$ defined in Prop. 1.4.12 and obtain

$$\|u_n - u\|_p = \left(\int_0^1 |u_n(x) - u(x)|^p \, dx\right)^{1/p} = \left(\int_0^1 x^{np} \, dx\right)^{1/p} = \left(\frac{1}{np+1}\right)^{1/p}$$

for all $n \in \mathbb{N}$, which implies the limit relation $\lim_{n\to\infty} \|u_n - u\|_p = 0$. Therefore, the sequence $(u_n)_{n\in\mathbb{N}}$ converges to $u$ in the norm $\|\cdot\|_p$. On the other hand, let us also try the norm $\|\cdot\|_\infty$ on $B[0,1]$. Here, we get

$$\begin{aligned} \|u_n - u\|_\infty &= \sup_{x\in[0,1]} |u_n(x) - u(x)| = \max\left\{ \sup_{x\in[0,1)} |u_n(x) - u(x)|, |u_n(1) - u(1)| \right\} \\ &= \sup_{x\in[0,1)} |u_n(x)| = \sup_{x\in[0,1)} x^n = 1 \quad \text{for all } n \in \mathbb{N} \end{aligned}$$

and $(u_n)_{n\in\mathbb{N}}$ cannot converge to $u$ in the norm $\|\cdot\|_\infty$. From this we see that convergence in function spaces depends on the particular norm.

For more results on convergence we refer to [NS82, p. 69ff].

*Exercises* 2.1.5. Solve the following problems:

(1) Does the sequence $\left(\frac{1}{n}\right)_{n\in\mathbb{N}}$ in $\mathbb{R}$ converge to 0 w.r.t. the discrete metric $d$ introduced in Ex. 1.1.3? Note here that the discrete metric is not induced by a norm on $\mathbb{R}$. Why?

(2) Find the pointwise limit functions $u, v : [0,\pi] \to \mathbb{R}$ of the function sequences $u_n, v_n : [0,\pi] \to \mathbb{R}$, $n \in \mathbb{N}$, given by

$$u_n(x) := \sin\left(\frac{x}{n}\right), \qquad v_n(x) := \frac{1}{n}\sin(nx).$$

Moreover, investigate their convergence properties w.r.t. the norms on $C[0,\pi]$, $C^1[0,\pi]$ and $L^1[0,\pi]$.

## 2.2 Continuity

From elementary calculus we are familiar with the concept of continuity. A function $f:\Omega\subseteq\mathbb{R}^n\to\mathbb{R}^m$ was called continuous in $x\in\Omega$, if the limit relation

$$\lim_{n\to\infty} f(x_n) = f\left(\lim_{n\to\infty} x_n\right)$$

holds for every sequence $(x_n)_{n\in\mathbb{N}}$ in $\Omega$ with limit $x$. The subsequent definition explains how to extend this to the general context of metric or function spaces.

**Definition 2.2.1** (continuous mapping)**.** Let $(X,d_X)$ and $(Y,d_Y)$ be metric spaces. A mapping $f:X\to Y$ is called *continuous in a point* $x\in X$, if

$$\lim_{n\to\infty} d_Y(f(x_n),f(x)) = 0$$

holds for every convergent sequence $(x_n)_{n\in\mathbb{N}}$ in $(X,d_X)$ with limit $x$. A mapping $f$ is called *continuous* (on $X$), if it is continuous in every $x\in X$.

*Remark* 2.2.2 (properties of continuous mappings). (1) If $Z$ is a further metric space and $f:X\to Y$, $g:Y\to Z$ are continuous mappings, then also the composition $g\circ f:X\to Z$, $g\circ f(x):=g(f(x))$ is continuous.

(2) If $Y$ is a normed space over $\mathbb{K}$ and $f,g:X\to Y$ are continuous, then also the mappings $\alpha f+\beta g:X\to Y$, $\alpha,\beta\in\mathbb{K}$, are continuous.

For later use we remark that metrics and norms are continuous functions:

**Lemma 2.2.3.** *On metric spaces* $(X,d)$ *the metric* $d:X\times X\to\mathbb{R}$ *is continuous.*

*Proof.* Let $x,y\in X$ and suppose $(x_n)_{n\in\mathbb{N}}$, $(y_n)_{n\in\mathbb{N}}$ are sequences in $X$ with respective limits $x$, $y$. Then the quadrangle inequality from Cor. 1.1.4(c) guarantees

$$0\le\left|d(x_n,y_n)-d(x,y)\right|\le d(x_n,x)+d(y_n,y)\xrightarrow[n\to\infty]{}0$$

and consequently $d:X\times X\to\mathbb{R}$ is continuous. Note here that on the product $X\times X$ we use the metric from Rem. 1.1.2. □

A particularly important class are linear mappings between normed spaces. As the reader might expect, their continuity properties crucially depend on the norms involved:

*Example* 2.2.4. We know from our previous Ex. 1.5.3(3) that the differentiation operator $D : C^1[0,\pi] \to C[0,\pi]$, $Du := u'$ is linear. In order to understand its continuity properties, we consider the sequence $u_n : [0,\pi] \to \mathbb{R}$ of continuously differentiable functions $u_n(x) := \frac{1}{n}\sin(nx)$ and obtain

$$\begin{aligned} \|u_n\|_\infty &= \max_{x\in[0,\pi]} |u_n(x)| = \max_{x\in[0,\pi]} \left|\tfrac{1}{n}\sin(nx)\right| = \tfrac{1}{n}, \\ \left\|u_n'\right\|_\infty &= \max_{x\in[0,\pi]} \left|u_n'(x)\right| = \max_{x\in[0,\pi]} |\cos(nx)| = 1 \quad \text{for all } n \in \mathbb{N}. \end{aligned}$$

Consequently, $(u_n)_{n\in\mathbb{N}}$ converges to the null function $0 : [0,\pi] \to \mathbb{R}$ in $\|\cdot\|_\infty$. First we consider $C^1[0,\pi]$ as a subspace of $C[0,\pi]$ equipped with $\|\cdot\|_\infty$ and deduce

$$\|Du_n - D0\|_\infty = \left\|u_n'\right\|_\infty = 1 \quad \text{for all } n \in \mathbb{N}.$$

This shows that the differentiation operator $D$ is not continuous in 0, if we equip its domain $C^1[0,\pi]$ with the supremum norm. Second, we use another norm on $C^1[0,\pi]$, namely the one defined in Def. 1.4.7 by $\|u\| := \|u\|_\infty + \left\|u'\right\|_\infty$. In this norm we derive

$$\|u_n\| = \|u_n\|_\infty + \left\|u_n'\right\|_\infty = \tfrac{1}{n} + 1 \quad \text{for all } n \in \mathbb{N}$$

and the sequence $(u_n)_{n\in\mathbb{N}}$ does not converge to 0. Nevertheless, the norm $\|\cdot\|$ ensures continuity of $D : (C^1[0,\pi], \|\cdot\|) \to (C[0,\pi], \|\cdot\|_\infty)$ in 0, since we have

$$\lim_{n\to\infty} \|Dv_n - D0\|_\infty = \lim_{n\to\infty} \left\|v_n'\right\|_\infty = 0$$

for every sequence $(v_n)_{n\in\mathbb{N}}$ in $C^1[0,\pi]$ with $\lim_{n\to\infty} \|v_n\| = 0$.

There is a slight inconsistency in the mathematical literature with the notion of a bounded mapping. In fact, as we know from Subsection 1.4.1 a general mapping $f : X \to \mathbb{K}$ is called bounded, if its range $f(X) \subseteq \mathbb{K}$ is a bounded subset of $\mathbb{K}$. In contrast, for linear mappings we introduce

**Definition 2.2.5** (bounded linear mapping). Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be normed spaces. A linear mapping $T : X \to Y$ is called *bounded*, if there exist a real $C \geq 0$ such that

$$\|Tx\|_Y \leq C\|x\|_X \quad \text{for all } x \in X. \tag{2.2a}$$

*Remark* 2.2.6. (1) The set of all linear bounded mappings $T : X \to Y$ is denoted by $L(X,Y)$. It is a subspace of the linear space of all linear mappings with domain $X$ and codomain $Y$.

(2) For a bounded linear operator $T : X \to Y$ one has

$$\left\| T\frac{x}{\|x\|_X}\right\|_Y = \frac{\|Tx\|_Y}{\|x\|_X} \leq C \quad \text{for all } x \neq 0$$

and we can define the so-called *operator norm*

$$\|T\| := \sup_{\|x\|_X=1} \|Tx\|_Y \leq C$$

of $T$ as the smallest constant $C \geq 0$ such that (2.2a) holds true. Indeed, the operator norm is a norm on $L(X,Y)$ and we have $\|Tx\|_Y \leq \|T\|\,\|x\|_X$ for all $x \in X$.

*Example* 2.2.7 (matrix norm). Let $A \in \mathbb{R}^{m\times n}$ and denote by $A : \mathbb{R}^n \to \mathbb{R}^m$ also the corresponding linear mapping $x \mapsto Ax$. Then $A : \mathbb{R}^n \to \mathbb{R}^m$ is a bounded linear mapping. However, its operator norm depends on the particular vector space norm we are using on the domain $\mathbb{R}^m$ and on the codomain $\mathbb{R}^n$:

- The 1-norm induces the matrix norm

$$\|A\|_1 = \sup_{\|x\|_1=1} \|Ax\|_1 = \max_{1\leq j\leq n} \sum_{k=1}^{m} \left|a_{kj}\right|,$$

- the 2-norm yields

$$\|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2 = \sqrt{\max\left\{\lambda \in [0,\infty) : \lambda \text{ is an eigenvalue of } A^* A\right\}},$$

- finally, the $\infty$-norm yields

$$\|A\|_\infty = \sup_{\|x\|_\infty=1} \|Ax\|_\infty = \max_{1\leq k\leq m} \sum_{j=1}^{n} \left|a_{kj}\right|. \tag{2.2b}$$

*Example* 2.2.8. (1) Suppose $a, b \in \mathbb{R}$ with $a < b$ are given. The integral operator $T : C[a,b] \to C^1[a,b]$, $(Tu)(x) := \int_a^x u(t)\,dt$ is bounded. To prove this, we pick an arbitrary $u \in C[a,b]$ and obtain $|u(x)| \leq \|u\|_\infty$,

$$\left|\int_a^x u(t)\,dt\right| \leq \int_a^x |u(t)|\,dt \leq \int_a^x \|u\|_\infty\,dt = (x-a)\,\|u\|_\infty \leq (b-a)\,\|u\|_\infty$$

for all $x \in [a,b]$. Consequently, one has

$$\begin{aligned}\|Tu\|_{C^1[a,b]} &= \max_{x\in[a,b]} |(Tu)(x)| + \max_{x\in[a,b]} \left|(Tu)'(x)\right| \\ &= \max_{x\in[a,b]} \left|\int_a^x u(x)\,dx\right| + \max_{x\in[a,b]} |u(x)| \leq (b-a+1)\,\|u\|_\infty\end{aligned}$$

and we can choose $C = b-a+1$ in the above Def. 2.2.5.

(2) The differential operator $D : C^1[0,\pi] \to C[0,\pi]$ given by $Du := u'$ is not bounded, provided we equip its domain $C^1[a,b]$ with the supremum norm $\|\cdot\|_\infty$. In order to see this, it suffices to show that for every real $C > 1$ there exists a $u_C \in C^1[0,\pi]$ with $\|Du_C\|_\infty > C\,\|u_C\|_\infty$. Indeed, consider the continuously dif-

ferentiable functions $u_C : [0,\pi] \to \mathbb{R}$, $u_C(x) := \sin(C^2 x)$, for which we have

$$C\,\|u_C\|_\infty = C \max_{x\in[0,\pi]} |u_C(x)| = C < C^2 = \max_{x\in[0,\pi]} \left|C^2 \cos(C^2 x)\right| = \|Du_C\|_\infty .$$

This also results from Ex. 2.2.4 and the subsequent theorem.

For linear mappings the notions of continuity and boundedness are equivalent. This is the main result of this section:

**Theorem 2.2.9.** *Let $X, Y$ be normed spaces. If $T : X \to Y$ is a linear mapping, then the following assertions are equivalent:*

*(a)* *$T$ is bounded, i.e. $T \in L(X,Y)$,*
*(b)* *$T$ is continuous,*
*(c)* *$T$ is continuous at $0 \in X$.*

*Proof.* $(a) \Rightarrow (b)$ Let $T : X \to Y$ be bounded, i.e. there exists a $C \geq 0$ such that $\|Tx\| \leq C\,\|x\|$ for all $x \in X$. If $(x_n)_{n\in\mathbb{N}}$ is a sequence in $X$ with limit $x \in X$ we obtain from linearity of $T$ that

$$\|Tx_n - Tx\| = \|T(x_n - x)\| \leq C\,\|x_n - x\| \xrightarrow[n\to\infty]{} 0$$

and hence $T$ is continuous in $x \in X$. Since $x$ was arbitrary, $T$ is continuous on $X$.

$(b) \Rightarrow (c)$ This is trivially true.

$(c) \Rightarrow (a)$ Let $T : X \to Y$ be continuous in 0. We assume $T$ is not bounded, which means that there exists a sequence $(x_n)_{n\in\mathbb{N}}$ in $X$ of vectors $x_n \neq 0$ such that

$$\|Tx_n\| > n\,\|x_n\| \quad \text{for all } n \in \mathbb{N}. \tag{2.2c}$$

We define the sequence $\xi_n := \frac{1}{n\|x_n\|} x_n$ and from

$$0 \leq \|\xi_n\| = \left\| \frac{1}{n\,\|x_n\|} x_n \right\| = \frac{1}{n\,\|x_n\|}\,\|x_n\| = \frac{1}{n} \xrightarrow[n\to\infty]{} 0$$

one gets $\lim_{n\to\infty} \xi_n = 0$. Since $T$ is continuous in 0 this implies the contradiction

$$1 = \frac{1}{n} n \overset{(2.2c)}{<} \frac{1}{n} \frac{\|Tx_n\|}{\|x_n\|} = \left\| T \frac{x_n}{n\,\|x_n\|} \right\| = \|T\xi_n\| \xrightarrow[n\to\infty]{} 0.$$

Consequently, $T : X \to Y$ must be a bounded operator. □

A quite similar statement also holds for sesquilinear or bilinear forms, and in particular for inner products:

**Theorem 2.2.10.** *Let $X, Y$ be normed spaces. A sesquilinear form $a : X \times Y \to \mathbb{K}$ is continuous, if and only if there exists a constant $C \geq 0$ with*

$$\left|a(x,y)\right| \leq C \left\|x\right\|_X \left\|y\right\|_Y \quad \textit{for all } x \in X,\ y \in Y. \tag{2.2d}$$

*Remark* 2.2.11. By the Cauchy-Schwarz inequality (see Prop. 1.3.7) every inner product satisfies (2.2d) with $C = 1$ and therefore inner products are continuous mappings $\langle \cdot, \cdot \rangle : X \times X \to \mathbb{K}$.

*Proof.* We have to verify two directions.

($\Rightarrow$) Let $a : X \times Y \to \mathbb{K}$ be continuous, but suppose that (2.2d) does not hold. Then for each $n \in \mathbb{N}$ there exist sequences $(x_n)_{n\in\mathbb{N}}$ in $X \setminus \{0\}$, $(y_n)_{n\in\mathbb{N}}$ in $Y \setminus \{0\}$ with

$$\left|a(x_n, y_n)\right| > n \left\|x_n\right\| \left\|y_n\right\| \quad \text{for all } n \in \mathbb{N}.$$

The sequences $\xi_n := \frac{1}{\sqrt{n}\|x_n\|} x_n$ in $X$ and $\eta_n := \frac{1}{\sqrt{n}\|y_n\|} y_n$ in $Y$ converge to 0 in the limit $n \to \infty$, but the continuity of $a$ yields the contradiction

$$1 = \frac{1}{n} n < \frac{1}{n} \frac{\left|a(x_n, y_n)\right|}{\left\|x_n\right\| \left\|y_n\right\|} = \left|a\left(\tfrac{1}{\sqrt{n}\|x_n\|} x_n, \tfrac{1}{\sqrt{n}\|y_n\|} y_n\right)\right| = \left|a(\xi_n, \eta_n)\right| \xrightarrow[n\to\infty]{} 0.$$

Hence, the relation (2.2d) must hold.

($\Leftarrow$) Conversely, suppose that (2.2d) is true. Let $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$ be sequences with respective limits $x$ and $y$. Since every convergent sequence is bounded (see Rem. 2.1.2(2)), there exists a real $c \geq 0$ with $\left\|y_n\right\| \leq c$ for all $n \in \mathbb{N}$. Then we have from linearity properties of $a$ and the triangle inequality that

$$\begin{aligned}
\left|a(x_n, y_n) - a(x, y)\right| &\leq \left|a(x_n, y_n) - a(x, y_n)\right| + \left|a(x, y_n) - a(x, y)\right| \\
&\leq \left|a(x_n - x, y_n)\right| + \left|a(x, y_n - y)\right| \\
&\overset{(2.2d)}{\leq} C \left\|x_n - x\right\| \left\|y_n\right\| + C \left\|x\right\| \left\|y_n - y\right\| \\
&\leq cC \left\|x_n - x\right\| + C \left\|x\right\| \left\|y_n - y\right\| \xrightarrow[n\to\infty]{} 0
\end{aligned}$$

and consequently $a$ is continuous in $(x, y)$. Since the pair $(x, y) \in X \times Y$ was arbitrary, the mapping $a : X \times Y \to \mathbb{K}$ is continuous. □

A detailed approach to the concept of continuity is given in [NS82, p. 61ff].

*Exercises* 2.2.12. Solve the following problems:

(1) Prove the relation (2.2b).

(2) Let $a, b \in \mathbb{R}$ with $a < b$ and choose $x \in [a, b]$. Show that the following linear functionals are bounded and determine an upper bound for their operator norm, i.e. for the real constant $C$ from Def. 2.2.5:

(i) Evaluation mapping $E_x : B[a,b] \to \mathbb{K}$, $E_x u := u(x)$,
(ii) differentiation at $x$, given by $\partial_x : C^1[a,b] \to \mathbb{K}$, $\partial_x u := u'(x)$,
(iii) integration $I_a^b : C[a,b] \to \mathbb{K}$, $I_a^b u := \int_a^b u(x)\,dx$.

## 2.3 Completeness

The Def. 2.1.1 of a convergent sequence has an intrinsic disadvantage: In order to show convergence of a sequence $(x_n)_{n\in\mathbb{N}}$, one has to know the limit $x$ in advance. For the purpose of circumventing this problem, the following notion is helpful:

**Definition 2.3.1** (Cauchy sequence)**.** Let $(X,d)$ be a metric space. A sequence $(x_n)_{n\in\mathbb{N}}$ in $X$ is called *Cauchy sequence,* if for every $\varepsilon > 0$ there exists a $N = N_\varepsilon > 0$ such that $d(x_n, x_m) < \varepsilon$ for all $m, n \geq N$.

As we show next, every convergent sequence is Cauchy but not conversely.

**Proposition 2.3.2.** *Let $(X,d)$ be a metric space. Every convergent sequence $(x_n)_{n\in\mathbb{N}}$ in $X$ is a Cauchy sequence.*

*Proof.* Let $\varepsilon > 0$. If $(x_n)_{n\in\mathbb{N}}$ is convergent, then there exists a $x \in X$ and an $N > 0$ such that $d(x_n, x) < \frac{\varepsilon}{2}$ for all $n \geq N$. Using the triangle inequality this yields

$$d(x_n, x_m) \leq d(x_n, x) + d(x_m, x) < \tfrac{\varepsilon}{2} + \tfrac{\varepsilon}{2} = \varepsilon \quad \text{for all } n, m \geq N$$

and $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence. □

*Example* 2.3.3. Let $p \geq 1$. We again consider functions $u_n : [0,1] \to \mathbb{R}$, $n \in \mathbb{N}$, given by the monomials $u_n(x) := x^n$.

(1) Clearly, it is $u_n \in L^p[0,1]$ for all $n \in \mathbb{N}$, and in Ex. 2.1.4 we have seen that $(u_n)_{n\in\mathbb{N}}$ converges to the limit function $u \in L^p[0,1]$ w.r.t. the $L^p$-norm, with

$$u(x) := \begin{cases} 0, & x \in [0,1), \\ 1, & x = 1. \end{cases}$$

Thus, Prop. 2.3.2 guarantees that $(u_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $L^p[0,1]$.

(2) On the other side, one also has $u_n \in C[0,1]$ for all $n \in \mathbb{N}$. Instead of the canonical supremum norm, one could also use the $L^p$-norm on the space $C[0,1]$. We have seen above that $(u_n)_{n\in\mathbb{N}}$ is a Cauchy sequence w.r.t. the norm $\|\cdot\|_p$. However, the normed space $(C[0,1], \|\cdot\|_p)$ has the deficit that the limit function $u : [0,1] \to \mathbb{R}$ is not continuous, i.e. $u \notin C[0,1]$.

The second part of the above example shows that Cauchy sequences might not converge in a given normed space. This observation motivates the following

**Definition 2.3.4** (complete metric space)**.** A metric space $(X,d)$ is called *complete*, if every Cauchy sequence in $X$ converges to a limit in $X$.

*Example* 2.3.5. The closed subsets of $\mathbb{K}^d$ are complete metric spaces, where the metric is given by one of the norms from Ex. 1.2.17(2).

On the basis of complete metric spaces we can deduce an important tool for the solution of a large variety of nonlinear equations, which can be formulated as fixed point[2] problems. Beyond guaranteeing the existence of solutions, it also yields their uniqueness, an approximation scheme and an error estimate. Before stating it explicitly, we introduce a convenient notation for the iterates of a function $F : X \to X$: We define recursively

$$F^0(x) := x, \qquad\qquad F^{n+1}(x) := F(F^n(x)) \quad \text{for all } n \in \mathbb{N}_0,$$

i.e. we have $F^1(x) = F(x)$, $F^2(x) = F(F(x))$, $F^3(x) = F\big(F(F(x))\big)$ and so on. Note that $F^n(x) \neq F(x)^n = F(x) \cdot \ldots \cdot F(x)$ in general.

This result is known as the *Banach fixed point theorem* or as

**Theorem 2.3.6** (contraction mapping principle)**.** *Let $(X,d)$ be a complete metric space and $q \in [0,1)$. If $F : X \to X$ is a* contraction *i.e. a mapping with*

$$d(F(x),F(\bar{x})) \le q\,d(x,\bar{x}) \quad \textit{for all } x,\bar{x} \in X, \tag{2.3a}$$

*then the following holds true:*

*(a) There exists a unique solution $x^* \in X$ of the equation $F(x) = x$,*
*(b) one has the limit relation $\lim_{n\to\infty} F^n(x) = x^*$ for all $x \in X$ and more precisely the a priori error estimate*

$$d(F^n(x),x^*) \le \frac{q^n}{1-q} d(F(x),x) \quad \textit{for all } n \in \mathbb{N},\ x \in X. \tag{2.3b}$$

*Proof.* See also [NS82, p. 126, Thm. 3.15.2]. Let $x \in X$ and define the sequence $x_n := F^n(x)$, $n \in \mathbb{N}_0$, in $X$. We split the proof into several parts:

(I) We start with a preliminary estimate. Thereto, without loss of generality, we can assume $n > m$ and obtain from (2.3a) that

$$d(x_n,x_m) = d(F^n(x),F^m(x)) = d(F^m(x_{n-m}),F^m(x)) \le q\,d(F^{m-1}(x_{n-m}),F^{m-1}(x)).$$

[2] a point $x \in X$ is called *fixed point* of a mapping $F : X \to X$, if $x = F(x)$ holds

By mathematical induction this yields

$$d(x_n, x_m) \leq q^m d(x_{n-m}, x) \tag{2.3c}$$

and using the triangle inequality, this becomes

$$d(x_n, x_m) \leq q^m \sum_{k=0}^{n-m-1} d(x_{k+1}, x_k) \overset{(2.3.6)}{\leq} q^m \sum_{k=0}^{n-m-1} q^k d(x_1, x).$$

Due to $q \in [0,1)$ we have the geometric series formula $\sum_{k=0}^{\infty} q^k = \frac{1}{1-q}$ at hand and deduce

$$d(x_n, x_m) \leq q^m \sum_{k=0}^{\infty} q^k d(x_1, x) = \frac{q^m}{1-q} d(x_1, x) \quad \text{for all } n > m. \tag{2.3d}$$

(II) We will show that $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence. Thereto, given $\varepsilon > 0$ we choose $N \in \mathbb{N}$ so large that $\frac{q^N}{1-q} d(x_1, x) < \frac{\varepsilon}{2}$, which is possible due to $q \in [0,1)$. Consequently, one has

$$d(x_n, x_m) \overset{(2.3d)}{\leq} \tfrac{q^N}{1-q} d(x_1, x) < \varepsilon \quad \text{for all } n, m \geq N$$

and $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $X$. Since $X$ is complete, $(F^n(x))_{n\in\mathbb{N}}$ converges and we define $x^* := \lim_{n\to\infty} F^n(x)$.

(III) We show that the limit $x^*$ is a fixed point of $F$. For this, we observe

$$\begin{aligned} d(x^*, F(x^*)) &\leq d(x^*, x_{n+1}) + d(x_{n+1}, F(x^*)) \\ &\overset{(2.3a)}{\leq} d(x^*, x_{n+1}) + q d(x_n, x^*) \xrightarrow[n\to\infty]{} 0 \end{aligned}$$

and consequently $d(x^*, F(x^*)) = 0$, i.e. $x^* = F(x^*)$. In order to show the error estimate (2.3b), we remark that metrics are continuous by Lemma 2.2.3. Hence, passing over to the limit $n \to \infty$ in (2.3d) immediately implies for all $m \in \mathbb{N}$ that

$$d(x^*, x_m) = d(x^*, F^m(x)) \leq q^m \sum_{k=0}^{\infty} q^k d(x_1, x) = q^m \sum_{k=0}^{\infty} q^k d(F(x), x).$$

(IV) Is remains to show that the fixed point $x^*$ is unique. If also $y^* \in X$ is a fixed point of $F : X \to X$ we obtain

$$d(x^*, y^*) = d(F(x^*), F(y^*)) \overset{(2.3a)}{\leq} q d(x^*, y^*).$$

Because of $q \in [0,1)$ his relation can only hold, provided $d(x^*, y^*) = 0$, which means $x^* = y^*$. □

Of particular importance are normed spaces:

**Definition 2.3.7** (Banach space)**.** A complete normed space is called *Banach space.*

*Example* 2.3.8. (1) Finite-dimensional normed spaces and in particular $\mathbb{K}^d$ or $\mathbb{K}^{m\times n}$ are Banach spaces (cf. [NS82, p. 267, Thm. 5.10.2]).

(2) If $X$ is a normed space and $Y$ a Banach space, then also the bounded linear operators $L(X,Y)$ form a Banach space. In particular, the linear bounded functionals $L(X,\mathbb{K})$ are a Banach space.

*Example* 2.3.9 (bounded functions). The bounded functions $(B(\Omega),\|\cdot\|_\infty)$ with the supremum norm

$$\|u\|_\infty := \sup_{x\in\Omega}|u(x)|$$

are a Banach space. This can be seen as follows: Let $(u_n)_{n\in\mathbb{N}}$ be a Cauchy sequence of bounded functions $u_n:\Omega\to\mathbb{K}$ and we have to show that it converges to a bounded function $u:\Omega\to\mathbb{K}$ w.r.t. the norm $\|\cdot\|_\infty$. Thereto, let $\varepsilon>0$ and we know that there exists an $N>0$ with

$$|u_n(x)-u_m(x)| \le \sup_{x\in\Omega}|u_n(x)-u_m(x)| = \|u_n-u_m\|_\infty < \varepsilon \tag{2.3e}$$

for all $m,n\ge N$ and $x\in\Omega$. This inequality guarantees that $(u_n(x))_{n\in\mathbb{N}}$, $x\in\Omega$, is a Cauchy sequence in $\mathbb{K}$. Since $\mathbb{K}$ is complete (see the above Ex. 2.3.8(1)), every sequence $(u_n(x))_{n\in\mathbb{N}}$ converges to an element $u(x)\in\mathbb{K}$ and we define the function $u(x):=\lim_{n\to\infty}u_n(x)$ for all $x\in\Omega$. It remains to show that $u:\Omega\to\mathbb{K}$ is a bounded function. For this purpose, we pass over to the limit $n\to\infty$ in the inequality

$$|u_n(x)-u_N(x)| \le \|u_n-u_N\|_\infty \overset{(2.3e)}{<} \varepsilon \quad \text{for all } n\ge N$$

and obtain $|u(x)-u_N(x)|\le\varepsilon$ for all $x\in\Omega$. On the other hand, due to $u_N\in B(\Omega)$ there exists a $C>0$ such that $|u_N(x)|\le C$ for all $x\in\Omega$. Combining these two fact, we obtain from the triangle inequality that

$$|u(x)| \le |u(x)-u_N(x)| + \|u_N(x)\| \le \tfrac{\varepsilon}{2}+C \quad \text{for all } x\in\Omega,$$

i.e., the desired inclusion $u\in B(\Omega)$.

*Example* 2.3.10 ($C^m$-spaces). (1) The continuous functions $(C(\Omega),\|\cdot\|_\infty)$ on a compact set $\Omega\subseteq\mathbb{R}^d$ are a Banach space (see [NS82, p. 219, Ex. 5]).

(2) The $m$-times continuously differentiable functions $C_b^m(\Omega)$ equipped with the norm from Prop. 1.4.7 are a Banach space. The same holds for $C_0^m(\Omega)$.

*Example* 2.3.11 ($p$-integrable functions). The linear space of $p$-integrable functions $(L^p(\Omega),\|\cdot\|_p)$ is a Banach space; this follows with [NS82, p. 589, Thm. D.11.2].

The contraction mapping principle immediately implies the following criterion for existence and uniqueness to solutions for linear equations:

**Corollary 2.3.12** (Neumann series)**.** *Let $X$ be a Banach space and suppose $T \in L(X)$. If $\|T\| < 1$, then for every $b \in X$ the linear equation*

$$x = Tx + b \tag{2.3f}$$

*has a unique solution $x^* \in X$ given by the* Neumann series $x^* = \sum_{n=0}^{\infty} T^n b$.

*Proof.* For given $b \in X$ we define the mapping $F : X \to X$, $F(x) := Tx + b$. Due to

$$\|F(x) - F(\bar{x})\| = \|Tx - T\bar{x}\| \le \|T\| \, \|x - \bar{x}\| \quad \text{for all } x, \bar{x} \in X$$

and $\|T\| < 1$ we see that $F$ is a contraction on the complete metric space $X$. Thus, by Thm. 2.3.6 we know that $F$ has a unique fixed point $x^* \in X$, which also solves the equation (2.3f). On the other hand, using mathematical induction one easily shows $\|T^n x\| \le \|T\|^n \|x\|$ for all $n \in \mathbb{N}_0$. This and $\|T\| < 1$ implies the limit relation $\lim_{n\to\infty} T^n x = 0$ for all $x \in X$. Mathematical induction also yields the iterates

$$F^n(x) = T^n x + \sum_{k=0}^{n-1} T^k b \quad \text{for all } x \in X,\ n \in \mathbb{N}_0$$

and passing over to the limit this ensures $x^* = \lim_{n\to\infty} F^n(x) = \sum_{k=0}^{\infty} T^k b$. □

*Example* 2.3.13 (Fredholm integral equation). Let $\Omega \subseteq \mathbb{R}^d$ be compact and suppose $k : \Omega \times \Omega \to \mathbb{K}$ is a continuous function satisfying

$$\sup_{x\in\Omega} \int_\Omega |k(x,y)| \, dy < 1 \tag{2.3g}$$

Then the Fredholm integral equation

$$\phi(x) = b(x) + \int_\Omega k(x,y)\phi(y) \, dy$$

has a unique solution $\phi \in C(\Omega)$ for every inhomogeneity $b \in C(\Omega)$. In order to show this, we define the operator $T\phi := \int_\Omega k(\cdot, y)\phi(y) \, dy$. It is well-defined, i.e. given a continuous function $\phi \in C(\Omega)$ one also has$T\phi \in C(\Omega)$. Moreover, it is

$$\begin{aligned} |(T\phi)(x)| &\le \int_\Omega |k(x,y)| |\phi(y)| \, dy \le \int_\Omega |k(x,y)| \, \|\phi\|_\infty \, dy \\ &\le \sup_{x\in\Omega} \int_\Omega |k(x,y)| \, dy \, \|\phi\|_\infty \quad \text{for all } x \in \Omega,\ \phi \in C(\Omega) \end{aligned}$$

and passing over to the maximum for $x \in \Omega$ guarantees the relation $\|T\phi\|_\infty \le \sup_{x\in\Omega} \int_\Omega |k(x,y)| \, dy \, \|\phi\|_\infty$. Thanks to (2.3g) we get $\|T\| < 1$ and thus $T \in L(X)$ satisfies the assumptions of Thm. 2.3.12 on the Banach space $X = C(\Omega)$.

**Definition 2.3.14** (Hilbert space)**.** An inner product space is called *Hilbert space*, if it is complete w.r.t. its natural norm.

*Example* 2.3.15 (square-integrable functions). The 2-integrable functions $L^2(\Omega)$ equipped with the inner product

$$\langle u, v\rangle := \int_\Omega u(x)\overline{v(x)}\,dx$$

are a Hilbert space (cf. [NS82, p. 279, Ex. 3]). However, $C(\Omega)$ equipped with the above inner product is not complete. This can be seen using the sequence of monomials $(u_n)_{n\in\mathbb{N}}$ in Ex. 2.1.4 or [NS82, p. 279, Exs. 5–6].

*Exercises* 2.3.16. Solve the following problems:

(1) Use the contraction mapping principle from Thm. 2.3.6 in order to solve the equation $\cos x - x = 0$ in $\mathbb{R}$ up to an absolute error of $10^{-10}$.
*Hint:* Use the relation $|\cos x - \cos\bar{x}| \le \frac{\sqrt{2}}{2}|x-\bar{x}|$ for all $x, \bar{x} \in [\frac{2\pi}{9}, \frac{\pi}{4}]$. Can you prove it using the mean value theorem?

(2) Let $b \in \mathbb{R}^d$ and suppose that $A \in \mathbb{R}^{d\times d}$ satisfies $\max_{1\le j\le d}\sum_{k=1}^d |a_{kj}| < 1$. Prove that there exists a unique solution to $x = Ax + b$.

## 2.4 Representation results

First, we give an affirmative answer to the "orthogonal projection" problem mentioned in the introduction to Sect. 1.3. Here we give a more general version valid for convex sets[3] (see also [NS82, p. 297, Thm. 5.15.6]).

**Theorem 2.4.1** (projection theorem)**.** *Let $X$ be a Hilbert space. If $Y \subseteq X$ denotes a nonempty, closed and convex set, then there exists a unique mapping $P : X \to Y$, the so-called* orthogonal projection *onto $Y$, such that*

*(a)* $\|x - P(x)\| = \min_{y\in Y}\|x - y\|$ *for all $x \in X$,*
*(b)* $\Re\langle x - P(x), y - P(x)\rangle \le 0$ *for all $x \in X$, $y \in Y$.*

*Proof.* Let $x \in X$ be fixed. There exists a sequence $(y_k)_{k\in\mathbb{N}}$ in $Y$ such that

[3] a subset $Y \subseteq X$ of a linear space $X$ is called *convex*, if for arbitrary $y_1, y_2 \in Y$ one has the inclusion $y_1 + t(y_2 - y_1) \in Y$ for all $t \in [0,1]$, i.e. the whole segment connecting $x_1$ with $x_2$ is contained in $Y$

$$\lim_{k\to\infty}\left\|x-y_k\right\| = \min_{y\in Y}\left\|x-y\right\| =: d.$$

Using the parallelogram identity from Prop. 1.3.8 we deduce

$$\left\|(x-y_k)-(x-y_l)\right\|^2 + \|(x-x_k)+(x-x_l)\|^2 \overset{(1.3c)}{=} 2\left(\left\|x-y_k\right\|^2 + \left\|x-y_l\right\|^2\right)$$

and consequently

$$\begin{aligned}\left\|y_l-y_k\right\|^2 &= 2\left(\left\|x-y_k\right\|^2 + \left\|x-y_l\right\|^2 - 2\left\|x-\tfrac{y_k+y_l}{2}\right\|^2\right)\\ &\leq 2\left(\left\|x-y_k\right\|^2 + \left\|x-y_l\right\|^2 - 2d^2\right) \xrightarrow[k,l\to\infty]{} 0,\end{aligned}$$

since $\frac{y_k+y_l}{2}\in Y$ holds due to the convexity of $Y$. Hence, $(y_k)_{k\in\mathbb{N}}$ is a Cauchy sequence with limit in the complete space $X$. Moreover, the closedness of $Y$ ensures $y := \lim_{k\to\infty} y_k \in Y$ and the continuity of the norm $\|\cdot\|$ (cf. Lemma 2.2.3) implies $\left\|x-y\right\| = d$. If another point $y^*\in Y$ has this property, we deduce as above that

$$\left\|y-y^*\right\|^2 \leq 2\left(\left\|x-y\right\|^2 + \left\|x-y^*\right\|^2 - 2d^2\right) = 0$$

and therefore $y = y^*$, i.e. $P(x) := y$ is uniquely determined.

Now choose $y\in Y$, $t\in[0,1]$ and due to $(1-t)P(x)+ty\in Y$ one has

$$\begin{aligned}\|x-P(x)\|^2 = d^2 \leq \left\|x-\left[(1-t)P(x)+ty\right]\right\|^2\\ = \|x-P(x)\|^2 - 2t\Re\left\langle x-P(x), y-P(x)\right\rangle + O(t^2),\end{aligned}$$

which implies $\Re\langle x-P(x), a-P(x)\rangle \leq 0$. Conversely, in case this holds, we deduce

$$\begin{aligned}\left\|x-y\right\|^2 &= \left\|x-P(x)+P(x)-y\right\|^2\\ &= \|x-P(x)\|^2 + 2\Re\left\langle x-P(x), P(x)-y\right\rangle + \left\|P(x)-y\right\|^2 \leq \|x-P(x)\|^2\end{aligned}$$

and obtain the claim. □

In an inner product space $X$ it is an immediate consequence of the Cauchy-Schwarz inequality from Prop. 1.3.7 that $\langle\cdot,y\rangle : X\to\mathbb{K}$ defines a bounded functional for every $y\in X$. Now we are interested in the converse situation, whether bounded functionals can be written as inner products. There is an affirmative answer to this problem in complete spaces. Indeed, the following important result states that every bounded functional $T: X\to\mathbb{K}$ on a Hilbert space can be represented by an inner product in a bijective[4] way. We also refer to [NS82, p. 345, Thm. 5.21.1 and p. 350, Ex. 7].

[4] a mapping $f: X\to Y$ between sets $X, Y$ is called *bijective*, if it is *one-to-one* (i.e. $f(x_1) = f(x_2)$ implies $x_1 = x_2$ for all $x_1, x_2\in X$) and *onto* (i.e. $f(X) = Y$)

**Theorem 2.4.2** (Riesz representation theorem)**.** *Let* $(X, \langle\cdot,\cdot\rangle)$ *be a Hilbert space. For every functional* $T \in L(X,\mathbb{K})$ *there exists a unique* $y_T \in X$ *with* $\langle x, y_T\rangle = Tx$ *for all* $x \in X$*. The mapping* $T \mapsto y_T$ *is bijective with*

$$y_{\alpha_1 T_1+\alpha_2 T_2} = \overline{\alpha_1} y_{T_1} + \overline{\alpha_2} y_{T_2} \quad \textit{for all } \alpha_1,\alpha_2 \in \mathbb{K},\ T_1, T_2 \in L(X,\mathbb{K})$$

*and finally* $\|y_T\| = \|T\|$*.*

*Proof.* We define the mapping $(Jx)(y) := \langle y, x\rangle$ and proceed in three steps:

(I) Due to the Cauchy-Schwarz inequality from Prop. 1.3.7 it is

$$\left|(Jx)(y)\right| \leq \|x\| \left\|y\right\| \quad \text{for all } y \in X$$

and thus $Jx \in L(X,\mathbb{K})$ with $\|Jx\| \leq \|x\|$. On the other hand, one has $|(Jx)(x)| = \|x\|^2$ yielding $\|Jx\| \geq \|x\|$ for all $x \neq 0$. This ensures $\|Jx\| = \|x\|$ and $J : X \to L(X,\mathbb{K})$ is one-to-one.

(II) Let $T \in L(X,\mathbb{K})$ be nonzero. Then $Y := \{x \in X : Tx = 0\}$ is nonempty closed and, as a subspace (cf. Ex. 1.5.12(1)), also convex. Referring to Thm. 2.4.1 there exists an orthogonal projection $P : X \to Y$, we choose $e \in X$ with $Te = 1$ and define $x_0 := e - Pe$. This implies $Tx_0 = Te - TPe = 1$ and in particular $x_0 \neq 0$. Using Thm. 2.4.1(b) it is $\Re\langle y - Pe, x_0\rangle = \Re\langle y - Pe, e - Pe\rangle \leq 0$ for all $y \in Y$ and thus $\langle y, x_0\rangle = 0$ for all $y \in Y$, since $Pe \in Y$ and $Y$ is a subspace. For $x \in X$ it is

$$x = \underbrace{x - Tx \cdot x_0}_{\in Y} + Tx \cdot x_0,$$

hence $\langle x, x_0\rangle = \langle Tx \cdot x_0, x_0\rangle = \|x_0\|^2\, Tx$ and finally $Tx = \left\langle x, \frac{x_0}{\|x_0\|^2}\right\rangle = \left(J\frac{x_0}{\|x_0\|^2}\right)(x)$. This establishes that $J : X \to L(X,\mathbb{K})$ is also onto.

(III) The above steps ensure that $J : X \to L(X,\mathbb{K})$ is bijective and by construction, $y_T := J^{-1}T$ fulfills the above assertions. □

Next we generalize the Riesz representation Thm. 2.4.2 from inner products to general bilinear forms:

**Theorem 2.4.3** (of Lax-Milgram)**.** *Let* $(X, \langle\cdot,\cdot\rangle)$ *be a real Hilbert space and let* $a : X \times X \to \mathbb{R}$ *be a continuous bilinear form, which is also* coercive*, i.e. there exists a* $c > 0$ *with*

$$c\,\|x\|^2 \leq a(x,x) \quad \textit{for all } x \in X. \tag{2.4a}$$

*For every bounded functional* $T \in L(X,\mathbb{R})$ *there exists a unique* $x_T \in X$ *with* $a(x, x_T) = Tx$ *for all* $x \in X$*, and one has* $\|x_T\| \leq \frac{1}{c}\,\|T\|$*.*

*Remark* 2.4.4 (variational problem). If the bilinear form $a$ is symmetric, then the vector $x_T \in X$ is the (unique) global minimum of the quadratic functional $F : X \to \mathbb{R}$, $F(x) := \frac{1}{2}a(x,x) - Tx$; this will be shown in Exercise 2.4.5(3). In this sense, the Lax-Milgram theorem is a tool to solve variational problems.

*Proof.* Throughout the proof, let $\alpha_1, \alpha_2 \in \mathbb{R}$. For every fixed $x \in X$ we define a functional $T_x : X \to \mathbb{R}$ by $T_x := a(x,\cdot)$ with the following properties:

- $T_x$ is linear, since for $y_1, y_2 \in X$ we obtain

  $$T_x(\alpha_1 y_1 + \alpha_2 y_2) = a(x, \alpha_1 y_1 + \alpha_2 y_2) = \alpha_1 a(x, y_1) + \alpha_2 a(x, y_2) = \alpha_1 T_x y_1 + \alpha_2 T_x y_2.$$

- $T_x$ is bounded: In order to verify this, we remark that $a : X \times X \to \mathbb{R}$ is assumed to be continuous. Thus, Thm. 2.2.10 guarantees that there exists a $C \geq 0$ with $\left|a(x,y)\right| \leq C\left\|x\right\|\left\|y\right\|$. This implies the estimate

  $$\left|T_x(y)\right| = \left|a(x,y)\right| \leq C\left\|x\right\|\left\|y\right\| \quad \text{for all } y \in X$$

  and consequently $T_x \in L(X,\mathbb{R})$.
- One has the linearity relation $T_{\alpha_1 x_1 + \alpha_2 x_2} = \alpha_1 T_{x_1} + \alpha_2 T_{x_2}$ for all $x_1, x_2 \in X$ (note that $X$ is a real Hilbert space).

Hence, we can apply the Riesz representation Thm. 2.4.2 and obtain that there exist unique $y_x, y \in X$ with[5] $\left\langle \xi, y_x \right\rangle = T_x\xi$, $\left\langle \xi, y \right\rangle = T\xi$ for all $\xi \in X$. With this we are in a position to define the mapping $F : X \to X$, $F(x) := x - \rho(y_x - y)$ for some $\rho > 0$ to be specified later. For all $x_1, x_2 \in X$ we obtain

$$\left\|y_{x_1 - x_2}\right\| = \left\|T_{x_1 - x_2}\right\| = \left\|a(x_1 - x_2, \cdot)\right\| \leq C\left\|x_1 - x_2\right\| \tag{2.4b}$$

and this implies

$$\begin{aligned}
\left\|F(x_1) - F(x_2)\right\|^2 &= \left\|x_1 - x_2 - \rho(y_{x_1} - y_{x_2})\right\|^2 \\
&= \left\|x_1 - x_2\right\|^2 - 2\rho\left\langle x_1 - x_2, y_{x_1 - x_2}\right\rangle + \rho^2\left\|y_{x_1 - x_2}\right\|^2 \\
&= \left\|x_1 - x_2\right\|^2 - 2\rho T_{x_1 - x_2}(x_1 - x_2) + \rho^2 T_{x_1 - x_2} y_{x_1 - x_2} \\
&= \left\|x_1 - x_2\right\|^2 - 2\rho a(x_1 - x_2, x_1 - x_2) + \rho^2 a(x_1 - x_2, y_{x_1 - x_2}) \\
&\overset{(2.4a)}{\leq} \left\|x_1 - x_2\right\|^2 - 2\rho c\left\|x_1 - x_2\right\|^2 + \rho^2 C\left\|x_1 - x_2\right\|\left\|y_{x_1 - x_2}\right\| \\
&\overset{(2.4b)}{\leq} \left(1 - 2\rho c + \rho^2 C^2\right)\left\|x_1 - x_2\right\|^2.
\end{aligned}$$

If we choose $\rho \in \left(0, \frac{2c}{C^2}\right)$ one obtains $1 - 2\rho c + \rho^2 C^2 \in (0,1)$. Consequently, if we take the square root in the above inequality, one arrives at

$$\left\|F(x_1) - F(x_2)\right\| \leq \underbrace{\sqrt{1 - 2\rho c + \rho^2 C^2}}_{\in(0,1)}\left\|x_1 - x_2\right\| \quad \text{for all } x_1, x_2 \in X.$$

[5] in the notation of Thm. 2.4.2 we have $y_x = y_{T_x}$and $y = y_T$, but we avoid this cumbersome notation

This guarantees that the mapping $F : X \to X$ is a contraction on the complete space $X$ and the contraction mapping principle from Thm. 2.3.6 guarantees that there exists a unique $x^* \in X$ with $x^* = F(x^*) = x^* - \rho(y_{x^*} - y)$. The last relation is equivalent to $y_{x^*} = y$ and consequently

$$a(x, x^*) = T_{x^*}x = \langle x, y_{x^*} \rangle = \langle x, y \rangle = Tx \quad \text{for all } x \in X.$$

Moreover, we have the estimate

$$c \left\| x^* \right\| \overset{(2.4a)}{\leq} \frac{|a(x,x^*)|}{\|x\|} = \frac{|Tx|}{\|x\|} \leq \sup_{\|x\|=1} |Tx| = \|T\|.$$

This finishes our present proof, if we set $x_T := x^*$. □

See [NS82, pp. 112] for further results on complete metric spaces. More information on Banach spaces can be found in [NS82, pp. 215ff] and Hilbert spaces are treated in [NS82, pp. 272].

Let us also conclude this chapter with a resumé: The abstract concepts from our first chapter became vitalized by equipping them with the notion of continuity. Here, it turned out that continuity of linear operators is equivalent to their boundedness (cf. Thm. 2.2.9). The prime example of unbounded mappings have been differential operators, while integral operators are typically bounded. This can be interpreted as deeper reason why the numerical approximation of integrals is more stable than of derivatives — in particular when it comes to higher order derivatives. Moreover, this is one reason why we introduce a variational formulation, which is based on integration, for boundary value problems of differential equations (see Section 3.2).

The most important theoretical results of this course appeared in Section 2.3. This underlines the importance of complete spaces like $B(\Omega)$, $C(\Omega)$ or $L^p(\Omega)$ equipped with their canonical norms. The contraction mapping principle from Thm. 2.3.6 is of eminent importance:

- In pure mathematics it yields various existence and uniqueness results. We only mention the Picard-Lindelöf theorem on ordinary differential equations $\dot{u} = f(t, u)$ with a Lipschitz-continuous right hand side $f$.
- Since Thm. 2.3.6 also provides an approximation algorithm (the so-called *fixed-point iteration*) and an error estimate, it is also an essential tool in numerical analysis. So, the convergence of various iterative schemes can be shown using Thm. 2.3.6.

Finally, the crucial Lax-Milgram Thm. 2.4.3 enables us to prove the existence and uniqueness of solutions to elliptic boundary value problems (see Chapter 3).

*Exercises* 2.4.5. Under the assumptions of Thm. 2.4.3 suppose that $a : X \times X \to \mathbb{R}$ is symmetric, i.e. $a(x, y) = a(y, x)$ for all $x, y \in X$. Show that the quadratic functional $F : X \to \mathbb{R}$, $F(x) := \frac{1}{2}a(x, x) - Tx$ achieves its absolute minimum at $x = x_T$.
*Hint:* Try to establish the relation $F(x) - F(x_T) \geq \frac{c}{2} \|x - x_T\|^2$.

# Chapter 3
# Boundary value problems

In this chapter, we will introduce yet another class of function spaces, namely so-called Sobolev spaces. The basic reason for their importance is that they allow to apply ideas and methods from functional analysis to boundary value problems in mechanics and general physics.

## 3.1 Sobolev spaces

The solutions of differential equations are differentiable functions. In this spirit, a function $u : \Omega \to \mathbb{K}$ is said to be a *classical solution* of an $m$th order[1] differential equation, if it satisfies the differential equation and is of class $C^m(\Omega)$, $m \in \mathbb{N}$.

However, classical function spaces as $C^m(\Omega)$ have limitations when dealing with various (partial) differential equations and/or nonsmooth coefficients. This might be illustrated by the following examples:

*Example* 3.1.1. Consider the simple 2nd order differential equation

$$xu''(x) = |x|, \tag{3.1a}$$

equipped with the boundary conditions $u(-1) = -1$, $u(1) = 1$. We assume that there exists a classical solution $u \in C^2[-1,1]$. However, from (3.1a) we deduce that $u$ satisfies[2] $u''(x) = \frac{|x|}{x} = \operatorname{sgn} x$ for $x \neq 0$. This shows that $u$ cannot have a 2nd order derivative, which is continuous in 0. In particular, the function $u(x) := x\,|x|$ is not of class $C^2$.

[1] the *order* of a differential equation (or of an operator) is the highest occurring derivative

[2] the *sign function* $\operatorname{sgn} : \mathbb{R} \to \mathbb{R}$ is defined by

$$\operatorname{sgn} x := \begin{cases} 1, & x > 0, \\ 0, & x = 0, \\ -1, & x < 0 \end{cases}$$

To prepare our next example, we remind the reader of the *integration by parts* formula (also know as *Green's first identity*) from vector analysis. Thereto, suppose that $\Omega$ is an open bounded subset of $\mathbb{R}^d$ with a piecewise smooth boundary $\partial\Omega$. If $u, \phi$ are two continuously differentiable functions on the closure $\Omega \cup \partial\Omega$, then

$$\int_\Omega \frac{\partial u(x)}{\partial x_i}\phi(x)\,dx = \int_{\partial\Omega} u(x)\phi(x)d\nu_i(x) - \int_\Omega u(x)\frac{\partial\phi(x)}{\partial x_i}\,dx \tag{3.1b}$$

for all $i = 1, \dots, d$, where $d\nu_i(x)$ denotes the $i$th component of the outward unit surface normal to the boundary $\partial\Omega$. Especially for functions $\phi \in C_0^1(\Omega)$ the boundary term vanishes and we obtain

$$\int_\Omega \frac{\partial u(x)}{\partial x_i}\phi(x)\,dx = -\int_\Omega u(x)\frac{\partial\phi(x)}{\partial x_i}\,dx \quad \text{for all } i = 1, \dots, d. \tag{3.1c}$$

*Example* 3.1.2. Let $\Omega \subseteq \mathbb{R}^2$ be a domain with piecewise smooth boundary. We are interested in the elliptic boundary value problem

$$\begin{cases} -\Delta u(x) = f(x), & \text{if } x \in \Omega, \\ u(x) = 0, & \text{if } x \in \partial\Omega, \end{cases} \tag{3.1d}$$

with the Laplace operator $\Delta u := \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}$. In order to tackle the problem (3.1d) we first establish a related formulation: Thereto, we multiply $f(x) = -\Delta u(x)$ with a function $\phi \in C_0^1(\Omega)$ and integrate over $\Omega$. Using integration by parts this yields[3]

$$\int_\Omega f\phi = -\int_\Omega \Delta u\phi = -\int_\Omega \frac{\partial^2 u}{\partial x_1^2}\phi + \frac{\partial^2 u}{\partial x_2^2}\phi \overset{(3.1c)}{=} \int_\Omega \frac{\partial u}{\partial x_1}\frac{\partial\phi}{\partial x_1} + \int_\Omega \frac{\partial u}{\partial x_2}\frac{\partial\phi}{\partial x_2}$$

for all $\phi \in C_0^1(\Omega)$. This, in turn, is equivalent to

$$a(u, \phi) = T\phi \quad \text{for all } \phi \in C_0^1(\Omega) \tag{3.1e}$$

with a bilinear form $a : C^1(\Omega) \times C^1(\Omega) \to \mathbb{R}$ and a functional $T : C^1(\Omega) \to \mathbb{R}$,

$$a(u, \phi) := \int_\Omega \frac{\partial u}{\partial x_1}\frac{\partial\phi}{\partial x_1} + \int_\Omega \frac{\partial u}{\partial x_2}\frac{\partial\phi}{\partial x_2}, \qquad T\phi := \int_\Omega f\phi.$$

However, an attempt to solve (3.1e) using the Lax-Milgram Thm. 2.4.3 immediately fails since $C^1(\Omega)$ is not a Hilbert space.

This backlash motivates the subsequent notion for a weak derivative. Thereto, let $\Omega \subseteq \mathbb{R}^d$ be an open set. A *multiindex* $\alpha = (\alpha_1, \dots, \alpha_d)$ is a $d$-tuple with entries $\alpha_i \in \mathbb{N}_0$ and $|\alpha| := \alpha_1 + \dots + \alpha_d$ denotes its *length*. This yields a convenient notation for partial derivatives of a smooth function $u : \Omega \to \mathbb{R}$, namely

[3] as a matter of convenience we sometimes neglect the dependence of the integrands on the variable $x$ and the symbol $dx$

$$D^\alpha u := \frac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}};$$

for instance $D^{(2,0)} u = \frac{\partial^2 u}{\partial x_1^2}$, $D^{(3,2,1)} u = \frac{\partial^6 u}{\partial x_1^3 \partial x_2^2 \partial x_3^1}$ or $D^{(0,\dots,0)} u = u$.

**Definition 3.1.3.** Let $\alpha \in \mathbb{N}_0^d$ be a multiindex and $u \in L^2(\Omega)$. We say that $v \in L^2(\Omega)$ is the *$\alpha$th weak partial derivative* of $u$, written $v = d^\alpha u$, if

$$\int_\Omega v(x)\phi(x)\,dx = (-1)^{|\alpha|} \int_\Omega u(x) D^\alpha \phi(x)\,dx \quad \text{for all } \phi \in C_0^{|\alpha|}(\Omega).$$

*Remark* 3.1.4. (1) Whereas a classical derivative $u'$ of a differentiable function $u : \Omega \to \mathbb{R}$ is defined point wise via $u'(x) = \frac{du}{dx}(x)$ for all $x \in \Omega$, weak derivatives are defined globally on $\Omega$ right from the beginning.

(2) Given $m \in \mathbb{N}_0$, a function $u \in L^2(\Omega)$ is called $m$-times weakly differentiable, if all the $\alpha$th weak partial derivatives $d^\alpha u \in L^2(\Omega)$ exist for $|\alpha| \le m$.

(3) In case $\Omega = (a, b) \subseteq \mathbb{R}$ is an interval one can obtain weak derivatives as follows: If $u \in L^2(a, b)$ has a weak derivative $du \in L^2(a, b)$, then $u$ is continuous, its classical derivative $u' : (a, b) \to \mathbb{K}$ exists everywhere besides a set $I$ of measure zero and one has $du(x) = u'(x)$ for all $x \in (a, b) \setminus I$ (see [Eva98, p. 280, Thm. 5]).

In order to compute weak derivatives we restrict to domains $\Omega \subseteq \mathbb{R}$.

*Example* 3.1.5. Let $a < b$ be given real numbers and $u \in C^1[a, b]$. Then $u$ is also an $L^2$-function and its derivative satisfies

$$\int_a^b u'(x)\phi(x)\,dx = u(x)\phi(x)|_a^b - \int_a^b u(x)\phi'(x)\,dx \quad \text{for all } \phi \in C_0^1[a, b].$$

Due to $\phi(a) = \phi(b) = 0$ this means $\int_a^b u'(x)\phi(x)\,dx = -\int_a^b u(x)\phi'(x)\,dx$ and consequently the classical derivative $u' : [a, b] \to \mathbb{R}$ is the weak derivative of $u$. Using mathematical induction it follows that a function $u \in C^m[a, b]$, $m \in \mathbb{N}$, has an $m$th weak derivative $d^m u = u^{(m)}$. Indeed, one also has

$$\int_a^b u^{(m)}(x)\phi(x)\,dx = (-1)^m \int_a^b u(x)\phi^{(m)}(x)\,dx \quad \text{for all } \phi \in C_0^m[a, b].$$

*Example* 3.1.6. Let $\Omega = [-1, 1]$ and obviously $u : [-1, 1] \to \mathbb{R}$, $u(x) := |x|$ is not differentiable. However, $u$ is weakly differentiable with derivative $v : [-1, 1] \to \mathbb{R}$, $v(x) = \operatorname{sgn} x$, since we have

$$\begin{aligned} \int_{-1}^1 v(x)\phi(x)\,dx &= \int_{-1}^0 v(x)\phi(x)\,dx + \int_0^1 v(x)\phi(x)\,dx \\ = \quad -\int_{-1}^0 \phi(x)\,dx + \int_0^1 \phi(x)\,dx &= \int_{-1}^0 u'(x)\phi(x)\,dx + \int_0^1 u'(x)\phi(x)\,dx \end{aligned}$$

$$\begin{aligned}
&\overset{(3.1b)}{=} \quad u(x)\phi(x)|_{-1}^{0} - \int_{-1}^{0} u(x)\phi'(x)\,dx + u(x)\phi(x)|_{0}^{1} - \int_{0}^{1} u(x)\phi'(x)\,dx \\
&= \quad u(0)\phi(0) - \int_{-1}^{0} u(x)\phi'(x)\,dx - u(0)\phi(0) - \int_{0}^{1} u(x)\phi'(x)\,dx \\
&= \quad -\int_{-1}^{1} u(x)\phi'(x)\,dx \quad \text{for all } \phi \in C_0^1[-1,1].
\end{aligned}$$

*Example* 3.1.7. Let $\Omega = (0,2)$ and define the discontinuous $L^2$-function

$$u : (0,2) \to \mathbb{R}, \qquad\qquad u(x) := \begin{cases} x, & 0 < x \le 1, \\ 2, & 1 < x < 2. \end{cases}$$

This function is not weakly differentiable. To see this, we suppose the contrary, i.e., that there exists a function $v \in L^2(0,2)$ satisfying

$$\int_0^2 u(x)\phi'(x)\,dx = -\int_0^2 v(x)\phi(x)\,dx \quad \text{for all } \phi \in C_0^1(0,2).$$

Now choose a sequence $\phi_n \in C_0^1(0,2)$ with $\phi_n(x) \in [0,1]$, $\phi_n(1) = 1$ for all $n \in \mathbb{N}$ and $\lim_{n\to\infty} \phi_n(x) = 0$ for all $x \neq 1$, and we obtain

$$\begin{aligned}
-\int_0^2 v(x)\phi_n(x)\,dx = \int_0^2 u(x)\phi_n'(x)\,dx &= \int_0^1 x\phi_n'(x)\,dx + 2\int_1^2 \phi_n'(x)\,dx \\
&= x\phi_n(x)|_0^1 - \int_0^1 \phi_n(x)\,dx + 2\phi_n(2) - 2\phi_n(1) \\
&= -\phi_n(1) - \int_0^1 \phi_n(x)\,dx \quad \text{for all } n \in \mathbb{N}.
\end{aligned}$$

This relation is equivalent to $\phi_n(1) = \int_0^2 v(x)\phi_n(x)\,dx - \int_0^1 \phi_n(x)\,dx$ for $n \in \mathbb{N}$. By the dominated convergence theorem (see [NS82, p. 579, Thm. D.8.4]) we can pass over to the limit $n \to \infty$ and obtain the contradiction

$$1 = \lim_{n\to\infty} \phi_n(1) = \lim_{n\to\infty} \int_0^2 v(x)\phi_n(x)\,dx - \lim_{n\to\infty} \int_0^1 \phi_n(x)\,dx = 0.$$

Thus, $u$ is not weakly differentiable.

Next we show that weak derivatives are uniquely determined and linear:

**Proposition 3.1.8** (properties of weak derivatives)**.** *Let $m \in \mathbb{N}_0$, $\alpha \in \mathbb{N}_0^d$ be a multiindex and $u_1, u_2 \in L^2(\Omega)$.*

*(a) If $d^\alpha u_1$ exists, then it is uniquely determined (up to a set of measure zero).*
*(b) If $d^\alpha v_1, d^\alpha v_2 \in L^2(\Omega)$ exist, then also the $\alpha$th weak derivative of the linear combination $\beta_1 v_1 + \beta_2 v_2 \in L^2(\Omega)$ exists and*

$$d^\alpha(\beta_1 v_1 + \beta_2 v_2) = \beta_2 d^\alpha v_1 + \beta_2 d^\alpha v_2 \quad \textit{for all } \beta_1, \beta_2 \in \mathbb{K}.$$

*Proof.* (a) Suppose that the functions $v_1, v_2 \in L^2(\Omega)$ both satisfy

$$\int_\Omega u_1(x) D^\alpha \phi(x)\, dx = (-1)^{|\alpha|} \int_\Omega v_1(x)\phi(x)\, dx = (-1)^{|\alpha|} \int_\Omega v_2(x)\phi(x)\, dx$$

and we obtain $\int_\Omega (v_1(x) - v_2(x))\phi(x)\, dx = 0$ for all $\phi \in C_0^1(\Omega)$. This, however, implies $v_1(x) = v_2(x)$ for all $x \in \Omega$ except from a set of measure zero.

(b) For the proof of the linearity we refer to [Eva98, p. 247, Theorem 1(ii)]. □

In order to close this section, we denote the set of all functions whose $\alpha$th weak derivatives $d^\alpha u \in L^2(\Omega)$ exist for all multiindices $|\alpha| \le m$, by $H^m(\Omega)$.

**Proposition 3.1.9** (Sobolev spaces). *Let $m \in \mathbb{N}_0$. The* Sobolev spaces $H^m(\Omega)$ *are Hilbert spaces with the inner product*

$$\langle u, v\rangle := \sum_{|\alpha|\le m} \int_\Omega d^\alpha u(x) \overline{d^\alpha v(x)}\, dx.$$

*Remark* 3.1.10. Let $m \in \mathbb{N}_0$.

(1) One has $H^0(\Omega) = L^2(\Omega)$. For $m \le n$ the inclusion $H^n(\Omega) \subseteq H^m(\Omega)$ holds with the norm inequality $\|u\|_{H^m(\Omega)} \le \|u\|_{H^n(\Omega)}$.

(2) Explicitly, the natural norm on $H^m(\Omega)$ reads as

$$\|u\|_{H^m(\Omega)} = \sqrt{\sum_{|\alpha|\le m} \int_\Omega |d^\alpha u(x)|^2\, dx} = \sqrt{\sum_{|\alpha|\le m} \|d^\alpha u\|^2_{L^2(\Omega)}}. \tag{3.1f}$$

(3) For simplicity we also define the subspaces $H_0^m(\Omega) \subseteq H^m(\Omega)$ as follows

$$H_0^m(\Omega) := \left\{u \in H^m(\Omega) : d^\alpha u(x) = 0 \text{ for } x \in \partial\Omega, |\alpha| < m\right\}$$

and refer to [Eva98, p. 245] for a more precise definition.

*Proof.* See [Eva98, p. 249, Theorem 3]. □

In Ex. 3.1.5 we have seen the inclusion $C^1[a,b] \subset H^1[a,b]$, while Ex. 3.1.6 shows $C^1[a,b] \ne H^1[a,b]$. For more general domains $\Omega \subseteq \mathbb{R}^d$ one has

*Remark* 3.1.11 (Sobolev embedding). Suppose that $\Omega \subseteq \mathbb{R}^d$ is an open, bounded set with a $C^1$-boundary[4]. If $m > \frac{d}{2}$, then the inclusion $H^m(\Omega) \subseteq C^{m-\left[\frac{d}{2}\right]-1}(\Omega)$

[4] this means the boundary $\partial\Omega$ can be parametrized by a $C^1$-function $\gamma : S \subseteq \mathbb{R}^{d-1} \to \mathbb{R}^d$, i.e. one has $\partial\Omega = \gamma(S)$

holds true[5] (cf. [Eva98, p. 270, Thm. 6(ii)]). In particular, weakly differentiable functions are also differentiable with a lower smoothness.

*Exercises* 3.1.12. Solve the following problems:

(1) Let $\Omega_1, \Omega_2 \subseteq \mathbb{R}^2$ be given by

$$\Omega_1 := \left\{(x_1,x_2) \in \mathbb{R}^2 : x_1^2 + x_2^2 < 1\right\}, \quad \Omega_2 := \left\{(x_1,x_2) \in \mathbb{R}^2 : \max\{|x_1|,|x_2|\} < 1\right\}$$

Given the functions $\phi_1 : \Omega_1 \to \mathbb{R}$, $\phi_1(x_1,x_2) := x_1^2 + x_2^2 - 1$ and $\phi_2 : \Omega_2 \to \mathbb{R}$, $\phi_2(x_1,x_2) := 1 - \max\{|x_1|,|x_2|\}$ compute the integrals

$$\int_{\partial\Omega_i} \arctan e^{x_1 x_2} \phi_i(x_1,x_2) d\nu_j(x_1,x_2) \quad \text{for } i,j = 1,2.$$

(2) Compute the weak derivative of $u : (0,2) \to \mathbb{R}$, $u(x) := \begin{cases} x, & 0 < x \leq 1, \\ 1, & 1 < x < 2. \end{cases}$

(3) Compute the 2nd order weak derivative of $u : (-1,1) \to \mathbb{R}$, $u(x) := x|x|$.

## 3.2 Variational formulation

A mathematically profound introduction to the theory of partial differential equations (PDEs for short) and their boundary value problems can be found in, for instance, [Eva98]. In order to demonstrate how fruitful our abstract functional analytical tools are, we restrict to Dirichlet boundary value problems. Their so-called weak or variational formulation not only allows a mathematically elegant existence and uniqueness proof, but also allows a satisfying treatment of simple boundary value problems as in Ex. 3.1.1.

*Example* 3.2.1 (Poisson problem). Let us consider a membrane stretched over a bounded region $\Omega \subset \mathbb{R}^2$ exposed to an external force $f(x_1,x_2)$. If the displacement of the membrane is denoted by $u(x_1,x_2)$, then $u$ and $f$ are related by

$$\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = f \quad \text{in } \Omega, \qquad\qquad u = 0 \quad \text{in } \partial\Omega.$$

This is a boundary value problem for a 2nd order partial differential equation.

We follow the explanations from [NS82, pp. 486ff] and suppose that $\Omega \subseteq \mathbb{R}^d$ is a bounded and open set.

[5] here, $[\cdot] : \mathbb{R} \to \mathbb{Z}$ denotes the *greatest integer function* given by $[x] := \max\{k \in \mathbb{Z} : k \leq x\}$. For instance, $[\frac{1}{2}] = 0$, $[\pi] = 3$ or $[k] = k$ for every $k \in \mathbb{Z}$

**Definition 3.2.2** (differential operator)**.** Suppose that $\alpha, \beta \in \mathbb{N}_0^d$ are multiindices and $a^{\alpha\beta} : \Omega \to \mathbb{K}$ are continuously differentiable functions with a continuous extension to the closure $\bar{\Omega}$. Then

$$L : C^2(\Omega) \to C(\Omega), \quad (Lu)(x) := \sum_{0 \le |\alpha|, |\beta| \le 1} (-1)^{|\alpha|} D^\alpha \left( a^{\alpha\beta}(x) D^\beta u(x) \right) \quad (3.2a)$$

is called *2nd order differential operator.*

*Remark* 3.2.3. (1) Alternatively, with certain coefficient functions $a_{ij} \in C^1(\Omega)$, $b_i, c \in C(\Omega)$ one can write $L$ as

$$(Lu)(x) := -\sum_{i,j=1}^d \frac{\partial}{\partial x_j} \left( a_{ij}(x) \frac{\partial u(x)}{\partial x_i} \right) + \sum_{i=1}^d b_i(x) \frac{\partial u(x)}{\partial x_i} + c(x) u(x). \qquad (3.2b)$$

The interested reader can easily derive a relation between the coefficients $a^{\alpha\beta}$ in (3.2a) and $a_{ij}, b_i, c$ in (3.2b).

(2) We sometimes interpret the symbol $D^\alpha u$ also as weak derivative $d^\alpha u$ and therefore obtain an operator $L : H^2(\Omega) \to L^2(\Omega)$.

**Definition 3.2.4** (symmetric, elliptic differential operator)**.** Suppose that $\alpha, \beta \in \mathbb{N}_0^d$ are multiindices and $a^{\alpha\beta} : \Omega \to \mathbb{K}$ are continuously differentiable functions. A 2nd order differential operator $L : C^2(\Omega) \to C(\Omega)$ is called

(a) *symmetric*, if $a^{\alpha\beta}(x) = \overline{a^{\beta\alpha}(x)}$ for all $x \in \Omega$,

(b) *uniformly elliptic*, if $L$ is symmetric and there exists a real $K > 0$ such that

$$\Re \sum_{|\alpha|=|\beta|=1} a^{\alpha\beta}(x) \xi^\alpha \xi^\beta \ge K \sum_{i=1}^d \xi_i^2 \quad \text{for all } \xi \in \mathbb{R}^d, x \in \Omega$$

with $\xi^\alpha := \xi_1^{\alpha_1} \cdot \ldots \cdot \xi_d^{\alpha_d}$.

*Remark* 3.2.5. (1) If the differential operator $L : C_0^2(\Omega) \to C_0(\Omega)$ is symmetric, then one has (cf. [NS82, p. 511])

$$\langle u, Lv \rangle_{L^2(\Omega)} = \langle Lu, v \rangle_{L^2(\Omega)} \quad \text{for all } u, v \in C_0^2(\Omega).$$

(2) If the coefficient functions $a^{\alpha\beta} \in C(\Omega)$ are real-valued, then the square matrix $(a^{\alpha\beta}(x))_{|\alpha|=|\beta|=1} \in \mathbb{R}^{d \times d}$ is symmetric and positively definite. In particular, the associated quadratic form $y \mapsto \langle y, Ay \rangle$ has ellipses (if $d = 2$) or ellipsoids (if $d = 3$) as level sets for all $x \in \Omega$.

*Example* 3.2.6. A prototypical uniformly elliptic 2nd order differential operator is the negative Laplace operator given by $Lu := -\Delta u$ with $\Delta u := \sum_{k=1}^{d} \frac{\partial^2 u}{\partial x_k^2}$. Indeed, it fits into the framework of Def. 3.2.4 with constant coefficients

$$a^{\alpha\beta}(x) \equiv \begin{cases} 1, & \text{if } \alpha = \beta \text{ with } |\alpha| = |\beta| = 1, \\ 0, & \text{else,} \end{cases}$$

as well as in the notation of (3.2b) with

$$a_{ij}(x) \equiv \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{else,} \end{cases}, \qquad b_i(x) \equiv 0, \qquad c(x) \equiv 0 \quad \text{on } \Omega.$$

From this we see that $L = -\Delta$ is symmetric and uniformly elliptic with $K = 1$.

For simplicity we now **restrict to real-valued boundary value problems**. This means all the functions $u, f, a^{\alpha\beta}$ etc. have values in $\mathbb{R}$.

The *classical Dirichlet problem* for the operator $L$ is as follows: Let $f$ be a continuous function on $\Omega$. Find a function $u \in C^2(\bar{\Omega})$ with the property that

$$Lu = f \quad \text{in } \Omega, \qquad\qquad u = 0 \quad \text{in } \partial\Omega; \tag{3.2c}$$

then $u$ would be called a *classical solution* of (3.2c).

In order to motivate the concept of a weak solution, we multiply $Lu = f$ with a function $\phi \in C_0^1(\Omega)$ and integrate over $\Omega$ yielding the relation

$$\langle \phi, Lu \rangle_{L^2(\Omega)} = \int_\Omega \phi(x) Lu(x)\, dx = \int_\Omega \phi(x) f(x)\, dx = \langle \phi, f \rangle_{L^2(\Omega)}$$

Moreover, integration by parts of the left hand side using (3.1c) implies

$$\langle \phi, Lu \rangle_{L^2(\Omega)} = \int_\Omega \phi(x) Lu(x)\, dx = a(\phi, u)$$

with the bilinear form

$$\begin{aligned} a(\phi, u) &:= \sum_{0 \le |\alpha|, |\beta| \le 1} \int_\Omega a^{\alpha\beta}(x) d^\alpha \phi(x) d^\beta u(x)\, dx \\ &\overset{(3.2b)}{=} \int_\Omega \sum_{i,j=1}^{d} a_{ij}(x) \frac{\partial \phi(x)}{\partial x_j} \frac{\partial u(x)}{\partial x_i} + \sum_{i=1}^{d} b_i(x) \frac{\partial \phi(x)}{\partial x_i} u(x) + c(x) u(x) v(x)\, dx. \end{aligned}$$

Having this at our disposal, the *generalized Dirichlet problem* for $L$ reads as follows: Given $f \in L^2(\Omega)$ find a function $u \in H_0^1(\Omega)$ satisfying the relation

$$a(\phi, u) = \langle \phi, f \rangle_{L^2(\Omega)} \quad \text{for all } \phi \in C_0^1(\Omega). \tag{3.2d}$$

Such a function $u$ would be called a *weak solution* for (3.2c), while $\phi \in C_0^1(\Omega)$ is said to be a *test function*. As opposed to the classical formulation, $f$ can be from the larger function space $L^2(\Omega)$.

In other words, solving (3.2d) means to tackle the following variational problem: Find the unique global minimum of the functional $F : H_0^1(\Omega) \to \mathbb{R}$,

$$F(u) := \tfrac{1}{2} a(u,u) - \langle u, f \rangle_{L^2(\Omega)}$$

(see Rem. 2.4.4). For this reason, (3.2d) is also said to be a *variational formulation* of the boundary value problem (3.2c). In order solve the problem (3.2d) we have to provide some crucial properties of the bilinear form $a$:

**Proposition 3.2.7.** *The bilinear form $a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ fulfills:*

*(a) $a$ is continuous, i.e. there exists a constant $C \ge 0$ such that*

$$|a(u,v)| \le C \, \|u\|_{H^1(\Omega)} \, \|v\|_{H^1(\Omega)} \quad \text{for all } u, v \in H_0^1(\Omega).$$

*(b) If $L$ is uniformly elliptic, then there exist reals $c_1 > 0$ and $c_0 \in \mathbb{R}$ such that* Gårdings inequality *holds:*

$$c_1 \, \|u\|^2_{H^1(\Omega)} - c_0 \, \|u\|^2_{L^2(\Omega)} \le a(u,u) \quad \text{for all } u \in H_0^1(\Omega). \tag{3.2e}$$

*Proof.* (a) By assumption the closure $\bar{\Omega} \subseteq \mathbb{R}^d$ is compact and therefore the coefficient functions $a^{\alpha\beta} : \bar{\Omega} \to \mathbb{R}$ are bounded. Using Hölder's inequality from Lemma 1.4.11 (with $p = q = 2$) this implies

$$\begin{aligned}
\left|a(\phi,u)\right| &\le \sum_{0\le|\alpha|,|\beta|\le 1} \int_\Omega \left|a^{\alpha\beta}(x)\right| \left|d^\alpha \phi(x)\right| \left|d^\beta u(x)\right| dx \\
&\le \sum_{0\le|\alpha|,|\beta|\le 1} \max_{x\in\bar{\Omega}} \left|a^{\alpha\beta}(x)\right| \int_\Omega \left|d^\alpha \phi(x)\right| \left|d^\beta u(x)\right| dx \\
&\overset{(1.4c)}{\le} \sum_{0\le|\alpha|,|\beta|\le 1} \max_{x\in\bar{\Omega}} \left|a^{\alpha\beta}(x)\right| \left\|d^\alpha \phi\right\|_{L^2(\Omega)} \left\|d^\beta u\right\|_{L^2(\Omega)} \\
&\overset{(3.1f)}{\le} \sum_{0\le|\alpha|,|\beta|\le 1} \max_{x\in\bar{\Omega}} \left|a^{\alpha\beta}(x)\right| \left\|\phi\right\|_{H^1(\Omega)} \|u\|_{H^1(\Omega)}
\end{aligned}$$

for all $\phi, u \in H_0^1(\Omega)$, and thus the claim with some appropriate constant $C < \infty$.

(b) This is shown in [Eva98, p. 300, Thm. 2(ii)]. □

**Theorem 3.2.8.** *Let $L$ be uniformly elliptic. If $f \in L^2(\Omega)$, then there exists a $c_0 \in \mathbb{R}$ such that the generalized Dirichlet problem for $L+\gamma I$ has a unique solution $u_\gamma \in H_0^1(\Omega)$ for all $\gamma \geq c_0$ and $\|u\|_{H^1(\Omega)} \leq \frac{1}{c_1}\|f\|_{L^2(\Omega)}$ with the constant $c_1$ from* (3.2e).

*Proof.* Take the real constant $c_0$ from Prop. 3.2.7(b) and define the bilinear form

$$a_\gamma : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}, \qquad a_\gamma(\phi, u) := a(\phi, u) + \gamma\langle\phi, u\rangle_{L^2(\Omega)}$$

for reals $\gamma \geq c_0$. Note that $a_\gamma$ is the bilinear form associated to the differential operator $L_\gamma u = Lu + \gamma u$. Thanks to Prop. 3.2.7(a) and the Cauchy-Schwarz inequality (see Prop. 1.3.7) it clearly satisfies

$$\left|a_\gamma(\phi, u)\right| \leq \left|a(\phi, u)\right| + |\gamma|\left|\langle\phi, u\rangle_{L^2(\Omega)}\right| \leq (C + |\gamma|)\|\phi\|_{H^1(\Omega)}\|u\|_{H^1(\Omega)}$$

for all $u, \phi \in H_0^1(\Omega)$ and by Thm. 2.2.10 we know that $a$ is continuous. In addition, Prop. 3.2.7(b) guarantees

$$a_\gamma(u, u) = a(u, u) + \gamma\|u\|_{L^2(\Omega)}^2 \overset{(3.2e)}{\geq} c_1\|u\|_{H^1(\Omega)}^2 + (\gamma - c_0)\|u\|_{L^2(\Omega)}^2 \geq c_1\|u\|_{H^1(\Omega)}^2$$

for all functions $u \in H_0^1(\Omega)$. Hence, $a$ is also coercive. Finally, we point out that $T : H_0^1(\Omega) \to \mathbb{R}$, $T\phi := \langle\phi, f\rangle_{L^2(\Omega)}$ is a bounded linear functional, since the Cauchy-Schwarz inequality from Prop. 1.3.7 implies

$$|T\phi| \leq \|\phi\|_{L^2(\Omega)}\|f\|_{L^2(\Omega)} \overset{(3.1f)}{\leq} \|f\|_{L^2(\Omega)}\|\phi\|_{H^1(\Omega)} \quad \text{for all } \phi \in H_0^1(\Omega)$$

and therefore $\|T\| \leq \|f\|_{L^2(\Omega)}$. We have verified all the assumptions of the Lax-Milgram Thm. 2.4.3 in the real Hilbert space $H_0^1(\Omega)$. This allows the conclusion that the generalized Dirichlet problem $a_\gamma(\phi, u) = \langle\phi, f\rangle_{L^2(\Omega)}$ for all $\phi \in H_0^1(\Omega)$ has a unique solution $u_\gamma \in H_0^1(\Omega)$ satisfying the claimed norm estimate. □

**Corollary 3.2.9.** *If a differential operator*

$$(Lu)(x) = \sum_{|\alpha|=|\beta|=1} (-1)^{|\alpha|} D^\alpha u(x)\left(a^{\alpha\beta} D^\beta u(x)\right)$$

*with constant $a^{\alpha\beta} \in \mathbb{R}$ is uniformly elliptic, then the generalized Dirichlet problem for $L$ has a unique solution.*

*Proof.* A closer inspection of the proof for Prop. 3.2.7(b) shows that $c_0$ can be chosen to be 0. Thus, the claim follows from Thm. 3.2.8 for $\gamma = 0$. □

To conclude this section we briefly address the question if a weak solution of a differential equation

$$Lu = f \quad \text{in } \Omega \tag{3.2f}$$

is in fact smooth. The corresponding field of mathematics is known as *regularity theory* and the following results hold:

- Suppose that the coefficient functions satisfy $a^{\alpha\beta} \in C^\infty(\Omega)$ and that we have a right-hand side $f \in C^\infty(\Omega)$. Then every weak solution $u \in H^1(\Omega)$ to (3.2f) satisfies $u \in C^\infty(\Omega)$ (see [Eva98, p. 316, Thm. 3]). The same results also holds for weak solutions $u \in H_0^1(\Omega)$ of the boundary value problem (3.2c) and $C^\infty$-boundaries $\partial\Omega$, if one replaces the above function spaces by $C^\infty(\bar{\Omega})$.
- For coefficient functions $a^{\alpha\beta} \in C^m(\bar{\Omega})$, right-hand sides $f \in H^m(\Omega)$ and a domain $\Omega$ with $C^{m+2}$-boundary $\partial\Omega$, every weak solution $u \in H_0^1(\Omega)$ of (3.2c) fulfills $u \in H^{m+2}(\Omega)$ (see [Eva98, p. 323, Thm. 5]). Then the Sobolev embedding from Rem. 3.1.11 yields a criterion for strong differentiability.

*Exercises* 3.2.10. State the variational formulation of the classical Poisson problem from Ex. 3.2.1.

## 3.3 Approximation methods

Throughout the whole section, let $\Omega \subseteq \mathbb{R}^d$ be open, bounded and connected.

The most direct method to solve classical boundary value problems (3.2c) is the *finite difference method*, where derivatives are replaced by differences as indicated in Ex. 1.5.6. This approach is successful, provided $\Omega$ has a simple geometry.

Yet, more modern approximation techniques for boundary value problems are based on their weak formulation. In order to illuminate this, assume $\alpha, \beta \in \mathbb{N}_0^d$ are multiindices and $a^{\alpha\beta} : \Omega \to \mathbb{R}$ are continuously differentiable functions with a continuous extension to the closure $\bar{\Omega}$. We investigate a uniformly elliptic 2nd order differential operator

$$(Lu)(x) := \sum_{0 \le |\alpha|,|\beta| \le 1} (-1)^{|\alpha|} D^\alpha \left( a^{\alpha\beta}(x) D^\beta u(x) \right) \quad \text{for all } x \in \Omega$$

as in Def. 3.2.2 and the associated generalized Dirichlet problem: Given $f \in L^2(\Omega)$, find a function $u \in H_0^1(\Omega)$ satisfying the relation

$$a(\phi, u) = \langle \phi, f \rangle_{L^2(\Omega)} \quad \text{for all } \phi \in H_0^1(\Omega). \tag{3.3a}$$

In order to solve (3.3a) numerically, i.e. to find an approximate weak solution, we have to determine appropriate finite-dimensional approximations of the weak solution $u$, as well as of the test functions $\phi$. Thus, the basic idea of so-called *Galerkin methods* is to replace the infinite-dimensional function space $H_0^1(\Omega)$ by a finite-dimensional subspace $V_n$, $n \in \mathbb{N}$.

For some fixed $n \in \mathbb{N}$, let us suppose that the space $V_n$ has a basis $\{e_k\}_{k=1}^n$ of functions $e_k \in H_0^1(\Omega)$ and consequently $n = \dim V_n$. We solve the generalized Dirichlet problem (3.3a) for test functions $\phi \in V_n$ instead of $\phi \in H_0^1(\Omega)$. Indeed, instead of using all test functions in $V_n$ we can restrict to the basis elements $e_k \in V_n$. This means we solve

$$a(e_k, u) = \langle e_k, f \rangle_{L^2(\Omega)} \quad \text{for all } 1 \le k \le n.$$

To find an approximation $u_n \in V_n$ of the solution $u \in L^2(\Omega)$ we make the ansatz

$$u_n = \sum_{j=1}^n \xi_j e_j$$

with unknown reals $\xi_1, \ldots, \xi_n \in \mathbb{R}$. With this our variational formulation becomes

$$\sum_{j=1}^n \xi_j a(e_k, e_j) = a\left(e_k, \sum_{j=1}^n \xi_j e_j\right) = a(e_k, u_n) = \langle e_k, f \rangle_{L^2(\Omega)} \quad \text{for all } 1 \le j, k \le n.$$

The alert reader realizes that this is an $n$-dimensional system of linear equations for $n$ unknown variables $\xi_1, \ldots, \xi_n$, i.e.

$$A\xi = F \tag{3.3b}$$

with

$$A = \begin{pmatrix} a(e_1, e_1) & \ldots & a(e_1, e_n) \\ \vdots & & \vdots \\ a(e_n, e_1) & \ldots & a(e_n, e_n) \end{pmatrix} \in \mathbb{R}^{n \times n}, \quad \xi = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix} \in \mathbb{R}^n, \quad F = \begin{pmatrix} \langle e_1, f \rangle_{L^2(\Omega)} \\ \vdots \\ \langle e_n, f \rangle_{L^2(\Omega)} \end{pmatrix} \in \mathbb{R}^n.$$

The matrix $A \in \mathbb{R}^{n \times n}$ is called *stiffness matrix*. For a symmetric differential operator $L$ the matrix $A$ is symmetric, and uniform ellipticity of $L$ guarantees that $A$ is symmetric and positively definite. Sometimes it is common to replace the function $f \in L^2(\Omega)$ by an approximation $f_n = \sum_{j=1}^n \eta_j e_j$ with known coefficients $\eta_j \in \mathbb{R}$. Under this premise the approximate variational formulation becomes

$$\sum_{j=1}^n \xi_j a(e_k, e_j) = \langle e_k, f_n \rangle_{L^2(\Omega)} = \sum_{j=1}^n \eta_k \langle e_k, e_j \rangle_{L^2(\Omega)} \quad \text{for all } 1 \le j, k \le n$$

or in a brief notation

$$A\xi = M\eta \tag{3.3c}$$

with the so-called *mass matrix* $M \in \mathbb{R}^{n \times n}$ given by

$$M = \begin{pmatrix} \langle e_1, e_1 \rangle_{L^2(\Omega)} & \ldots & \langle e_1, e_n \rangle_{L^2(\Omega)} \\ \vdots & & \vdots \\ \langle e_n, e_1 \rangle_{L^2(\Omega)} & \ldots & \langle e_n, e_n \rangle_{L^2(\Omega)} \end{pmatrix} \in \mathbb{R}^{n \times n}.$$

Thus, we reduced the classical Dirichlet problem (3.2c), or more detailed its variational formulation (3.3a) to finite-dimensional linear equations (3.3b) or (3.3c).

In order to solve the problems (3.3b) or (3.3c) though, one has to choose a subspace $V_n \subseteq H_0^1(\Omega)$ and a corresponding basis $\{e_k\}_{k=1}^n$. We present two possible approaches:

### 3.3.1 Spectral Galerkin method

The spectral Galerkin method to solve elliptic boundary value problems is based on the notions of eigenvalues and eigenfunctions.

**Definition 3.3.1.** A point $\lambda \in \mathbb{C}$ is called *eigenvalue* of $L$, if the boundary value problem

$$Lu = \lambda u \quad \text{in } \Omega, \qquad\qquad u = 0 \quad \text{in } \partial\Omega; \tag{3.3d}$$

has a nonzero weak solution $u \in L^2(\Omega)$, which is called an *eigenfunction* corresponding to $\lambda$. The set of all eigenvalues is said to be the spectrum $\sigma(L) \subseteq \mathbb{C}$ of $L$.

In absence of first and zeroth order derivatives in $L$ we will show that $\sigma(L)$ has a relatively simple structure, which resembles the spectrum of symmetric positive definite matrices.

**Theorem 3.3.2.** *For the* $2$*nd order differential operator*

$$(Lu)(x) := \sum_{|\alpha|=|\beta|=1} (-1)^{|\alpha|} D^\alpha u(x) \left( a^{\alpha\beta}(x) D^\beta u(x) \right) \quad \text{for all } x \in \Omega,$$

*the following holds true:*

*(a)* $\sigma(L) = \{\lambda_k\}_{k\in\mathbb{N}}$ *with eigenvalues* $\lambda_k \in \mathbb{R}$ *satisfying*

$$0 < \lambda_1 \le \lambda_2 \le \ldots \le \lambda_k \xrightarrow[k\to\infty]{} \infty, \tag{3.3e}$$

*(b) there exists an orthonormal basis* $\{e_k\}_{k\in\mathbb{N}}$ *of* $L^2(\Omega)$ *consisting of eigenfunctions* $e_k \in H_0^1(\Omega)$ *for* $L$ *corresponding to* $\lambda_k$.

*Proof.* See [Eva98, p. 335, Thm. 1]. □

The following examples are taken from [Her06, pp. 9ff, Sect. 1.2.5], which also treats the case where $\Omega$ is a planar disk.

*Example* 3.3.3 ($\Omega$ is an interval). In case $\Omega = (0, a)$ for some $a > 0$, the eigenvalues and eigenfunctions of the negative Laplacian $-\Delta u = -u''$ equipped with Dirichlet boundary conditions $u(0) = u(a) = 0$ read as

$$\lambda_k = \left(\tfrac{\pi k}{a}\right)^2, \qquad e_k(x) = \sin\left(\tfrac{\pi k}{a} x\right) \quad \text{for all } k \in \mathbb{N}.$$

*Example* 3.3.4 ($\Omega$ is a rectangle). Let $\Omega = (0, a) \times (0, b)$ with $a, b > 0$ be a planar rectangle and consider the negative Laplacian

$$Lu := -D^{(2,0)}u - D^{(0,2)}u = -\tfrac{\partial^2 u}{\partial x_1^2} - \tfrac{\partial^2 u}{\partial x_2^2}$$

subject to Dirichlet boundary conditions $u = 0$ on $\partial\Omega$. Then the eigenvalues resp. eigenfunctions are given by

$$\lambda_{jk} = \pi^2 \left[\left(\tfrac{\pi j}{a}\right)^2 + \left(\tfrac{\pi k}{b}\right)^2\right], \quad e_{jk}(x) = \frac{2}{\sqrt{ab}} \sin\left(\tfrac{\pi j}{a} x_1\right) \sin\left(\tfrac{\pi k}{b} x_2\right) \quad \text{for all } j, k \in \mathbb{N}.$$

Keeping this in mind, we suppose the operator $L$ has the known spectrum $\sigma(L) = \{\lambda_k\}_{k \in \mathbb{N}}$ satisfying (3.3e) and known eigenfunctions $e_k \in H_0^1(\Omega)$ forming an orthonormal basis of $L^2(\Omega)$. As finite-dimensional subspace of $H_0^1(\Omega)$ we choose

$$V_n := \left\{\sum_{k=1}^n \xi_k e_k \in H_0^1(\Omega) : \xi_1, \dots, \xi_d \in \mathbb{R}\right\}$$

for a given $n \in \mathbb{N}$. Due to the orthonormality of $e_j$ and $e_k$ we have

$$a(e_k, e_j) = \int_\Omega e_k(x) L e_j(x)\, dx = \lambda_j \int_\Omega e_k(x) e_j(x)\, dx = \begin{cases} \lambda_j, & j = k, \\ 0, & j = k \end{cases}$$

and therefore the stiffness matrix $A$ becomes diagonal, i.e. $A = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, while the mass matrix $M$ is the identity. This yields the approximate solution

$$u_n = \sum_{k=1}^n \frac{\langle f, e_k \rangle_{L^2(\Omega)}}{\lambda_k} e_k = \sum_{k=1}^n \frac{1}{\lambda_k} \int_\Omega f(x) e_k(x)\, dx e_k. \tag{3.3f}$$

*Example* 3.3.5. On the rectangular domain $\Omega = (0, a) \times (0, b)$ we consider the Dirichlet boundary value problem

$$-\frac{\partial^2 u(x_1, x_2)}{\partial x_1^2} - \frac{\partial^2 u(x_1, x_2)}{\partial x_2^2} = x_1 x_2 \quad \text{for } (x_1, x_2) \in \Omega,$$
$$u(x_1, x_2) = 0 \quad \text{for } (x_1, x_2) \in \partial\Omega$$

and an inhomogeneity $f \in L^2(\Omega)$. It allows the abstract formulation (3.2c) with the differential operator $L$ defined in the above Ex. 3.3.4. Consequently, spectral Galerkin approximations to its solution are given by

$$u_{n_1,n_2}(x_1,x_2) = \frac{4}{ab}\sum_{j=1}^{n_1}\sum_{k=1}^{n_2}\left(\pi^2\left[\left(\tfrac{\pi j}{a}\right)^2+\left(\tfrac{\pi k}{b}\right)^2\right]\right)^{-1}\cdot$$
$$\cdot\int_0^a\int_0^b f(y_1,y_2)\sin\left(\tfrac{\pi j}{a}y_1\right)\sin\left(\tfrac{\pi k}{b}y_2\right)\,dy_2dy_1\sin\left(\tfrac{\pi j}{a}x_1\right)\sin\left(\tfrac{\pi k}{b}x_2\right)$$

for arbitrarily large $n_1, n_2 \in \mathbb{N}$.

### 3.3.2 Finite element method

The above spectral Galerkin method has the disadvantage that the eigenfunctions $e_k \in H_0^1(\Omega)$ and corresponding eigenvalues $\lambda_k$ have to be known in advance. In general, this is only possible for simply geometries of the domain $\Omega$, i.e. if $\Omega$ is rectangular or e.g. a circle.

For the finite element method one chooses a *triangulation* of the domain $\Omega$, i.e. a subdivision

$$\Omega = \bigcup_{k=1}^{m}\Delta_k \tag{3.3g}$$

into intervals (if $d=1$), triangles (if $d=2$) or general $d$-simplices $\Delta_k \subseteq \Omega$, which intersect only on their boundary. As finite-dimensional subspace $V_n$ one typically chooses the set of all $H_0^1$-functions, which are linear over every $\Delta_k$. Formally, this means

$$V_n := \left\{v \in H_0^1(\Omega) : v|_{\Delta_k} \text{ is affine-linear for } k \in \{1,\dots,m\}\right\}$$

and one has to find an appropriate basis of $V_n$. We will illustrate this in the simplest case $\Omega = (a,b)$:

*Example* 3.3.6. Choose real numbers $a = x_0 < x_1 < \ldots < x_m < x_{m+1} = b$ and define the subintervals $\Delta_k := [x_k, x_{k+1}]$ for $k = 0,\dots,m$, which guarantees that (3.3g) holds. As subspace $V_n$ we define

$$V_n := \left\{v \in C_0[a,b] : v|_{\Delta_k} \text{ is linear for } k = 0,\dots,m\right\}$$

and as basis one makes use of the tent functions (see Figure 3.1)

$$e_k(x) := \begin{cases} \frac{x-x_{k-1}}{x_k-x_{k-1}}, & x \in [x_{k-1}, x_k],\\ \frac{x_{k+1}-x}{x_{k+1}-x_k}, & x \in [x_k, x_{k+1}],\\ 0, & \text{otherwise}\end{cases}$$

for $1 \le k \le m$. Every $e_k : [a,b] \to \mathbb{R}$ is piecewise linear, achieves its maximum $e_k(x_k) = 1$ and vanishes at $x_j$ for $j \neq k$.

*Exercises* 3.3.7. Determine $\dim V_n$ for the subspaces $V_n \subseteq H_0^1(\Omega)$ from Ex. 3.3.6.

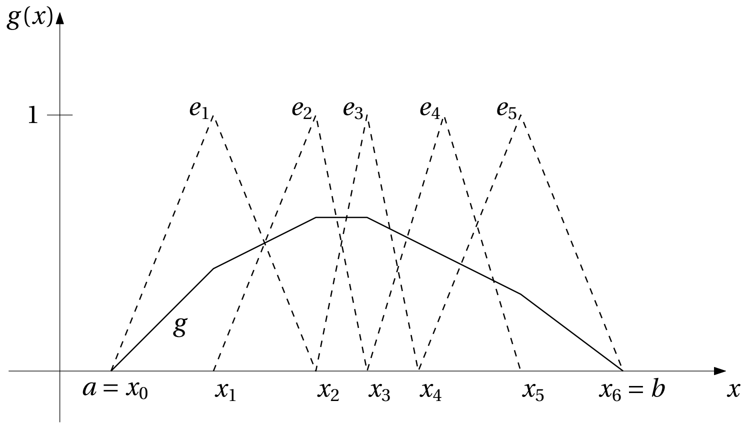

**Fig. 3.1** Basis functions $e_k$ and a piecewise linear function $g$

# References

[Ban32] S. Banach, *Théorie des opérations linéaires*, Monogr. Mat. Vol. 1, Subwncji Funduszu Narodowej, Warszawa, 1932.

[Coh80] D.L. Cohn, *Measure theory*, Birkhäuser, Boston-Basel-Stuttgart, 1980.

[Col66] L. Collatz, *Functional analysis and numerical mathematics*, Academic Press, New York-London, 1966.

[Con90] J.B. Conway, *A course in functional analysis*, 2nd ed., Graduate Texts in Mathematics 96, Springer, New York, 1990.

[Eva98] L.C. Evans, *Partial differential equations*, Graduate Studies in Mathematics 19, American Mathematical Society, Providence, RI, 1998.

[Her06] A. Hernot, *Extremum problems for eigenvalues of elliptic operators*, Birkhäuser, 2006.

[Lax02] P. Lax, *Functional analysis*, Interscience Series in Pure and Applied Mathematics, Wiley, Chichester, 2002.

[LV03] L.P. Lebedev & I.I. Vorovich, *Functional analysis in mechanics*, revised and extended translation of the russian ed., Monographs in Mathematics, Springer, New York, etc., 2003.

[NS82] A.W. Naylor & G.R. Sell, *Linear operator theory in engineering and science*, Applied Mathematical Sciences 40, Springer, Berlin etc., 1982.

[Yos80] K. Yosida, *Functional analysis*, Grundlehren der mathematischen Wissenschaften 123, Springer, Berlin etc., 1980.

# Index