Applied mathematics – lecture notes

J. Stebel

January 27, 2015

Contents

1 Introduction. Auxiliary tools

1.1 Relations and mappings

Definition 1 *Any subset of the cartesian product $A \times B$ is called a* binary relation*, or just relation, between sets A and B.*

The fact that $(a, b) \in R$, where R is a relation between A and B, is usually denoted aRb. E.g. for the sets $A = \{$Adam, Socrates, David Beckham$\}$ and $B = \{$Eve, Xanthippe, Maria Theresa$\}$ one can introduce the partnership relation $P = \{$(Adam,Eve), (Socrates,Xanthippe)$\}$. Another example of a relation is the ordering on $\mathbb{R}$, represented by the symbol $\leqslant$, or the equality of elements in $\mathbb{R}$.

Definition 2 *Let R be a relation on the set A (i.e. $R \subset A \times A$). R is called*

(i) symmetric *iff*
$$(a, b) \in R \Leftrightarrow (b, a) \in R;$$

(ii) transitive *iff*
$$(a, b) \in R \& (b, c) \in R \Rightarrow (a, c) \in R;$$

(iii) reflexive *iff*
$$\forall a \in A : \ (a, a) \in R.$$

E.g. equality of real numbers is a symmetric, transitive and reflexive relation. The relations $<$, $\leqslant$ on $\mathbb{R}$ are transitive, $\leqslant$ is in addition reflexive.

Mapping f of a set A into B is a relation $f \subset A \times B$ which satisfies:
$$\forall a \in A \ \exists! b \in B : \ (a, b) \in f.$$
We also write $f : A \to B$, $f : a \mapsto b$, $f(a) = b$.

Image of the set A' under $f : A \to B$ is the set
$$f(A') := \{f(a); \ a \in A'\}.$$

Injective mapping is characterized by the property
$$f(a_1) = f(a_2) \Rightarrow a_1 = a_2.$$

Inverse mapping to an injective mapping $f : A \to B$ is the mapping $f^{-1} : f(A) \to A$ defined by
$$f^{-1}(b) = a \Leftrightarrow f(a) = b.$$

Surjective mapping satisfies
$$f(A) = B.$$
We also say that f maps A *onto* B.

Bijective mapping, or isomorphism, is an injective and surjective mapping. If there exists an isomorphism between A and B, then we say that A and B are isomorphic.

Using isomorphisms we can categorize sets: We say that a set is finite if it is isomorphic with the set $\{1, \ldots, n\}$ for some natural number $n \in \mathbb{N}$. Countable set is isomorphic to $\mathbb{N}$. A set is infinite if it is not finite. A set is uncountable if it is neither finite nor countable.

E.g. $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ are countable, $\mathbb{R}$ is uncountable.

Compound mapping. Let $f : A \to B$ and $g : B \to C$. Then $g \circ f : A \to C$ is a mapping composed of f and g, defined by

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

1.2 Linear spaces

Definition 3 *A non-empty set V is called a* linear space *if there are relations $+ : V \times V \to V$ (addition of vectors) and $\cdot : \mathbb{R} \times V \to V$ (multiplication by scalars), if for any $\vec{x}, \vec{y}, \vec{z} \in V$ and $\alpha, \beta \in \mathbb{R}$ the following holds:*

1. $\forall \vec{x}, \vec{y} \in V : \ \vec{x} + \vec{y} = \vec{y} + \vec{x}$;
2. $\forall \vec{x}, \vec{y}, \vec{z} \in V : \ (\vec{x} + \vec{y}) + \vec{z} = \vec{x} + (\vec{y} + \vec{z})$;
3. $\forall \alpha, \beta \in \mathbb{R} \ \forall \vec{x} \in V : \ \alpha \cdot (\beta \cdot \vec{x}) = (\alpha\beta) \cdot \vec{x}$;
4. $\forall \alpha \in \mathbb{R} \ \forall \vec{x}, \vec{y} \in V : \ \alpha \cdot (\vec{x} + \vec{y}) = \alpha \cdot \vec{x} + \alpha \cdot \vec{y}$;
5. $\forall \alpha, \beta \in \mathbb{R} \ \forall \vec{x} \in V : \ (\alpha + \beta) \cdot \vec{x} = \alpha \cdot \vec{x} + \beta \cdot \vec{x}$;
6. $\forall \vec{x} \in \mathbb{R} : \ 1 \cdot \vec{x} = \vec{x}$;
7. $\exists \vec{0} \in V \ \forall \vec{x} \in V : \ 0 \cdot \vec{x} = \vec{0}$.

The elements of a linear space are called **vectors**. Real numbers in the context of the multiplication $\cdot : \mathbb{R} \times V \to V$ are called **scalars**. The element $\vec{0}$ is called the **zero element** or the **zero vector**. From the properties 1.–7. one can deduce the following properties of the zero vector $\vec{0} \in V$:

- $\forall \vec{x} \in V$: $\vec{x} + \vec{0} = \vec{x}$,
- $\forall \alpha \in \mathbb{R}$: $\alpha \cdot \vec{0} = \vec{0}$,
- $\forall \vec{x} \in V \ \forall \alpha \in \mathbb{R}, \alpha \neq 0$: $\alpha \cdot \vec{x} = \vec{0} \Rightarrow \vec{x} = \vec{0}$.

Example 1 *The following are linear spaces:*

- *Euclidean spaces $\mathbb{R}$, $\mathbb{R}^2$, $\mathbb{R}^n$ (with componentwise addition and scalar multiplication);*
- *trivial space $\{\vec{0}\}$ ($\alpha\vec{0} = \vec{0} + \vec{0} = \vec{0}$);*

- *the space $\mathcal{F}$ of real functions ($(\alpha f)(x) := f(\alpha x)$, $(f+g)(x) := f(x)+g(x)$);*
- *the space of polynomials $\mathcal{P}$;*
- *the space of polynomials $\mathcal{P}_n$ of degree $\leqslant n$.*

Definition 4 *We say that W is a* subspace *of a linear space V iff $W \subset V$ and W with the relations $+$, $\cdot$ adopted from V is a linear space.*

E.g. $\mathcal{P}_n$ is a subspace of $\mathcal{P}$ and both are subspaces of $\mathcal{F}$.

To decide whether a set is a subspace of V it may be inconvenient to verify all 7 properties of Definition 3. The following assertion simplified this process.

Theorem 1 *Let V be a linear space and $\emptyset \neq W \subset V$. Then W is a subspace of V if and only if*

(i) $\forall \vec{x}, \vec{y} \in W$*:* $\vec{x} + \vec{y} \in W$*,*

(ii) $\forall \vec{x} \in W$*,* $\alpha \in \mathbb{R}$*:* $\alpha\vec{x} \in W$*.*

Intersection of linear spaces is a linear space. On the other hand, union of linear spaces is in general not a linear space. For example, let $A = \{(\alpha, 0);\ \alpha \in R\}$ and $B = \{(0, \beta);\ \beta \in \mathbb{R}\}$ be subspaces of $\mathbb{R}^2$. Then $A \cap B = \{\vec{0}\}$ is the trivial space, while $A \cup B = \{(\alpha, \beta);\ \alpha = 0 \text{ or } \beta = 0\}$ is not a linear space since e.g. $(1, 0) + (0, 1) = (1, 1) \notin A \cup B$.

Definition 5 *Let $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n$ be elements of a linear space V and $\alpha_1, \alpha_2, \ldots, \alpha_n$ be real numbers. Any vector*

$$\alpha_1\vec{x}_1 + \alpha_2\vec{x}_2 + \ldots + \alpha_n\vec{x}_n$$

is called a linear combination *of the vectors $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n$. The numbers $\alpha_1, \alpha_2, \ldots, \alpha_n$ are called the* coeficients *of the linear combination.*

Definition 6 *If all coefficients are zero then a linear combination is said to be* trivial. *A* nontrivial *linear combination is such that at least one of its coefficients is nonzero.*

Note that a trivial linear combination always equals zero vector.

Definition 7 *A finite set of vectors $\{\vec{x}_1, \ldots, \vec{x}_n\}$ is called* linearly dependent *if there exists their nontrivial linear combination which equals zero vector. Briefly, we say that vectors $\vec{x}_1, \ldots, \vec{x}_n$ are linearly dependent.*

A set of vectors $\{\vec{x}_1, \ldots, \vec{x}_n\}$ is linearly independent *if it is not linearly dependent.*

Two and more vectors are linearly dependent if and only if one of the vectors is a linear combination of the remaining ones. E.g., functions $\cos^2 x$, $\sin^2 x$, $\cos 2x$ are linearly dependent elements of the space $\mathcal{F}$, because for any $x \in \mathbb{R}$ it holds: $\cos 2x = \cos^2 x + (-1)\sin^2 x$.

Linear (in)dependence can be generalized to infinite sets of vectors as follows.

Definition 8 *A set M of vectors is called* linearly dependent *if there exists a finite linearly dependent subset of M.*

A set M is called linearly independent *if it is not linearly dependent.*

An example of a linearly independent set in $\mathcal{P}$ is the set $\{1, x, x^2, x^3, x^4, \ldots\}$.

Definition 9 Linear span *of a finite set $\{\vec{x}_1, \ldots, \vec{x}_n\}$ is the set of all linear combinations of these vectors. Linear span of an infinite set M is the union of linear spans of all finite subsets of M.*

Linear span of $\{\vec{x}_1, \ldots, \vec{x}_n\}$ is denoted $\langle \vec{x}_1, \ldots, \vec{x}_n \rangle$, linear span of a set M is denoted $\langle M \rangle$. If M is a subset of a linear space V then $\langle M \rangle$ is the smallest linear space containing M.

Definition 10 *A set $\subset V$ is called a* basis *of a linear space V if:*

(i) B is linearly independent,

(ii) $\langle B \rangle = V$.

Theorem 2 *Every linear space has a basis. If B_1 and B_2 are bases of V then they both are infinite or have the same number of elements.*

Theorem 3 *Let vectors $\vec{x}_1, \ldots, \vec{x}_n$ form a basis of a linear space V. For any $\vec{x} \in V$ there exists exactly one n-tuple of numbers $(c_1, \ldots, c_n)$ such that*

$$\vec{x} = c_1\vec{x}_1 + \ldots + c_n\vec{x}_n.$$

Definition 11 *The numbers $c_1, \ldots, c_n$ from the previous theorem are called the* coordinates *of the vector $\vec{x}$ in the basis $\vec{x}_1, \ldots, \vec{x}_n$.*

Definition 12 *The number of elements in any basis of a linear space V is called its* dimension *and denoted by* $\dim V$. *Special cases are:*

- *The trivial space, whose dimension is 0.*
- *Spaces with infinite bases, in which case we define* $\dim V = +\infty$.

If M is a subspace of a linear space V then $\dim M \leqslant \dim V$.

Theorem 4 *Let V be a linear space with* $\dim V = n$, *and $M = \{\vec{x}_1, \ldots, \vec{x}_m\}$. The following is true:*

1. *If M is linearly independent then $m \leqslant n$.*
2. *If $m > n$ then M is linearly dependent.*
3. *Let $m = n$. Then M is linearly independent iff $\langle M \rangle = V$.*

1.3 Linear mappings and matrices

Definition 13 *A mapping f of a linear space V into a linear space W is called* linear *if it satisfies for every $\alpha \in \mathbb{R}$ and $\vec{x}, \vec{y} \in V$:*

$$f(\alpha\vec{x} + \vec{y}) = \alpha f(\vec{x}) + f(\vec{y}).$$

Kernel *of f is the set*

$$\ker f := \{\vec{x} \in V;\ f(\vec{x}) = \vec{0}\}.$$

Range *of f is the set*

$$\mathcal{R}(f) := f(V).$$

A linear mapping f is injective if and only if $\ker f = \{\vec{0}\}$.

Theorem 5 *If $\vec{x}_1, \ldots, \vec{x}_k$ are linearly dependent then $f(\vec{x}_1), \ldots, f(\vec{x}_k)$ are also linearly dependent. If in addition f is injective then these properties are equivalent:*

$\vec{x}_1, \ldots, \vec{x}_k$ are linearly dependent $\Leftrightarrow$ $f(\vec{x}_1), \ldots, f(\vec{x}_k)$ are linearly dependent.

Definition 14 *A* real (complex) matrix *of type (m, n) is a symbol*

$$\mathbf{A} = \begin{pmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,n} \\ a_{2,1} & a_{2,2} & \ldots & a_{2,n} \\ & & \vdots & \\ a_{m,1} & a_{m,2} & \ldots & a_{m,n} \end{pmatrix} = (a_{ij})_{i=1,\ldots,m}^{j=1,\ldots,n},$$

where for $i = 1, \ldots, m$ and $j = 1, \ldots, n$, the symbols a_{ij} are real (complex) numbers.

The set of all (real) matrices of type (m, n) will be denoted $\mathbb{R}^{m\times n}$. Addition of matrices and multiplication of a matrix by a real number is defined componentwise. Consequently, the set $\mathbb{R}^{m\times n}$ is a linear space.

Definition 15 *We say that $\mathbf{B} \in \mathbb{R}^{n\times m}$ is the* transposed matrix *to the matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ iff*

$$\forall i = 1, \ldots, m\ \forall j = 1, \ldots, n:\ a_{ij} = b_{ji}.$$

The transposed matrix is denoted $\mathbf{B} = \mathbf{A}^\top$.

Definition 16 *We say that $\mathbf{A}$ is* symmetric *if $\mathbf{A} = \mathbf{A}^\top$.*

Product of $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times p}$ is a matrix $\mathbf{C} \in \mathbb{R}^{m\times p}$ whose components satisfies:

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj},\ i = 1, \ldots, m,\ j = 1, \ldots, p.$$

Matrix multiplication is non-commutative, i.e. in general $\mathbf{AB} \neq \mathbf{BA}$. For all matrices of proper types (such that multiplication is possible) it holds:

$$\begin{aligned}
&\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}, \\
&(\mathbf{A}+\mathbf{B})\mathbf{C} = \mathbf{AC}+\mathbf{BC}, \\
&(\alpha\mathbf{A})\mathbf{B} = \mathbf{A}(\alpha\mathbf{B}) = \alpha(\mathbf{AB}),\ \alpha \in \mathbb{R}.
\end{aligned}$$

Definition 17 *A square matrix* $\mathbf{I} = (e_{i,j}) \in \mathbb{R}^{n\times n}$ *is called the* unit matrix *if its components satisfy:* $e_{i,j} = 0$ *for* $i \neq j$ *and* $e_{i,j} = 1$ *for* $i = j$.

Definition 18 *We say that* $\mathbf{B} \in \mathbb{R}^{n\times n}$ *is the* inverse matrix *to* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *iff* $\mathbf{AB} = \mathbf{BA} = \mathbf{I}$. *The inverse matrix is denoted* $\mathbf{B} = \mathbf{A}^{-1}$. *If* $\mathbf{A}^{-1}$ *exists then* $\mathbf{A}$ *is called* nonsingular. *Otherwise* $\mathbf{A}$ *is called a* singular *matrix.*

Definition 19 Rank *of a matrix* $\mathbf{A}$, *denoted* $\operatorname{rank}\mathbf{A}$, *is the number of its linearly independent rows.*

It also holds that the rank is equal to the number of linearly independent columns, i.e. $\operatorname{rank}\mathbf{A}^\top = \operatorname{rank}\mathbf{A}$. A matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is nonsingular if and only if $\operatorname{rank}\mathbf{A} = n$ (we say that it has full rank).

Definition 20 Permutation *of* n *elements is an ordered* n*-tuple of numbers* $1, 2, \ldots, n$ *such that it contains every number just once.*

Note: there exist $n!$ distinct permutations of n elements.

Definition 21 *Let* $(i_1, i_2, \ldots, i_n)$ *be a permutation of* n *elements. The* number of inversions *of this permutation is the number of pairs* (i_k, i_l) *such that* $i_k > i_l$ *and* $k < l$.

Definition 22 *For every permutation* $\pi = (i_1, \ldots, i_n)$ *we define its* sign $\operatorname{sgn}\pi$ *as follows:*

$$\operatorname{sgn}\pi = \begin{cases} +1 & \text{if } \pi \text{ has an even number of inversions,} \\ -1 & \text{if } \pi \text{ has an odd number of inversions.} \end{cases}$$

Interchanging two elements in a permutation causes the change of its sign.

Definition 23 *Let* $\mathbf{A} = (a_{i,j}) \in \mathbb{R}^{n\times n}$. Determinant *of* $\mathbf{A}$ *is the number*

$$\det\mathbf{A} = \sum_{\pi=(i_1,i_2,\ldots,i_n)} (\operatorname{sgn}\pi) a_{1,i_1} a_{2,i_2} \cdots a_{n,i_n}.$$

In the above formula, the summation is done over all permutations of n *elements, i.e. there are* $n!$ *addends.*

Theorem 6 *Let* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$. *Then*

1. $\det \mathbf{A} = 0$ *if and only if* $\mathbf{A}$ *is singular,*
2. $\det \mathbf{A}^\top = \det \mathbf{A}$*,*
3. $\det(\mathbf{AB}) = (\det \mathbf{A})(\det \mathbf{B})$*,*
4. $\det(\mathbf{A}^{-1}) = 1/\det \mathbf{A}$*.*
5. *If* $\mathbf{B}$ *can be obtained from* $\mathbf{A}$ *by interchanging two rows then* $\det \mathbf{B} = -\det \mathbf{A}$*.*
6. *If* $\mathbf{A}$ *has two identical rows then* $\det \mathbf{A} = 0$*.*

Note that for $\mathbf{A} \in \mathbb{R}^{2\times 2}$ the determinant can be expressed as follows:

$$\det \mathbf{A} = a_{11}a_{22} - a_{12}a_{21}.$$

For $\mathbf{A} \in \mathbb{R}^{3\times 3}$ the formula for determinant reads:

$$\det \mathbf{A} = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31}.$$

Definition 24 *A number* $\lambda \in \mathbb{C}$ *is called an* eigenvalue *of a (complex) matrix* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *if there exists a nonzero vector* $\vec{u} \in \mathbb{C}^n$ *such that*

$$\mathbf{A}\vec{u} = \lambda\vec{u}.$$

The vector $\vec{u}$ *is called the* eigenvector *of* $\mathbf{A}$ *associated with the eigenvalue* λ*. The set of all eigenvalues of* A *is called the* spectrum *of* $\mathbf{A}$ *and is denoted* $\sigma(\mathbf{A})$*.*

The number λ is an eigenvalue of $\mathbf{A}$ if and only if the system $(\mathbf{A} - \lambda\mathbf{I})$ has a nontrivial solution, i.e. when $\mathbf{A} - \lambda\mathbf{I}$ is singular, which is equivalent to the condition $\det(\lambda\mathbf{I} - \mathbf{A}) = 0$. The polynomial $\chi_{\mathbf{A}}(\lambda) := \det(\lambda\mathbf{I} - \mathbf{A})$ is called the *characteristic polynomial* of the matrix $\mathbf{A}$. Hence, λ is an eigenvalue of $\mathbf{A}$ if it is a root of $\chi_{\mathbf{A}}$. We remark that a real polynomial can have complex roots, thus a real matrix can have complex eigenvalues. But if a real matrix is symmetric then all its eigenvalues are real.

1.4 Systems of linear equations

In what follows we shall identify vectors from $\mathbb{R}^n$ with matrices of type $(n, 1)$, i.e. $\vec{a} \in \mathbb{R}^n$ means the same as $\vec{a} \in \mathbb{R}^{n\times 1}$.

Definition 25 *Let* $\mathbf{A} \in \mathbb{R}^{m\times n}$*,* $\vec{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$ *and* $\vec{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \in \mathbb{R}^m$*. Then the matrix equality*

$$\mathbf{A}\vec{x} = \vec{b}$$

is called a system of m linear equations for n unknowns. *The matrix* $\mathbf{A}$ *is called the* system matrix *and the vector* $\vec{b}$ *is the* vector of right hand sides. *Extending the system matrix by the vector of right hand sides (for clarity separated by a vertical line) we obtain the matrix* $(\mathbf{A}|\vec{b}) \in \mathbb{R}^{m\times(n+1)}$*, called the* extended system matrix.

Definition 26 *A vector* $\vec{a} = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$ *is said to be a* solution *of the system* $\mathbf{A}\vec{x} = \vec{b}$ *if it satisfies:*

$$\mathbf{A}\begin{pmatrix}\alpha_1\\ \vdots\\ \alpha_n\end{pmatrix} = \begin{pmatrix}b_1\\ \vdots\\ b_m\end{pmatrix}.$$

Theorem 7 (Frobenius) *A system* $\mathbf{A}\vec{x} = \vec{b}$ *has a solution if and only if*

$$\operatorname{rank}\mathbf{A} = \operatorname{rank}\mathbf{A}|\vec{b},$$

i.e. when the extended system matrix has the same rank as the system matrix.

Definition 27 *Let* $\mathbf{A}\vec{x} = \vec{b}$ *be a system of* m *linear equations with* n *uknowns and* $\mathbf{C}\vec{x} = \vec{d}$ *is a system of* k *linear equations with the same number* n *of unknowns. We say that these systems are* equivalent *iff they have identical sets of solutions.*

Theorem 8 *To every system* $\mathbf{A}\vec{x} = \vec{b}$ *there exists an equivalent system* $\mathbf{C}\vec{x} = \vec{d}$ *whose matrix* $\mathbf{C}$ *is upper triangular.*

Definition 28 *If at least one element of* $\vec{b}$ *is nonzero, then we say that the system* $\mathbf{A}\vec{x} = \vec{b}$ *is* nonhomogeneous. *If vector* $\vec{b}$ *is identically zero, we call the system* homogeneous *and write*

$$\mathbf{A}\vec{x} = \vec{0}.$$

Theorem 9 *The set* M_0 *of all solutions to a homogeneous system* $\mathbf{A}\vec{x} = \vec{0}$ *with* n *unknowns is a linear subspace of* $\mathbb{R}^n$*.*

Theorem 10 *Let* $\mathbf{A}\vec{x} = \vec{0}$ *be a homogeneous system of linear equations with* n *unknowns and denote* $k := n - \operatorname{rank}\mathbf{A}$*. Then there exist* k *linearly independent vectors* $\vec{u}_1, \ldots, \vec{u}_k \in \mathbb{R}^n$ *which form a basis of the set* M_0 *of all solutions to* $\mathbf{A}\vec{x} = \vec{0}$*, i.e.*

$$M_0 = \langle \vec{u}_1, \ldots, \vec{u}_k \rangle.$$

Consequently, $\dim M_0 = n - \operatorname{rank}\mathbf{A}$.

Definition 29 *Any solution* $\vec{v} \in \mathbb{R}^n$ *of a nonhomogeneous system* $\mathbf{A}\vec{x} = \vec{b}$ *is called a* particular solution *of this system.*

The system $\mathbf{A}\vec{x} = \vec{0}$ *is called the* associated homogeneous system *to the system* $\mathbf{A}\vec{x} = \vec{b}$*.*

Theorem 11 *1. Let $\vec{v}$ be a particular solution to the nonhomogeneous system $\mathbf{A}\vec{x} = \vec{b}$ and $\vec{u}$ be any solution of the associated homogeneous system $\mathbf{A}\vec{x} = \vec{0}$. Then $\vec{v} + \vec{u}$ is also a solution to the system $\mathbf{A}\vec{x} = \vec{b}$.*

2. Let $\vec{v}$ and $\vec{w}$ be two particular solutions to the nonhomogeneous system $\mathbf{A}\vec{x} = \vec{b}$. Then $\vec{v} - \vec{w}$ is a solution to the associated homogeneous system $\mathbf{A}\vec{x} = \vec{0}$.

Theorem 12 *Let $\vec{v}$ be a particular solution of the system $\mathbf{A}\vec{x} = \vec{b}$ and M_0 be the set of all solutions to the associated homogeneous system $\mathbf{A}\vec{x} = \vec{0}$. Then the set M of all solutions to $\mathbf{A}\vec{x} = \vec{b}$ is given as follows:*

$$M = \{\vec{v} + \vec{u};\ \vec{u} \in M_0\}.$$

Theorem 13 (Cramer's rule) *Let $\mathbf{A}$ be a nonsingular matrix. Then the i-th component of the solution to $\mathbf{A}\vec{x} = \vec{b}$ satisfies:*

$$\alpha_i = \frac{\det \mathbf{B}_i}{\det \mathbf{A}},$$

where the matrix $\mathbf{B}_i$ is identical to $\mathbf{A}$ up to i-th column, which is replaced by the column of right hand sides.

2 Iterative methods for systems of linear equations

In this section we shall deal with the numerical solution of the system

$$\mathbf{A}\vec{x} = \vec{b}.$$

We assume that the reader is familiar with the Gaussian elimination method, an example of the so-called direct methods. Its main advantage is the universality — the method can solve in exact arithmetics any system with a nonsingular matrix. A disadvantage is its computational complexity ($O(n^3)$) and also that during the computation the user has no information about the result. For systems with large sparse matrices $\mathbf{A}$, which arise in many practical problems, or for problems where the matrix is not given explicitely or it is expensive to assemble, it can be advantageous to use iterative methods. These methods use in principle only multiplication of vectors by $\mathbf{A}$ and throughout the computation they improve the approximation of the exact solution. The convergence of iterative methods can be either asymptotic or in finite number of iterations.

2.1 Classical iterative methods

Classical iterative methods are based on the splitting $\mathbf{A} = \mathbf{M} + \mathbf{N}$ such that the matrix $\mathbf{M}$ is nonsingular and easily invertible and $\mathbf{M}$ and $\mathbf{N}$ are chosen in a suitable way. Using the identity $\mathbf{A}\vec{x} = \vec{b}$ we obtain:

$$\vec{x} = \vec{x} + \mathbf{M}^{-1}(\vec{b} - \mathbf{A}\vec{x}).$$

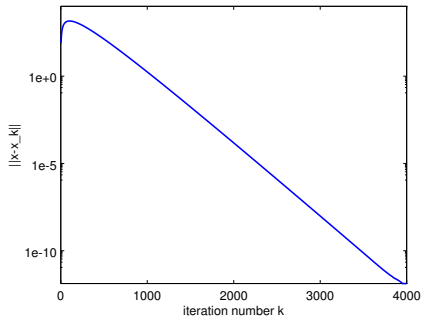

Figure 1: Transition effect of a classical iterative method.

Given an initial approximation $\vec{x}_0$ of the solution, one can define the iterative process as follows:

$$\vec{x}_k = \vec{x}_{k-1} + \mathbf{M}^{-1}(\vec{b} - \mathbf{A}\vec{x}_{k-1}) = (\mathbf{I} - \mathbf{M}^{-1}\mathbf{A})\vec{x}_{k-1} + \mathbf{M}^{-1}\vec{b}.$$

From these identities one can show that the approximation error satisfies the estimate

$$\frac{\|\vec{x} - \vec{x}_k\|}{\|\vec{x} - \vec{x}_0\|} \leqslant \|(\mathbf{I} - \mathbf{M}^{-1}\mathbf{A})^k\| \leqslant \|\mathbf{I} - \mathbf{M}^{-1}\mathbf{A}\|^k,$$

where for k large $\|(\mathbf{I} - \mathbf{M}^{-1}\mathbf{A})^k\| \approx \rho(\mathbf{I} - \mathbf{M}^{-1}\mathbf{A})^k$ (the symbol $\rho(\mathbf{A})$ denotes the so-called spectral radius of $\mathbf{A}$, i.e. $\max\{|\lambda|;\ \lambda \in \sigma(\mathbf{A})\}$). We therefore see that the methods converge to the exact solution if

$$\rho(\mathbf{I} - \mathbf{M}^{-1}\mathbf{A}) < 1.$$

Even if this condition is satisfied, it can hold that $\|\mathbf{I} - \mathbf{M}^{-1}\mathbf{A}\|^k > 1$. In such a case one can observe the so-called transition effect, i.e. the approximation error first grows and only after certain number of iterations it starts decreasing (see Fig. 1).

Examples of classical iterative methods. The following methods are based on the splitting $\mathbf{A} = \mathbf{D} - \mathbf{L} - \mathbf{U}$, where $\mathbf{D}$ is the diagonal part, $-\mathbf{L}$ is the strict lower triangle and $-\mathbf{U}$ is the strict upper triangle of $\mathbf{A}$. From the equation

$$(\mathbf{D} - \mathbf{L} - \mathbf{U})\vec{x} = \vec{b}$$

one can derive particular methods.

Jacobi's method is defined by the iteration

$$\mathbf{D}\vec{x}_k = \mathbf{L}\vec{x}_{k-1} + \mathbf{U}\vec{x}_{k-1} + \vec{b}.$$

Writing this identity componentwise (x_i^k denotes the i-th component of $\vec{x}_k$), we obtain for $i = 1, \ldots, n$:

$$x_i^k = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1, j\neq i}^{n} a_{ij}x_j^{k-1}\right).$$

A disadvantage of this method can be that during the computation one needs to store two successive approximations $\vec{x}_{k-1}$, $\vec{x}_k$. **Gauss-Seidel's method** differs from the previous one in that it immediately uses the newly computed components of $\vec{x}_k$, i.e.

$$x_i^k = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^k - \sum_{j=i+1}^{n} a_{ij} x_j^{k-1} \right).$$

The computed components can thus replace the old ones within the same computational array. In the matrix form this method reads:

$$\mathbf{D}\vec{x}_k = \mathbf{L}\vec{x}_k + \mathbf{U}\vec{x}_{k-1} + \vec{b}.$$

From the Gauss-Seidel method one can derive the **Successive over-relaxation method** (SOR), which works with a relaxation parameter $\omega \in [0, 2]$ and is defined by the relation

$$\mathbf{D}\vec{x}_k = \omega(\mathbf{L}\vec{x}_k + \mathbf{U}\vec{x}_{k-1} + \vec{b}) + (1 - \omega)\mathbf{D}\vec{x}_{k-1},$$

i.e. it combines the Gauss-Seidel method with the previous approximation.

Example 2 *Consider the matrix*

$$\mathbf{A} = \begin{pmatrix} 0.01 & -0.4 \\ 0 & 0.01 \end{pmatrix}.$$

The convergence of the Jacobi method for this matrix depends on the properties of the matrix

$$\mathbf{I} - \mathbf{D}^{-1}\mathbf{A} = \begin{pmatrix} 0 & -40 \\ 0 & 0 \end{pmatrix}.$$

Since this matrix has only one eigenvalue $\lambda = 0$, the condition $\rho(\mathbf{I} - \mathbf{D}^{-1}\mathbf{A}) < 1$ is satisfied and hence Jacobi's method converges. On the other hand, $\|\mathbf{I} - \mathbf{D}^{-1}\mathbf{A}\| = 40 > 1$, hence during the computation one can observe the transition effect.

2.2 Krylov subspace methods

An important class of iterative methods is based on the idea of projecting the system $\mathbf{A}\vec{x} = \vec{b}$ onto a sequence of the Krylov spaces and so obtain successive approximations.

Definition 30 *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\vec{v} \in \mathbb{R}^n$ and $k \leqslant n$. The k-th Krylov space is the set*

$$\mathcal{K}_k(\mathbf{A}, \vec{v}) := \langle \vec{v}, \mathbf{A}\vec{v}, \mathbf{A}^2\vec{v}, \ldots, \mathbf{A}^{k-1}\vec{v} \rangle.$$

The methods which will be mentioned in the sequel, belong to the general class of the so-called projection methods which construct approximations in the form

$$\vec{x}_k \in \vec{x}_0 + \mathcal{S}_k, \quad \vec{r}_k \perp \mathcal{C}_k,$$

where $\vec{r}_k := \vec{b} - \mathbf{A}\vec{x}_k$ is the residual and $\mathcal{S}_k$ and $\mathcal{C}_k$ are suitable subspaces. The space $\mathcal{S}_k$ is usually the Krylov subspace $\mathcal{K}_k(\mathbf{A}, \vec{r}_0)$, but other choices are also possible, e.g. $\mathbf{A}\mathcal{K}_k(\mathbf{A}, \vec{r}_0)$. By the choice of the space $\mathcal{C}_k$ one can achieve the optimality of the approximation in the sense that the approximation error $\vec{x} - \vec{x}_k$ is minimal in some norm. If the dimension of $\mathcal{S}_k$, $\mathcal{C}_k$ increases, then for $k = n$ we obtain $\mathcal{C}_n = \mathbb{R}^n$ and from the condition $\vec{r}_k \perp \mathbb{R}^n$ it follows that $\vec{r}_n = \vec{0}$, i.e. $\vec{x}_n = \vec{x}$ is the exact solution. In other words, if the dimensions of $\mathcal{S}_k$, $\mathcal{C}_k$ increase then the projection methods find the exact solution of the system $\mathbf{A}\vec{x} = \vec{b}$ in at most n steps.

2.2.1 Conjugate gradient method (CG)

This method is intended for symmetric positive definite matrices.

Definition 31 *A matrix* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *is called* positive definite *if for every nonzero vector* $\vec{x} \in \mathbb{R}^n$ *it holds:*

$$\mathbf{A}\vec{x} \cdot \vec{x} > 0.$$

The expression

$$\|\vec{x}\|_{\mathbf{A}} := \sqrt{\vec{x} \cdot \mathbf{A}\vec{x}}$$

is called the energetic norm *or* $\mathbf{A}$*-norm. We say that the vectors* $\vec{u}, \vec{v} \in \mathbb{R}^n$ *are* $\mathbf{A}$*-orthogonal if*

$$\vec{u} \cdot \mathbf{A}\vec{v} = 0.$$

The approximations in CG are constructed according to the formula

$$\vec{x}_k := \vec{x}_{k-1} + \gamma_{k-1}\vec{p}_{k-1},$$

where $\vec{p}_{k-1}$ is called the conjugate vector and γ_{k-1} is the step length. These parameters are determined as follows:

- $\vec{p}_k$ is chosen in the form $\vec{p}_k := \vec{r}_k + \delta_k\vec{p}_{k-1}$ in such a way that it is $\mathbf{A}$-orthogonal to $\vec{p}_{k-1}$, i.e. $\vec{p}_k \cdot \mathbf{A}\vec{p}_{k-1} = 0$, which can be achieved for

$$\delta_k := \frac{\vec{r}_k \cdot \vec{r}_k}{\vec{r}_{k-1} \cdot \vec{r}_{k-1}}.$$

- γ_{k-1} is such that the energetic norm $\|\vec{x} - \vec{x}_k\|_{\mathbf{A}}$ is minimal. That happens if

$$\gamma_{k-1} := \frac{\vec{r}_{k-1} \cdot \vec{p}_{k-1}}{\vec{p}_{k-1} \cdot \mathbf{A}\vec{p}_{k-1}} = \frac{\vec{r}_{k-1} \cdot \vec{r}_{k-1}}{\vec{p}_{k-1} \cdot \mathbf{A}\vec{p}_{k-1}}.$$

Using the above properties, one can show that CG is a Krylov subspace method since

$$\vec{x}_k \in \vec{x}_0 + \mathcal{K}_k(\mathbf{A}, \vec{r}_0), \quad \vec{r}_k \perp \mathcal{K}_k(\mathbf{A}, \vec{r}_0).$$

The CG method can also be interpreted as a method for finding the minimum of the quadratic functional $\frac{1}{2}\vec{x} \cdot \mathbf{A}\vec{x} - \vec{x} \cdot \vec{b}$. The following algorithm represents the standard implementation of CG.

Algorithm A1 Conjugate gradient method

input $\mathbf{A}$, $\vec{b}$, $\vec{x}_0$
$\vec{r}_0 := \vec{b} - \mathbf{A}\vec{x}_0$
$\vec{p}_0 := \vec{r}_0$
for $k = 1, 2, \ldots$
  $\gamma_{k-1} := \frac{\vec{r}_{k-1} \cdot \vec{r}_{k-1}}{\vec{p}_{k-1} \cdot \mathbf{A}\vec{p}_{k-1}}$
  $\vec{x}_k := \vec{x}_{k-1} + \gamma_{k-1}\vec{p}_{k-1}$
  $\vec{r}_k := \vec{r}_{k-1} - \gamma_{k-1}\mathbf{A}\vec{p}_{k-1}$
  $\delta_k := \frac{\vec{r}_k \cdot \vec{r}_k}{\vec{r}_{k-1} \cdot \vec{r}_{k-1}}$
  $\vec{p}_k := \vec{r}_k + \delta_k \vec{p}_{k-1}$
end

We see that each iteration involves 1 multiplication of the matrix $\mathbf{A}$ with a vector and during the process it is necessary to store only 4 vectors. CG is therefore very efficient particularly for large sparse matrices. If the matrix is symmetric positive definite then CG in exact arithmetics finds the solution after at most n iterations. In practice, however, due to rounding errors the vectors $\{\vec{p}_k\}$ lose their $\mathbf{A}$-orthogonality (and $\{\vec{r}_k\}$ lose their orthogonality), which causes the delay of convergence, i.e. even after n steps $\vec{x}_n \neq \vec{x}$. This deficiency can be fixed by multiple orthogonalization of $\vec{r}_k$ with respect to all $\{\vec{r}_i\}_{i=0}^{k-1}$ (double orthogonalization is usually sufficient).

Now we mention some facts about the convergence rate of CG. For this reason we introduce the *condition number*.

Definition 32 *Let* $\mathbf{A}$ *be a symmetric positive definite matrix. The* condition number *of* $\mathbf{A}$ *is defined as*

$$\varkappa(\mathbf{A}) := \frac{\lambda_{max}(\mathbf{A})}{\lambda_{min}(\mathbf{A})},$$

where $\lambda_{max}(\mathbf{A})$, $\lambda_{min}(\mathbf{A})$ *stands for the largest and smallest eigenvalue of* $\mathbf{A}$, *respectively.*

Denoting $\vec{e}_k := \vec{x}_k - \vec{x}$ the error of k-th approximation then the following estimate holds:

$$\frac{\|\vec{e}_k\|_{\mathbf{A}}}{\|\vec{e}_0\|_{\mathbf{A}}} \leqslant 2\left(\frac{\sqrt{\varkappa(\mathbf{A})} - 1}{\sqrt{\varkappa(\mathbf{A})} + 1}\right)^k.$$

Notice that the fraction in brackets in the previous inequality is always less than 1. If $\varkappa(\mathbf{A})$ is close to 1 then the estimate says that the error decreases very rapidly. For ill-conditioned matrices (i.e. $\varkappa(\mathbf{A})$ is large) the number in brackets is close to 1 and the estimate can be too pesimistic (it overvalues the real size of the error). Ill-conditioning however ofter causes a slow convergence of the method. This fact can be overcome by preconditioning (see Section 2.3), i.e. by replacing the system $\mathbf{A}\vec{x} = \vec{b}$ by an equivalent system $\hat{\mathbf{A}}\hat{\vec{x}} = \hat{\vec{b}}$ with a matrix $\hat{\mathbf{A}}$ that is better conditioned than $\mathbf{A}$.

2.2.2 Generalized minimal residual method (GMRES)

The GMRES method can be characterized as a projection method which satisfies:

$$\vec{x}_k \in \vec{x}_0 + \mathcal{K}_k(\mathbf{A}, \vec{r}_0), \quad \vec{r}_k \perp \mathbf{A}\mathcal{K}_k(\mathbf{A}, \vec{r}_0).$$

Its property is that in each iteration it minimizes the residual norm $\|\vec{r}_k\|$. It leads to the least squares problem whose efficient implementation is technically demanding. For that reason we do not present its algorithm here. An inconvenient property of GMRES is that it produces a sequence of orthogonal vectors $\{\vec{v}_k\}$ that have to be stored (we say that the method generates long recurrences) and so a large amount of memory may be required. For this price however the method can solve any system with a nonsingular matrix.

Similarly as for CG, due to rounding errors the convergence of GMRES is delayed since the system $\{\vec{v}_k\}$ loses its orthogonality. For GMRES it is hence also usefull to perform reorthogonalization. The memory requirements are usually reduced by the so-called restart — the program stores only the last m vectors $\{\vec{v}_i\}_{i=k-m+1}^{k}$ instead of the whole sequence.

2.2.3 Biconjugate gradient method (BiCG)

The last example of a frequently used Krylov subspace method is BiCG, which in contrast to the previous methods solves simultaneously two systems: $\mathbf{A}\vec{x} = \vec{b}$ ans $\mathbf{A}^\top\vec{y} = \vec{c}$. Denoting $\vec{s}_k := \vec{c} - \mathbf{A}^\top\vec{y}_k$, then BiCG is characterized by the relations

$$\vec{x}_k \in \vec{x}_0 + \mathcal{K}_k(\mathbf{A}, \vec{r}_0), \qquad \vec{r}_k \perp \mathcal{K}_k(\mathbf{A}^\top, \vec{s}_0),$$
$$\vec{y}_k \in \vec{y}_0 + \mathcal{K}_k(\mathbf{A}^\top, \vec{s}_0), \qquad \vec{s}_k \perp \mathcal{K}_k(\mathbf{A}, \vec{r}_0).$$

Vectors $\{\vec{r}_k\}$ and $\{\vec{s}_k\}$ are mutually biorthogonal: $\vec{s}_i \cdot \vec{r}_j = 0$ for $i \neq j$.

Algorithm A2 Biconjugate gradient method (BiCG)

input $\mathbf{A}$, $\vec{b}$, $\vec{c}$, $\vec{x}_0$, $\vec{y}_0$
$\vec{r}_0 := \vec{p}_0 := \vec{b} - \mathbf{A}\vec{x}_0$
$\vec{s}_0 := \vec{q}_0 := \vec{c} - \mathbf{A}^\top\vec{y}_0$
for $k = 1, 2, \ldots$
  $\gamma_{k-1} := \frac{\vec{s}_{k-1}\cdot\vec{r}_{k-1}}{\vec{q}_{k-1}\cdot\mathbf{A}\vec{p}_{k-1}}$
  $\vec{x}_k := \vec{x}_{k-1} + \gamma_{k-1}\vec{p}_{k-1}$
  $\vec{r}_k := \vec{r}_{k-1} - \gamma_{k-1}\mathbf{A}\vec{p}_{k-1}$
  $\vec{y}_k := \vec{y}_{k-1} + \gamma_{k-1}\vec{q}_{k-1}$
  $\vec{r}_k := \vec{s}_{k-1} - \gamma_{k-1}\mathbf{A}^\top\vec{q}_{k-1}$
  $\delta_k := \frac{\vec{s}_k\cdot\vec{r}_k}{\vec{s}_{k-1}\cdot\vec{r}_{k-1}}$
  $\vec{p}_k := \vec{r}_k + \delta_k\vec{p}_{k-1}$
  $\vec{q}_k := \vec{s}_k + \delta_k\vec{q}_{k-1}$
end

This method generates short recurrences, i.e. it is efficient in terms of memory usage, and can be applied to any nonsingular matrix. However, in comparison with CG and GMRES the convergence of BiCG is not guaranteed. Actually,

if $\mathbf{A}$ is nonsymmetric then it may happen that $\vec{r}_k \cdot \vec{s}_k = 0$ and the algorithm terminates.

2.3 Preconditioning

As it was shown in Section 2.2.1, the convergence of Krylov subspace methods is closely related to the condition number of the matrix $\mathbf{A}$. We shall demonstrate the idea of preconditioning for CG (one can proceed similarly with other methods). Let $\mathbf{C}$ be any nonsingular matrix. Then the system $\mathbf{A}\vec{x} = \vec{b}$ with a symmetric positive definite matrix can be written as follows:

$$(\mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-\top})(\mathbf{C}^{\top}\vec{x}) = \mathbf{C}^{-1}\vec{b}.$$

Denoting $\hat{\mathbf{A}} := \mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-\top}$, $\hat{\vec{x}} := \mathbf{C}^{\top}\vec{x}$ and $\hat{\vec{b}} := \mathbf{C}^{-1}\vec{b}$, the new system can be written as $\hat{\mathbf{A}}\hat{\vec{x}} = \hat{\vec{b}}$, where $\hat{\mathbf{A}}$ is again symmetric positive definite. This system can be solved by CG and the approximations of the new and original system satisfy the relation $\vec{x}_k = \mathbf{C}^{-\top}\hat{\vec{x}}_k$. For completeness we present here the algorithm of the preconditioned CG method:

Algorithm A3 Preconditioned conjugate gradient method (PCG)

input $\mathbf{A}$, $\vec{b}$, $\vec{x}_0$
$\vec{r}_0 := \vec{b} - \mathbf{A}\vec{x}_0$
$\vec{z}_0 := \mathbf{C}^{-\top}\mathbf{C}^{-1}\vec{r}_0$
$\vec{p}_0 := \vec{z}_0$
for $k = 1, 2, \ldots$
  $\hat{\gamma}_{k-1} := \frac{\vec{z}_{k-1}\cdot\vec{r}_{k-1}}{\vec{p}_{k-1}\cdot\mathbf{A}\vec{p}_{k-1}}$
  $\vec{x}_k := \vec{x}_{k-1} + \hat{\gamma}_{k-1}\vec{p}_{k-1}$
  $\vec{r}_k := \vec{r}_{k-1} - \hat{\gamma}_{k-1}\mathbf{A}\vec{p}_{k-1}$
  $\vec{z}_k := \mathbf{C}^{-\top}\mathbf{C}^{-1}\vec{r}_k$
  $\hat{\delta}_k := \frac{\vec{z}_k\cdot\vec{r}_k}{\vec{z}_{k-1}\cdot\vec{r}_{k-1}}$
  $\vec{p}_k := \vec{z}_k + \hat{\delta}_k\vec{p}_{k-1}$
end

We remark that in this algorithm we never compute the inverse matrix $\mathbf{C}^{-1}$, but the operation $\vec{z}_k := \mathbf{C}^{-\top}\mathbf{C}^{-1}\vec{r}_k$ is transformed to successive solution of two systems

$$\mathbf{C}\vec{y} = \vec{r}_k, \quad \mathbf{C}^{\top}\vec{z}_k = \vec{y}.$$

For an efficient solution of the new system it is necessary to choose the matrix $\mathbf{C}$ according to the following rules:

- The matrix $\mathbf{C}$ is chosen in such a way that CG converges as fast as possible. In ideal case, $\hat{\mathbf{A}} = \mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-\top} \approx \mathbf{I}$.
- It has to be possible to solve the systems $\mathbf{C}\vec{y} = \vec{r}_k$ and $\mathbf{C}^{\top}\vec{z}_k = \vec{y}$ fast.
- If $\mathbf{A}$ is sparse then also $\mathbb{C}$ should be sparse. Otherwise the memory and computational requirements increase substantially.

An efficient choice of the preconditioning matrix often arises from the given (e.g. physical) problem or from a specific structure of the matrix $\mathbf{A}$. Commonly used general preconditioning techniques include e.g.:

- incomplete Cholesky decomposition (for symmetric positive definite matrices), which constructs a lower triangular matrix $\mathbf{C}$ such that $\mathbf{A} \approx \mathbf{C}\mathbf{C}^\top$,
- incomplete LU decomposition (for general nonsingular matrices): $\mathbf{A} \approx \mathbf{L}\mathbf{U}$, where $\mathbf{L}$ is lower triangular and $\mathbf{U}$ is upper triangular. The preconditioned system then has the form

$$(\mathbf{L}^{-1}\mathbf{A}\mathbf{U}^{-1})(\mathbf{U}\vec{x}) = \mathbf{L}^{-1}\vec{b}.$$

3 Introduction to functional analysis

In this section we shall study some abstract terms such as the metric, the norm and the scalar product. Before introducing the general notion we start by the example of space of continuous functions.

3.1 Space of continuous functions

In what follows, Ω denotes an open connected set in $\mathbb{R}$, $\mathbb{R}^2$ or $\mathbb{R}^3$. We remind that an open connected set is called domain. For simplicity we shall assume that Ω is bounded. The boundary of Ω will be denoted $\partial\Omega$, for the closure we use the symbol $\overline{\Omega} := \Omega \cup \partial\Omega$.

Let $C(\overline{\Omega})$ denote the linear space of all continuous functions in $\overline{\Omega}$. For any functions $u, v \in C(\overline{\Omega})$ we define the following operations:

Definition 33 *Scalar product of functions $u, v \in C(\overline{\Omega})$ is the (real) number*

$$(u, v) := \int_\Omega u(x)v(x)\ dx.$$

Norm of a function $u \in C(\overline{\Omega})$ is the number

$$\|u\|_2 := \sqrt{(u, u)} = \sqrt{\int_\Omega u^2(x)\ dx}.$$

Distance of two functions $u, v \in C(\overline{\Omega})$ is the number

$$\varrho_2(u, v) := \|u - v\|_2.$$

The distance ϱ_2 is also called the *metric*. One can easily show some basic properties of the above defined scalar product, norm and metric:

Theorem 14 *Let $u, v, w \in C(\overline{\Omega})$ and $\alpha, \beta \in \mathbb{R}$. Then*

(i) $(u, u) \geqslant 0$,

(ii) $(u, v) = (v, u)$,

(iii) $(\alpha u + \beta v, w) = \alpha(u, w) + \beta(v, w)$,

(iv) $(u, u) = 0 \Leftrightarrow u \equiv 0 \ v\ \Omega$.

Theorem 15 *Let* $u, v \in C(\overline{\Omega})$ *and* $\alpha \in \mathbb{R}$. *Then*

(i) $\|u\|_2 = 0 \Leftrightarrow u \equiv 0 \ v\ \Omega$,

(ii) $\|\alpha u\|_2 = |\alpha| \|u\|_2$,

(iii) $\|u + v\|_2 \leqslant \|u\|_2 + \|v\|_2$.

Theorem 16 *Let* $u, v, w \in C(\overline{\Omega})$. *Then*

(i) $\varrho_2(u, v) = 0 \Leftrightarrow u = v \ v\ \Omega$,

(ii) $\varrho_2(u, v) = \varrho_2(v, u)$,

(iii) $\varrho_2(u, w) \leqslant \varrho_2(u, v) + \varrho_2(v, w)$.

In addition to the above introduced norm and metric, it is possible to define the following norm and metric in the space of continuous functions:

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

$$\varrho_\infty(u, v) := \|u - v\|_\infty.$$

One can show that Theorem 15 and 16 hold if we replace $\|\cdot\|_2$ by $\|\cdot\|_\infty$ and ϱ_2 by ϱ_∞.

Example 3 *Consider the functions*

$$u(x) := \begin{cases} 10 \sin(1000\pi x) & \text{for } x \in [0, \frac{1}{1000}] \\ 0 & \text{otherwise} \end{cases} \quad \text{and } v(x) = 0.$$

Let us compute the distance of u *and* v*:*

$$\varrho_2(u, v) = \sqrt{\int_0^{1/1000} 100 \sin^2(1000\pi x)\ dx}$$

$$= \sqrt{100 \left[\frac{x}{2} - \frac{\sin(2000\pi x)}{4000\pi} \right]_{x=0}^{1/1000}} = \frac{1}{2\sqrt{5}} \doteq 0,224.$$

$$\varrho_\infty(u, v) = \max_{x \in [0, 1/1000]} |10 \sin(1000\pi x)| = 10.$$

The metric ϱ_∞ appears more natural for the measurement of the deviation of two functions since it measures the maximal pointwise difference. Anyway, there exist reasons for using ϱ_2 instead of ϱ_∞. First, the metric ϱ_2 is introduced using the scalar product, which plays an important role in certain problems. On the set of continuous functions it is not possible to define a scalar product with reasonable properties, which would induce the norm $\|\cdot\|_\infty$ and the metric ϱ_∞. Another reason is that many applications deal with discontinuous functions. It is not easy to extend ϱ_∞ to a more general class, while the extension of ϱ_2 to a sufficiently general class of functions is very straightforward and leads to the space L^2.

3.2 Spaces $L^p(\Omega)$

Definition 34 *Let $p \in [1, \infty)$. The space $L^p(\Omega)$ is the set*

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

equipped by the norm

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

If it is evident what domain is considered then we shall write only $\|u\|_p$. From the definition it follows that every continuous function in $\overline{\Omega}$ belongs to $L^p(\Omega)$ for all $p \in [1, \infty)$. These spaces contain also many other functions — e.g. discontinuous or unbounded ones. We remind that for correctness the integrals used in Definition 34 have to be considered in the so called Lebesgue sense.

Example 4 *Consider the function*

$$u(x) := \frac{1}{\sqrt{x}}$$

defined in $\Omega := (0, 1)$. It holds:

$$\|u\|_1 = \int_0^1 \frac{1}{\sqrt{x}}\ dx = \left[2\sqrt{x}\right]_{x=0}^1 = 2,$$

and thus $u \in L^1(0, 1)$. On the other hand

$$\|u\|_2 = \sqrt{\int_0^1 \frac{1}{x}\ dx} = +\infty,$$

hence $u \notin L^2(0, 1)$. The function

$$v(x) := \frac{1}{\sqrt[3]{x}}$$

belongs both to $L^1(0,1)$ *and* $L^2(0,1)$*, since*

$$\int_0^1 v(x)\ dx = \int_0^1 \frac{1}{\sqrt[3]{x}}\ dx = \frac{3}{2}, \quad \int_0^1 |v(x)|^2\ dx = \int_0^1 \frac{1}{\sqrt[3]{x^2}}\ dx = 3.$$

If the domain Ω *is bounded then always* $L^2(\Omega) \subset L^1(\Omega)$*, or more generally, for* $1 \leqslant p \leqslant q < \infty$ *it holds that* $L^q(\Omega) \subset L^p(\Omega)$*.*

Example 5 *The function*

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

is an element of the space $L^p(-1,1)$ *for any* $p \in [1,\infty)$*, because*

$$\int_{-1}^1 |\operatorname{sgn} x|^p\ dx = \int_{-1}^0 |\operatorname{sgn} x|^p\ dx + \int_0^1 |\operatorname{sgn} x|^p\ dx = \int_{-1}^0 1\ dx + \int_0^1 1\ dx = 2,$$

and thus $\|\operatorname{sgn} x\|_p = \sqrt[p]{2}$*. Similarly, the function*

$$u(x) := \begin{cases} 0 & \text{for } x \neq 0 \\ 10 & \text{for } x = 0 \end{cases}$$

belongs to $L^p(-1,1)$*,* $p \in [1,\infty)$*, and its norm is* $\|u\|_p = 0$*. We see that the norm is independent of the function value at* $x = 0$*. It is even not needed that the function is defined at 0.*

We have shown that the spaces $L^p(\Omega)$ contain also some discontinuous and unbounded functions. A function that is zero everywhere except for a single point has zero norm, i.e. it is in some sense equialent to the zero function. More generally, if a function equals zero everywhere in Ω except for a *set of measure zero*, then its norm is zero. For example, every finite and every countable set has zero measure.

Definition 35 *Let the functions* $u, v \in L^p(\Omega)$*,* $p \in [1,\infty)$ *be* equal almost everywhere *in the domain* Ω*, i.e. everywhere except for a set of zero measure (where their values are different or some of the functions is not defined). The we say that* u *and* v *are* equivalent *in* $L^p(\Omega)$ *and we write* $u = v$ *in* $L^p(\Omega)$*.*

Functions u and v from the above definition are considered identical in the space $L^p(\Omega)$. Equivalent functions u, v in $L^p(\Omega)$ are characterized by the property

$$\int_\Omega |u(x) - v(x)|^p\ dx = 0.$$

In $L^p(\Omega)$ one can introduce the metric $\varrho_{p,\Omega}(u,v) := \|u - v\|_{p,\Omega}$. In addition, the space $L^2(\Omega)$ is endowed with the scalar product from Definition 33:

$$(u,v)_\Omega := \int_\Omega u(x)v(x)\ dx,\ u, v \in L^2(\Omega).$$

The finiteness of the scalar product of any $u, v \in L^2(\Omega)$ is a consequence of the Schwarz inequality:

$$\forall u, v \in L^2(\Omega) : \ |(u, v)_\Omega| \leqslant \|u\|_{2,\Omega} \|v\|_{2,\Omega},$$

which is a special case of the following more general statement:

Theorem 17 (Hölder's inequality) *Let $p, q \in (1, \infty)$ satisfy the relation*

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Then for all functions $u \in L^p(\Omega)$ and $v \in L^q(\Omega)$ it holds:

$$\left| \int_\Omega u(x) v(x) \ dx \right| \leqslant \left(\int_\Omega |u(x)|^p \ dx \right)^{1/p} \left(\int_\Omega |v(x)|^q \ dx \right)^{1/q}.$$

3.3 Metric spaces

In this section we introduce the metric, a generalization of the notion of distance. We shall state some basic properties of metric spaces and name some particular examples.

Definition 36 *Let X be a nonempty set. The function $\varrho : X \times X \to \mathbb{R}$ is called a* metric, *if it satisfies for all $x, y, z \in X$:*

(i) $\varrho(x, y) = 0 \Leftrightarrow x = y$,

(ii) $\varrho(x, y) = \varrho(y, x)$,

(iii) $\varrho(x, z) \leqslant \varrho(x, y) + \varrho(y, z)$.

The pair (X, ϱ) is called a metric space.

The properties of the metric it follows that ϱ attains only nonnegative values (try to prove it!).

Example 6 *Let X be the set of all cities in the Czech Republic. A metric on X can be introduced e.g. as the direct euclidean distance, the shortest distance vie roads or the travel time of a car.*

Example 7 *Let $X = \mathbb{R}^n$, $n \in \mathbb{N}$. Let us define the functions*

$$d_p(\vec{x}, \vec{y}) := \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{1/p} \quad \text{for } p \in [1, \infty),$$

$$d_\infty(\vec{x}, \vec{y}) := \max_{i=1,\dots,n} |x_i - y_i|.$$

It can be shown that these functions are metrics on $\mathbb{R}^n$. The metric d_1 is called the counting metric, d_2 is the euclidean metric and d_∞ is the maximum metric.

Example 8 *Levenshtein's metric mearures the similarity of text strings. It is defined as the minimal number of character substitutions, insertions or deletions necessary to transform one string to another. E.g.* $lev(dog,frog) = 2$, *because the transformation can be done as follows:*

$$dog \to \underline{f}og \to f\underline{r}og.$$

It can be shown that lev satisfies the axioms of the metric.

Example 9 *Functions* ϱ_2 *and* ϱ_∞ *are metrics on the space* $C(I)$ *of continuous functions on a closed bounded interval* I. *The same holds on the space* $C(K)$, *where* K *is a compact (i.e. closed and bounded) set in* $\mathbb{R}^n$, $n \in \mathbb{N}$.

Example 10 *The pair* $(L^p(\Omega), \varrho_p)$, $p \in [1, \infty)$ *is a metric space. Since continuous function on* $\overline{\Omega}$ *belong to* $L^p(\Omega)$, *the pair* $(C(\overline{\Omega}), \varrho_p)$ *is a subspace of the metric space* $(L^p(\Omega), \varrho_p)$.

3.3.1 Sets in metric spaces

Using the notion of metric, it is possible to generalize many objects introduced through the euclidean distance in $\mathbb{R}^n$ such as a ball, a neighbourhood or an open set.

Definition 37 *Let* (X, ϱ) *be a metric space.*

- *A* ball *centered at* $x \in X$ *with the radius* $r > 0$ *is the set*

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

- *A set* $O \subset X$ *is called a* neighbourhood *of a point* x, *if there exists a radius* $r > 0$, *such that* O *contains the ball* $B_r(x)$.

- *If* O *is a neighbourhood of the point* x, *then the set* $O \backslash \{x\}$ *is called a* ring neighbourhood *of* x.

- *A set* M *is called* open, *if for every point* $x \in M$ *there exists a ball centered at* x *and contained in* M.

- *A set is called* closed, *if its complement in* X *is open.*

Example 11 *A ball in the space* $\mathbb{R}^2$ *centered at the origin has the following shape:*

a) square whose vertices lie on the axes and the barycenter at the origin, if the counting metric d_1 *is considered;*

b) circle centered at the origin, if the euclidean metric d_2 *is considered;*

c) square whose sides are parallel to the axes and the barycenter lies at the origin, if the maximum metric d_∞ *is considered.*

Definition 38 *Let (X, ϱ) be a metric space, $x \in X$ and $M \subset M$.*

- *A point x is an* interior point of the set M*, if there exists a radius $r > 0$ such that $B_r(x) \subset M$. The set of all interior points of M is denoted* $\operatorname{Int} M$.
- *A point x is a* boundary point of the set M*, if every neighbourhood of x contains some point from M and some point from $X \backslash M$. The set of all boundary points of M is called the* boundary *of M and is denoted ∂M.*
- *The closure of the set M is the set $\overline{M} := M \cup \partial M$.*
- *A point x is an* accumulation point of the set M*, if every its ring neighbourhood contains some point from M. The set of all accumulation points of M is denoted* $\operatorname{Hr} M$.
- *A point x is an* isolated point of the set M*, if $x \in M$, but x is not an accumulation point of M. The set of all isolated points of M is denoted* $\operatorname{Iz} M$.

There are many relations between the previously defined sets. We give a few examples:

$$\operatorname{Int} M \subset M \subset \overline{M}, \quad \operatorname{Int} M \cap \partial M = \varnothing,$$
$$\overline{M} = \operatorname{Hr} M \cup \operatorname{Iz} M, \quad \operatorname{Hr} M \cap \operatorname{Iz} M = \varnothing,$$
$$\operatorname{Iz} M \subset \partial M, \quad \operatorname{Int} M \subset \operatorname{Hr} M.$$

3.3.2 Convergence

Definition 39 *A sequence $\{x_n\}_{n\in\mathbb{N}}$ in a metric space (X, ϱ) is called* convergent*, if there exists an element $x \in X$ such that*

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

We say that x is the limit of the sequence $\{x_n\}$ and write

$$x = \lim_{n\to\infty} x_n \text{ in } (X, \varrho), \text{ or } x_n \to x \text{ in } (X, \varrho).$$

The limit in metric space shares many properties of the classical limit in $\mathbb{R}^n$. E.g. every sequence has at most one limit. For the metric ϱ_p, $p \in [1, \infty]$, on function spaces it holds that

$$(\lim_{n\to\infty} u_n)(x) = \lim_{n\to\infty} (u_n(x)),$$

i.e. the limit in ϱ_p coinsides with the pointwise limit. When examining the convergence of a sequence of functions, it is therefore suitable to check whether the sequence has a pointwise limit.

Example 12 *Consider the sequence of functions* $\{u_n\}$,

$$u_n(x) := \begin{cases} 10\sin(n\pi x) & \text{for } x \in [0, \frac{1}{n}] \\ 0 & \text{else} \end{cases}$$

in the space $C([0,1])$. *To decide whether the sequence is convergent or not, we first need to find a suitable "candidate" for the limit. We compute the pointwise limit at every* $x \in [0,1]$*: Clearly* $\lim u_n(0)$. *For every* $x \in (0,1]$ *one can find a number* $n_0 \in \mathbb{N}$ *such that* $x > \frac{1}{n_0}$, *so that for* $n \geqslant n_0$ *we have* $u_n(x) = 0$, *and thus* $\lim u_n(x) = 0$. *The pointwise limit of our sequence is hence the zero function. It is possible to show that*

$$\varrho_2(u_n, 0) = \frac{1}{\sqrt{2n}}, \text{ which implies } \lim \varrho_2(u_n, 0) = 0,$$

and therefore

$$\lim u_n = 0 \text{ in } (C([0,1]), \varrho_2).$$

Further,

$$\varrho_\infty(u_n, 0) = 10,$$

which implies that in the space $(C([0,1]), \varrho_\infty)$ *the zero function is not the limit of the sequence* $\{u_n\}$ *(in fact, the sequence does not have any limit in this space, so it is not convergent).*

The above example reveals that the existence of a limit depends on the considered metric.

For bounded domains we can characterize relations between various convergences.

Theorem 18 *Let* Ω *be a bounded domain in* $\mathbb{R}^d$, $d \in \{1,2,3\}$, *and* $\{u_n\}$ *be a sequence of functions.*

(i) If $u_n \to u$ *in* $(C(\overline{\Omega}), \varrho_\infty)$, *then* $u_n \to u$ *also in* $(L^p(\Omega), \varrho_p)$, $p \in [1, \infty)$.

(ii) If $u_n \to u$ *in* $(L^p(\Omega), \varrho_p)$, *then* $u_n(x) \to u(x)$ *for almost all* $x \in \Omega$.

Definition 40 *Let* ϱ^1 *and* ϱ^2 *be metrics on the set* X. *If there exist constants* $\alpha, \beta > 0$ *such that for every* $x, y \in X$*:*

$$\alpha\varrho^1(x,y) \leqslant \varrho^2(x,y) \leqslant \beta\varrho^1(x,y),$$

then we say that ϱ^1 *and* ϱ^2 *are* equivalent *on* X.

If ϱ^1 and ϱ^2 are equivalent metrics then

$$x_n \to x \text{ in } (X, \varrho^1) \quad \Leftrightarrow \quad x_n \to x \text{ in } (X, \varrho^2).$$

Equivalent metrics also induce the same open and closed sets.

3.3.3 Complete metric space

Definition 41 *A sequence $\{x_n\}$ in a metric space (X, ϱ) is called a Cauchy sequence if*

$$\forall \varepsilon > 0 \ \exists N \in \mathbb{N} \ \forall m, n \in \mathbb{N} : \ m, n > N \Rightarrow \varrho(x_m, x_n) < \varepsilon.$$

The distance of elements of a Cauchy sequence tends to zero when the indices of these elements increase. Every convergent sequence is Cauchy, however the reciprocal assertion does not hold true in general.

Definition 42 *A metric space (X, ϱ) is called* complete*, if every Cauchy sequence is convergent in this space.*

Example 13 *The spaces $(\mathbb{R}^n, d_p)$, $n \in \mathbb{N}$, $p \in [1, \infty]$ are complete (due to Bolzano-Cauchy theorem). The space $(\mathbb{Q}, d_2)$ is not complete (e.g. the sequence $\{(1 + \frac{1}{n})^n\}$ is Cauchy but its limit e satisfies $e \notin \mathbb{Q}$).*

Example 14 *Consider the sequence of functions $\{u_n\}$,*

$$u_n(x) := \sqrt[2n+1]{x},$$

in the space $(C([-1, 1]), \varrho_2)$. Its pointwise limit is sgn x*, a discontinuous function. However it holds that*

$$\varrho_2(u_n, \mathrm{sgn}) = \sqrt{\frac{2}{(n+1)(2n+3)}},$$

so that $\varrho_2(u_n, \mathrm{sgn}) \to 0$, and consequently $\{u_n\}$ converges to sgn *in the space $(L^2(-1, 1), \varrho_2)$. The sequence is hence Cauchy, but not convergent in $(C([-1, 1]), \varrho_2)$. This shows that the space $(C([-1, 1]), \varrho_2)$ is not complete. Similarly one can show that the space $(C(\overline{\Omega}), \varrho_p)$ is incomplete for any $p \in [1, \infty)$.*

Theorem 19 *The space $(C(\overline{\Omega}), \varrho_\infty)$ is complete.*

Theorem 20 *The space $(L^p(\Omega), \varrho_p)$, $p \in [1, \infty)$, is complete.*

3.3.4 Dense set, separable space

Metric spaces in general do not have linear structure like linear spaces. Thus it is not possible to define a basis. Every metric space, however, contains a subset whose elements can approximate every element of the space.

Definition 43 *We say that a set $M \subset X$ is* dense *in the metric space (X, ϱ), if $\overline{M} = X$.*

If M is a dense set then for every $x \in X$ there exists a sequence $\{x_n\}$ of elements of M such that

$$x_n \to x.$$

Theorem 21 *The set of all polynomials is dense in $L^p(\Omega)$, $p \in [1, \infty)$.*

A consequence of the previous theorem is that for every function $f \in L^p(\Omega)$ and every number $\varepsilon > 0$ there exists a polynomial f_ε satisfying

$$\|f - f_\varepsilon\|_p < \varepsilon.$$

The set of all polynomials is however quite large (namely uncountable).

Definition 44 *We say that a metric space is* separable*, if it contains a dense set which is at most countable.*

If a space is separable then there exists a sequence of its elements which form a dense set.

Theorem 22 *The space $L^p(\Omega)$, $p \in [1, \infty)$, is separable.*

An example of a countable dense set in $L^p(\Omega)$ is the set of all polynomials with rational coefficients.

3.4 Normed linear spaces

Many sets have both properties of metric and linear spaces. In particular, in $L^p(\Omega)$ we can multiply by a scalar and add functions, measure their distance and evaluate the norm. In such case we speak about a normed linear space.

Definition 45 *Let X be a linear space. A function $\|\cdot\| : X \to \mathbb{R}$ is called a* norm *in X, if $\forall x, y \in X,\ \alpha \in \mathbb{R}$:*

(i) $\|x\| = 0 \ \Leftrightarrow \ x = \vec{0}$,

(ii) $\|\alpha x\| = |\alpha| \|x\|$,

(iii) $\|x + y\| \leqslant \|x\| + \|y\|$.

If there exists a norm in X then X is called a normed linear space.

With the help of a norm one can always define a metric:

$$\varrho(x, y) := \|x - y\|,$$

thus every normed linear space is also a metric space.

Definition 46 *A complete normed linear space is called a* Banach space.

Example 15 *Examples of norms and normed linear spaces:*

- *The set $\mathbb{R}$ with the absolute value $\|x\| := |x|$;*
- *$\mathbb{R}^n$, $n \in \mathbb{N}$, with the norm $\|(x_1, \ldots, x_n)\|_p := \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$, $p \in [1, \infty)$, or with the norm $\|(x_1, \ldots, x_n)\|_\infty := \max_{i=1,\ldots,n} |x_i|$;*

- $C(\overline{\Omega})$ *with the norm* $\|\cdot\|_p$, $p \in [1,\infty]$;
- $L^p(\Omega)$ *with the norm* $\|\cdot\|_p$, $p \in [1,\infty)$;
- $C^1(\overline{\Omega}) := \{f \in C(\overline{\Omega});\ \forall i = 1,\dots,n\ \frac{\partial f}{\partial x_i} \in C(\overline{\Omega})\}$ *with the norm* $\|f\|_{C^1(\overline{\Omega})} := \|f\|_{\infty,\overline{\Omega}} + \sum_{i=1}^n \|\frac{\partial f}{\partial x_i}\|_{\infty,\overline{\Omega}}$.

A special class of norms are matrix norms.

Definition 47 *Let* $\|\cdot\|_X$ *denote a norm in* R^n *and* $\|\cdot\|_Y$ *a norm in* $\mathbb{R}^m$. *An induced norm in the space of matrices* $\mathbb{R}^{m\times n}$ *is defined by the relation*

$$\|\mathbf{A}\|_{XY} := \max_{\vec{x}\in\mathbb{R}^n\setminus\{\vec{0}\}} \frac{\|\mathbf{A}\vec{x}\|_Y}{\|\vec{x}\|_X} = \max_{\vec{x}\in\mathbb{R}^n, \|\vec{x}\|_X=1} \|\mathbf{A}\vec{x}\|_Y.$$

Induced norms have the following properties:

$$\|\mathbf{AB}\|_{XY} \leqslant \|\mathbf{A}\|_{XY}\|\mathbf{B}\|_{XY}, \quad \rho(\mathbf{A}) \leqslant \|\mathbf{A}\|_{XY}, \quad \|\mathbf{I}\|_{XY} = 1.$$

Example 16 *Examples of induced matrix norms:*

- $\|\mathbf{A}\|_1 := \max_{\|\vec{x}\|_1=1} \|\mathbf{A}\vec{x}\|_1 = \max_{j=1,\dots,n} \sum_{i=1}^m |a_{ij}|$;
- $\|\mathbf{A}\|_2 := \max_{\|\vec{x}\|_2=1} \|\mathbf{A}\vec{x}\|_2 = \sqrt{\rho(\mathbf{A}^\top\mathbf{A})}$;
- $\|\mathbf{A}\|_\infty := \max_{\|\vec{x}\|_\infty=1} \|\mathbf{A}\vec{x}\|_\infty = \max_{i=1,\dots,m} \sum_{j=1}^n |a_{ij}|$.

There exist many other matrix norms that are not induced by any vector norm. A frequently used is the Frobenius norm

$$\|\mathbf{A}\|_F := \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2}.$$

One can show that the Frobenius norm is not induced but still it is multiplicative:

$$\|\mathbf{AB}\|_F \leqslant \|\mathbf{A}\|_F\|\mathbf{B}\|_F.$$

3.5 Spaces $H^1(\Omega)$

It is sometimes necessary to work with derivatives of functions from the space $L^p(\Omega)$. Before introducing the new type of derivative, we mention a motivating example. Consider a function $u \in C^1([0,1])$. If $v \in C^1([0,1])$, $v(0) = v(1) = 0$, then by integration by parts we have:

$$\int_0^1 u'(x)v(x)\ dx = [u(x)v(x)]_{x=0}^1 - \int_0^1 u(x)v'(x)\ dx = -\int_0^1 u(x)v'(x)\ dx.$$

Here we see that, while the left integral requires the existence of u', the expression on the right is defined also for $u \in L^1(0,1)$. This leads to the following definition.

Definition 48 *Let $u \in L^p(\Omega)$. A function $g \in L^p(\Omega)$ is called the* generalized partial derivative *of the function u with respect to i-th variable, if for every $v \in C^1(\overline{\Omega})$, $v_{\partial\Omega} = 0$, it holds:*

$$\int_\Omega g(x)v(x)\ dx = -\int_\Omega u(x)\frac{\partial v}{\partial x_i}(x)\ dx.$$

We write $g = \frac{\partial u}{\partial x_i}$ in $L^p(\Omega)$.

Example 17 *Let us compute the generalized derivative of $u(x) := |x|$ in the interval $(-1,1)$. For $v \in C^1([-1,1])$, $v(-1) = v(1) = 0$ it holds:*

$$-\int_{-1}^1 |x|v'(x)\ dx = -\int_{-1}^0 (-x)v'(x)\ dx - \int_0^1 xv'(x)\ dx$$
$$= [xv(x)]_{x=-1}^0 - \int_{-1}^0 v(x)\ dx - [xv(x)]_{x=0}^1 + \int_0^1 v(x)\ dx = \int_{-1}^1 \operatorname{sgn} x v(x)\ dx.$$

Hence $u' = \operatorname{sgn}$ in $L^p(-1,1)$ for every $p \in [1,\infty)$.

Example 18 *Consider the function $u(x) = \operatorname{sgn} x$. For $v \in C^1([-1,1])$, $v(-1) = v(1) = 0$ it holds:*

$$-\int_{-1}^1 u(x)v'(x)\ dx = -\int_{-1}^0 (-v'(x))\ dx - \int_0^1 v'(x)\ dx$$
$$= [v(x)]_{x=-1}^0 - [v(x)]_{x=0}^1 = 2v(0) = 2\int_{-1}^1 \delta_0(x)v(x)\ dx,$$

where δ_0 is the so-called Dirac δ-function (in fact it is not a function but a distribution*). In certain sense it holds that $\operatorname{sgn}' = 2\delta_0$, however $\delta_0 \notin L^p(\Omega)$.*

Not every function from $L^p(\Omega)$ has a generalized derivative in $L^p(\Omega)$.

Definition 49 *The space $H^1(\Omega)$ is defined as follows:*

$$H^1(\Omega) := \{u \in L^2(\Omega);\ \forall i = 1,\dots,n : \frac{\partial u}{\partial x_i} \in L^2(\Omega)\}.$$

It is equipped with the norm

$$\|u\|_{H^1(\Omega)} := \left(\|u\|_2^2 + \sum_{i=1}^n \|\frac{\partial u}{\partial x_i}\|_2^2\right)^{1/2}$$

and the scalar product

$$((u,v)) := (u,v) + \sum_{i=1}^n (\frac{\partial u}{\partial x_i}, \frac{\partial v}{\partial x_i}).$$

Theorem 23 *The space $H^1(\Omega)$ is a separable Banach space.*

3.6 Spaces with scalar product

Scalar product plays an important role in many physical and engineering problems. We known the properties of the scalar product in Euclidean spaces $\mathbb{R}^n$, in $L^2(\Omega)$ or in $H^1(\Omega)$. Now we present the general definition and properties of spaces with scalar product.

Definition 50 *Let X be (real) linear space. A mapping $(\cdot,\cdot) : X \times X \to \mathbb{R}$ is called a scalar product, if for every $x, y, z \in X$ and $\alpha \in \mathbb{R}$ it holds:*

(i) $(x,x) \geqslant 0$, $(x,x) = 0 \Leftrightarrow x = \vec{0}$,

(ii) $(x,y) = (y,x)$,

(iii) $(\alpha x, y) = \alpha(x,y)$,

(iv) $(x+y,z) = (x,z) + (y,z)$.

A linear space with a scalar product is called a space with scalar product.

Every scalar product induces a norm $|||x||| := \sqrt{(x,x)}$. If X endowed with this norm is complete then X is called a *Hilbert space*. There holds the so-called *Cauchy-Schwarz inequality*:

$$\forall x, y \in X : \ |(x,y)| \leqslant |||x||| \cdot |||y|||.$$

Definition 51 *A set $M \subset X$ in a Hilbert space X is called* orthogonal, *if all its elements are mutually orthogonal, i.e.*

$$\forall x, y \in M, x \neq y : \ (x,y) = 0.$$

If in addition

$$\forall x \in M : \ |||x||| = 1,$$

then M is called an orthonormal system.

An example of an orthonormal system is the canonical basis in $\mathbb{R}^n$ or the set

$$\left\{\frac{1}{\sqrt{2\pi}}\right\} \cup \left\{\frac{1}{\sqrt{\pi}}\sin nx;\ n \in \mathbb{N}\right\} \cup \left\{\frac{1}{\sqrt{\pi}}\cos nx;\ n \in \mathbb{N}\right\} \text{ in } L^2(-\pi,\pi).$$

3.7 Weak solution of a boundary value problem and the Galerkin method

At the end we demonstrate the functional-analytic approach to the solution of boundary value problems for ODE with discontinuous right hand side.

Let us consider the problem

$$-u'' + u = f \text{ v } (0,1), \quad u(0) = u(1) = 0.$$

If $f \in C[0,1]$, then it makes sense to look for the classical solution, i.e. a function $u \in C^2(0,1) \cap C[0,1]$ such that the above identities hold in the whole interval $(0,1)$. If the right hand side is less regular then the classical solution need not exist. For this case we will demonstrate the derivation of the so-called weak (generalized) solution.

Let us assume that u is a classical solution. Then for every $v \in V := \{v \in C^1[0,1];\ v(0) = v(1) = 0\}$ it holds:

$$(-u'' + u, v) = (f, v).$$

Integrating by parts we obtain:

$$\begin{aligned}(-u'' + u, v) &= \int_0^1 (-u''(x) + u(x))v(x)\ dx \\ &= [-u'(x)v(x)]_{x=0}^1 + \int_0^1 u'(x)v'(x) + u(x)v(x)\ dx = ((u, v)).\end{aligned}$$

Instead of a classical solution we can therefore look for a function u such that

$$((u, v)) = (f, v)$$

for all $v \in V$. Since V is not complete in the norm of $H^1(0,1)$, it is not suitable for the definition of the generalized solution u. By completing V in the norm of $H^1(0,1)$ we obtain the space

$$H_0^1(0,1) := \{v \in H^1(0,1);\ v(0) = v(1) = 0^1\}.$$

Weak (generalized) solution of the boundary value problem thus can be defined as a function $u \in H_0^1(0,1)$, which satisfies

$$\forall v \in H_0^1(0,1) :\ ((u, v)) = (f, v).$$

Note that this formulation has sense for every $f \in L^2(0,1)$.

Let $\{v_i\}_{i=1}^\infty$ be a basis of the space $H_0^1(0,1)$. Galerkin approximation of the weak solution is defined as a function

$$u^n(x) := \sum_{i=1}^n \alpha_i^n v_i(x),$$

which satisfies

$$\forall j = 1, \ldots, n :\ ((u^n, v_j)) = (f, v_j).$$

Expressing u^n in the above identity we obtain a system of linear algebraic equations

$$\sum_{i=1}^n \alpha_i^n ((v_i, v_j)) = (f, v_j),\ j = 1, \ldots, n,$$

[1] The expressions $v(0)$, $v(1)$ here denote the so-called trace of a function v.

for the unknown coefficients $\vec{u} := (\alpha_1^n, \dots, \alpha_n^n)^\top$. Defining the matrix $\mathbf{A} = (a_{ij})_{i,j=1}^n$, where $a_{ij} := ((v_j, v_i))$, and the vector $\vec{b} = (b_i)_{i=1}^n$, where $b_i := (f, v_i)$, we can rewrite this system in the compact form

$$\mathbf{A}\vec{u} = \vec{b}.$$

Due to the properties of the scalar product, the matrix $\mathbf{A}$ is symmetric positive definite, hence the system has a unique solution for every $\vec{b} \in \mathbb{R}^n$.

One can also show that the sequence of functions $\{u^n\}$ is in certain sense convergent and that its limit is the weak solution u.

4 Notation

Below are listed and explained some symbols frequently used in this text.

symbol	**meaning**
$\mathbb{N}$	set of all natural numbers (1, 2, 3, ...)
$\mathbb{Z}$	set of all integers
$\mathbb{Q}$	set of all rational numbers
$\mathbb{R}$	set of all real numbers
$\mathbb{C}$	set of all complex numbers
$A \subset B$	A is a subset of B
$A \cap B$	intersection
$A \cup B$	union
$A \backslash B$	set difference
$A \times B$	cartesian product
$(a_1, \dots, a_n)$	ordered n-tuple
(a, b)	open interval
$[a, b]$	closed interval