1 Linear Algebraic Systems and Eigenvalue Problems

In this chapter, we introduce the basic notation and facts about the normed real or complex vector spaces $\mathbb{K}^n$ of n-dimensional vectors and $\mathbb{K}^{n\times n}$ of corresponding $n \times n$-matrices. The emphasis is on square matrices as representations of linear mappings in $\mathbb{K}^n$ and their spectral properties.

1.1 The normed Euclidean space $\mathbb{K}^n$

1.1.1 Vector norms and scalar products

We recall some basic topological properties of the finite dimensional "normed" (vector) space $\mathbb{K}^n$, where depending on the concrete situation $\mathbb{K} = \mathbb{R}$ (real space) or $\mathbb{K} = \mathbb{C}$ (complex space). In the following each point $x \in \mathbb{K}^n$ is expressed by its canonical coordinate representation $x = (x_1, \ldots, x_n)$ in terms of a (fixed) Cartesian basis $\{e^1, \ldots, e^n\}$ of $\mathbb{K}^n$,

$$x = \sum_{i=1}^{n} x_i e^i.$$

Definition 1.1: *A mapping $\|\cdot\| : \mathbb{K}^n \to \mathbb{R}$ is a "(vector) norm" if it has the following properties:*

(N1) Definiteness: $\|x\| \geq 0, \quad \|x\| = 0 \;\Rightarrow\; x = 0, \quad x \in \mathbb{K}^n.$

(N2) Homogeneity: $\|\alpha x\| = |\alpha|\,\|x\|, \quad \alpha \in \mathbb{K},\ x \in \mathbb{K}^n.$

(N3) Triangle inequality: $\|x + y\| \leq \|x\| + \|y\|, \quad x, y \in \mathbb{K}^n.$

The notion of a "norm" can be defined on any vector space V over $\mathbb{K}$, finite or infinite dimensional. The resulting pair $\{V, \|\cdot\|\}$ is called "normed space".

Remark 1.1: The property $\|x\| \geq 0$ is a consequence of the other conditions. With (N2), we obtain $0 = \|0\|$ and then with (N3) and (N2) $0 = \|x - x\| \leq \|x\| + \|-x\| = 2\|x\|$. With the help of (N3) we obtain the useful inequality

$$\big|\|x\| - \|y\|\big| \leq \|x - y\|, \quad x, y \in \mathbb{K}^n. \tag{1.1.1}$$

Example 1.1: The standard example of a vector norm is the "Euclidian norm"

$$\|x\|_2 := \Big(\sum_{i=1}^{n} |x_i|^2\Big)^{1/2}.$$

The first two norm properties, (N1) and (N2), are obvious, while the triangle inequality is a special case of the "Minkowski inequality" provided below in Lemma 1.4. Other examples of useful norms are the "maximum norm" (or "l_∞ norm") and the "l_1 norm"

$$\|x\|_\infty := \max_{i=1,\dots,n} |x_i|, \qquad \|x\|_1 := \sum_{i=1}^{n} |x_i|.$$

The norm properties of $\|\cdot\|_\infty$ and $\|\cdot\|_1$ are immediate consequences of the corresponding properties of the modulus function. Between l_1 norm and maximum norm there are the so-called "l_p norms" for $1 < p < \infty$:

$$\|x\|_p := \Big(\sum_{i=1}^{n} |x_i|^p\Big)^{1/p}.$$

Again the first two norm properties, (N1) and (N2), are obvious and the triangle inequality is the Minkowski inequality provided in Lemma 1.4, below.

With the aid of a norm $\|\cdot\|$ the "distance" $d(x, x') := \|x - x'\|$ of two vectors in $\mathbb{K}^n$ is defined. This allows the definition of the usual topological terms "open", "closed", "compact", "diameter", and "neighborhood" for point sets in $\mathbb{K}^n$ in analogy to the corresponding situation in $\mathbb{K}$. We use the maximum norm $\|\cdot\|_\infty$ in the following discussion, but we will see later that this is independent of the chosen norm. For any $a \in \mathbb{K}^n$ and $r > 0$, we use the ball

$$K_r(a) := \{x \in \mathbb{K}^n : \ \|x - a\|_\infty < r\}$$

as standard neighborhood of a with radius r. This neighborhood is "open" since for each point $x \in K_r(a)$ there exists a neighborhood $K_\delta(x) \subset K_r(a)$; accordingly the complement $K_r(a)^c$ is "closed". The "closure" of $K_r(a)$ is defined by $\overline{K}_r(a) := K_r(a) \cup \partial K_r(a)$ with the "boundary" $\partial K_r(a) = \{x \in \mathbb{K}^n : \ \|x - a\|_\infty = r\}$ of $K_r(a)$.

Definition 1.2: *A sequence of vectors $(x^k)_{k\in\mathbb{N}}$ in $\mathbb{K}^n$ is called*
- "bounded" if all its elements are contained in a ball $K_R(0)$, i. e., $\|x^k\|_\infty < R$, $k \in \mathbb{N}$,
- "Cauchy sequence" if for each $\varepsilon \in \mathbb{R}_+$ there is an $N_\varepsilon \in \mathbb{N}$, such that $\|x^k - x^l\|_\infty < \varepsilon$ for $k, l \geq N_\varepsilon$,
- "convergent" towards an $x \in \mathbb{K}^n$ if $\|x^k - x\|_\infty \to 0$ $(k \to \infty)$.

For a convergent sequence $(x^k)_{k\in\mathbb{N}}$, we also write $\lim_{k\to\infty} x^k = x$ or $x^k \to x$ $(k \to \infty)$. Geometrically this means that any standard neighborhood $K_\varepsilon(x)$ of x contains almost all (i. e., all but finitely many) of the elements x^k. This notion of "convergence" is obviously equivalent to the componentwise convergence:

$$\|x^k - x\|_\infty \to 0 \ \ (k \to \infty) \quad \Leftrightarrow \quad x_i^k \to x_i \ \ (k \to \infty), \ i = 1, \dots, n.$$

This allows the reduction of the convergence of sequences of vectors in $\mathbb{K}^n$ to that of sequences of numbers in $\mathbb{K}$. As basic results, we obtain n-dimensional versions of the Cauchy criterion for convergence and the theorem of Bolzano-Weierstraß.

Theorem 1.1 (Theorems of Cauchy and Bolzano-Weierstraß):

i) Each Cauchy sequence in $\mathbb{K}^n$ is convergent, i. e., the normed space $(\mathbb{K}^n, \|\cdot\|_\infty)$ is complete (a so-called "Banach space").

ii) Each bounded sequence in $\mathbb{K}^n$ contains a convergent subsequence.

Proof. i) For any Cauchy sequence $(x^k)_{k\in\mathbb{N}}$, in view of $|x_i| \le \|x\|_\infty$, $i = 1, \ldots, n$, for $x \in \mathbb{K}^n$, also the component sequences $(x_i^k)_{k\in\mathbb{N}}$, $i = 1, \ldots, n$, are Cauchy sequences in $\mathbb{K}$ and therefore converge to limits $x_i \in \mathbb{K}$. Then, the vector $x := (x_1, \ldots, x_n) \in \mathbb{K}^n$ is limit of the vector sequence $(x^k)_{k\in\mathbb{N}}$ with respect to the maximum norm.

ii) For any bounded vector sequence $(x^k)_{k\in\mathbb{N}}$ the component sequences $(x_i^k)_{k\in\mathbb{N}}$, $i = 1, \ldots, n$, are likewise bounded. By successively applying the theorem of Bolzano-Weierstraß in $\mathbb{K}$, in the first step, we obtain a convergent subsequence $(x_1^{k_{1j}})_{j\in\mathbb{N}}$ of $(x_1^k)_{k\in\mathbb{N}}$ with $x_1^{k_{1j}} \to x_1$ $(j \to \infty)$, in the next step a convergent subsequence $(x_2^{k_{2j}})_{j\in\mathbb{N}}$ of $(x_2^{k_{1j}})_{j\in\mathbb{N}}$ with $x_2^{k_{2j}} \to x_2$ $(j \to \infty)$, and so on. After n selection steps, we eventually obtain a subsequence $(x^{k_{nj}})_{j\in\mathbb{N}}$ of $(x^k)_{k\in\mathbb{N}}$, for which all component sequences $(x_i^{k_{nj}})_{j\in\mathbb{N}}$, $i = 1, \ldots, n$, converge. Then, with the limit values $x_i \in \mathbb{K}$, we set $x := (x_1, \ldots, x_n) \in \mathbb{K}^n$ and have the convergence $x^{k_{nj}} \to x$ $(j \to \infty)$. Q.E.D.

The following important result states that on the (finite dimensional) vector space $\mathbb{K}^n$ the notion of convergence, induced by any norm $\|\cdot\|$, is equivalent to the convergence with respect to the maximum norm, i. e., to the componentwise convergence.

Theorem 1.2 (Equivalence of norms): *All norms on the finite dimensional vector space $\mathbb{K}^n$ are equivalent to the maximum norm, i. e., for each norm $\|\cdot\|$ there are positive constants m, M such that*

$$m\,\|x\|_\infty \le \|x\| \le M\,\|x\|_\infty\,, \quad x \in \mathbb{K}^n\,. \tag{1.1.2}$$

Proof. Let $\|\cdot\|$ be a vector norm. For any vector $x = \sum_{i=11}^n x_i e^i \in \mathbb{K}^n$ there holds

$$\|x\| \le \sum_{k=1}^n |x_k|\,\|e^k\| \le M\,\|x\|_\infty\,, \quad M := \sum_{k=1}^n \|e^k\|\,.$$

We set

$$S_1 := \{x \in \mathbb{K}^n :\ \|x\|_\infty = 1\}, \qquad m := \inf\{\|x\|,\ x \in S_1\} \ge 0.$$

We want to show that $m > 0$ since then, in view of $\|x\|_\infty^{-1} x \in S_1$, it follows that also $m \le \|x\|_\infty^{-1}\|x\|$ for $x \ne 0$, and consequently,

$$0 < m\|x\|_\infty \le \|x\|, \quad x \in \mathbb{K}^n.$$

Suppose $m = 0$. Then, there is a sequence $(x^k)_{k\in\mathbb{N}}$ in S_1 with $\|x^k\| \to 0$ $(k \to \infty)$. Since this sequence is bounded in the maximum norm, by the theorem of Bolzano-Weierstrass it possesses a subsequence, likewise denoted by x^k, which converges in the maximum norm

to some $x \in \mathbb{K}^n$. Since

$$\big|1 - \|x\|_\infty\big| = \big|\|x^k\|_\infty - \|x\|_\infty\big| \le \|x^k - x\|_\infty \to 0 \quad (k \to \infty),$$

we have $x \in S_1$. On the other hand, for all $k \in \mathbb{N}$, there holds

$$\|x\| \le \|x - x^k\| + \|x^k\| \le M\|x - x^k\|_\infty + \|x^k\|.$$

This implies for $k \to \infty$ that $\|x\| = 0$ and therefore $x = 0$, which contradicts $x \in S_1$. Q.E.D.

Remark 1.2: i) For the two foregoing theorems, the theorem of Bolzano-Weierstrass and the theorem of norm equivalence, the *finite* dimensionality of $\mathbb{K}^n$ is decisive. Both theorems do not hold in *infinite*-dimensional normed spaces such as the space l_2 of (infinite) l_2-convergent sequences or the space $C[a,b]$ of continuous functions on $[a,b]$.

ii) A subset $M \subset \mathbb{K}^n$ is called "compact" (or more precisely "sequentially compact"), if each sequence of vectors in M possesses a convergent subsequence with limit in M. Then, the theorem of Bolzano-Weierstrass implies that the compact subsets in $\mathbb{K}^n$ are exactly the bounded and closed subsets in $\mathbb{K}^n$.

iii) A point $x \in \mathbb{K}^n$ is called "accumulation point" of a set $M \subset \mathbb{K}^n$ if each neighborhood of x contains at least one point from $M \setminus \{x\}$. The set of accumulation points of M is denoted by $\mathcal{H}(M)$ (closed "hull" of M). A point $x \in M \setminus \mathcal{H}(M)$ is called "isolated".

Remark 1.3: In many applications there occur pairs $\{x, y\}$ (or more generally tuples) of points $x, y \in \mathbb{K}^n$. These form the so-called "product space" $V = \mathbb{K}^n \times \mathbb{K}^n$, which may be equipped with the generic norm $\|\{x,y\}\| := \big(\|x\|^2 + \|y\|^2\big)^{1/2}$. Since this space may be identified with the $2n$-dimensional Euclidian space $\mathbb{K}^{2n}$ all results on subsets of $\mathbb{K}^n$ carry over to subsets of $\mathbb{K}^n \times \mathbb{K}^n$. This can be extended to more general product spaces of the form $V = \mathbb{K}^{n_1} \times \cdots \times \mathbb{K}^{n_m}$.

The basic concept in the geometry of $\mathbb{K}^n$ is that of "orthogonality" of vectors or subspaces. For its definition, we use a "scalar product".

Definition 1.3: *A mapping $(\cdot,\cdot) : \mathbb{K}^n \times \mathbb{K}^n \to \mathbb{K}$ is called "scalar product" if it has the following properties:*

(S1) Conjugate Symmetry: $(x,y) = \overline{(y,x)}, \quad x, y \in \mathbb{K}^n.$

(S2) Linearity: $(\alpha x + \beta y, z) = \alpha(x,z) + \beta(y,z), \quad x, y, z \in \mathbb{K}^n,\ \alpha, \beta \in \mathbb{K}.$

(S3) Definiteness: $(x,x) \in \mathbb{R},\ (x,x) > 0, \ \ x \in \mathbb{K}^n \setminus \{0\}.$

In the following, we will mostly use the "euclidian" scalar product

$$(x,y)_2 = \sum_{j=1}^{n} x_j \overline{y_j}, \quad (x,x)_2 = \|x\|_2^2.$$

Remark 1.4: i) If the *strict* definiteness (S3) is relaxed, $(x,x) \in \mathbb{R},\ (x,x) \geq 0$, the sesquilinear form becomes a so-called "semi-scalar product".

ii) From property (S2) (linearity in the first argument) and (S1) (conjugate symmetry), we obtain the conjugate linearity in the second argument. Hence, a scalar product is a special kind of "sesquilinear form" (if $\mathbb{K} = \mathbb{C}$) or "bilinear form" (if $\mathbb{K} = \mathbb{R}$)

Lemma 1.1: *For a scalar product on $\mathbb{K}^n$ there holds the "Cauchy-Schwarz inequality"*

$$|(x,y)|^2 \leq (x,x)(y,y), \quad x,y \in \mathbb{K}^n. \tag{1.1.3}$$

Proof. The assertion is obviously true for $y = 0$. Hence, we can now assume that $y \neq 0$. For arbitrary $\alpha \in \mathbb{K}$ there holds

$$0 \leq (x + \alpha y, x + \alpha y) = (x,x) + \alpha(y,x) + \overline{\alpha}(x,y) + \alpha\overline{\alpha}(y,y).$$

With $\alpha := -(x,y)(y,y)^{-1}$ this implies

$$\begin{aligned} 0 &\leq (x,x) - (x,y)(y,y)^{-1}(y,x) - \overline{(x,y)}(y,y)^{-1}(x,y) + (x,y)\overline{(x,y)}(y,y)^{-1} \\ &= (x,x) - |(x,y)|^2(y,y)^{-1} \end{aligned}$$

and, consequently, $0 \leq (x,x)(y,y) - |(x,y)|^2$. This is the asserted inequality. Q.E.D.

The Cauchy-Schwarz inequality in $\mathbb{K}^n$ is a special case of the "Hölder[1] inequality".

Corollary 1.1: *Any scalar product $(\cdot,\cdot)$ on $\mathbb{K}^n$ generates a norm $\|\cdot\|$ on $\mathbb{K}^n$ by*

$$\|x\| := (x,x)^{1/2}, \quad x \in \mathbb{K}^n.$$

The "Euclidian" scalar product $(\cdot,\cdot)_2$ corresponds to the "Euclidian" norm $\|x\|_2$.

Proof. The norm properties (N1) and (N2) are obvious. It remains to show (N3). Using the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} \|x+y\|^2 &= (x+y, x+y) = (x,x) + (x,y) + (y,x) + (y,y) \\ &\leq \|x\|^2 + 2|(x,y)| + \|y\|^2 \leq \|x\|^2 + 2\|x\|\|y\| + \|y\|^2 = (\|x\| + ||y||)^2, \end{aligned}$$

what was to be shown. Q.E.D.

Next, we provide a useful inequality, which is a special case of so-called "Young[2] inequalities".

[1]Ludwig Otto Hölder (1859–1937): German mathematician; Prof. in Tübingen; contributions first to the theory of Fourier series and later to group theory; found 1884 the inequality named after him.

[2]William Henry Young (1863–1942): English mathematician; worked at several universities worldwide, e. g., in Calcutta, Liverpool and Wales; contributions to differential and integral calculus, topological set theory and geometry.

Lemma 1.2 (Young inequality): *For $p, q \in \mathbb{R}$ with $1 < p, q < \infty$ and $1/p + 1/q = 1$, there holds the inequality*

$$|xy| \leq \frac{|x|^p}{p} + \frac{|y|^q}{q}, \quad x, y \in \mathbb{K}. \tag{1.1.4}$$

Proof. The logarithm $\ln(x)$ is on $\mathbb{R}_+$, in view of $\ln''(x) = -1/x^2 < 0$, a concave function. Hence, for $x, y \in \mathbb{K}$ there holds:

$$\ln\left(\tfrac{1}{p}|x|^p + \tfrac{1}{q}|y|^q\right) \geq \tfrac{1}{p}\ln(|x|^p) + \tfrac{1}{q}\ln(|y|^q) = \ln(|x|) + \ln(|y|).$$

Because of the monotonicity of the exponential function e^x it further follows that for $x, y \in \mathbb{K}$:

$$\tfrac{1}{p}|x|^p + \tfrac{1}{q}|y|^q \geq \exp\big(\ln(|x|) + \ln(|y|)\big) = \exp\big(\ln(|x|)\big)\exp\big(\ln(|y|)\big) = |x||y| = |xy|,$$

what was to be proven. Q.E.D.

Lemma 1.3 (Hölder inequality): *For the Euclidian scalar product there holds, for arbitrary $p, q \in \mathbb{R}$ with $1 < p, q < \infty$ and $1/p + 1/q = 1$, the so-called "Hölder inequality"*

$$|(x, y)_2| \leq \|x\|_p \, \|y\|_q, \quad x, y \in \mathbb{K}^n. \tag{1.1.5}$$

This inequality also holds for the limit case $p = 1$, $q = \infty$.

Proof. For $x = 0$ or $y = 0$ the asserted estimate is obviously true. Hence, we can assume that $\|x\|_p \neq 0$ and $\|y\|_q \neq 0$. First, there holds

$$\frac{|(x, y)_2|}{\|x\|_p \|y\|_q} = \frac{1}{\|x\|_p \|y\|_q}\Big|\sum_{i=1}^n x_i \overline{y}_i\Big| \leq \sum_{i=1}^n \frac{|x_i||y_i|}{\|x\|_p \|y\|_q}.$$

Using the Young inequality it follows that

$$\frac{|(x, y)_2|}{\|x\|_p \|y\|_q} \leq \sum_{i=1}^n \Big\{\frac{|x_i|^p}{p\|x\|_p^p} + \frac{|y_i|^q}{q\|y\|_q^q}\Big\} = \frac{1}{p\|x\|_p^p}\sum_{i=1}^n |x_i|^p + \frac{1}{q\|y\|_q^q}\sum_{i=1}^n |y_i|^q = \frac{1}{p} + \frac{1}{q} = 1.$$

This implies the asserted inequality. Q.E.D.

As consequence of the Hölder inequality, we obtain the so-called "Minkowski[3] inequality", which is the triangle inequality for the l_p norm.

[3]Hermann Minkowski (1864–1909): Russian-German mathematician; Prof. in Göttingen; several contributions to pure mathematics; introduced the non-euclidian 4-dimensional space-time continuum ("Minkowski space") for describing the theory of relativity of Einstein.

Lemma 1.4 (Minkowski inequality): *For arbitrary* $p \in \mathbb{R}$ *with* $1 \le p < \infty$ *as well as for* $p = \infty$ *there holds the "Minkowski inequality"*

$$\|x + y\|_p \le \|x\|_p + \|y\|_p, \quad x, y \in \mathbb{K}^n. \tag{1.1.6}$$

Proof. For $p = 1$ and $p = \infty$ the inequality follows from the triangle inequality on $\mathbb{R}$:

$$\|x + y\|_1 = \sum_{i=1}^{n} |x_i + y_i| \le \sum_{i=1}^{n} |x_i| + \sum_{i=1}^{n} |y_i| = \|x\|_1 + \|y\|_1,$$
$$\|x + y\|_\infty = \max_{1\le i\le n} |x_i + y_i| \le \max_{1\le i\le n} |x_i| + \max_{1\le i\le n} |y_i| = \|x\|_\infty + \|y\|_\infty.$$

Let now $1 < p < \infty$ and q be defined by $1/p + 1/q = 1$, i. e., $q = p/(p-1)$. We set

$$\xi_i := |x_i + y_i|^{p-1}, \quad i = 1, \dots, n, \quad \xi := (\xi_i)_{i=1}^n.$$

This implies that

$$\|x + y\|_p^p = \sum_{i=1}^{n} |x_i + y_i||x_i + y_i|^{p-1} \le \sum_{i=1}^{n} |x_i|\xi_i + \sum_{i=1}^{n} |y_i|\xi_i$$

and further by the Hölder inequality

$$\|x + y\|_p^p \le \|x\|_p\|\xi\|_q + \|y\|_p\|\xi\|_q = (\|x\|_p + \|y\|_p)\|\xi\|_q.$$

Observing $q = p/(p-1)$, we conclude

$$\|\xi\|_q^q = \sum_{i=1}^{n} |\xi_i|^q = \sum_{i=1}^{n} |x_i + y_i|^p = \|x + y\|_p^p,$$

and consequently,

$$\|x + y\|_p^p \le (\|x\|_p + \|y\|_p)\|x + y\|_p^{p/q} = (\|x\|_p + \|y\|_p)\|x + y\|_p^{p-1}.$$

This implies the asserted inequality. Q.E.D.

Using the Euclidian scalar product, we can introduce a canonical notion of "orthogonality", i. e., two vectors $x, y \in \mathbb{K}^n$ are called "orthogonal" (in symbols $x \perp y$) if

$$(x, y)_2 = 0.$$

Two subspaces $N, M \subset \mathbb{K}^n$ are called "orthogonal" (in symbols $N \perp M$) if

$$(x, y)_2 = 0, \ x \in N, \ y \in M.$$

Accordingly to each subspace $M \in \mathbb{K}^n$, we can assign its "orthogonal complement" $M^\perp := \{x \in \mathbb{K}^n, \ \mathrm{span}(x) \perp M\}$, which is uniquely determined. Then, $\mathbb{K}^n = M \oplus M^\perp$, the "direct sum" of M and $M^\perp$. Let $M \subset \mathbb{K}^n$ be a (nontrivial) subspace. Then, for any vector $x \in \mathbb{K}^n$ the "orthogonal projection" $P_M x \in M$ is determined by the relation

$$\|x - P_M x\|_2 = \min_{y \in M} \|x - y\|. \tag{1.1.7}$$

This "best approximation" property is equivalent to the relation

$$(x - P_M x, y)_2 = 0 \quad \forall y \in M, \tag{1.1.8}$$

which can be used to actually compute $P_M x$.

For arbitrary vectors there holds the "parallelogram identity" (exercise)

$$\|x + y\|_2^2 + \|x - y\|_2^2 = 2\|x\|_2^2 + 2\|y\|_2^2, \quad x, y \in \mathbb{K}^n, \tag{1.1.9}$$

and for orthogonal vectors the "Theorem of Pythagoras" (exercise):

$$\|x + y\|_2^2 = \|x\|_2^2 + \|y\|_2^2, \quad x, y \in \mathbb{K}^n,\ x \perp y. \tag{1.1.10}$$

A set of vectors $\{a^1, \ldots, a^m\}$, $a^i \neq 0$, of $\mathbb{K}^n$, which are mutually orthogonal, $(a^k, a^l) = 0$, for $k \neq l$, is necessarily linearly independent. Because for $\sum_{k=1}^m c_k a^k = 0$, successively taking the scalar product with a^l, $l = 1, \ldots, m$, yields

$$0 = \sum_{k=1}^m c_k (a^k, a^l)_2 = c_l (a^l, a^l)_2 \quad \Rightarrow \quad c_l = 0.$$

Definition 1.4: *A set of vectors $\{a^1, \ldots, a^m\}$, $a^i \neq 0$ of $\mathbb{K}^n$, which are mutually orthogonal, $(a^k, a^l)_2 = 0$, $k \neq l$, is called "orthogonal system" (in short "ONS") and in the case $m = n$ "orthogonal basis" (in short "ONB"). If $(a^k, a^k) = 1$, $k = 1, \ldots, m$, one speaks of an "orthonormal system" and an "orthonormal basis", respectively. The cartesian basis $\{e^1, \ldots, e^n\}$ is obviously an orthonormal basis of $\mathbb{R}^n$ with respect to the Euclidian scaler product. However, there are many other (actually infinitely many) of such orthonormal bases in $\mathbb{R}^n$.*

Lemma 1.5: *Let $\{a^i,\ i = 1, \ldots, n\}$ be an orthonormal basis of $\mathbb{K}^n$ (with respect to the canonical Euclidian scalar product). Then, each vector $x \in \mathbb{K}^n$ possesses a representation of the form (in analogy to the "Fourier expansion" with trigonometric functions)*

$$x = \sum_{i=1}^n (x, a^i)_2\, a^i, \tag{1.1.11}$$

and there holds the "Parseval[4] identity"

$$\|x\|_2^2 = \sum_{i=1}^n |(x, a^i)_2|^2, \quad x \in \mathbb{K}^n. \tag{1.1.12}$$

[4] *Marc-Antoine Parseval des Chênes (1755–1836): French mathematician; worked on partial differential equations in physics (only five mathematical publications); known by the identity named after him, which he stated without proof and connection to Fourier series.*

Proof. From the representation $x = \sum_{j=1}^n \alpha_j a^j$ taking the product with a^i it follows that

$$(x, a^i)_2 = \sum_{j=1}^n \alpha_j (a^j, a^i)_2 = \alpha_i, \quad i = 1, \ldots, n,$$

and consequently the representation (1.1.11). Further there holds:

$$\|x\|_2^2 = (x,x)_2 = \sum_{i,j=1}^n (x, a^i)_2 \overline{(x, a^j)_2}\, (a^i, a^j)_2 = \sum_{i=1}^n |(x, a^i)_2|^2,$$

what was to be proven. Q.E.D.

By the following Gram[5]-Schmidt[6] algorithm, we can orthonormalize an arbitrary basis of $\mathbb{K}^n$, i. e., construct an *orthonormal* basis.

Theorem 1.3 (Gram-Schmidt algorithm): *Let $\{a^1, \ldots, a^n\}$ be any basis of $\mathbb{K}^n$. Then, the following so-called "Gram-Schmidt orthonormalization algorithm",*

$$\begin{aligned} b^1 &:= \|a^1\|_2^{-1} a^1, \\ \tilde{b}^k &:= a^k - \sum_{j=1}^{k-1} (a^k, b^j)_2 b^j, \quad b^k := \|\tilde{b}^k\|_2^{-1} \tilde{b}^k, \quad k = 2, \ldots, n, \end{aligned} \tag{1.1.13}$$

yields an orthonormal basis $\{b^1, \ldots, b^n\}$ of $\mathbb{K}^n$.

Proof. First, we show that the construction process of the b^k does not stop with $k < n$. The vectors b^k are linear combinations of the $a^1, \ldots, a^k$. If for some $k \le n$

$$a^k - \sum_{j=1}^{k-1} (a^k, b^j)_2 b^j = 0,$$

the vectors $\{a^1, \ldots, a^k\}$ would be linearly dependent contradicting the a priori assumption that $\{a^1, \ldots, a^n\}$ is a basis. Now, we show by induction that the Gram-Schmidt process yields an orthonormal basis. Obviously $\|b^1\|_2 = 1$. Let now $\{b^1, \ldots, b^k\}$, for $k \le n$, be an already constructed orthonormal system. Then, for $l = 1, \ldots, k$, there holds

$$(b^{k+1}, b^l)_2 = (a^{k+1}, b^l)_2 - \sum_{j=1}^k (a^{k+1}, b^j)_2 \underbrace{(b^j, b^l)_2}_{= \delta_{jl}} = 0$$

and $\|b^{k+1}\|_2 = 1$, i. e., $\{b^1, \ldots, b^{k+1}\}$ is also an orthonormal system. Q.E.D.

[5] Jørgen Pedersen Gram (1850–1916): Danish mathematician, employee and later owner of an insurance company, contributions to algebra (invariants theory), probability theory, numerics and forestry; the orthonormalization algorithm named after him had already been used before by Cauchy 1836.

[6] Erhard Schmidt (1876–1959): German mathematician, Prof. in Berlin, there founder of the Institute for Applied Mathematics 1920, after the war Director of the Mathematical Institute of the Academy of Sciences of DDR; contributions to the theory of integral equations and Hilbert spaces and later to general topology.

The Gram-Schmidt algorithm in its "classical" form (1.1.13) is numerically unstable due to accumulation of round-off errors. Below, in Section 4.3.1, we will consider a stable version, the so-called "modified Gram-Schmidt algorithm", which for *exact* arithmetic yields the same result.

1.1.2 Linear mappings and matrices

We now consider linear mappings from the n-dimensional vector space $\mathbb{K}^n$ into the m-dimensional vector space $\mathbb{K}^m$, where not necessarily $m = n$. However, the special case $m = n$ plays the most important role. A mapping $\varphi = (\varphi_1, \ldots, \varphi_m) : \mathbb{K}^n \to \mathbb{K}^m$ is called "linear", if for $x, y \in \mathbb{K}^n$ and $\alpha, \beta \in \mathbb{K}$ there holds

$$\varphi(\alpha x + \beta y) = \alpha\varphi(x) + \beta\varphi(y). \tag{1.1.14}$$

The action of a linear mapping φ on a vector space can be described in several ways. It obviously suffices to prescribe the action of φ on the elements of a basis of the space, e. g., a Cartesian basis $\{e^i, i = 1, \ldots, n\}$,

$$x = \sum_{i=1}^{n} x_i e^i \quad \to \quad \varphi(x) = \varphi\Big(\sum_{i=1}^{n} x_i e^i\Big) = \sum_{i=1}^{n} x_i \varphi(e^i).$$

Thereby, to each vector (or point) $x \in \mathbb{K}^n$ a "coordinate vector" $\hat{x} = (x_i)_{i=1}^n$ is uniquely associated. If the images $\varphi(x)$ are expressed with respect to a Cartesian basis of $\mathbb{K}^m$,

$$\varphi(x) = \sum_{j=1}^{m} \varphi_j(x) e^j = \sum_{j=1}^{m} \Big(\sum_{i=1}^{n} \underbrace{\varphi_j(e^i)}_{=:\, a_{ji}} x_i \Big) e^j,$$

with the coordinate vector $\hat{\varphi}(x) = (\varphi_j(x))_{j=1}^m$, we can write the action of the mapping φ on a vector $x \in \mathbb{K}^n$ in "matrix form" using the usual rules of matrix-vector multiplication as follows:

$$\varphi_j(x) = (A\hat{x})_j := \sum_{i=1}^{n} a_{ji} x_i, \quad j = 1, \ldots, m,$$

with the $n \times m$-array of numbers $A = (a_{ij})_{i,j=1}^{n,m} \in \mathbb{K}^{m\times n}$, a "matrix",

$$\begin{pmatrix} \varphi_1(e^1) & \cdots & \varphi_1(e^n) \\ \vdots & \ddots & \vdots \\ \varphi_m(e^1) & \cdots & \varphi_m(e^n) \end{pmatrix} =: \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} = A \in \mathbb{K}^{m\times n}.$$

By this matrix $A \in \mathbb{K}^{m\times n}$ the linear mapping φ is uniquely described with respect to the chosen bases of $\mathbb{K}^n$ and $\mathbb{K}^m$. In the following discussion, for simplicity, we identify the point $x \in \mathbb{K}^n$ with its special cartesian coordinate representation $\hat{x}$. Here, we follow the convention that in the notation $\mathbb{K}^{m\times n}$ for matrices the first parameter m stands for the dimension of the target space $\mathbb{K}^m$, i. e., the number of rows in the matrix, while the

second one n corresponds to the dimension of the initial space $\mathbb{K}^n$, i. e., the number of columns. Accordingly, for a matrix entry a_{ij} the first index refers to the row number and the second one to the column number of its position in the matrix. We emphasize that this is only one of the possible concrete representations of the linear mapping $\varphi : \mathbb{K}^n \to \mathbb{K}^m$. In this sense each quadratic matrix $A \in \mathbb{K}^{n\times n}$ represents a linear mapping in $\mathbb{K}^n$. The identity map $\varphi(x) = x$ is represented by the "identity matrix" $I = (\delta_{ij})_{i,j=1}^n$ where $\delta_{ij} := 1$ for $i = j$ and $\delta_{ij} = 0$ else (the usual "Kronecker symbol").

Clearly, two matrices $A, A' \in \mathbb{K}^{m\times n}$ are identical, i. e., $a_{ij} = a'_{ij}$ if and only if $Ax = A'x$, $x \in \mathbb{K}^n$. To a general matrix $A \in \mathbb{K}^{m\times n}$, we associate the "adjoint transpose" $\bar{A}^T = (a^T_{i,j})_{i,j=1}^{n\times m}$ by setting $a^T_{ij} := \bar{a}_{ji}$. A quadratic matrix $A \in \mathbb{K}^{n\times n}$ is called "regular", if the corresponding linear mapping is injective and surjective, i. e., bijective, with "inverse" denoted by $A^{-1} \in \mathbb{K}^{n\times n}$. Further, to each matrix $A \in \mathbb{K}^{n\times n}$, we associate the following quantities, which are uniquely determined by the corresponding linear mapping φ:

- "determinant" of A: $\det(A)$.
- "adjugate" of A: $adj(A) := C^T$, $c_{ij} := (-1)^{i+j} A_{ij}$ (A_{ij} the cofactors of A).
- "trace" of A: $\mathrm{trace}(A) := \sum_{i=1}^n a_{ii}$.

The following property of the determinant will be useful below: $\det(\bar{A}^T) = \overline{\det(A)}$.

Lemma 1.6: *For a square matrix $A = (a_{ij})_{i,j=1}^n \in \mathbb{K}^{n\times n}$ the following statements are equivalent:*

i) A is regular with inverse A^{-1}.

ii) The equation $Ax = 0$ has only the zero solution $x = 0$ (injectivity).

iii) The equation $Ax = b$ has for any $b \in \mathbb{K}^n$ a solution (surjectivity).

iv) $\det(A) \neq 0$.

v) The adjoint transpose $\bar{A}^T$ is regular with inverse $(\bar{A}^T)^{-1} = (\overline{A^{-1}})^T$.

Proof. For the proof, we refer to the standard linear algebra literature. Q.E.D.

Lemma 1.7: *For a general matrix $A \in \mathbb{K}^{m\times n}$, we introduce its "range" and its "kernel" (or "null space")*

$$\mathrm{range}(A) := \{y \in \mathbb{K}^m \,|\, y = Ax \text{ for some } x \in \mathbb{K}^n\},$$
$$\mathrm{kern}(A) := \{x \in \mathbb{K}^n \,|\, Ax = 0\}.$$

There holds

$$\mathrm{range}(A) = \mathrm{kern}(\bar{A}^T)^T, \qquad \mathrm{range}(\bar{A}^T) = \mathrm{kern}(A)^T, \tag{1.1.15}$$

i. e., the equation $Ax = b$ has a solution if and only if $(b, y)_2 = 0$ for all $y \in \mathrm{kern}(\bar{A}^T)$.

Proof. For the proof, we refer to the standard linear algebra literature. Q.E.D.

In many practical applications the governing matrices have special properties, which require the use of likewise special numerical methods. Some of the most important properties are those of "symmetry" or "normality" and "definiteness".

Definition 1.5: *i) A quadratic matrix* $A \in \mathbb{K}^{n\times n}$ *is called "Hermitian" if it satisfies*

$$A = \bar{A}^T \quad (\Leftrightarrow \quad a_{ij} = \bar{a}_{ji},\ i,j = 1,\ldots,n), \tag{1.1.16}$$

or equivalently,

$$(Ax, y)_2 = (x, Ay)_2, \quad x, y \in \mathbb{K}^n. \tag{1.1.17}$$

ii) It is called "normal" if $\bar{A}^T A = A\bar{A}^T$.

iii) It is called "positive semi-definite" if

$$(Ax, x)_2 \in \mathbb{R}, \quad (Ax, x)_2 \geq 0, \ \ x \in \mathbb{K}^n. \tag{1.1.18}$$

and "positive definite" if

$$(Ax, x)_2 \in \mathbb{R}, \quad (Ax, x)_2 > 0, \ \ x \in \mathbb{K}^n \setminus \{0\}. \tag{1.1.19}$$

iv) A real Hermitian matrix $A \in \mathbb{R}^{n\times n}$ *is called "symmetric" .*

Lemma 1.8: *For a Hermitian positive definite matrix* $A \in \mathbb{K}^{n\times n}$ *the main diagonal elements are real and positive,* $a_{ii} > 0$ *, and the element with largest modulus lies on the main diagonal.*

Proof. i) From $a_{ii} = \bar{a}_{ii}$ it follows that $a_{ii} \in \mathbb{R}$. The positiveness follows via testing by the Cartesian unit vector e^i yielding $a_{ii} = (Ae^i, e^i)_2 > 0$.

ii) Let $a_{ij} \neq 0$ be an element of A with maximal modulus and suppose that $i \neq j$. Testing now by $x = e^i - \operatorname{sign}(a_{ij})e^j \neq 0$, we obtain the following contradiction to the definiteness of A:

$$\begin{aligned} 0 < (Ax, x)_2 &= (Ae^i, e^i)_2 - 2\operatorname{sign}(a_{ij})(Ae^i, e^j)_2 + \operatorname{sign}(a_{ij})^2(Ae^j, e^j)_2 \\ &= a_{ii} - 2\operatorname{sign}(a_{ij})a_{ij} + a_{jj} = a_{ii} - 2|a_{ij}| + a_{jj} \leq 0. \end{aligned}$$

This completes the proof. Q.E.D.

Remark 1.5 (Exercises): i) If a matrix $A \in \mathbb{K}^{n\times n}$ is positive definite (or more generally just satisfies $(Ax, x)_2 \in \mathbb{R}$ for $x \in \mathbb{C}^n$), then it is necessarily Hermitian. This does not need to be true for *real* matrices $A \in \mathbb{R}^{n\times n}$.

ii) The general form of a scalar product $(\cdot,\cdot)$ on $\mathbb{K}^n$ is given by $(x, y) = (Ax, y)_2$ with a (Hermitian) positive definite matrix $A \in \mathbb{K}^{n\times n}$.

Definition 1.6 (Orthonormal matrix): *A matrix $Q \in \mathbb{K}^{m\times n}$ is called "orthogonal" or "orthonormal" if its column vectors form an orthogonal or orthonormal system in $\mathbb{K}^n$, respectively. In the case $n = m$ such a matrix is called "unitary".*

Lemma 1.9: *A unitary matrix $Q \in \mathbb{K}^{n\times n}$ is regular and its inverse is $Q^{-1} = \bar{Q}^T$. Further, there holds:*

$$(Qx, Qy)_2 = (x, y)_2, \quad x, y \in \mathbb{K}^n, \tag{1.1.20}$$

$$\|Qx\|_2 = \|x\|_2, \quad x \in \mathbb{K}^n. \tag{1.1.21}$$

Proof. First, we show that $\overline{Q}^T$ is the inverse of Q. Let $q_i \in \mathbb{K}^n$ denote the column vectors of Q satisfying by definition $(q_i, q_j)_2 = q_i^T \overline{q}_j = \delta_{ij}$. This implies:

$$\overline{Q}^T Q = \begin{pmatrix} \overline{q}_1^T q_1 & \dots & \overline{q}_1^T q_n \\ \vdots & \ddots & \vdots \\ \overline{q}_n^T q_1 & \dots & \overline{q}_n^T q_n \end{pmatrix} = \begin{pmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{pmatrix} = I.$$

From this it follows that

$$(Qx, Qy)_2 = (x, \bar{Q}^T Q x)_2 = (x, y)_2, \quad x, y \in \mathbb{K}^n,$$

and further

$$\|Qx\|_2 = (Qx, Qx)_2^{1/2} = \|x\|_2, \quad x \in \mathbb{K}^n,$$

which completes the proof. Q.E.D.

Example 1.2: The real unitary matrix

$$Q_\theta^{(ij)} = \begin{array}{c} \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccc} & & i & & j \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & \cos(\theta) & 0 & -\sin(\theta) & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & \sin(\theta) & 0 & \cos(\theta) & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array} \begin{array}{l} \\ \\ i \\ \\ j \end{array}$$

describes a rotation in the (x_i, x_j)-plane about the origin $x = 0$ with angle $\theta \in [0, 2\pi)$.

Remark 1.6: i) In view of the relations (1.1.20) and (1.1.21) Euclidian scalar product and norm of vectors are invariant under unitary transformations. This explains why it is the Euclidian norm, which is used for measuring length or distance of vectors in $\mathbb{R}^n$.

ii) The Schwarz inequality (1.1.3) allows the definition of an "angle" between two vectors

in $\mathbb{R}^n$. For any number $\alpha \in [-1,1]$ there is exactly one $\theta \in [0,\pi]$ such that $\alpha = \cos(\theta)$. By

$$\cos(\theta) = \frac{(x,y)_2}{\|x\|_2\|y\|_2}, \quad x,y \in \mathbb{K}^n \setminus \{0\},$$

a $\theta \in [0,\pi]$ is uniquely determined. This is then the "angle" between the two vectors x and y. The relation (1.1.20) states that the Euclidian scalar product of two vectors in $\mathbb{K}^n$ is invariant under rotations. By some rotation Q in $\mathbb{R}^n$, we can achieve that $Qx, Qy \in \operatorname{span}\{e^{(1)}, e^{(2)}\}$ and $Qx = \|x\|_2 e^{(1)}$. Then, there holds

$$(x,y)_2 = (Qx,Qy)_2 = \|x\|_2(e^{(1)},Qy)_2 = \|x\|_2(Qy)_1 = \|x\|_2\|Qy\|_2\cos(\theta) = \|x\|_2\|y\|_2\cos(\theta),$$

i. e., θ is actually the "angle" between the two vectors in the sense of elementary geometry.

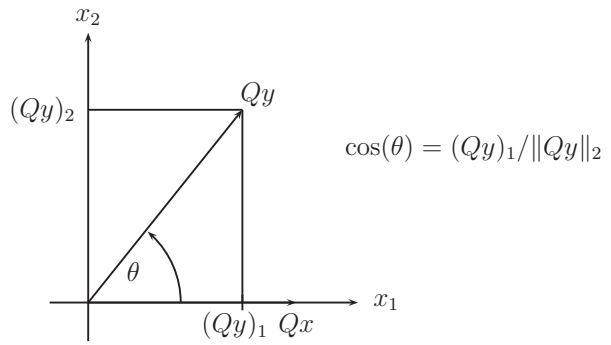

Figure 1.1: *Angle between two vectors* $x = \|x\|e^1$ *and* y *in* $\mathbb{R}^2$.

1.1.3 Non-quadratic linear systems

Let $A \in \mathbb{R}^{m\times n}$ be a not necessarily quadratic coefficient matrix and $b \in \mathbb{R}^m$ a given vector. We concentrate in the case $m \neq n$ and consider the non-quadratic linear system

$$Ax = b, \tag{1.1.22}$$

for $x \in \mathbb{R}^n$. Here, $\operatorname{rank}(A) < \operatorname{rank}[A,b]$ is allowed, i. e., the system does not need to possess a solution in the normal sense. In this case an appropriately extended notion of "solution" is to be used. In the following, we consider the so-called "method of least error-squares", which goes back to Gauss. In this approach a vector $\bar{x} \in \mathbb{R}^n$ is seeked with minimal defect norm $\|d\|_2 = \|b - A\bar{x}\|_2$. Clearly, this extended notion of "solution" coincides with the traditional one if $\operatorname{rank}(A) = \operatorname{rank}([A,b])$.

Theorem 1.4 ("Least error-squares" solution): *There exists always a "solution"* $\bar{x} \in \mathbb{R}^n$ *of (1.1.22) in the sense of least error-squares ("least error-squares" solution)*

$$\|A\bar{x} - b\|_2 = \min_{x \in \mathbb{R}^n} \|Ax - b\|_2. \tag{1.1.23}$$

This is equivalent to $\bar{x}$ *being solution of the so-called "normal equation"*

$$A^T A\bar{x} = A^T b. \tag{1.1.24}$$

If $m \geq n$ *and* $\operatorname{rank}(A) = n$ *the "least error-squares" solution* $\bar{x}$ *is uniquely determined. Otherwise each other solution has the form* $\bar{x} + y$ *with* $y \in \operatorname{kern}(A)$*. In this case, there always exists such a solution with minimal Euclidian norm, i. e., a "minimal" solution with least error-squares,*

$$\|x^{\min}\|_2 = \min\{\|\bar{x} + y\|_2,\ y \in \operatorname{kern}(A)\}. \tag{1.1.25}$$

Proof. i) Let $\bar{x}$ be a solution of the normal equation. Then, for arbitrary $x \in \mathbb{R}^n$ there holds

$$\begin{aligned} \|b - Ax\|_2^2 &= \|b - A\bar{x} + A(\bar{x} - x)\|_2^2 \\ &= \|b - A\bar{x}\|_2^2 + 2\underbrace{(b - A\bar{x}}_{\in \operatorname{kern}(A^T)}, \underbrace{A[\bar{x} - x])}_{\in \operatorname{range}(A)} + \|A(\bar{x} - x)\|_2^2 \geq \|b - A\bar{x}\|_2^2, \end{aligned}$$

i. e., $\bar{x}$ has least error-squares. In turn, for such a least error-squares solution $\bar{x}$ there holds

$$\begin{aligned} 0 = \frac{\partial}{\partial x_i} \|Ax - b\|_{2|x=\bar{x}}^2 &= \frac{\partial}{\partial x_i} \Big(\sum_{j=1}^n \Big| \sum_{k=1}^n a_{jk} x_k - b_j \Big|^2 \Big)_{|x=\bar{x}} \\ = 2 \sum_{j=1}^n a_{ji} \Big(\sum_{k=1}^n a_{jk} \bar{x}_k - b_j \Big) &= 2(A^T A\bar{x} - A^T b)_i, \end{aligned}$$

i. e., $\bar{x}$ solves the normal equation.

ii) We now consider the solvability of the normal equation. The orthogonal complement of $\operatorname{range}(A)$ in $\mathbb{R}^m$ is $\operatorname{kern}(A^T)$. Hence the element b has a unique decomposition

$$b = s + r\,, \quad s \in \operatorname{range}(A)\,, \quad r \in \operatorname{kern}(A^T).$$

Then, for any $\bar{x} \in \mathbb{R}^n$ satisfying $A\bar{x} = s$ there holds

$$A^T A\bar{x} = A^T s \ = \ A^T s + A^T r = A^T b,$$

i. e., $\bar{x}$ solves the normal equation. In case that $\operatorname{range}(A) = n$ there holds $\operatorname{kern}(A) = \{0\}$ and $\operatorname{range}(A) = \mathbb{R}^n$. Observing $A^T Ax = 0$ and $\operatorname{kern}(A^T) \perp \operatorname{range}(A)$, we conclude $Ax = 0$ and $x = 0$. The matrix $A^T A \in \mathbb{R}^{n \times n}$ is regular and consequently $\bar{x}$ uniquely determined. In case that $\operatorname{range}(A) < n$, for any other solution x_1 of the normal equation, we have

$$b = Ax_1 + (b - Ax_1) \ \in \ \operatorname{range}(A) + \operatorname{kern}(A^T) = \operatorname{range}(A) + \operatorname{range}(A)^T.$$

In view of the uniqueness of the orthogonal decomposition, we necessarily obtain $Ax_1 = A\bar{x}$ and $\bar{x} - x_1 \in \mathrm{kern}(A)$.

iii) We finally consider the case $\mathrm{rank}(A) < n$. Among the solutions $\bar{x} + \mathrm{kern}(A)$ of the normal equation, we can find one with minimal euclidian norm,

$$\|x^{\min}\|_2 = \min\{\|\bar{x} + y\|_2,\ y \in \mathrm{kern}(A)\}.$$

This follows from the non-negativity of the function $F(y) := \|\bar{x} + y\|_2$ and its uniform strict convexity, which also implies uniqueness of the minimal solution. Q.E.D.

For the computation of the "solution with smallest error-squares" of a non-quadratic system $Ax = b$, we have to solve the normal equation $A^T Ax = A^T b$. Efficient methods for this task will be discussed in the next chapter.

Lemma 1.10: *For any matrix $A \in \mathbb{K}^{m\times n}$ the matrices $\bar{A}^T A \in \mathbb{K}^{n\times n}$ and $A\bar{A}^T \in \mathbb{K}^{m\times m}$ are Hermitian (symmetric) and positive semi-definite. In the case $m \geq n$ and if $\mathrm{rank}(A) = n$ the matrix $\bar{A}^T A$ it is even positive definite.*

Proof. Following the rules of matrix arithmetic there holds

$$(\bar{A}^T A)^T = A^T \bar{A} = \overline{\bar{A}^T A}, \qquad \bar{x}^T (\bar{A}^T A)x = \overline{(Ax)}^T Ax = \|Ax\|_2^2 \geq 0,$$

i. e., $\bar{A}^T A$ is Hermitian and positive semi-definite. The argument for $A\bar{A}^T$ is analogous. In case that $m \geq n$ and $\mathrm{rank}(A) = n$ the matrix viewed as mapping $A : \mathbb{R}^n \to \mathbb{R}^m$ is injective, i. e., $\|Ax\|_2 = 0$ implies $x = 0$. Hence, the matrix $\bar{A}^T A$ is positive definite. Q.E.D.

1.1.4 Eigenvalues and eigenvectors

In the following, we consider square matrices $A = (a_{ij})_{i,j=1}^n \in \mathbb{K}^{n\times n}$.

Definition 1.7: *i) A number $\lambda \in \mathbb{C}$ is called "eigenvalue" of A, if there is a corresponding "eigenvector" $w \in \mathbb{C}^n$, $w \neq 0$, such that the "eigenvalue equation" holds:*

$$Aw = \lambda w. \tag{1.1.26}$$

ii) The vector space of all eigenvectors of an eigenvalue λ is called "eigenspace" and denoted by E_λ. Its dimension is the "geometric multiplicity" of λ. The set of all eigenvalues of a matrix $A \in \mathbb{K}^{n\times n}$ is called its "spectrum" and denoted by $\sigma(A) \subset \mathbb{C}$. The matrix function $R_A(z) := zI - A$ is called the "resolvent" of A and $\mathrm{Res}(A) := \{z \in \mathbb{C} \,|\, zI - A$ is regular$\}$ the corresponding "resolvent set".

iii) The eigenvalues are just the zeros of the "characteristic polynomial" $\chi_A \in P_n$ of A,

$$\chi_A(z) := \det(zI - A) = z^n + b_1 z^{n-1} + \ldots + b_n.$$

Hence, by the fundamental theorem of algebra there are exactly n eigenvalues counted accordingly to their multiplicity as zeros of χ_A, their so-called "algebraic multiplicities". The algebraic multiplicity is always greater or equal than the geometric multiplicity. If it is strictly greater, then the eigenvalue is called "deficient" and the difference the "defect" of the eigenvalue.

iv) The eigenvalues of a matrix can be determined independently of each other. One speaks of the "partial eigenvalue problem" if only a small number of the eigenvalues (e. g., the largest or the smallest one) and the corresponding eigenvectors are to be determined. In the "full eigenvalue problem" one seeks all eigenvalues with corresponding eigenvectors. For a given eigenvalue $\lambda \in \mathbb{C}$ (e. g., obtained as a zero of the characteristic polynomial) a corresponding eigenvector can be determined as any solution of the (singular) problem

$$(A - \lambda I)w = 0. \tag{1.1.27}$$

Conversely, for a given eigenvector $w \in \mathbb{K}^n$ (e. g., obtained by the "power method" described below), one obtains the corresponding eigenvalue by evaluating any of the quotients (choosing $w_i \neq 0$)

$$\lambda = \frac{(Aw)_i}{w_i},\ i = 1, \ldots, n, \qquad \lambda = \frac{(Aw, w)_2}{\|w\|_2^2}.$$

The latter quotient is called the "Rayleigh[7] quotient".

The characteristic polynomial of a matrix $A \in \mathbb{K}^{n\times n}$ has the following representation with its mutually distinct zeros λ_i:

$$\chi_A(z) = \prod_{i=1}^{m} (z - \lambda_i)^{\sigma_i}, \quad \sum_{i=1}^{m} \sigma_i = n,$$

where σ_i is the algebraic multiplicity of eigenvalue λ_i. Its geometric multiplicity is $\rho_i := \dim(\mathrm{kern}(A - \lambda_i I))$. We recall that generally $\rho_i \leq \sigma_i$, i. e., the defect satisfies $\alpha_i := \sigma_i - \rho_i \geq 0$. The latter corresponds to the largest integer $\alpha = \alpha(\lambda)$ such that

$$\mathrm{kern}(A - \lambda I)^{\alpha+1}) \neq \mathrm{kern}((A - \lambda I)^{\alpha}). \tag{1.1.28}$$

Since

$$\det(\bar{A}^T - \bar{z}I) = \det(\overline{A^T - zI}) = \det(\overline{A - zI)^T}) = \overline{\det(A - zI)}$$

the eigenvalues of the matrices A and $\bar{A}^T$ are related by

$$\lambda(\bar{A}^T) = \overline{\lambda(A)}. \tag{1.1.29}$$

[7] *John William Strutt (Lord Rayleigh) (1842–1919): English mathematician and physicist; worked at the beginning as (aristocratic) private scholar, 1879–1884 Professor for Experimental Physics in Cambridge; fundamental contributions to theoretical physics: scattering theory, acoustics, electro-magnetics, gas dynamics.*

Hence, associated to a normalized "primal" (right) eigenvector $w \in \mathbb{K}^n$, $\|w\|_2 = 1$, corresponding to an eigenvalue λ of A there is a "dual" (left) eigenvector $w^* \in \mathbb{K}^n \backslash \{0\}$ corresponding to the eigenvalue $\bar{\lambda}$ of $\bar{A}^T$ satisfying the "adjoint" eigenvalue equation

$$\bar{A}^T w^* = \bar{\lambda}\, w^* \quad (\Leftrightarrow \quad \bar{w}^{*T} A = \lambda \bar{w}^{*T}). \tag{1.1.30}$$

The dual eigenvector w^* may also be normalized by $\|w^*\|_2 = 1$ or, what is more suggested by numerical purposes, by $(w, w^*)_2 = 1$. In the "degenerate" case $(w, w^*)_2 = 0$, and only then, the problem

$$Aw^1 - \lambda w^1 = w \tag{1.1.31}$$

has a solution $w^1 \in \mathbb{K}^n$. This follows from the relations $w^* \in \mathrm{kern}(\bar{A}^T - \bar{\lambda} I)$, $w \perp \mathrm{kern}(\bar{A}^T - \bar{\lambda} I)$, and $\mathrm{range}(A - \lambda I) = \mathrm{kern}(\bar{A}^T - \bar{\lambda} I)^T$, the latter following from the result of Lemma 1.7. The vector w^1 is called "generalized eigenvector (of level one)" of A (or "Hauptvektor erster Stufe" in German) corresponding to the eigenvalue λ. Within this notion, eigenvectors are "generalized eigenvectors" of level zero. By definition, there holds

$$(A - \lambda I)^2 w^1 = (A - \lambda I) w = 0,$$

i. e., $w^1 \in \mathrm{kern}((A - \lambda I)^2)$ and, consequently, in view of the above definition, the eigenvalue λ has "defect" $\alpha(\lambda) \geq 1$. If this construction can be continued, i. e., if $(w^1, w^*)_2 = 0$, such that also the problem $Aw^2 - \lambda w^2 = w^1$ has a solution $w^2 \in \mathbb{K}^n$, which is then a "generalized eigenvector" of level two, by construction satisfying $(A - \lambda I)^3 w^2 = 0$. In this way, we may obtain "generalize eigenvectors" $w^m \in \mathbb{K}^n$ of level m for which $(A - \lambda I)^{m+1} w^m = 0$ and $(w^m, w^*)_2 \neq 0$. Then, the eigenvalue λ has defect $\alpha(\lambda) = m$.

Example 1.3: The following special matrices $C_m(\lambda)$ occur as building blocks, so-called "Jordan blocks", in the "Jordan[8] normal form" of a matrix $A \in \mathbb{K}^{n \times n}$ (see below):

$$C_m(\lambda) = \begin{bmatrix} \lambda & 1 & & & 0 \\ & \lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda & 1 \\ 0 & & & & \lambda \end{bmatrix} \in \mathbb{K}^{m \times m}, \quad \text{eigenvalue} \quad \lambda \in \mathbb{C}$$

$$\chi_{C_m(\lambda)}(z) = (z - \lambda)^m \quad \Rightarrow \quad \sigma = m, \quad \mathrm{rank}(C_m(\lambda) - \lambda I) = m - 1 \quad \Rightarrow \quad \rho = 1\,.$$

[8] Marie Ennemond Camille Jordan (1838–1922): French mathematician; Prof. in Paris; contributions to algebra, group theory, calculus and topology.

1.1.5 Similarity transformations

Definition 1.8: *Two matrices $A, B \in \mathbb{K}^{n\times n}$ are called "similar (to each other)", if there is a regular matrix $T \in \mathbb{K}^{n\times n}$ such that*

$$B = T^{-1}AT. \tag{1.1.32}$$

The transition $A \to B$ is called "similarity transformation".

Suppose that the matrix $A \in \mathbb{K}^{n\times n}$ is the representation of a linear mapping $\varphi : \mathbb{K}^n \to \mathbb{K}^n$ with respect to a basis $\{a^1, \ldots, a^n\}$ of $\mathbb{K}^n$. Then, using the regular matrix $T \in \mathbb{K}^{n\times n}$, we obtain a second basis $\{Ta^1, \ldots, Ta^n\}$ of $\mathbb{K}^n$ and B is the representation of the mapping φ with respect to this new basis. Hence, similar matrices are representations of the same linear mapping and any two representations of the same linear mapping are similar. In view of this fact, we expect that two similar matrices, representing the same linear mapping, have several of their characteristic quantities as matrices in common.

Lemma 1.11: *For any two similar matrices $A, B \in \mathbb{K}^{n\times n}$ there holds:*

a) $\det(A) = \det(B)$.

b) $\sigma(A) = \sigma(B)$.

c) $\operatorname{trace}(A) = \operatorname{trace}(B)$.

Proof. i) The product theorem for determinants implies that $\det(AB) = \det(A)\det(B)$ and further $\det(T^{-1}) = \det(T)^{-1}$. This implies that

$$\det(B) = \det(T^{-1}AT) = \det(T^{-1})\det(A)\det(T) = \det(T)^{-1}\det(A)\det(T) = \det(A).$$

ii) Further, for any $z \in \mathbb{C}$ there holds

$$\begin{aligned}\det(zI - B) &= \det(zT^{-1}T - T^{-1}AT) = \det(T^{-1}(zI - A)T)\\ &= \det(T^{-1})\det(zI - A)\det(T) = \det(zI - A),\end{aligned}$$

which implies that A and B have the same eigenvalues.

iii) The trace of A is just the coefficient of the monom z^{n-1} in the characteristic polynomial $\chi_A(z)$. Hence by (i) the trace of A equals that of B. Q.E.D.

Any matrix $A \in \mathbb{K}^{n\times n}$ is similar to its "canonical form" (Jordan normal form) which has the eigenvalues λ_i of A on its main diagonal counted accordingly to their algebraic multiplicity. Hence, in view of Lemma 1.11 there holds

$$\det(A) = \prod_{i=1}^{n} \lambda_i, \qquad \operatorname{trace}(A) = \sum_{i=1}^{n} \lambda_i. \tag{1.1.33}$$

Definition 1.9 (Normal forms): *i) Any matrix $A \in \mathbb{K}^{n\times n}$ is similar to its "canonical normal form" J_A ("Jordan normal form") which is a block diagonal matrix with main*

diagonal blocks, the "Jordan blocks", of the form as shown in Example 1.3. Here, the "algebraic" multiplicity of an eigenvalue corresponds to the number of occurrences of this eigenvalue on the main diagonal of J_A, while its "geometric" multiplicity corresponds to the number of Jordan blocks containing λ.

ii) A matrix $A \in \mathbb{K}^{n\times n}$, which is similar to a diagonal matrix, then having its eigenvalues on the main diagonal, is called "diagonalizable" ,

$$WAW^{-1} = \Lambda = \mathrm{diag}(\lambda_i) \quad (\lambda_i \ \textit{eigenvalues of}\ A).$$

This relation implies that the transformation matrix $W = [w^1, \ldots, w^n]$ has the eigenvectors w^i corresponding to the eigenvalues λ_i as column vectors. This means that orthogonalizability of a matrix is equivalent to the existence of a basis of eigenvectors.

iii) A matrix $A \in \mathbb{K}^{n\times n}$ is called "unitarily diagonalizable" if it is diagonalizable with a unitary transformation matrix. This is equivalent to the existence of an orthonormal basis of eigenvectors.

Positive definite Hermitian matrices $A \in \mathbb{K}^{n\times n}$ have very special spectral properties. These are collected in the following lemma and theorem, the latter one being the basic result of matrix analysis ("spectral theorem").

Lemma 1.12: *i) A Hermitian matrix has only real eigenvalues and eigenvectors to different eigenvalues are mutually orthogonal.*

ii) A Hermitian matrix is positive definite if and only if all its (real) eigenvalues are positive.

iii) Two normal matrices $A, B \in \mathbb{K}^{n\times n}$ commute, $AB = BA$, if and only if they possess a common basis of eigenvectors.

Proof. For the proofs, we refer to the standard linear algebra literature. Q.E.D.

Theorem 1.5 (Spectral theorem): *For square Hermitian matrices, $A = \bar{A}^T$, or more general for "normal" matrices, $\bar{A}^T A = A\bar{A}^T$, algebraic and geometric multiplicities of eigenvalues are equal, i. e., these matrices are diagonalizable. Further, they are even unitarily diagonalizable, i. e., there exists an orthonormal basis of eigenvectors.*

Proof. For the proof, we refer to the standard linear algebra literature. Q.E.D.

1.1.6 Matrix analysis

We now consider the vector space of all $m\times n$-matrices $A \in \mathbb{K}^{m\times n}$. This vector space may be identified with the vector space of mn-vectors, $\mathbb{K}^{n\times n} \cong \mathbb{K}^{mn}$. Hence, all statements for vector norms carry over to norms for matrices. In particular, all norms for $m\times n$-matricen

are equivalent and the convergence of sequences of matrices is again the componentwise convergence

$$A^k \to A \,(k \to \infty) \quad \Longleftrightarrow \quad a_{ij}^k \to a_{ij} \,(k \to \infty)\,, \quad i = 1, \ldots, m\,, j = 1, \ldots, n\,.$$

Now, we restrict the further discussion to square matrices $A \in \mathbb{K}^{n\times n}$. For an arbitrary vector norm $\|\cdot\|$ on $\mathbb{K}^n$ a norm for matrices $A \in \mathbb{K}^{n\times n}$ is generated by

$$\|A\| := \sup_{x\in\mathbb{K}^n\setminus\{0\}} \frac{\|Ax\|}{\|x\|} = \sup_{x\in\mathbb{K}^n, \|x\|=1} \|Ax\|.$$

The definiteness and homogeneity are obvious and the triangle inequality follows from that holding for the given vector norm. This matrix norm is called the "natural matrix norm" corresponding to the vector norm $\|\cdot\|$. In the following for both norms, the matrix norm and the generating vector norm, the same notation is used. For a natural matrix norm there always holds $\|I\| = 1$. Such a "natural" matrix norm is automatically "compatible" with the generating vector norm, i. e., it satisfies

$$\|Ax\| \le \|A\|\,\|x\|\,, \quad x \in \mathbb{K}^n,\ A \in \mathbb{K}^{n\times n}\,. \tag{1.1.34}$$

Further it is "submultiplicative",

$$\|AB\| \le \|A\|\,\|B\|\,, \quad A, B \in \mathbb{K}^{n\times n}. \tag{1.1.35}$$

Not all matrix norms are "natural" in the above sense. For instance, the square-sum norm (also called "Frobenius[9]-norm")

$$\|A\|_F := \Big(\sum_{j,k=1}^{n} |a_{jk}|^2\Big)^{1/2}$$

is compatible with the Euclidian norm and submultiplicative but cannot be a *natural* matrix norm since $\|I\|_F = \sqrt{n}$ (for $n \ge 2$). The natural matrix norm generated from the Euclidian vector norm is called "spectral norm". This name is suggested by the following result.

Lemma 1.13 (Spectral norm): *For an arbitrary square matrix $A \in \mathbb{K}^{n\times n}$ the product matrix $\overline{A}^T A \in \mathbb{K}^{n\times n}$ is always Hermitian and positive semi-definitsemi-definite. For the spectral norm of A there holds*

$$\|A\|_2 = \max\{|\lambda|^{1/2},\, \lambda \in \sigma(\bar{A}^T A)\}. \tag{1.1.36}$$

If A is Hermitian (or symmetric), then,

$$\|A\|_2 = \max\{|\lambda|,\, \lambda \in \sigma(A)\}. \tag{1.1.37}$$

[9]Ferdinand Georg Frobenius (1849–1917): German mathematician; Prof. in Zurich and Berlin; contributions to the theory of differential equations, to determinants and matrices as well as to group theory.

Proof. i) Let the matrix $A \in \mathbb{K}^{n\times n}$ be Hermitian. For any eigenvalue λ of A and corresponding eigenvector x there holds

$$|\lambda| = \frac{\|\lambda x\|_2}{\|x\|_2} = \frac{\|Ax\|_2}{\|x\|_2} \le \|A\|_2.$$

Conversely, let $\{a^i, i = 1, \cdots, n\} \subset \mathbb{C}^n$ be an ONB of eigenvectors of A and $x = \sum_i x_i a^i \in \mathbb{C}^n$ be arbitrary. Then,

$$\|Ax\|_2 = \big\|A\big(\sum\nolimits_i x_i a^i\big)\big\|_2 = \big\|\sum\nolimits_i \lambda_i x_i a^i\big\|_2 \le \max_i |\lambda_i| \, \big\|\sum\nolimits_i x_i a^i\big\|_2 = \max_i |\lambda_i| \, \|x\|_2,$$

and consequently,

$$\frac{\|Ax\|_2}{\|x\|_2} \le \max_i |\lambda_i|.$$

ii) For a general matrix $A \in \mathbb{K}^{n\times n}$ there holds

$$\|A\|_2^2 = \max_{x\in C^n\setminus 0} \frac{\|Ax\|_2^2}{\|x\|_2^2} = \max_{x\in C^n\setminus 0} \frac{(\bar{A}^T Ax, x)_2}{\|x\|_2^2} \le \max_{x\in C^n\setminus 0} \frac{\|\bar{A}^T Ax\|_2}{\|x\|_2} = \|\bar{A}^T A\|_2.$$

and $\|\bar{A}^T A\|_2 \le \|\bar{A}^T\|_2 \|A\|_2 = \|A\|_2^2$ (observe that $\|A\|_2 = \|\bar{A}^T\|_2$ due to $\|Ax\|^2 = \|\bar{A}^T \bar{x}\|^2$). This completes the proof. Q.E.D.

Lemma 1.14 (Natural matrix norms): *The natural matrix norms generated by the l_∞ norm $\|\cdot\|_\infty$ and the l_1 Norm $\|\cdot\|_1$ are the so-called "maximal-row-sum norm" and the "maximal-column-sum norm" , respectively,*

$$\|A\|_\infty := \max_{1\le i\le n} \sum_{j=1}^n |a_{ij}|, \qquad \|A\|_1 := \max_{1\le j\le n} \sum_{i=1}^n |a_{ij}| \,. \tag{1.1.38}$$

Proof. We give the proof only for the l_∞ norm. For the l_1 norm the argument is analogous. i) The maximal row sum $\|\cdot\|_\infty$ is a matrix norm. The norm properties (N1) - (N3) follow from the corresponding properties of the modulus. For the matrix product AB there holds

$$\begin{aligned}\|AB\|_\infty &= \max_{1\le i\le n} \Big| \sum_{j=1}^n \Big(\sum_{k=1}^n a_{ik} b_{kj} \Big) \Big| \le \max_{1\le i\le n} \sum_{k=1}^n \Big(|a_{ik}| \sum_{j=1}^n |b_{kj}| \Big) \\ &\le \max_{1\le i\le n} \sum_{k=1}^n |a_{ik}| \max_{1\le k\le n} \sum_{j=1}^n |b_{kj}| = \|A\|_\infty \|B\|_\infty.\end{aligned}$$

ii) Further, in view of

$$\|Ax\|_\infty = \max_{1\le j\le n} |\sum_{k=1}^n a_{jk} x_k| \le \max_{1\le j\le n} \sum_{k=1}^n |a_{jk}| \max_{1\le k\le n} |x_k| = \|A\|_\infty \|x\|_\infty$$

the maximal row-sum is compatible with the maximum norm $\|\cdot\|_\infty$ and there holds

$$\sup_{\|x\|_\infty=1} \|Ax\|_\infty \le \|A\|_\infty.$$

iii) In the case $\|A\|_\infty = 0$ also $A = 0$, i. e.,

$$\|A\|_\infty = \sup_{\|x\|_\infty=1} \|Ax\|_\infty.$$

Therefore, let $\|A\|_\infty > 0$ and $m \in \{1, \ldots, n\}$ an index such that

$$\|A\|_\infty = \max_{1\le j\le n} \sum_{k=1}^{n} |a_{jk}| = \sum_{k=1}^{n} |a_{mk}|.$$

For $k = 1, \ldots, n$, we set

$$z_k := \begin{cases} |a_{mk}|/a_{mk} & \text{für } a_{mk} \neq 0, \\ 0, & \text{sonst}, \end{cases}$$

i. e., $z = (z_k)_{k=1}^n \in \mathbb{K}^n$, $\|z\|_\infty = 1$. For $v := Az$ it follows that

$$v_m = \sum_{k=1}^{n} a_{mk} z_k = \sum_{k=1}^{n} |a_{mk}| = \|A\|_\infty.$$

Consequently,

$$\|A\|_\infty = v_m \le \|v\|_\infty = \|Az\|_\infty \le \sup_{\|y\|_\infty=1} \|Ay\|_\infty,$$

what was to be shown. Q.E.D.

Let $\|\cdot\|$ be an arbitrary vector norm and $\|\cdot\|$ a corresponding compatible matrix norm. Then, with a normalized eigenvector $\|w\| = 1$ corresponding to the eigenvalue λ there holds

$$|\lambda| = |\lambda|\,\|w\| = \|\lambda w\| = \|Aw\| \le \|A\|\,\|w\| = \|A\|, \tag{1.1.39}$$

i. e., all eigenvalues of A are contained in a circle in $\mathbb{C}$ with center at the origin and radius $\|A\|$. Especially with $\|A\|_\infty$, we obtain the eigenvalue bound

$$\max_{\lambda\in\sigma(A)} |\lambda| \le \|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|. \tag{1.1.40}$$

Since the eigenvalues of $\bar{A}^T$ and A are related by $\lambda(\bar{A}^T) = \bar{\lambda}(A)$, using the bound (1.1.40) simultaneously for $\bar{A}^T$ and A yields the following refined bound:

$$\begin{aligned}\max_{\lambda\in\sigma(A)} |\lambda| &\le \min\{\|A\|_\infty, \|\bar{A}^T\|_\infty\} \\ &= \min\Big\{ \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|, \max_{1\le j\le n} \sum_{i=1}^{n} |a_{ij}| \Big\}.\end{aligned} \tag{1.1.41}$$

The following lemma contains a useful result on the regularity of small perturbations of the unit matrix.

Lemma 1.15 (Perturbation of unity): *Let $\|\cdot\|$ be any natural matrix norm on $\mathbb{K}^{n\times n}$ and $B \in \mathbb{K}^{n\times n}$ a matrix with $\|B\| < 1$. Then, the perturbed matrix $I+B$ is regular and its inverse is given as the (convergent) "Neumann[10] series"*

$$(I+B)^{-1} = \sum_{k=0}^{\infty} B^k. \tag{1.1.42}$$

Further, there holds

$$\|(I+B)^{-1}\| \le \frac{1}{1-\|B\|}\,. \tag{1.1.43}$$

Proof. i) First, we show the regularity of $I+B$ and the bound (1.1.43). For all $x \in \mathbb{K}^n$ there holds

$$\|(I+B)x\| \ge \|x\| - \|Bx\| \ge \|x\| - \|B\|\|x\| = (1-\|B\|)\|x\|\,.$$

In view of $1-\|B\| > 0$ this implies that $I+B$ is injective and consequently regular. Then, the following estimate implies (1.1.43):

$$\begin{aligned}1 = \|I\| \;&=\; \|(I+B)(I+B)^{-1}\| = \|(I+B)^{-1} + B(I+B)^{-1}\| \\ &\ge \|(I+B)^{-1}\| - \|B\|\,\|(I+B)^{-1}\| = \|(I+B)^{-1}\|(1-\|B\|) > 0\,.\end{aligned}$$

ii) Next, we define

$$S := \lim_{k\to\infty} S_k, \quad S_k = \sum_{s=0}^{k} B^s.$$

S is well defined due to the fact that $\{S_n\}_{n\in\mathbb{N}}$ is a Cauchy sequence with respect to the matrix norm $\|\cdot\|$ (and, by the norm equivalence in finite dimensional normed spaces,

[10] *Carl Gottfried Neumann (1832–1925): German mathematician; since 1858 "Privatdozent" and since 1863 apl. Prof. in Halle. After holding professorships in Basel and Tübingen he moved 1868 to Leipzig where he worked for more than 40 years. He contributed to the theory of (partial) differential and integral equations, especially to the Dirichlet problem. The "Neumann boundary condition" and the "Neumann series" are named after him. In mathematical physics he worked on analytical mechanics and potential theory. together with A. Clebsch he founded the journal "Mathematische Annalen".*

with respect to any matrix norm). By employing the triangle inequality, using the matrix norm property and the limit formula for the geometric series, we see that

$$\|S\| = \lim_{k\to\infty} \|S_k\| = \lim_{k\to\infty} \Big\| \sum_{s=0}^{k} B^s \Big\| \le \lim_{n\to\infty} \sum_{s=0}^{k} \|B\|^s = \lim_{k\to\infty} \frac{1-\|B\|^{k+1}}{1-\|B\|} = \frac{1}{1-\|B\|}.$$

Furthermore, $S_k\,(I-B) = I - B^{k+1}$ and due to the fact that multiplication with $I-B$ is continuous,

$$I = \lim_{k\to\infty} \big(S_k(I-B)\big) = \big(\lim_{k\to\infty} S_k\big)(I-B) = S(I-B).$$

Hence, $S = (I-B)^{-1}$ and the proof is complete. Q.E.D.

Corollary 1.2: *Let $A \in \mathbb{K}^{n\times n}$ be a regular matrix and $\tilde{A}$ another matrix such that*

$$\|\tilde{A} - A\| < \frac{1}{\|A^{-1}\|}. \tag{1.1.44}$$

Then, also $\tilde{A}$ is regular. This means that the "resolvent set" $\mathrm{Res}(A)$ of a matrix $A \in \mathbb{K}^{n\times n}$ is open in $\mathbb{K}^{n\times n}$ and the only "singular" points are just the eigenvalues of A, i. e., there holds $\mathbb{C} = \mathrm{Res}(A) \cup \sigma(A)$.

Proof. Notice that $\tilde{A} = A + \tilde{A} - A = A(I + A^{-1}(\tilde{A} - A))$. In view of

$$\|A^{-1}(\tilde{A} - A)\| \le \|A^{-1}\|\,\|\tilde{A} - A\| < 1$$

by Lemma 1.15 the matrix $I + A^{-1}(\tilde{A} - A)$ is regular. Then, also the product matrix $A(I + A^{-1}(\tilde{A} - A))$ is regular, which implies the regularity of $\tilde{A}$. Q.E.D.

1.2 Spectra and pseudo-spectra of matrices

1.2.1 Stability of dynamical systems

We consider a finite dimensional dynamical system of the form

$$u'(t) = F(t, u(t)), \quad t \ge 0, \quad u(0) = u^0, \tag{1.2.45}$$

where $u : [0,\infty) \to \mathbb{R}^n$ is a continuously differentiable vector function and the system function $F(\cdot,\cdot)$ is assumed (for simplicity) to be defined on all of $\mathbb{R} \times \mathbb{R}^n$ and twice continuously differentiable. The system (1.2.45) may originate from the discretization of an infinite dimensional dynamical system such as the nonstationary Navier-Stokes equations mentioned in the introductory Chapter 0. Suppose that u is a particular solution of (1.2.45). We want to investigate its stability against small perturbations $u(t_0) \to u(t_0) + w^0 =: v(t_0)$ at any time $t_0 \ge 0$. For this, we use the strongest concept of

stability, which is suggested by the corresponding properties of solutions of the Navier-Stokes equations.

Definition 1.10: *The solution* $u \in C^1[0,\infty;\mathbb{R}^n)$ *of (1.2.45) is called "exponentially stable" if there are constants* $\delta, K, \kappa \in \mathbb{R}_+$ *such that for any perturbation* $w^0 \in \mathbb{R}^n$, $\|w^0\|_2 \le \delta$, *at any time* $t_0 \ge 0$, *there exists a secondary solution* $v \in C^1(t_0,\infty;\mathbb{R}^n)$ *of the perturbed system*

$$v'(t) = F(t, v(t)), \quad t \ge 0, \quad v(t_0) = u(t_0) + w^0, \tag{1.2.46}$$

and there holds

$$\|v(t) - u(t)\|_2 \le K e^{-\kappa(t-t_0)} \|w^0\|_2, \quad t \ge t_0. \tag{1.2.47}$$

For simplicity, we restrict the following discussion to the special situation of an autonomous system, i. e., $F(t,\cdot) \equiv F(\cdot)$, and a stationary particular solution $u(t) \equiv u \in \mathbb{R}^n$, i. e., to the solution of the nonlinear system

$$F(u) = 0. \tag{1.2.48}$$

The investigation of the stability of u leads us to consider the so-called "perturbation equation" for the perturbation $w(t) := v(t) - u$,

$$w'(t) = F(v(t)) - F(u) = F'(u)w(t) + \mathcal{O}(\|w(t)\|_2^2), \;\; t \ge 0, \quad w(0) = w^0, \tag{1.2.49}$$

where the higher-order term depends on bounds on u and u' as well as on the smoothness properties of $F(\cdot)$.

Theorem 1.6: *Suppose that the Jacobian* $A := F'(u)$ *is diagonalizable and that all its eigenvalues have negative real part. Then, the solution* u *of (1.2.48) is exponentially stable in the sense of Definition 1.10 with the constants* $\kappa = |\mathrm{Re}\lambda_{\max}|$ *and* $K = \mathrm{cond}_2(W)$, *where* $\lambda_{\max}$ *is the eigenvalue of* A *with largest (negative) real part and* $W = [w^1, \ldots, w^n]$ *the column matrix formed by the (normalized) eigenbasis of* A. *If* A *is normal then* $K = \mathrm{cond}_2(W) = 1$.

Proof. i) Consider the linearized system (linearized perturbation equation)

$$w'(t) = Aw(t), \;\; t \ge t_0, \quad w(0) = w^0. \tag{1.2.50}$$

Since the Jacobian A is diagonalizable there exists an ONB $\{w^1, \ldots, w^n\}$ of eigenvectors of A:

$$Aw^i = \lambda_i w^i, \quad i = 1, \ldots, n.$$

With the matrices $W := [w^1, \ldots, w^n]$ and $\Lambda := \mathrm{diag}(\lambda_i)$ there holds

$$W^{-1}AW = \Lambda, \quad A = W\Lambda W^{-1}.$$

Using this notation the perturbation equation can be rewritten in the form

$$w'(t) = Aw(t) \quad \Leftrightarrow \quad w'(t) = W\Lambda W^{-1}w(t) \quad \Leftrightarrow \quad (W^{-1}w)'(t) = \Lambda W^{-1}w(t),$$

or for the transformed variable $v := W^{-1}w$ componentwise:

$$v_i'(t) = \lambda_i v_i(t), \;\; t \geq 0, \quad v_i(0) = (W^{-1}w)_i(0).$$

The solution behavior is (observe that $e^{i\mathrm{Im}\lambda_i t} = 1$)

$$|v_i(t)| \leq e^{\mathrm{Re}\lambda_i t}|(W^{-1}w)_i(0)|, \;\; t \geq 0.$$

This implies:

$$\|v(t)\|_2^2 \leq \sum_{i=1}^n |v_i(t)|^2 \leq \sum_{i=1}^n e^{2\mathrm{Re}\lambda_i t}|(W^{-1}w)_i(0)|^2 \leq e^{2\mathrm{Re}\lambda_{\min} t}\|(W^{-1}w)(0)\|_2^2,$$

and consequently,

$$\begin{aligned} \|w(t)\|_2 &\leq \|Wv(t)\|_2 \leq \|W\|_2\|v(t)\|_2 \leq \|W\|_2\, e^{\mathrm{Re}\lambda_{\min} t}\, \|(W^{-1}w)(0)\|_2 \\ &\leq \|W\|_2\, e^{\mathrm{Re}\lambda_{\min} t}\, \|W^{-1}\|_2\|w(0)\|_2 \\ &= \mathrm{cond}_2(W)\, e^{\mathrm{Re}\lambda_{\min} t}\|w(0)\|_2. \end{aligned} \tag{1.2.51}$$

The condition number of W can become arbitrarily large depending on the "non-orthogonality" of the eigenbasis of the Jacobian A.

ii) The assertion now follows by combining (1.2.51) and (1.2.49) within a continuation argument. The proof is complete. Q.E.D.

Following the argument in the proof of Theorem 1.6, we see that the occurrence of just one eigenvalue with $\mathrm{Re}\,\lambda > 0$ inevitably causes dynamic instability of the solution u, i. e., arbitrarily small perturbations may grow in time without bound. Denoting by $S : \mathbb{R}^n \to C^1[0,\infty;\mathbb{R}^n)$ the "solution operator" of the *linearized* perturbation equation (1.2.50), i. e., $w(t) = S(t)w^0$, this can be formulated as

$$\max_{\lambda\in\sigma(A)} \mathrm{Re}\,\lambda > 0 \quad \Rightarrow \quad \sup_{t\geq 0} \|S(t)\|_2 = \infty. \tag{1.2.52}$$

The result of Theorem 1.6 can be extended to the case of a non-diagonalizable Jacobian $A = F'(u)$. In this case, one obtains a stability behavior of the form

$$\|S(t)\|_2 \approx K(1 + t^\alpha)e^{\mathrm{Re}\lambda_{\max} t}, \quad t \geq 0, \tag{1.2.53}$$

where $\alpha \geq 1$ is the defect of the most critical eigenvalue $\lambda_{\max}$, i. e., that eigenvalue with largest real part $\mathrm{Re}\lambda_{\max} < 0$. This implies that

$$\sup_{t>0} \|S(t)\| \approx \Big(\frac{\alpha}{e}\Big)^\alpha \frac{1}{|\,\mathrm{Re}\,\lambda_{\max}|^\alpha}, \tag{1.2.54}$$

i. e., for $-1 \ll \operatorname{Re}\lambda_{\min} < 0$ initially small perturbations may grow beyond a value at which nonlinear instability is triggered. Summarizing, we are interested in the case that all eigenvalues of $A = F'(u)$ have negative real part, suggesting stability in the sense of Theorem 1.6, and especially want to compute the most "critical" eigenvalue, i. e., that $\lambda \in \sigma(A)$ with maximal $\operatorname{Re}\lambda < 0$ to detect whether the corresponding solution operator $S(t)$ may behave in a critical way.

The following result, which is sometimes addressed as the "easy part of the Kreiss[11] matrix theorem" indicates in which direction this analysis has to go.

Lemma 1.16: *Let $A := F'(u)$ and $z \in \mathbb{C} \setminus \sigma(A)$ with $\operatorname{Re} z > 0$. Then, for the solution operator $S(t)$ of the linearized perturbation equation (1.2.50), there holds*

$$\sup_{t\geq 0} \|S(t)\|_2 \geq |\operatorname{Re} z|\, \|(zI - A)^{-1}\|_2. \tag{1.2.55}$$

Proof. We continue using the notation from the proof of Theorem 1.6. If $\|S(t)\|_2$ is unbounded over $[0, \infty)$, the asserted estimate holds trivially. Hence, let us assume that

$$\sup_{t\geq 0} \|w(t)\|_2 = \sup_{t\geq 0} \|S(t)w^0\|_0 \leq \sup_{t\geq 0} \|S(t)\|_2 \|w^0\|_2 < \infty.$$

For $z \notin \sigma(A)$ the resolvent $R_A(z) = zI - A$ is regular. Let $w^0 \in \mathbb{K}^n$ be an arbitrary but nontrivial initial perturbation and $w(t) = S(t)w^0$. We rewrite equation (1.2.50) in the form

$$\partial_t w - zw + (zI - A)w = 0,$$

and multiply by e^{-tz}, to obtain

$$\partial_t(e^{-tz}w) + e^{-tz}(zI - A)w = 0.$$

Next, integrating this over $0 \leq t < T$ and observing $\operatorname{Re} z > 0$ and $\lim_{t\to\infty} e^{-tz}w = 0$ yields

$$-(zI - A)^{-1}w^0 = \Big(\int_0^\infty e^{-tz} S(t)\, dt \Big)\, w^0.$$

From this, we conclude

$$\|(zI - A)^{-1}\|_2 \leq \Big(\int_0^\infty e^{-t|\operatorname{Re} z|}\, dt \Big) \sup_{t>0} \|S(t)\|_2 \leq |\operatorname{Re} z|^{-1} \sup_{t>0} \|S(t)\|_2,$$

which implies the asserted estimate. Q.E.D.

The above estimate (1.2.55) for the solution operator $S(t)$ can be interpreted as follows: Even if all eigenvalues of the matrix A have negative real parts, which in view

[11] Heinz-Otto Kreiss (1930–2015): Swedish/US-American mathematician; worked in Numerical Analysis and in the new field Scientific Computing in the early 1960s; born in Hamburg, Germany, he studied and worked at the Kungliga Tekniska Hgskolan in Stockholm, Sweden; he published a number of books; later he became Prof. at the California Institute of Technology and University of California, Los Angeles (UCLA).

of Theorem 1.6 would indicate stability of solutions to (1.2.50), there may be points z in the right complex half plane for which $\|(zI - A)^{-1}\|_2 \gg |\operatorname{Re} z|^{-1}$ and consequently,

$$\sup_{t\geq 0} \|S(t)\|_2 \gg 1. \tag{1.2.56}$$

Hence, even small perturbations of the particular solution u may be largely amplified eventually triggering nonlinear instability.

1.2.2 Pseudo-spectrum of a matrix

The estimate (1.2.55) makes us search for points $z \in \mathbb{C} \setminus \sigma(A)$ with $\operatorname{Re} z > 0$ and

$$\|(zI - A)^{-1}\|_2 \gg |\operatorname{Re} z|^{-1}.$$

This suggests the concept of the "pseudo-spectrum" of the matrix A, which goes back to Landau [9] and has been extensively described and applied in the stability analysis of dynamical systems, e. g., in Trefethen [20] and Trefethen & Embree [22].

Definition 1.11 (Pseudo-spectrum): *For $\varepsilon \in \mathbb{R}_+$ the "ε-pseudo-spectrum" $\sigma_\varepsilon(A) \subset \mathbb{C}$ of a matrix $A \in \mathbb{K}^{n\times n}$ is defined by*

$$\sigma_\varepsilon(A) := \left\{ z \in \mathbb{C} \setminus \sigma(A) \,\middle|\, \|(A - zI)^{-1}\|_2 \geq \varepsilon^{-1} \right\} \cup \sigma(A). \tag{1.2.57}$$

Remark 1.7: The concept of a pseudo-spectrum is interesting only for non-normal operators, since for a normal operator $\sigma_\varepsilon(A)$ is just the union of ε-circles around its eigenvalues. This follows from the estimate (see Dunford & Schwartz [8] or Kato [12])

$$\|(A - zI)^{-1}\|_2 \geq \operatorname{dist}(z, \sigma(A))^{-1}, \quad z \notin \sigma(A), \tag{1.2.58}$$

where equality holds if A is normal.

Remark 1.8: The concept of the "pseudo-spectrum" can be introduced in much more general situations, such as that of closed linear operators in abstract Hilbert or Banach spaces (see Trefethen & Embree [22]). Typically hydrodynamic stability analysis concerns differential operators defined on bounded domains. This situation fits into the Hilbert-space framework of "closed unbounded operators with compact inverse".

Using the notion of the pseudo-spectrum the estimate (1.2.55) can be expressed in the following form

$$\sup_{t\geq 0} \|S(t)\|_2 \geq \sup \left\{ \frac{|\operatorname{Re} z|}{\varepsilon} \,\middle|\, \varepsilon > 0, z \in \sigma_\varepsilon(A), \operatorname{Re} z > 0 \right\}, \tag{1.2.59}$$

or

$$\max_{\lambda\in\sigma_\varepsilon(A)} \operatorname{Re}\lambda > K\varepsilon \quad \Rightarrow \quad \sup_{t\geq 0} \|S(t)\|_2 > K. \tag{1.2.60}$$

Below, we will present methods for computing estimates for the pseudo-spectrum of a matrix. This will be based on related methods for solving the partial eigenvalue problem. To this end, we provide some results on several basic properties of the pseudo-spectrum.

Lemma 1.17: *i) For a matrix* $A \in \mathbb{K}^{n\times n}$ *the following definitions of an* ε*-pseudo-spectrum are equivalent:*

a) $\sigma_\varepsilon(A) := \{z \in \mathbb{C} \setminus \sigma(A) \,|\, \|(A - zI)^{-1}\|_2 \geq \varepsilon^{-1}\} \cup \sigma(A).$

b) $\sigma_\varepsilon(A) := \{z \in \mathbb{C} \,|\, z \in \sigma(A + E) \text{ for some } E \in \mathbb{K}^{n\times n} \text{ with } \|E\|_2 \leq \varepsilon\}.$[2mm]

c) $\sigma_\varepsilon(A) := \{z \in \mathbb{C} \,|\, \|(A - zI)v\|_2 \leq \varepsilon \text{ for some } v \in \mathbb{K}^n \text{ with } \|v\|_2 = 1\}.$

ii) Let $0 \notin \sigma(A)$*. Then, the* ε*-pseudo-spectra of* A *and that of its inverse* A^{-1} *are related by*

$$\sigma_\varepsilon(A) \subset \{z \in \mathbb{C} \setminus \{0\} \,|\, z^{-1} \in \sigma_{\delta(z)}(A^{-1})\} \cup \{0\}, \tag{1.2.61}$$

where $\delta(z) := \varepsilon\|A^{-1}\|_2/|z|$ *and, for* $0 < \varepsilon < 1$*, by*

$$\sigma_\varepsilon(A^{-1}) \cap B_1(0)^c \subset \{z \in \mathbb{C} \setminus \{0\} \,|\, z^{-1} \in \sigma_\delta(A)\}, \tag{1.2.62}$$

where $B_1(0) := \{z \in \mathbb{C},\ |z| \leq 1\}$ *and* $\delta := \varepsilon/(1 - \varepsilon)$*.*

Proof. The proof of part (i) can be found in Trefethen & Embree [22]. For completeness, we recall a sketch of the argument. The proof of part (ii) is taken from Gerecht et al. [35].

ia) In all three definitions, we have $\sigma(A) \subset \sigma_\varepsilon(A)$. Let $z \in \sigma_\varepsilon(A)$ in the sense of definition (a). There exists a $w \in \mathbb{K}^n$ with $\|w\|_2 = 1$, such that $\|(zI - A)^{-1}w\|_2 \geq \varepsilon^{-1}$. Hence, there is a $v \in \mathbb{K}^n$ with $\|v\|_2 = 1$, and $s \in (0, \varepsilon)$, such that $(zI - A)^{-1}w = s^{-1}v$ or $(zI - A)v = sw$. Let $Q(v, w) \in \mathbb{K}^{n\times n}$ denote the unitary matrix, which rotates the unit vector v into the unit vector w, such that $sw = sQ(v, w)v$. Then, $z \in \sigma(A+E)$ where $E := sQ(v, w)$ with $\|E\|_2 \leq \varepsilon$, i. e., $z \in \sigma_\varepsilon(A)$ in the sense of definition (b). Let now be $z \in \sigma_\varepsilon(A)$ in the sense of definition (b), i. e., there exists $E \in \mathbb{K}^{n\times n}$ with $\|E\|_2 \leq \varepsilon$ such that $(A + E)w = zw$, with some $w \in \mathbb{K}^n$, $w \neq 0$. Hence, $(A - zI)w = -Ew$, and therefore,

$$\begin{aligned}
\|(A - zI)^{-1}\|_2 &= \sup_{v\in\mathbb{K}^n\setminus\{0\}} \frac{\|(A - zI)^{-1}v\|_2}{\|v\|_2} = \sup_{v\in\mathbb{K}^n\setminus\{0\}} \frac{\|v\|_2}{\|(A - zI)v\|_2} \\
&= \Big(\inf_{v\in\mathbb{K}^n\setminus\{0\}} \frac{\|(A - zI)v\|_2}{\|v\|_2}\Big)^{-1} \geq \Big(\frac{\|(A - zI)w\|_2}{\|w\|_2}\Big)^{-1} \\
&= \Big(\frac{\|Ew\|_2}{\|w\|_2}\Big)^{-1} \geq \|E\|_2^{-1} \geq \varepsilon^{-1}.
\end{aligned}$$

Hence, $z \in \sigma_\varepsilon(A)$ in the sense of (a). This proves the equivalence of definitions (a) and (b).

ib) Next, let again $z \in \sigma_\varepsilon(A) \setminus \sigma(A)$ in the sense of definition (a). Then,

$$\varepsilon \geq \|(A - zI)^{-1}\|_2^{-1} = \Big(\sup_{w \in \mathbb{K}^n \setminus \{0\}} \frac{\|(A - zI)^{-1} w\|_2}{\|w\|_2} \Big)^{-1} = \inf_{v \in \mathbb{K}^n \setminus \{0\}} \frac{\|(A - zI)v\|_2}{\|v\|_2}.$$

Hence, there exists a $v \in \mathbb{K}^n$ with $\|v\|_2 = 1$, such that $\|(A - zI)v\| \leq \varepsilon$, i. e., $z \in \sigma_\varepsilon(A)$ in the sense of definition (c). By the same argument, now used in the reversed direction, we see that $z \in \sigma_\varepsilon(A)$ in the sense of definition (c) implies that also $z \in \sigma_\varepsilon(A)$ in the sense of definition (a). Thus, definition (a) is also equivalent to condition (c).

iia) We use the definition (c) from part (i) for the ε-pseudo-spectrum. Let $z \in \sigma_\varepsilon(A)$ and accordingly $v \in \mathbb{K}^n$, $\|v\|_2 = 1$, satisfying $\|(A - zI)v\|_2 \leq \varepsilon$. Then,

$$\|(A^{-1} - z^{-1} I)v\|_2 = \|z^{-1} A^{-1}(zI - A)v)\|_2 \leq |z|_2^{-1} \|A^{-1}\|_2 \, \varepsilon.$$

This proves the asserted relation (1.2.61).

iib) To prove the relation (1.2.62), we again use the definition (c) from part (i) for the ε-pseudo-spectrum. Accordingly, for $z \in \sigma_\varepsilon(A^{-1})$ with $|z| \geq 1$ there exists a unit vector $v \in \mathbb{K}^n$, $\|v\|_2 = 1$, such that

$$\varepsilon \geq \|(zI - A^{-1})v\|_2 = |z| \|(A - z^{-1} I) A^{-1} v\|_2.$$

Then, setting $w := \|A^{-1}v\|_2^{-1} A^{-1} v$ with $\|w\|_2 = 1$, we obtain

$$\|(A - z^{-1} I)w\|_2 \leq |z|^{-1} \|A^{-1} v\|_2^{-1} \varepsilon.$$

Hence, observing that

$$\|A^{-1}v\|_2 = \|(A^{-1} - zI)v + zv\|_2 \geq \|zv\|_2 - \|(A^{-1} - zI)v\|_2 \geq |z| - \varepsilon,$$

we conclude that

$$\|(A - z^{-1} I)w\|_2 \leq \frac{\varepsilon}{|z|(|z| - \varepsilon)} \leq \frac{\varepsilon}{1 - \varepsilon}.$$

This completes the proof. Q.E.D.

The next proposition relates the size of the resolvent norm $\|(zI - A)^{-1}\|_2$ to easily computable quantities in terms of the eigenvalues and eigenvectors of the matrix $A = F'(u)$.

Theorem 1.7: *Let $\lambda \in \mathbb{C}$ be a non-deficient eigenvalue of the matrix $A := F'(u)$ with corresponding primal and dual eigenvectors $v, v^* \in \mathbb{K}^n$ normalized by $\|v\|_2 = (v, v^*)_2 = 1$. Then, there exists a continuous function $\omega : \mathbb{R}_+ \to \mathbb{C}$ with $\lim_{\varepsilon \searrow 0+} \omega(\varepsilon) = 1$, such that for $\lambda_\varepsilon := \lambda - \varepsilon \omega(\varepsilon) \|v^*\|_2$, there holds*

$$\|(A - \lambda_\varepsilon I)^{-1}\|_2 \geq \varepsilon^{-1}, \tag{1.2.63}$$

i. e., the point λ_ε lies in the ε-pseudo-spectrum of the matrix A.

Proof. The argument of the proof is recalled from Gerecht et. al. [35] where it is developed within a function space setting and has therefore to be simplified here for the finite dimensional situation.

i) Let $B \in \mathbb{K}^{n\times n}$ be a matrix with $\|B\|_2 \le 1$. We consider the perturbed eigenvalue problem

$$(A+\varepsilon B)v_\varepsilon = \lambda_\varepsilon\, v_\varepsilon. \tag{1.2.64}$$

Since this is a regular perturbation and λ non-deficient, there exist corresponding eigenvalues $\lambda_\varepsilon \in \mathbb{C}$ and eigenvectors $v_\varepsilon \in \mathbb{K}^n$, $\|v_\varepsilon\|_2 = 1$, such that

$$|\lambda_\varepsilon - \lambda| = \mathcal{O}(\varepsilon), \quad \|v_\varepsilon - v\|_2 = \mathcal{O}(\varepsilon).$$

Furthermore, from the relation

$$(Av - \lambda_\varepsilon I)v_\varepsilon = -\varepsilon B v_\varepsilon, \quad \varphi \in \mathbf{J}_1,$$

we conclude that

$$\|(A-\lambda_\varepsilon I)v_\varepsilon\|_2 \le \varepsilon\|B\|_2\|v_\varepsilon\|_2 \le \varepsilon\|v_\varepsilon\|_2,$$

and from this, if λ_ε is not an eigenvalue of A,

$$\begin{aligned}\|(A-\lambda_\varepsilon I)^{-1}\|_2^{-1} &= \Big(\sup_{y\in\mathbb{K}^n}\frac{\|(A-\lambda_\varepsilon I)^{-1}y\|_2}{\|y\|_2}\Big)^{-1} = \Big(\sup_{x\in\mathbb{K}^n}\frac{\|x\|_2}{\|(A-\lambda_\varepsilon I)x\|_2}\Big)^{-1}\\ &= \inf_{x\in\mathbb{K}^n}\frac{\|(A-\lambda_\varepsilon I)x\|_2}{\|x\|_2} \le \frac{\|(A-\lambda_\varepsilon I)v_\varepsilon\|_2}{\|v_\varepsilon\|_2} \le \varepsilon.\end{aligned}$$

This implies the asserted estimate

$$\|(A-\lambda_\varepsilon I)^{-1}\|_2 \ge \varepsilon^{-1}. \tag{1.2.65}$$

ii) Next, we analyze the dependence of the eigenvalue λ_ε on ε in more detail. Subtracting the equation for v from that for v_ε, we obtain

$$A(v_\varepsilon - v) + \varepsilon B v_\varepsilon = (\lambda_\varepsilon - \lambda)v_\varepsilon + \lambda(v_\varepsilon - v).$$

Multiplying this by v^* yields

$$(A(v_\varepsilon - v), v^*)_2 + \varepsilon(Bv_\varepsilon, v^*)_2 = (\lambda_\varepsilon - \lambda)v_\varepsilon, v^*)_2 + \lambda(v_\varepsilon - v, v^*)_2$$

and, using the equation satisfied by v^*,

$$\varepsilon(Bv_\varepsilon, v^*)_2 = (\lambda_\varepsilon - \lambda)(v_\varepsilon, v^*)_2.$$

This yields $\lambda_\varepsilon = \lambda + \varepsilon\omega(\varepsilon)(Bv, v^*)_2$, where, observing $v_\varepsilon \to v$ and $(v, v^*) = 1$,

$$\omega(\varepsilon) := \frac{(Bv_\varepsilon, v^*)_2}{(v_\varepsilon, v^*)_2(Bv, v^*)_2} \to 1 \quad (\varepsilon \to 0).$$

iii) It remains to construct an appropriate perturbation matrix B. For convenience, we consider the renormalized dual eigenvectors $\tilde{v}^* := v^*\|v^*\|_2^{-1}$, satisfying $\|\tilde{v}^*\|_2 = 1$. With the vector $w := (v - \tilde{v}^*)\|v - \tilde{v}^*\|_2^{-1}$, we set for $\psi \in \mathbb{K}^n$:

$$S\psi := \psi - 2\,\mathrm{Re}\,(\psi, w)_2 w, \qquad B := -S.$$

The unitary matrix S acts like a Householder transformation mapping v into $\tilde{v}^*$ (s. the discussion in Section 2.3.1, below). In fact, observing $\|v\|_2 = \|\tilde{v}^*\|_2 = 1$, there holds

$$\begin{aligned} Sv &= v - \frac{2\,\mathrm{Re}\,(v, v - \tilde{v}^*)_2}{\|v - \tilde{v}^*\|_2^2}(v - \tilde{v}^*) = \frac{\{2 - 2\,\mathrm{Re}\,(v, \tilde{v}^*)_2\}v - 2\,\mathrm{Re}\,(v, v - \tilde{v}^*)_2(v - \tilde{v}^*)}{2 - 2\,\mathrm{Re}\,(v, \tilde{v}^*)_2} \\ &= \frac{2v - 2\,\mathrm{Re}\,(v, \tilde{v}^*)_2 v - 2v + 2\,\mathrm{Re}\,(v, \tilde{v}^*)_2 v + (2 - 2\,\mathrm{Re}\,(v, \tilde{v}^*)_2)\tilde{v}^*}{2 - 2\,\mathrm{Re}\,(v, \tilde{v}^*)_2} = \tilde{v}^*. \end{aligned}$$

This implies that

$$(Bv, v^*)_2 = -(Sv, v^*)_2 = -(\tilde{v}^*, v^*)_2 = -\|v^*\|_2.$$

Further, observing $\|w\|_2 = 1$ and

$$\|Sv\|_2^2 = \|v\|_2^2 - 2\,\mathrm{Re}\,(v, w)_2(v, w)_2 - 2\,\mathrm{Re}\,(v, w)_2(w, v)_2 + 4\,\mathrm{Re}\,(v, w)_2^2\|w\|_2^2 = \|v\|_2^2,$$

we have $\|B\|_2 = \|S\|_2 = 1$. Hence, for this particular choice of the matrix B, we have

$$\lambda_\varepsilon = \lambda - \varepsilon\omega(\varepsilon)\|v^*\|_2, \qquad \lim_{\varepsilon \to 0} \omega(\varepsilon) = 1,$$

as asserted. Q.E.D.

Remark 1.9: i) We note that the statement of Theorem 1.7 becomes trivial if the matrix A is *normal*. In this case primal and dual eigenvectors coincide and, in view of Remark 1.7, $\sigma_\varepsilon(A)$ is the union of ε-circles around its eigenvalues λ. Hence, observing $\|w^*\|_2 = \|w\|_2 = 1$ and setting $\omega(\varepsilon) \equiv 1$, we trivially have $\lambda_\varepsilon := \lambda - \varepsilon \in \sigma_\varepsilon(A)$ as asserted.

ii) If A is non-normal it may have a nontrivial pseudo-spectrum. Then, a large norm of the dual eigenfunction $\|w^*\|_2$ corresponding to a critical eigenvalue λ_{crit} with $-1 \ll \mathrm{Re}\,\lambda_{\mathrm{crit}} < 0$, indicates that the ε-pseudo-spectrum $\sigma_\varepsilon(A)$, even for small ε, reaches into the right complex half plane.

iii) If the eigenvalue $\lambda \in \sigma(A)$ considered in Theorem 1.7 is deficient, the normalization $(w, w^*)_2 = 1$ is not possible. In this case, as discussed above, there is still another mechanism for triggering nonlinear instability.

1.3 Perturbation theory and conditioning

First, we analyze the "conditioning" of quadratic linear systems. There are two main sources of errors in solving an equation $Ax = b$:

a) errors in the "theoretical" solution caused by errors in the data, i. e,, the elements of A and b,
b) errors in the "numerical" solution caused by round-off errors in the course of the solution process.

1.3.1 Conditioning of linear algebraic systems

We give an error analysis for linear systems

$$Ax = b \tag{1.3.66}$$

with regular coefficient matrix $A \in \mathbb{K}^{n\times n}$. The matrix A and the vector b are faulty by small errors δA and δb, so that actually the perturbed system

$$\tilde{A}\tilde{x} = \tilde{b}, \tag{1.3.67}$$

is solved with $\tilde{A} = A + \delta A$, $\tilde{b} = b + \delta b$ and $\tilde{x} = x + \delta x$. We want to estimate the error δx in dependence of δA and δb. For this, we use an arbitrary vector norm $\|\cdot\|$ and the associated natural matrix norm likewise denoted by $\|\cdot\|$.

Theorem 1.8 (Perturbation theorem): *Let the matrix $A \in \mathbb{K}^{n\times n}$ be regular and the perturbation satisfy $\|\delta A\| < \|A^{-1}\|^{-1}$. Then, the perturbed matrix $\tilde{A} = A + \delta A$ is also regular and for the resulting relative error in the solution there holds*

$$\frac{\|\delta x\|}{\|x\|} \le \frac{\mathrm{cond}(A)}{1 - \mathrm{cond}(A)\|\delta A\|/\|A\|} \left\{ \frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right\}, \tag{1.3.68}$$

with the so-called "condition number" $\mathrm{cond}(A) := \|A\|\,\|A^{-1}\|$ of the matrix A.

Proof. The assumptions imply

$$\|A^{-1}\delta A\| \le \|A^{-1}\|\,\|\delta A\| < 1,$$

such that also $A + \delta A = A[I + A^{-1}\delta A]$ is regular by Lemma 1.15. From

$$(A + \delta A)\tilde{x} = b + \delta b\,, \quad (A + \delta A)x = b + \delta A x$$

it follows that then for $\delta x = \tilde{x} - x$

$$(A + \delta A)\delta x = \delta b - \delta A x,$$

and consequently using the estimate of Lemma 1.15,

$$\begin{aligned}
\|\delta x\| &\le \|(A+\delta A)^{-1}\| \left\{\|\delta b\| + \|\delta A\|\,\|x\|\right\} \\
&= \|\big(A(I+A^{-1}\delta A)\big)^{-1}\| \left\{\|\delta b\| + \|\delta A\|\,\|x\|\right\} \\
&= \|(I+A^{-1}\delta A)^{-1}A^{-1}\| \left\{\|\delta b\| + \|\delta A\|\,\|x\|\right\} \\
&\le \|(I+A^{-1}\delta A)^{-1}\|\,\|A^{-1}\| \left\{\|\delta b\| + \|\delta A\|\,\|x\|\right\} \\
&\le \frac{\|A^{-1}\|}{1-\|A^{-1}\delta A\|} \left\{\|\delta b\| + \|\delta A\|\,\|x\|\right\} \\
&\le \frac{\|A^{-1}\|\,\|A\|\,\|x\|}{1-\|A^{-1}\|\,\|\delta A\|\,\|A\|\,\|A\|^{-1}} \left\{\frac{\|\delta b\|}{\|A\|\,\|x\|} + \frac{\|\delta A\|}{\|A\|}\right\}.
\end{aligned}$$

Since $\|b\| = \|Ax\| \le \|A\|\,\|x\|$ it eventually follows that

$$\|\delta x\| \le \frac{\mathrm{cond(A)}}{1-\mathrm{cond(A)}\|\delta \mathrm{A}\|\|\mathrm{A}\|^{-1}} \left\{\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|}\right\} \|x\|,$$

what was to be shown. Q.E.D.

The condition number $\mathrm{cond(A)}$ depends on the chosen vector norm in the estimate (1.3.68). Most often the max-norm $\|\cdot\|_\infty$ or the euclidian norm $\|\cdot\|_2$ are used. In the first case there holds

$$\mathrm{cond}_\infty(\mathrm{A}) := \|\mathrm{A}\|_\infty \|\mathrm{A}^{-1}\|_\infty$$

with the maximal row sum $\|\cdot\|_\infty$. Especially for Hermitian matrices Lemma 1.13 yields

$$\mathrm{cond}_2(\mathrm{A}) := \|\mathrm{A}\|_2 \|\mathrm{A}^{-1}\|_2 = \frac{|\lambda_{\max}|}{|\lambda_{\min}|}$$

with the eigenvalues $\lambda_{\max}$ and $\lambda_{\min}$ of A with largest and smallest modulus, respectively. Accordingly, the quantity $\mathrm{cond}_2(\mathrm{A})$ is called the "spectral condition (number)" of A. In the case $\mathrm{cond(A)}\|\delta \mathrm{A}\|\,\|\mathrm{A}\|^{-1} \ll 1$, the stability estimate (1.3.68) takes the form

$$\frac{\|\delta x\|}{\|x\|} \;\approx\; \mathrm{cond(A)} \left\{\frac{\|\delta \mathrm{b}\|}{\|\mathrm{b}\|} + \frac{\|\delta \mathrm{A}\|}{\|\mathrm{A}\|}\right\},$$

i. e., $\mathrm{cond(A)}$ is the amplification factor by which relative errors in the data A and b affect the relative error in the solution x.

Corollary 1.3: *Let the condition of A be of size $\mathrm{cond(A)} \sim 10^{\mathrm{s}}$. Are the elements of A and b faulty with a relative error if size*

$$\frac{\|\delta A\|}{\|A\|} \approx 10^{-k}, \quad \frac{\|\delta b\|}{\|b\|} \approx 10^{-k} \quad (k > s),$$

then the relative error in the solution can be at most of size of A and b faulty with a relative error if size

$$\frac{\|\delta x\|}{\|x\|} \approx 10^{s-k}.$$

In the case $\|\cdot\| = \|\cdot\|_\infty$, one may lose s decimals in accuracy.

Example 1.4: Consider the following coefficient matrix A and its inverse A^{-1}:

$$A = \begin{bmatrix} 1.2969 & 0.8648 \\ 0.2161 & 0.1441 \end{bmatrix}, \quad A^{-1} = 10^8 \begin{bmatrix} 0.1441 & -0.8648 \\ -0.2161 & 1.2969 \end{bmatrix}$$

$$\|A\|_\infty = 2.1617\,, \quad \|A^{-1}\|_\infty = 1.513 \cdot 10^8 \Rightarrow \mathrm{cond(A)} \approx 3.3 \cdot 10^8.$$

In solving the linear system $Ax = b$, one may lose 8 decimals in accuracy by which the elements a_{jk} and b_j are given. Hence, this matrix is very ill-conditioned.

Finally, we demonstrate that the stability estimate (1.3.68) is essentially sharp. Let A be a positive definite $n \times n$-matrix with smallest and largest eigenvalues λ_1 and λ_n and corresponding normalized eigenvectors w_1 and w_n, respectively. We choose

$$\delta A \equiv 0, \quad b \equiv w_n, \quad \delta b \equiv \varepsilon w_1 \, (\varepsilon \neq 0).$$

Then, the equations $Ax = b$ and $A\tilde{x} = b + \delta b$ have the solutions

$$x = \lambda_n^{-1} w_n\,, \quad \tilde{x} = \lambda_n^{-1} w_n + \varepsilon \lambda_1^{-1} w_1\,.$$

Consequently, for $\delta x = \tilde{x} - x$ there holds

$$\frac{\|\delta x\|_2}{\|x\|_2} = \varepsilon \frac{\lambda_n}{\lambda_1} \frac{\|w_1\|_2}{\|w_n\|_2} = \mathrm{cond}_2(\mathrm{A}) \frac{\|\delta \mathrm{b}\|_2}{\|\mathrm{b}\|_2},$$

i. e., in this very special case the estimate (1.3.68) is sharp.

1.3.2 Conditioning of eigenvalue problems

The most natural way of computing eigenvalues of a matrix $A \in \mathbb{K}^{n\times n}$ appears to go via its definition as zeros of the characteristic polynomial $\chi_A(\cdot)$ of A and to compute corresponding eigenvectors by solving the singular system $(A - \lambda I)w = 0$. This approach is not advisable in general since the determination of zeros of a polynomial may be highly ill-conditioned, at least if the polynomial is given in canonical form as sum of monomials. We will see that the determination of eigenvalues may be well- or ill-conditioned depending on the properties of A, i. e., its deviation from being "normal".

Example 1.5: A symmetric matrix $A \in \mathbb{R}^{20\times 20}$ with eigenvalues $\lambda_j = j$, $j = 1, \ldots, 20$, has the characteristic polynomial

$$\chi_A(z) = \prod_{j=1}^{20} (z - j) = z^{20} \underbrace{-210}_{b_1} z^{19} + \ldots + \underbrace{20!}_{b_{20}}.$$

The coefficient b_1 is perturbed: $\tilde{b}_1 = -210 + 2^{-23} \sim -210,000000119\ldots$, which results in

$$\text{relative error} \quad \left|\frac{\tilde{b}_1 - b_1}{b_1}\right| \sim 10^{-10}.$$

Then, the perturbed polynomial $\tilde{\chi}_A(z)$ has two roots $\lambda_\pm \sim 16.7 \pm 2.8i$, far away from the trues.

The above example shows that via the characteristic polynomial eigenvalues may be computed reliably only for very special matrices, for which $\chi_A(z)$ can be computed without determining its monomial from. Examples of some practical importance are, e. g., "tridiagonal matrices" or more general "Hessenberg[12] matrices".

$$\begin{bmatrix} a_1 & b_1 & & \\ c_2 & \ddots & \ddots & \\ & \ddots & & b_{n-1} \\ & & c_n & a_n \end{bmatrix} \qquad \begin{bmatrix} a_{11} & \cdots & & a_{1n} \\ a_{21} & \ddots & & \vdots \\ & \ddots & & a_{n-1,n} \\ 0 & & a_{n,n-1} & a_{nn} \end{bmatrix}$$

tridiagonal matrix Hessenberg matrix

Next, we provide a useful estimate which will be the basis for estimating the conditioning of the eigenvalue problem.

Lemma 1.18: *Let $A, B \in \mathbb{K}^{n\times n}$ be arbitrary matrices and $\|\cdot\|$ a natural matrix norm. Then, for any eigenvalue λ of A, which is not eigenvalue of B there holds*

$$\|(\lambda I - B)^{-1}(A - B)\| \geq 1. \tag{1.3.69}$$

Proof. If w is an eigenvector corresponding to the eigenvalue λ of A it follows that

$$(A - B)\, w = (\lambda I - B)\, w,$$

and for λ not being an eigenvalue of B,

$$(\lambda I - B)^{-1}(A - B)\, w = w.$$

Consequently

$$1 \leq \sup_{x\in\mathbb{K}^n\setminus\{0\}} \frac{\|(\lambda I - B)^{-1}(A - B)\, x\|}{\|x\|} = \|(\lambda I - B)^{-1}(A - B)\|,$$

what was to be shown. Q.E.D.

As consequence of Lemma 1.18, we obtain the following important inclusion theorem of Gerschgorin[13] (1931).

[12]Karl Hessenberg (1904–1959): German mathematician; dissertation "Die Berechnung der Eigenwerte und Eigenl"osungen linearer Gleichungssysteme", TU Darmstadt 1942.

[13]Semyon Aranovich Gershgorin (1901–1933): Russian mathematician; since 1930 Prof. in Leningrad (St. Petersburg); worked in algebra, complex function theory differential equations and numerics.

Theorem 1.9 (Theorem of Gerschgorin): *All eigenvalues of a matrix* $A \in \mathbb{K}^{n\times n}$ *are contained in the union of the corresponding "Gerschgorin circles"*

$$K_j := \Big\{ z \in \mathbb{C} : |z - a_{jj}| \le \sum_{k=1, k\neq j}^{n} |a_{jk}| \Big\}, \quad j = 1, \ldots, n. \tag{1.3.70}$$

If the sets $U \equiv \cup_{i=1}^{m} K_{j_i}$ *and* $V \equiv \overline{\cup_{j=1}^{n} K_j \setminus U}$ *are disjoint then* U *contains exactly* m *and* V *exactly* $n-m$ *eigenvalues of* A *(counted accordingly to their algebraic multiplicities).*

Proof. i) We set $B \equiv D = \mathrm{diag}(\mathrm{a_{jj}})$ in Lemma 1.18 and take the "maximal row sum" as natural matrix norm. Then, it follows that for $\lambda \neq a_{jj}$:

$$\|(\lambda I - D)^{-1}(A - D)\|_\infty = \max_{j=1,\ldots,n} \frac{1}{|\lambda - a_{jj}|} \sum_{k=1, k\neq j}^{n} |a_{jk}| \ge 1,$$

i. e., λ is contained in one of the Gerschgorin circles.

ii) For proving the second assertion, we set $A_t \equiv D + t(A - D)$. Obviously exactly m eigenvalues of $A_0 = D$ are in U and $n - m$ eigenvalues in V. The same then also follows for $A_1 = A$ since the eigenvalues of A_t (ordered accordingly to their algebraic multiplicities) are continuous functions of t. Q.E.D.

The theorem of Gerschgorin yields much more accurate information on the position of eigenvalues λ of A than the rough estimate $|\lambda| \le \|A\|_\infty$ derived above. The eigenvalues of the matrices A and $\bar{A}^T$ are related by $\lambda(\bar{A}^T) = \overline{\lambda(A)}$. By applying the Gerschgorin theorem simultaneously to A and $\bar{A}^T$, one may obtain a sharpening of the estimates for the eigenvalues.

Example 1.6: Consider the 3×3-matrix:

$$A = \begin{bmatrix} 1 & 0.1 & -0.2 \\ 0 & 2 & 0.4 \\ -0.2 & 0 & 3 \end{bmatrix} \quad \|A\|_\infty = 3.2\,, \quad \|A\|_1 = 3.6.$$

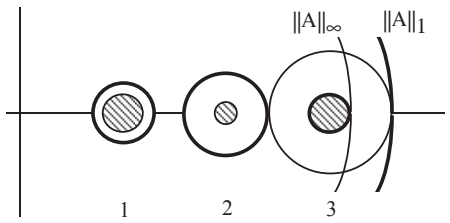

Figure 1.2: *Gerschgorin circles of* A *and* A^T

$$\begin{aligned} K_1 &= \{z \in \mathbb{C} : |z-1| \le 0.3\} & K_1^T &= \{z \in \mathbb{C} : |z-1| \le 0.2\} \\ K_2 &= \{z \in \mathbb{C} : |z-2| \le 0.4\} & K_2^T &= \{z \in \mathbb{C} : |z-2| \le 0.1\} \\ K_3 &= \{z \in \mathbb{C} : |z-3| \le 0.2\} & K_3^T &= \{z \in \mathbb{C} : |z-3| \le 0.6\} \end{aligned}$$

$$|\lambda_1 - 1| \le 0.2\,, \quad |\lambda_2 - 2| \le 0.1\,, \quad |\lambda_3 - 3| \le 0.2$$

Next, from the estimate of Lemma 1.18, we derive the following basic stability result for the eigenvalue problem.

Theorem 1.10 (Stability theorem): *Let $A \in \mathbb{K}^{n\times n}$ be a diagonalizable matrix, i. e., one for which n linearly independent eigenvectors $\{w^1, \ldots, w^n\}$ exist, and let $B \in \mathbb{K}^{n\times n}$ be an arbitrary second matrix. Then, for each eigenvalue $\lambda(B)$ of B there is a corresponding eigenvalue $\lambda(A)$ of A such that with the matrix $W = [w^1, \ldots, w^n]$ there holds*

$$|\lambda(A) - \lambda(B)| \;\le\; \mathrm{cond}_2(\mathrm{W})\, \|\mathrm{A} - \mathrm{B}\|_2\,. \tag{1.3.71}$$

Proof. The eigenvalue equation $Aw^i = \lambda_i(A)w^i$ can be rewritten in matrix form $AW = W\,\mathrm{diag}(\lambda_\mathrm{i}(\mathrm{A}))$ with the regular matrix $W = [w_1, \ldots, w_n]$. Consequently,

$$A = W\,\mathrm{diag}(\lambda_\mathrm{i}(\mathrm{A}))\,\mathrm{W}^{-1},$$

i. e., A is "similar" to the diagonal matrix $\Lambda = \mathrm{diag}(\lambda_\mathrm{i}(\mathrm{A}))$. Since $\lambda = \lambda(B)$ is not an eigenvalue of A,

$$\begin{aligned} \|(\lambda I - A)^{-1}\|_2 &= \|W(\lambda I - \Lambda)^{-1}W^{-1}\|_2 \\ &\le \|W^{-1}\|_2 \|W\|_2 \|(\lambda I - \Lambda)^{-1}\|_2 \\ &= \mathrm{cond}_2(\mathrm{W}) \max_{\mathrm{i}=1,\ldots,\mathrm{n}} |\lambda - \lambda_\mathrm{i}(\mathrm{A})|^{-1}. \end{aligned}$$

Then, Lemma 1.18 yields the estimate,

$$\begin{aligned} 1 &\le \|(\lambda I - A)^{-1}\,(B - A)\| \le \|(\lambda I - A)^{-1}\|\|(B - A)\| \\ &\le \mathrm{cond}_2(\mathrm{W}) \max_{\mathrm{i}=1,\ldots,\mathrm{n}} |\lambda - \lambda_\mathrm{i}(\mathrm{A})|^{-1}\|(\mathrm{B} - \mathrm{A})\|, \end{aligned}$$

from which the assertion follows. Q.E.D.

For Hermitian matrices $A \in \mathbb{K}^{n\times n}$ there exists an ONB in $\mathbb{K}^n$ of eigenvectors so that the matrix W in the estimate (1.3.71) can be assumed to be unitary, $W\bar{W}^T = I$. In this special case there holds

$$\mathrm{cond}_2(\mathrm{W}) = \|\bar{\mathrm{W}}^{\mathrm{T}}\|_2\, \|\mathrm{W}\|_2 = 1, \tag{1.3.72}$$

i. e., the eigenvalue problem of "Hermitian" (or more general "normal") matrices is well conditioned. For general "non-normal" matrices the conditioning of the eigenvalue problem may be arbitrarily bad, $\mathrm{cond}_2(\mathrm{W}) \gg 1$.

1.4 Exercises

Exercise 1.1 (Some useful inequalities):
Verify the following inequalities:

a) $ab \le \varepsilon a^2 + \frac{1}{4\varepsilon} b^2, \quad a, b \in \mathbb{R}, \ \ \varepsilon \in \mathbb{R}_+.$

b) $\left(\sum_{i=1}^n x_i \lambda_i \right)^{-1} \le \sum_{i=1}^n x_i^{-1} \lambda_i, \quad x_i \in \mathbb{R}_+, \ 0 \le \lambda_i \le 1, \ \sum_{i=1}^n \lambda_i = 1.$

c) $\max_{0 \le x \le 1} \left\{ x^2 (1-x)^{2n} \right\} \le (1+n)^{-2}.$

Exercise 1.2 (Some useful facts about norms and scalar products):
Verify the following claims for vectors $x, y \in \mathbb{R}^n$ and the Euclidean norm $\| \cdot \|_2$ and scalar product $(\cdot, \cdot)_2$:

a) $2\|x\|_2^2 + 2\|y\|_2^2 = \|x+y\|_2^2 + \|x-y\|_2^2$ (Parallelogram identity).

b) $|(x,y)_2| \le \|x\|_2 \|y\|_2$ (Schwarz inequality).

c) For any symmetric, positive definite matrix $A \in \mathbb{R}^{n \times n}$ the bilinear form $(x,y)_A := (Ax, y)_2$ is a scalar product. i) Can any scalar product on $\mathbb{R}^{n \times n}$ be written in this form? ii) How has this to be formulated for complex matrices $A \in \mathbb{C}^{n \times n}$?

Exercise 1.3 (Some useful facts about matrix norms):
Verify the following relations for matrices $A, B \in \mathbb{K}^{n \times n}$ and the Euclidean norm $\| \cdot \|_2$:

a) $\|A\|_2 := \max \left\{ \|Ax\|_2 / \|x\|_2, \ x \in \mathbb{K}^n, x \ne 0 \right\} = \max \left\{ \|Ax\|_2, \ x \in \mathbb{R}^n, \|x\|_2 = 1 \right\}.$

b) $\|Ax\|_2 \le \|A\|_2 \|x\|_2.$

c) $\|AB\|_2 \le \|A\|_2 \|B\|_2$ (Is this relation true for any matrix norm?).

d) For Hermitian matrices $A \in \mathbb{C}^{n \times n}$ there holds $\|A\|_2 = \max\{|\lambda|, \ \lambda \text{ eigenvalue of } A\}$.

e) For general matrices $A \in \mathbb{C}^{n \times n}$ there holds $\|A\|_2 = \max\{|\lambda|^{1/2}, \ \lambda \text{ eigenvalue of } \bar{A}^T A\}$.

Exercise 1.4 (Some useful facts about vector spaces and matrices):
a) Formulate the Gram-Schmidt algorithm for orthonormalizing a set of linearly independent vectors $\{x_1, \ldots, x_m\} \subset \mathbb{R}^n$:

b) How can one define the square root $A^{1/2}$ of a symmetric, positive definite matrix $A \in \mathbb{R}^{n \times n}$?

c) Show that a positive definite matrix $A \in \mathbb{C}^{n \times n}$ is automatically Hermitian, i. e., $A = \bar{A}^T$. This is not necessarily true for real matrices $A \in \mathbb{R}^{n \times n}$, i. e., for real matrices the definition of positiveness usually goes together with the requirement of symmetry.

Exercise 1.5: Recall the definitions of the following quantities:

a) The "maximum-norm" $\| \cdot \|_\infty$ and the "l_1-norm" $\| \cdot \|_1$ on $\mathbb{K}^n$.

b) The "spectrum" $\Sigma(A)$ of a matrix $A \in \mathbb{K}^{n \times n}$.

c) The "Gerschgorin circles" $K_i \subset \mathbb{C}$, $i = 1, \ldots, n$, of a matrix $A \in \mathbb{K}^{n \times n}$.

d) The "spectral radius" $\rho(A)$ of a matrix $A \in \mathbb{K}^{n\times n}$.

e) The "spectral condition number" $\kappa_2(A)$ of a matrix $A \in \mathbb{K}^{n\times n}$.

Exercise 1.6: Recall the proofs of the following facts about matrices:

a) The diagonal elements of a (Hermitian) positive definite matrix $A \in \mathbb{K}^{n\times n}$ are real and positive.

b) For the trace $\mathrm{tr}(A) := \sum_{i=1}^n a_{ii}$ of a Hermitian matrix $A \in \mathbb{K}^{n\times n}$ with eigenvalues $\lambda_i \in \Sigma(A)$ there holds

$$\mathrm{tr}(A) = \sum_{i=1}^n \lambda_i.$$

c) A strictly diagonally dominant matrix $A \in \mathbb{K}^{n\times n}$ is regular. If it is also Hermitian with (real) positive diagonal entries, then it is positive definite.

Exercise 1.7: Let $B \in \mathbb{K}^{n\times n}$ be a matrix, which for some matrix norm $\|\cdot\|$ satisfies $\|B\| < 1$. Prove that the matrix $I - B$ is regular with inverse satisfying

$$\|(I-B)^{-1}\| \le \frac{1}{1-\|B\|}\,.$$

Exercise 1.8: Prove that each connected component of k Gerschgorin circles (that are disjoined to all other $n-k$ circles) of a matrix $A \in \mathbb{C}^{n\times n}$ contains exactly k eigenvalues of A (counted accordingly to their algebraic multiplicities). This implies that such a matrix, for which all Gerschgorin circles are mutually disjoint, has exactly n simple eigenvalues and is therefore diagonalizable.

Exercise 1.9: Let $A, B \in \mathbb{K}^{n\times n}$ be two Hermitian matrices. Then, the following statements are equivalent:

i) A and B commute, i. e., $AB = BA$.

ii) A and B possess a common basis of eigenvectors.

iii) AB is Hermitian.

Does the above equivalence in an appropriate sense also hold for two general "normal" matrices $A, B \in \mathbb{K}^{n\times n}$, i. e., if $\bar{A}^T A = A\bar{A}^T$ and $\bar{B}^T B = B\bar{B}^T$?

Exercise 1.10: A "'sesquilinear form"' on $\mathbb{K}^n$ is a mapping $\varphi(\cdot,\cdot) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{K}$, which is bilinear in the following sense:

$$\varphi(\alpha x + \beta y, z) = \bar{\alpha}\varphi(x,z) + \bar{\beta}\varphi(y,z), \quad \varphi(z, \alpha x + \beta y) = \alpha\varphi(z,x) + \beta\varphi(z,y), \quad \alpha,\beta \in \mathbb{K}.$$

i) Show that for any regular matrix $A \in \mathbb{K}^{n\times n}$ the sesquilinear form $\varphi(x,y) := (Ax, Ay)_2$ is a scalar product on $\mathbb{K}^n$.

ii) In an earlier exercise, we have seen that each scalar product (x,y) on $\mathbb{K}^n$ can be written

in the form $(x, y) = (x, Ay)_2$ with a (Hermitian) positive definite matrix $A \in \mathbb{K}^{n\times n}$. Why does this statement not contradict (i)?

Exercise 1.11: Let $A \in \mathbb{K}^{n\times n}$ be Hermitian.

i) Show that eigenvectors corresponding to different eigenvalues $\lambda_1(A)$ and $\lambda_2(A)$ are orthogonal. Is this also true for (non-Hermitian) "normal" matrices, i. e., if $\bar{A}^T A = A\bar{A}^T$?

ii) Show that there holds

$$\lambda_{\min}(A) = \min_{x\in\mathbb{K}^n\setminus\{0\}} \frac{(Ax,x)_2}{\|x\|_2^2} \le \max_{x\in\mathbb{K}^n\setminus\{0\}} \frac{(Ax,x)_2}{\|x\|_2^2} = \lambda_{\max}(A),$$

where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimal and maximal (real) eigenvalues of A, respectively. (Hint: Use that a Hermitian matrix possesses an ONB of eigenvectors.)

Exercise 1.12: Let $A \in \mathbb{K}^{n\times n}$ and $0 \notin \sigma(A)$. Show that the ε-pseudo-spectra of A and that of its inverse A^{-1} are related by

$$\sigma_\varepsilon(A) \subset \big\{z \in \mathbb{C}\setminus\{0\} \,\big|\, z^{-1} \in \sigma_{\delta(z)}(A^{-1})\big\} \cup \{0\},$$

where $\delta(z) := \varepsilon\|A^{-1}\|/|z|$ and, for $0 < \varepsilon < 1$, by

$$\sigma_\varepsilon(A^{-1}) \setminus B_1(0) \subset \big\{z \in \mathbb{C}\setminus\{0\} \,\big|\, z^{-1} \in \sigma_\delta(A)\big\},$$

where $B_1(0) := \{z \in \mathbb{C},\ \|z\| \le 1\}$ and $\delta := \varepsilon/(1-\varepsilon)$.