March 26, 2018

Lecture 1 – Basic Notions

A Norms

Let $\mathcal{V}$ be a (complex) linear space.
A map $\|\cdot\| : \mathcal{V} \to [0,\infty)$ is a **norm** on $\mathcal{V}$ and $(\mathcal{V}, \|\cdot\|)$ is a **normed space** if

$$\begin{array}{lll} 1) & \|\mathbf{x}\| = 0 \quad \Leftrightarrow \quad \mathbf{x} = \mathbf{0} & (\mathbf{x} \in \mathcal{V}) \\ 2) & \|\alpha\mathbf{x}\| = |\alpha|\,\|\mathbf{x}\| & (\mathbf{x} \in \mathcal{V},\, \alpha \in \mathbb{C}) \\ 3) & \|\mathbf{x}+\mathbf{y}\| \le \|\mathbf{x}\| + \mathbf{y}\| & (\mathbf{x}, \mathbf{y} \in \mathcal{V}) \end{array} \tag{1.1}$$

Norms are used to measure errors (approximation errors as well as errors coming from rounded arithmetic).

A sequence $(\mathbf{x}_n)$ in $\mathcal{V}$ **converges** to $\mathbf{x} \in \mathcal{V}$, if

$$\|\mathbf{x}_n - \mathbf{x}\| \to 0 \quad \text{if} \quad n \to \infty.$$

Formally, we should say that the sequence converges with respect to the norm $\|\cdot\|$. However, for convergence, it does not matter what norm is used if $\mathcal{V}$ is finite dimensional.

Theorem 1.1 *If $\mathcal{V}$ is finite dimensional, then all norms on $\mathcal{V}$ are* **equivalent***, i.e., if $\|\cdot\|$ and $\left|\cdot\right|$ are norms on $\mathcal{V}$, then there are constants M, m, $M > m > 0$ such that*

$$m\left|\mathbf{x}\right| \le \|\mathbf{x}\| \le M\left|\mathbf{x}\right| \qquad (\mathbf{x} \in \mathcal{V}).$$

Exercise 1.1. ***Proof of Theorem*** 1.1**.** Let $\mathcal{V}$ be finite dimensional with norm $\|\cdot\|$. Let $\mathbf{v}_1, \ldots, \mathbf{v}_k$ be a basis.
Define $\left|\sum_j \alpha_j \mathbf{v}_j\right|_\infty \equiv \|(\alpha_1, \ldots, \alpha_j)^{\mathrm{T}}\|_\infty \equiv \max_j |\alpha_j|$.

(a) Show that $\left|\cdot\right|_\infty$ is a norm on $\mathcal{V}$.

(b) Prove that $\|\mathbf{x}\| \le M\left|\mathbf{x}\right|_\infty$ $(\mathbf{x} \in \mathcal{V})$ for some M $(\le \sum_j \|\mathbf{v}_j\|)$.

(c) Prove that $\mathcal{S} \equiv \{(\alpha_1, \ldots, \alpha_k)^{\mathrm{T}} \in \mathbb{C}^k \,\big|\, \|\sum_j \alpha_j \mathbf{v}_j\| = 1\}$ is a closed bounded subset of $\mathbb{C}^k$. Conclude that

$$0 < K \equiv \operatorname{argmin}\{\max_j |\alpha_j| \,\big|\, (\alpha_1, \ldots, \alpha_k)^{\mathrm{T}} \in \mathcal{S}\}$$

and, therefore, with $m \equiv 1/K$, we have $m\left|\mathbf{x}\right|_\infty \le \|\mathbf{x}\|$.

(d) Prove Theorem 1.1.

In practice, it is often said that a sequence $(\mathbf{x}_n)$ (of finitely many $\mathbf{x}_n$) is converging to $\mathbf{x}$ if for some (large) n, $\|\mathbf{x}_n - \mathbf{x}\| < \mathrm{tol}$, where tol is some prescribed tolerance (or accuracy). Now, the accuracy depends on the norm that is used (the M and m affect the actual value of the error bound).

Below $p, q \in [1,\infty]$. The important cases are $p, q \in \{1, 2, \infty\}$.
$\mathbf{x} = (x_1, \ldots, x_k)^{\mathrm{T}}$ is a k-vector. The p-**norm** $\|\mathbf{x}\|_p$ is defined by

$$\|\mathbf{x}\|_p \equiv \sqrt[p]{\sum |x_i|^p} \quad (p \in [1,\infty)), \qquad \|\mathbf{x}\|_\infty \equiv \max_i |x_i|.$$

Property 1.2 *$\|\cdot\|_p$ defines a norm on $\mathcal{V} = \mathbb{C}^k$ (whence, on $\mathcal{V} = \mathbb{C}^k$) for each $p \in [1,\infty]$.*

For a proof, see Exercise 1.2.(c).

The following "duality relation" between p-norms can be usefull:

Property 1.3 *For all $p \in [1,\infty]$ and $p' \in [1,\infty]$ such that $\frac{1}{p} + \frac{1}{p'} = 1$, we have that*

$$[\textbf{Hölder's inequality}] \qquad |(\mathbf{x},\mathbf{y})| \le \|\mathbf{x}\|_p \, \|\mathbf{y}\|_{p'} \quad (\mathbf{x},\mathbf{y} \in \mathbb{C}^k) \tag{1.2}$$

and

$$\|\mathbf{x}\|_p = \sup\{|(\mathbf{x},\mathbf{y})| \,\Big|\, \|\mathbf{y}\|_{p'} \le 1\} \quad (\mathbf{x} \in \mathbb{C}^k). \tag{1.3}$$

Since $p' = 2$ if $p = 2$, Hölder's inequality in (1.2) *can be viewed as a generalisation of*

$$[\textbf{Cauchy–Schwartz inequality}] \quad |(\mathbf{x},\mathbf{y})| \le \|\mathbf{x}\|_2 \, \|\mathbf{y}\|_2 \quad (\mathbf{x},\mathbf{y} \in \mathbb{C}^k). \tag{1.4}$$

The 2-norm is important for mathematical reasons: it is associated to the inner product $(\mathbf{x},\mathbf{y}) \equiv \mathbf{y}^{\mathrm{H}}\mathbf{x}$ and results as Cauchy-Schwartz (see, (1.4)) and Pythagoras (see, Exercise 1.2.(f)) can be used. Other norms (as for $p = 1$ and $p = \infty$) are easier to compute and are frequently used in error analysis. Then, precise values for the equivalence constants of Theorem 1.1 are of importance and the following result can be useful.

Property 1.4 *For $\mathbf{x} \in \mathbb{C}^k$ and $p, q \in [1,\infty]$ we have that*

$$\|\mathbf{x}\|_p \le k^{\frac{1}{q}-\frac{1}{q}} \, \|\mathbf{x}\|_q \quad \textit{if} \quad p \le q \quad \textit{and} \quad \|\mathbf{x}\|_p \le \|\mathbf{x}\|_q \quad \textit{if} \quad p \ge q. \tag{1.5}$$

In particular,

$$\|\mathbf{x}\|_1 \le \sqrt{k}\, \|\mathbf{x}\|_2, \quad \|\mathbf{x}\|_2 \le \sqrt{k}\, \|\mathbf{x}\|_\infty \quad \textit{and} \quad \|\mathbf{x}\|_\infty \le \|\mathbf{x}\|_2 \le \|\mathbf{x}\|_1. \tag{1.6}$$

Exercise 1.2. **Vector norms.**

(a) Prove (1.3). (Hint: for a given $\mathbf{x} \in \mathbb{C}^k$, argue that the supremum in (1.3) is a maximum and that, for the maximising $\mathbf{y}^{(m)}$, you may assume that, without loss of generality, $x_i, y_i^{(m)} \in (0,\infty)$ all $i = 1,\ldots,k$. Now use Lagragian multipliers: with $f(\mathbf{y}) \equiv (\mathbf{x},\mathbf{y})$ and $g(\mathbf{y}) \equiv \|\mathbf{y}\|_{p'}^{p'}$ there is a $\lambda \in \mathbb{R}$ such that $\nabla f(\mathbf{y}) = \lambda \nabla g(\mathbf{y})$ for $\mathbf{y} = \mathbf{y}^{(m)}$.)

An alternative proof of (1.4) can be obtained by using a decomposition of $\mathbf{y}$ into a sum of a multiple of $\mathbf{x}$ and of a vector orthogonal to $\mathbf{x}$.

(b) Prove (1.2).

(c) Use (1.3) to show that $\|\cdot\|_p$ is norm.

(d) Prove (1.5) and (1.6).

(e) Show that $\|\mathbf{x}\|_\infty = \lim_{p\to\infty} \|\mathbf{x}\|_p$

(f) Prove **Pythagoras' Theorem**:

$$\mathbf{x} \perp \mathbf{y} \quad \Rightarrow \quad \|\mathbf{x}+\mathbf{y}\|_2^2 = \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 \qquad (\mathbf{x},\mathbf{y} \in \mathbb{C}^k). \tag{1.7}$$

(g) Show that the estimates in (1.5) and (1.6) are sharp, that is, give for each of the inequalities, a non-trivial vector $\mathbf{x}$ for which the inequality is an equality.

(h) Note that $\|\mathbf{x}\|_p$ can also be defined for $p \in (0,1)$: $\|\mathbf{x}\|_p \equiv \sqrt[p]{\sum |x_i|^p}$.
Show that, for these p, $\|\cdot\|_p$ does **not** define a norm.

(i) Sometimes $\|\mathbf{x}\|_0$ is used to denote the number of non-zero coordinates of $\mathbf{x}$, i.e.,

$$\|\mathbf{x}\|_0 \equiv \#\{i \in \{1,\ldots,k\} \,\Big|\, x_i \neq 0\}.$$

Show that $\|\mathbf{x}\|_0 = \lim_{p>0,p\to 0} \|\mathbf{x}\|_p^p$. Do we have that $\|\mathbf{x}\|_0 = \lim_{p>0,p\to 0} \|\mathbf{x}\|_p$?

(j) Sketch, for $k = 2$, the unit balls $\{\mathbf{x} \in \mathbb{R}^2 \,\Big|\, \|\mathbf{x}\|_p \le 1\}$ for $p \in \{\infty, 2, 1, \frac{1}{2}\}$ (and for $p = 0$?). For what values of p is this ball a convex set?

Below $\mathbf{A} = (A_{ij})$ is an $n \times k$ matrix, $|\mathbf{A}|$ is the matrix $(|A_{ij}|)$.
Norms on $\mathbb{C}^n$ (or on $\mathbb{R}^n$ if $\mathbf{A}$ is real) induce norms on matrices. The **induced p-norm** is defined by[1]

$$\|\mathbf{A}\|_p \equiv \sup\{\|\mathbf{A}\mathbf{x}\|_p \,\big|\, \|\mathbf{x}\|_p \leq 1\}.$$

Convention 1.5 If we use the same notation $\|\cdot\|$ for a norm on vector spaces as $\mathbb{R}^n$ (or $\mathbb{C}^n$) and on matrices of matching size, then we assume that the norm on matrices is **induced** by the norm on vectors:

$$\|\mathbf{A}\| \equiv \sup \|\mathbf{A}\mathbf{x}\|$$

with supremum over all $\mathbf{x}$ with $\|\mathbf{x}\| \leq 1$.

Non-induced norms are also frequently used, as the **Frobenius norm** $\|\mathbf{A}\|_{\mathrm{F}}$ and the norm $\|\mathbf{A}\|_{\mathrm{M}}$

$$\|\mathbf{A}\|_{\mathrm{F}} \equiv \sqrt{\sum_{i,j} |A_{ij}|^2} = \sqrt{\sum_j \|\mathbf{A}\mathbf{e}_j\|_2^2}, \qquad \|\mathbf{A}\|_{\mathrm{M}} \equiv \max_{i,j} |A_{ij}|.$$

An $n \times k$ matrix $\mathbf{A}$ can be associated to an nk-vector $\mathbf{A}^{\downarrow}$ by putting the consecutive columns of $\mathbf{A}$ below each others. Note that $\|\mathbf{A}\|_{\mathrm{F}} = \|\mathbf{A}^{\downarrow}\|_2$, $\|\mathbf{A}\|_{\mathrm{M}} = \|\mathbf{A}^{\downarrow}\|_\infty$.

Exercise 1.3.

(a) **Norms.** Prove that $\|\cdot\|_p$, $\|\cdot\|_{\mathrm{F}}$ and $\|\cdot\|_{\mathrm{M}}$ are norms on the space of $n \times k$ matrices.

(b) **Multiplicativity.** Let $\mathbf{A}$ be an $n \times k$ matrix and $\mathbf{B}$ a $k \times m$ matrix. Prove that p-norms (or, more generally, induced norms) and the Frobenius norm are **multiplicative**:[2]

$$\|\mathbf{A}\mathbf{B}\|_p \leq \|\mathbf{A}\|_p \|\mathbf{B}\|_p \quad (p \in [1,\infty]), \qquad \|\mathbf{B}\|_{\mathrm{F}} \leq \|\mathbf{A}\|_2 \|\mathbf{B}\|_{\mathrm{F}} \leq \|\mathbf{A}\|_{\mathrm{F}} \|\mathbf{B}\|_{\mathrm{F}}. \tag{1.8}$$

Here, you can use that $\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_{\mathrm{F}}$ (see (1.14)). Is the M-norm $\|\cdot\|_M$ also multiplicative?

(c) Prove that, for any square matrix $\mathbf{A}$, and any induced norm, we have that

$$\rho(\mathbf{A}) \leq \|\mathbf{A}\|. \tag{1.9}$$

Here $\rho(\mathbf{A})$ is the spectral radius of $\mathbf{A}$ (cf., p.11).

From the Theorem 1.6 below, we can conclude that (1.9) even holds for any multiplicative matrix norm (in particular for the Frobenius norm).

Exercise 1.4.

(a) Prove that for any induced norm $\|\cdot\|$, we have that

$$\|\mathbf{A}\| = \|\mathbf{A}\mathbf{x}\| \quad \text{for some vector} \quad \mathbf{x} \quad \text{with} \quad \|\mathbf{x}\| = 1.$$

(Hint. Use the fact that matrix-vector multiplication is continuous (why?), and the unit ball $\{\mathbf{x} \in \mathbb{C}^k \,\big|\, \|\mathbf{x}\| \leq 1\}$ is closed and bounded (why?).)

(b) **Diagonal matrices.** Let $\mathbf{D} = \mathrm{diag}(D_i)$ be an $n \times n$ diagonal matrix.
Show that for any induced p-norm ($1 \leq p < \infty$) we have that

$$\|\mathbf{D}\|_p = \max_i |D_i|.$$

(c) **1- and ∞-norms.** Prove that

$$\|\mathbf{A}\|_1 = \max_j \sum_i |A_{ij}| = \max_j \|\mathbf{A}\mathbf{e}_j\|_1, \qquad \|\mathbf{A}\|_\infty = \max_i \sum_j |A_{ij}| = \max_i \|\mathbf{e}_i^* \mathbf{A}\|_1. \tag{1.10}$$

[1] More generally, a norm $\|\cdot\|$ on $\mathbb{C}^n$ and a norm $|\cdot|$ on $\mathbb{C}^k$ induce a norm on the space of $n \times k$ matrices by $\sup\{\|\mathbf{A}\mathbf{x}\| \,\big|\, \mathbf{x} \in \mathbb{C}^k, |\mathbf{x}| \leq 1\}$. However, to keep notation simple, we will not pursue this generalisation here.

[2] A norm $\|\cdot\|$ on the space of $n \times n$ matrices is multiplicative if $\|\mathbf{A}\mathbf{B}\| \leq \|\mathbf{A}\| \|\mathbf{B}\|$ for all $n \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$.

(d) **Duality.** Prove that for $p' \in [1,\infty]$ such that $\frac{1}{p} + \frac{1}{p'} = 1$, we have

$$\|\mathbf{A}\|_p = \|\mathbf{A}^{\mathrm{H}}\|_{p'}, \quad \text{in particular} \quad \|\mathbf{A}\|_2 = \|\mathbf{A}^{\mathrm{H}}\|_2. \tag{1.11}$$

Here, you can use (1.2).

(e) **Equivalence.** For $q \in [p,\infty]$, prove that

$$\|\mathbf{x}\|_q \le \|\mathbf{x}\|_p \quad \text{and} \quad \|\mathbf{x}\|_p \le \kappa\|\mathbf{x}\|_q \quad \text{where} \quad \kappa \equiv k^{\frac{1}{p}-\frac{1}{q}}$$

and for all $q \in [1,\infty]$,

$$\|\mathbf{A}\|_q \le \kappa\|\mathbf{A}\|_p \quad \text{where} \quad \kappa \equiv k^{|\frac{1}{p}-\frac{1}{q}|}.$$

Theorem 1.6 *Let* $\mathbf{A}$ *be an* $n \times n$ *matrix with spectral radius* $\rho(\mathbf{A})$ *(cf., p.11). If* $\mathbf{A}$ *is normal (*$\mathbf{A}^{\mathrm{H}}\mathbf{A} = \mathbf{A}\mathbf{A}^{\mathrm{H}}$*, in particular, if* $\mathbf{A}$ *Hermitian), then*

$$\|\mathbf{A}\|_2 = \rho(\mathbf{A}). \tag{1.12}$$

For any norm $\|\cdot\|$ *on the space of all* $n \times n$ *matrices, we have*

$$\rho(\mathbf{A}) = \lim_{j\to\infty} \sqrt[j]{\|\mathbf{A}^j\|}. \tag{1.13}$$

Exercise 1.5. ***Proof of Theorem*** 1.6**.**

(a) Prove (1.12) for normal matrices $\mathbf{A}$,

To prove (1.13), put $\widetilde{\rho}(\mathbf{A}) \equiv \lim_{j\to\infty} \sqrt[j]{\|\mathbf{A}^j\|}$ for the right hand side expression in (1.13). Let $\mathbf{A}$ be an $n \times n$ matrix.

(b) Use Theorem 1.1 to prove that $\widetilde{\rho}(\mathbf{A})$ does not depend on the norm. Take $\|\cdot\| = \|\cdot\|_\infty$.

(c) Let $\mathbf{x}$ be an eigenvector associated to the absolute largest eigenvalue, scaled such that $\|\mathbf{x}\|_\infty = 1$. Show that $\rho(\mathbf{A}) = \sqrt[j]{\|\mathbf{A}^j\mathbf{x}\|_\infty}$ and conclude that $\rho(\mathbf{A}) \le \widetilde{\rho}(\mathbf{A})$.

(d) Show that $\widetilde{\rho}(\mathbf{A}) = \widetilde{\rho}(\mathbf{T}^{-1}\mathbf{A}\mathbf{T})$ for any non-singular matrix $\mathbf{T}$.

(e) Let $\varepsilon > 0$. Show there is a non-singular matrix $\mathbf{T}$ such that $\mathbf{J} \equiv \mathbf{T}^{-1}\mathbf{A}\mathbf{T}$ is on Jordan normal form be it that the $(i, i+1)$ entries of the Jordan blocks J_λ are ε rather than 1 (cf. Theorem 0.7). Show that $\|J_\lambda\|_\infty \le |\lambda| + \varepsilon$. Conclude that $\widetilde{\rho}(\mathbf{A}) \le \rho(\mathbf{A}) + \varepsilon$. Prove (1.13).

Exercise 1.6. Let $\mathbf{A}$ be an $n \times k$ matrix.

(a) **2-norm.** Show (from the definitions) that

$$\begin{array}{ll} \text{(i)} & \|\mathbf{A}\|_2 = \sqrt{\|\mathbf{A}^{\mathrm{H}}\mathbf{A}\|_2}, \\ \text{(ii)} & \|\mathbf{A}\|_2 = \|\mathbf{A}^{\mathrm{H}}\|_2, \\ \text{(iii)} & \|\mathbf{A}\|_2 \le \|\mathbf{A}\|_{\mathrm{F}}, \\ \text{(iv)} & \|\mathbf{A}\|_{\mathrm{F}} \le \sqrt{k}\|\mathbf{A}\|_2, \\ \text{(v)} & \|\mathbf{A}\|_{\mathrm{F}} = \sqrt{\mathrm{trace}(\mathbf{A}^{\mathrm{H}}\mathbf{A})}. \end{array} \tag{1.14}$$

Show that the inequalities (iii) and (iv) are sharp, that is, give a non-trivial $\mathbf{A}$ (i.e., $\mathbf{A} \neq \mathbf{0}$) such that $\|\mathbf{A}\|_2 = \|\mathbf{A}\|_{\mathrm{F}}$ and a(nother) non-trivial $\mathbf{A}$ such $\|\mathbf{A}\|_{\mathrm{F}} = \sqrt{k}\,\|\mathbf{A}\|_2$.

Show also (here you can use (1.12)) that

$$\|\mathbf{A}\|_2 \le \sqrt{\|\mathbf{A}\|_1\,\|\mathbf{A}\|_\infty}. \tag{1.15}$$

(b) Prove that orthonormal transformations preserve the 2-norm, that is, if $\mathbf{A}$ is orthonormal, then $\|\mathbf{A}\mathbf{B}\|_2 = \|\mathbf{B}\|_2$. Similarly, if $\mathbf{B}^*$ is orthonormal then $\|\mathbf{A}\mathbf{B}\|_2 = \|\mathbf{A}\|_2$. Do we also have that $\|\mathbf{A}\mathbf{B}\|_2 = \|\mathbf{A}\|_2$ in case $\mathbf{B}$ is orthonormal?

Do orthonormal matrices also preserve the p-norm for $p \in [1,\infty], p \neq 2$?

In estimates of effects of rounding errors involving an $n \times k$ matrix $\mathbf{A} = (A_{ij})$, the matrix $|\mathbf{A}| \equiv (|A_{ij}|)$ shows up, cf., the Sections E and F below. For instance, the vector $\mathbf{b}$ that we obtain from the matrix vector multiplication $\mathbf{Ax}$, $\mathbf{b} = \mathbf{Ax}$, using rounded arithmetic (that is, the computer result) is equal to the *exact* matrix vector multiplication $(\mathbf{A} + \Delta)\mathbf{b}$ for some $n \times k$ 'perturbation' matrix Δ with such that $|\Delta| \leq p_A \mathbf{u} |\mathbf{A}|$, where the inequality is entry-wise, $\mathbf{u}$ is the relative machine precision (in Matlab $\mathbf{u} = 0.5 * \texttt{eps} = 0.5\,10^{-16}$; see Exercise 1.20 below), and p_A is the maximum non-zeros per row of $\mathbf{A}$. Δ is a **perturbation** of $\mathbf{A}$. Estimates of the size of Δ in terms of $\mathbf{A}$ will involve $\| \, |\mathbf{A}| \, \|$, e.g., $\|\Delta\|_2 \leq \| \, |\Delta| \, \|_2 \leq p_A \, \mathbf{u} \, \| \, |\mathbf{A}| \, \|_2$ if $|\Delta| \leq p_A \, \mathbf{u} \, |\mathbf{A}|$ (inequalities matrix entry wise). The following exercise relates $\|\mathbf{A}\|$ and $\| \, |\mathbf{A}| \, \|$.

Exercise 1.7. **The norm of $|\mathbf{A}|$.** Let $\mathbf{A}$ be an $n \times k$ matrix.

(a) Prove that

$$\|\mathbf{A}\|_p \leq \| \, |\mathbf{A}| \, \|_p \quad (p \in [1, \infty]), \qquad \| \, |\mathbf{A}| \, \|_{\mathrm{F}} = \|\mathbf{A}\|_{\mathrm{F}}, \quad \| \, |\mathbf{A}| \, \|_2 \leq \sqrt{k} \|\mathbf{A}\|_2. \tag{1.16}$$

Prove also, that for $p \in \{1, \infty\}$, $\|\mathbf{A}\|_p = \| \, |\mathbf{A}| \, \|_p$. May we expect that $\| \, |\mathbf{A}| \, \|_2 = \|\mathbf{A}\|_2$?

(b) Prove that

$$\| \, |\mathbf{A}^{\mathrm{H}}| \, |\mathbf{A}| \, \|_2 \leq \min(\sqrt{k}, \sqrt{n}) \|\mathbf{A}^{\mathrm{H}} \mathbf{A}\|_2. \tag{1.17}$$

In particular, if $\mathbf{A} = \mathbf{C}\mathbf{C}^{\mathrm{H}}$ for an $n \times n$ matrix $\mathbf{C}$ ($\mathbf{A}$ is $n \times n$ and positive definite), then $\| \, |\mathbf{C}| \, |\mathbf{C}^{\mathrm{H}}| \, \|_2 \leq \sqrt{n} \|\mathbf{A}\|_2$ (see Exercise 2.21).

(c) **If A is sparse.** Put $p_c \equiv \max_j \#\{i \, \big| \, A_{ij} \neq 0\}$ and $p_r \equiv \max_i \#\{j \, \big| \, A_{ij} \neq 0\}$, the maximum number of non-zeros per column, and per row, respectively. We will prove that

$$\| \, |\mathbf{A}| \, \|_2 \leq \sqrt{\min(p_r, p_c)} \, \|\mathbf{A}\|_2. \tag{1.18}$$

Prove that, for all k-vectors $\mathbf{x} = (x_1, \ldots, x_k)^{\mathrm{T}}$ with $\|\mathbf{x}\|_2 = 1$,

$$\| \, |\mathbf{A}|\mathbf{x}\|_2^2 = \sum_{i=1}^{n} |(\mathbf{x}, |\mathbf{A}|^{\mathrm{H}} \mathbf{e}_i)|^2 \leq p_c \max_i \|\mathbf{A}^{\mathrm{H}} \mathbf{e}_i\|_2^2$$

(Hint: first show that $\sum_{i=1}^{n} \sum_{j, A_{ij} \neq 0} |x_j|^2 \leq p_c \sum_{i=1}^{n} |x_i|^2 = p_c$). Conclude that

$$\| \, |\mathbf{A}| \, \|_2 \leq \sqrt{p_c} \max_i \|\mathbf{A}^{\mathrm{H}} \mathbf{e}_i\|_2 \quad \text{and} \quad \| \, |\mathbf{A}| \, \|_2 \leq \sqrt{p_r} \max_j \|\mathbf{A}\mathbf{e}_j\|_2$$

and that (1.18) is correct.

The theorem below summerizes the main results in this section.

Theorem 1.7 *Here, $\mathbf{A}$, $\mathbf{A}_1$ and $\mathbf{A}_2$ are $n \times k$ matrices, $\mathbf{B}$ is a $k \times m$ matrix, $\mathbf{x}$ is a k-vector, and α is a scalar. Let $\| \cdot \|$ be a norm on matrices induced by a vector norm $\| \cdot \|$.*
Then, $\| \cdot \|$ is a norm on the space of $n \times n$ matrices, i.e.,

1) $\|\mathbf{A}\| = 0 \quad \Leftrightarrow \quad \mathbf{A} = \mathbf{0}$,
2) $\|\alpha \mathbf{A}\| = |\alpha| \, \|\mathbf{A}\|$,
3) $\|\mathbf{A}_1 + \mathbf{A}_2\| \leq \|\mathbf{A}_1\| + \|\mathbf{A}_2\|$.

In addition,

4) $\|\mathbf{Ax}\| \leq \|\mathbf{A}\| \, \|\mathbf{x}\|$, *with equality for some* $\mathbf{x} \neq \mathbf{0}$.
5) $\|\mathbf{AB}\| \leq \|\mathbf{A}\| \, \|\mathbf{B}\|$.
6) $\rho(\mathbf{A}) \leq \|\mathbf{A}\|$, *with equality if* $\mathbf{A}$ *is normal and* $\| \cdot \| = \| \cdot \|_2$.

$\|\mathbf{A}\|_1$ is the maximum column absolute sum, $\|\mathbf{A}\|_\infty$ is the maximum row absolute sum.
$\|\mathbf{A}\|_2 = \sqrt{\|\mathbf{A}^{\mathrm{H}} \mathbf{A}\|_2} \leq \|\mathbf{A}\|_{\mathrm{F}} \leq \sqrt{k} \|\mathbf{A}\|_2, \quad \|\mathbf{A}\|_2 \leq \sqrt{\|\mathbf{A}\|_1 \, \|\mathbf{A}\|_\infty}, \quad \|\mathbf{AB}\|_{\mathrm{F}} \leq \|\mathbf{A}\|_2 \, \|\mathbf{B}\|_{\mathrm{F}}$.
If p_A is the maximum number of non-zeros per row of $\mathbf{A}$, then $\| \, |\mathbf{A}| \, \|_2 \leq \sqrt{p_A} \, \|\mathbf{A}\|_2$.

Note 1.8 In the discussion above involving induced matrix norms, we implicitly assumed that we used the same norm for k- as for n-vectors. But different norms can be employed as well. Notations as $\|\mathbf{A}\|_{p\leftarrow q}$ are used to indicate that for the $n \times k$ matrix $\mathbf{A}$ the q-norm is used on $\mathbb{C}^k$ and the p-norm on $\mathbb{C}^p$. We will not go into this type of details in this course.

For an $n \times n$ Hermitian, **semi-positive definite** matrix $\mathbf{A}$, i.e., $\mathbf{x}^*\mathbf{A}\mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{C}^n$, put

$$\|\mathbf{x}\|_A \equiv \sqrt{\mathbf{x}^*\mathbf{A}\mathbf{x}} \qquad (\mathbf{x} \in \mathbb{C}^n). \tag{1.19}$$

Exercise 1.8. **The A-norm.** Show that $\|\cdot\|_A$ defines a norm on $\mathbb{C}^n$ in case $\mathbf{A}$ is positive definite (i.e., $\mathbf{A}$ is semi-positive definite and $\mathbf{x}^*\mathbf{A}\mathbf{x} = 0$ only if $\mathbf{x} = \mathbf{0}$).

To relate inner products and norms, the following 'parallel law' can be usefull. For $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$, with ζ the **sign** of $\mathbf{y}^*\mathbf{A}\mathbf{x}$, that is, $\zeta \in \mathbb{C}$, $|\zeta| = 1$ such that $\bar{\zeta}\,\mathbf{y}^*\mathbf{A}\mathbf{x} \geq 0$, we have that

$$4\mathbf{y}^*\mathbf{A}\mathbf{x} = \zeta\,(\|\mathbf{x} + \zeta\,\mathbf{y}\|_A^2 - \|\mathbf{x} - \zeta\,\mathbf{y}\|_A^2) \tag{1.20}$$

Exercise 1.9. **The parallel law.**

(a) Prove **parallel law** (1.20).

(b) Conlcude that $\mathbf{A} = \mathbf{0}$ if $\mathbf{x}^*\mathbf{A}\mathbf{x} = 0$ for all $\mathbf{x} \in \mathbb{C}^n$.

Exercise 1.10. **Orthonormal matrices.** Let $\mathbf{V} \equiv [\mathbf{v}_1, \ldots, \mathbf{v}_k]$ be an $n \times k$ matrix.
We say that (a transformation by) $\mathbf{V}$ preserves the 2-norm if $\|\mathbf{V}x\|_2 = \|x\|_2$ for all $x \in \mathbb{C}^k$.
$\mathbf{V}$ preserves orthogonality if $\mathbf{V}x \perp \mathbf{V}y$ for all $x, y \in \mathbb{C}^k$ for which $x \perp y$.

(a) Assume $\mathbf{V}$ preserves the 2-norm. Prove that $\mathbf{V}$ preserves orthogonality.
(Hint: consider $\|\mathbf{V}(x + \zeta y)\|_2$ for $\zeta \in \mathbb{C}$, $|\zeta| = 1$.)

(b) Prove that the following four properties are equivalent:
- $\mathbf{V}$ is orthonormal
- $\mathbf{V}$ preserves the 2-norm.
- $\|\mathbf{V}X\|_2 = \|X\|_2$ for all $k \times \ell$ matrices X (with $\ell \geq 1$).
- $\|\mathbf{V}X\|_{\mathrm{F}} = \|X\|_{\mathrm{F}}$ for all $k \times \ell$ matrices X (with $\ell \geq 1$).

(c) Prove that the following two properties are equivalent:
- $\mathbf{V}$ preserves orthogonality.
- $\alpha\mathbf{V}$ is orthonormal for some scalar α.

B Perturbations and the conditioning of a problem

Let $\mathbf{A}$ be an $n \times n$ matrix.
For a given vector $\mathbf{b} \in \mathbb{C}^n$, we are interested in solving the **linear system**

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{1.21}$$

for $\mathbf{x} \in \mathbb{C}^n$, and we are interested in solving the **eigenvalue problem**

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x} \tag{1.22}$$

for a non-trivial $\mathbf{x} \in \mathbb{C}^n$ (eigenvector) and a scalar $\lambda \in \mathbb{C}$ (the eigenvalue associated with $\mathbf{x}$).

In practice, the problems will be perturbed (by rounding errors, model errors [from discretization], errors from measurements, etc.). For some (small) $n \times n$ matrix Δ, we will have $\mathbf{A} + \Delta$ rather than $\mathbf{A}$: Δ is a **perturbation** of $\mathbf{A}$ In Problem (1.21), $\mathbf{b}$ will be perturbed as well. Therefore, we actaully will be solving perturbed problems.

If small perturbations of the problem lead to large errors in the solution, then the problem is said to be **ill conditioned**, and we can not expect to be able to compute accurate solutions. Additional information is needed, i.e., the problem has to be modified (adapted model) to a well-conditioned one (for instance, in case of (1.21), a modification might be "find an $\mathbf{x}$ with

minimal norm for which $\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2 \leq \varepsilon$"). Of course, these modifications should render the problem into a realistic model of the practical underlying problem that is to be solved.

Condition numbers quantify how sensitive (the solution of) a problem is to perturbations. For Problem (1.21) we have (recall Conv. 1.5):

Theorem 1.9 *Consider the linear equation* $\mathbf{A}\mathbf{x}=\mathbf{b}$.
For some perturbations Δ *of* $\mathbf{A}$ *and* δ_b *of* $\mathbf{b}$, *let* $\widetilde{\mathbf{x}}$ *be the solution of the perturbed problem*

$$(\mathbf{A}+\Delta)\widetilde{\mathbf{x}}=\mathbf{b}+\delta_b.$$

Then, with respect to some norm $\|\cdot\|$ *on* $\mathbb{C}^n$, *we have that*

$$\frac{\|\widetilde{\mathbf{x}}-\mathbf{x}\|}{\|\mathbf{x}\|} \lesssim \mathcal{C}(\mathbf{A})\left(\frac{\|\Delta\|}{\|\mathbf{A}\|}+\frac{\|\delta_b\|}{\|\mathbf{b}\|}\right), \quad \textit{where} \quad \mathcal{C}(\mathbf{A}) \equiv \|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|. \tag{1.23}$$

For an exact upper bound divide the expression at the right-hand side by $1-\mathcal{C}(\mathbf{A})\frac{\|\Delta\|}{\|\mathbf{A}\|}$.

The quantity $\mathcal{C}(\mathbf{A})$ is called the **condition number of the the matrix A** with respect to the norm $\|\cdot\|$. However, since it tells us how perturbations on the problem, that is, on the matrix and on the right-hand side vector, affect the 'relative' accuracy of the solution, it also is the the **condition number of the linear problem** (1.21) with respect to the norm that is used to measure the size of the perturbations. Note that here the perturbations are also measured in some 'relative' or 'scaled' sense (and that the perturbations should not be too large: $\mathcal{C}(\mathbf{A})\frac{\|\Delta\|}{\|\mathbf{A}\|}<1$).

Exercise 1.11. ***Proof of Theorem*** 1.9***.***
(a) First consider the special case where $\Delta=\mathbf{0}$ and prove that

$$\frac{\|\widetilde{\mathbf{x}}-\mathbf{x}\|}{\|\widetilde{\mathbf{x}}\|} \leq \mathcal{C}(\mathbf{A})\frac{\|\delta_b\|}{\|\mathbf{b}\|}.$$

No consider the general situation of Theorem 1.9.
(b) Show that $\widetilde{\mathbf{x}}-\mathbf{x}=-\mathbf{A}^{-1}\Delta\widetilde{\mathbf{x}}+\mathbf{A}^{-1}\delta_b$.
(c) Note that $\|\mathbf{b}\| \leq \|\mathbf{A}\|\,\|\mathbf{x}\|$ and $\frac{\|\widetilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq 1+\frac{\|\widetilde{\mathbf{x}}-\mathbf{x}\|}{\|\mathbf{x}\|}$.
(d) Prove Theorem 1.9.

The estimate in Theorem 1.9 is sharp: that is, for any $\delta_1 \geq 0$ and $\delta_2 \geq 0$, there are perturbations Δ and δ_b such that $\|\Delta\|=\delta_1$ and $\|\delta_b\|=\delta_2$ and for which the inequality in (1.23) is an equality.

Exercise 1.12. Let $\delta>0$ (small).
(a) Prove there is a perturbation δ_b of $\mathbf{b}$ such that $\|\delta_b\|=\delta$ and $\|\widetilde{\mathbf{x}}-\mathbf{x}\|=\|\mathbf{A}^{-1}\|\,\|\delta_b\|$, where $\widetilde{\mathbf{x}}$ solves $\mathbf{A}\widetilde{\mathbf{x}}=\mathbf{b}+\delta_b$.
(b) There is perturbation Δ such that $\|\Delta\|=\delta$ and $\|\widetilde{\mathbf{x}}-\mathbf{x}\|=\|\mathbf{A}^{-1}\|\,\|\Delta\|\,\|\widetilde{\mathbf{x}}\|$, where $\widetilde{\mathbf{x}}$ solves $(\mathbf{A}+\Delta)\widetilde{\mathbf{x}}=\mathbf{b}$.
Prove the slightly weaker statement $\|\widetilde{\mathbf{x}}-\mathbf{x}\| = \|\mathbf{A}^{-1}\|\,\|\Delta\|\,\|\widetilde{\mathbf{x}}\|+\mathcal{O}(\delta^2)$. (Hint: consider a perturbation of the form $\delta\mathbf{z}\mathbf{y}^*$ and use the dual norm $|\!|\!|\cdot|\!|\!|$ on $\mathbb{C}^n$ defined by $|\!|\!|\mathbf{y}|\!|\!| \equiv \sup|\mathbf{y}^*\mathbf{u}|$, with supremum taken over all $\mathbf{u}$ with $\|\mathbf{u}\|=1$.)
(c) Discuss the sharpness of the estimate $\|\mathbf{b}\| \leq \|\mathbf{A}\|\,\|\mathbf{x}\|$.
(d) Discuss the sharpness of (1.23). (Note that, strictly speaking, the estimate in the Theorem is not sharp: it only is sharp if $\|\mathbf{b}\|=\|\mathbf{A}\|\,\|\mathbf{x}\|$. Unfortunately, $\mathbf{b}$ can not be selected: it is part of the linear equation.)

If the $n\times n$ matrix $\mathbf{A}$ is ill conditioned (large condition number) then small perturbations can lead to large errors in the solution of linear systems with $\mathbf{A}$. Therefore, it seems a good

idea to minimise the condition number with some simple manipulations as scaling the rows, i.e., work with the system $\mathbf{DAx} = \mathbf{Db}$ instead of $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{D}$ an appropriate diagonal matrix. Note that row scaling does not affect the sparsity structure of the matrix: the set of (i, j) of non-zero matrix entries is the same for $\mathbf{A}$ and $\mathbf{DA}$. Moreover, row scaling does not affect the solution $\mathbf{x}$.[3] Of course the row scaling has to be applied before perturbation (by computational steps) are being introduced. Otherwise the row scaling would only be cosmetic. Scaling such that all rows have equal norm, **row equilibration**, seems to be the best as we will see in the next exercise.

To preserve algebraic structure (symmetry, Hermitian), rows as well as columns have to be scaled.

Exercise 1.13. **Row scaling.** Let $\|\cdot\|$ be a multiplicative norm. $\mathcal{C}(\mathbf{A}) \equiv \|\mathbf{A}^{-1}\|\,\|\mathbf{A}\|$.

(a) Prove that $\|\mathbf{A}^{-1}\| \leq \|(\mathbf{DA})^{-1}\|\,\|\mathbf{D}\|$ for any $n \times n$ matrix $\mathbf{D}$.

(b) Conclude that $\mathcal{C}(\mathbf{A}) \leq \mathcal{C}(\mathbf{DA})$ if $\|\mathbf{D}\|\,\|\mathbf{A}\| = \|\mathbf{DA}\|$.

(c) If all rows of $\mathbf{A}$ have 1-norm equal to 1, $\|\mathbf{A}^*\mathbf{e}_j\|_1 = 1$ all j, then $\|\mathbf{DA}\|_\infty = \|\mathbf{D}\|_\infty$ for all diagonal $\mathbf{D}$.

(d) Show that for an arbitrary $n \times n$ matrix $\mathbf{A}$ the diagonal $\mathbf{D}$ with jth diagonal entry $1/\|\mathbf{A}^* e_j\|_1$ leads to smallest the condition number $\mathcal{C}_\infty(\mathbf{DA})$ with respect to the $\|\cdot\|_\infty$ norm, smallest w.r.t. to all diagonal scalings.

(e) If all rows of $\mathbf{A}$ have 2-norm equal to 1, $\|\mathbf{A}^*\mathbf{e}_j\|_2 = 1$ all j, then $\|\mathbf{DA}\|_{\mathrm{F}} = \|\mathbf{D}\|_{\mathrm{F}}$ for all diagonal $\mathbf{D}$.

(f) For an $n \times n$ matrix $\mathbf{A}$, let $\mathbf{D}_0$ be the $n \times n$ diagonal matrix with jth diagonal entry equal to $1/\|\mathbf{A}^* e_j\|_2$. Show that for any $n \times n$ diagonal matrix $\mathbf{D}$ we have that $\mathcal{C}_{\mathrm{F}}(\mathbf{D}_0\mathbf{A}) \leq \sqrt{n}\,\mathcal{C}_F(\mathbf{DA})$: except for a factor at most $\sqrt{n}$, row equilibration (w.r.t. the 2-norm) leads to the smallest condition number in Frobenius norm (here, denoted by $\mathcal{C}_F$).

C Forward and backward error

Let $\mathbf{x}$ be a solution of (1.21) or of (1.22) with eigenvalue λ.
If $\mathbf{u}$ is a vector in $\mathbb{C}^n$ that approximates $\mathbf{x}$ (an **approximate solution**), and, in case of (1.22), ϑ is an approximate eigenvalue, then

$\mathbf{x} - \mathbf{u}$ (and $\lambda - \vartheta$) is the **error**, also called **forward error**,[4]

$\mathbf{r} \equiv \mathbf{A}(\mathbf{x} - \mathbf{u}) = \mathbf{b} - \mathbf{Au}$ is the **residual** for the linear system and

$\mathbf{r} \equiv \mathbf{Au} - \vartheta\mathbf{u}$ is the **residual** for the eigenvalue problem.

If there is an $n \times n$ matrix Δ, a **perturbation** of $\mathbf{A}$ such that

$$(\mathbf{A} + \Delta)\mathbf{u} = \mathbf{b} \tag{1.24}$$

in case of the linear system and

$$(\mathbf{A} + \Delta)\mathbf{u} = \vartheta\mathbf{u} \tag{1.25}$$

in case of the eigenvalue problem, then Δ is a **backward error**.

With the backward error, the approximate solution is viewed as an exact solution of a (hopefully slightly) perturbed problem: the error in the solution is 'trowed back' to the problem. The challenge is, given an approximate solution $\mathbf{u}$, to find a perturbation Δ with $\|\Delta\|$, or rather $\|\Delta\|/\|\mathbf{A}\|$, as small as possible. The idea here is that if the (scaled) backward error is small, then the numerical method that produced the approximate solution $\mathbf{u}$ is stable, even if the error $\mathbf{x} - \mathbf{u}$ happens to be large. In such a case, the problem is unstable: the problem is to 'blame' for the inaccurate solution, rather than the numerical method.

[3] If we scale the columns, then we have to 'unscale' to solution of the scaled system to find the desired solution. With scaling we tried to minimise effect of perturbations, by unscaling we might reverse this beneficial action.

For some applications, for maintaining the point of view that the problem is to blame, it is more realistic to require the requested perturbation Δ to have a similar structure as the matrix $\mathbf{A}$: rather than having $\|\Delta\|/\|\mathbf{A}\|$ small, it is required to have $|\Delta| < \varepsilon|\mathbf{A}|$ for some small ε. Here the inequality is matrix entry wise. Or, if $\mathbf{A}$ is symmetric, then Δ should be symmetric as well.

For the linear system case, the backward error can also be formulated as a perturbation on $\mathbf{b}$.

Note that the error is not readily available. The following theorem shows that a backward error can be expressed in terms of residuals and approximate solutions. Since it is reasonable to assume that the problem is (reasonably) well-conditioned, it makes sense to try to design numerical algorithms that produce approximate solutions with small residuals. Moreover, if we can quantify (in terms of properties of $\mathbf{A}$, ...) how solutions respond to perturbations on the problem, then we can bound the error in terms of the residual (and these properties of $\mathbf{A}$). In summary,

- Design algorithms that lead to small residuals (small backward error).
- Analyse the effect of perturbations on the solution (forward error analysis).

Theorem 1.10 *Let Δ be the $n \times n$ matrix given by*

$$\Delta \equiv \frac{\mathbf{r}\,\mathbf{u}^*}{\mathbf{u}^*\mathbf{u}}.$$

Equation (1.24) *holds if* $\mathbf{r} \equiv \mathbf{b} - \mathbf{A}\mathbf{u}$. *Equation* (1.25) *holds if* $\mathbf{r} \equiv -(\mathbf{A}\mathbf{u} - \vartheta\mathbf{u})$. *Moreover, Δ has rank* 1 *and*

$$\frac{\|\Delta\|_2}{\|\mathbf{A}\|_2} \leq \frac{\|\mathbf{r}\|_2}{\|\mathbf{A}\|_2\,\|\mathbf{u}\|_2}. \tag{1.26}$$

From (1.26), we conclude that the size if the residual scaled by the norm of the matrix and the approximate solution appears to be an appropriate measure for the backward error.

Exercise 1.14. ***Proof of Theorem*** 1.10***.***

(a) Prove Theorem 1.10.

The perturbation term Δ for which, say, (1.24) holds is not unique. Put $\Delta_1 \equiv \frac{\mathbf{r}\,\mathbf{u}^*}{\mathbf{u}^*\mathbf{u}}$.

(b) Show that $\Delta = \Delta_2$, where

$$\Delta_2 \equiv \frac{\mathbf{r}\,\mathbf{r}^*}{\mathbf{r}^*\mathbf{u}},$$

also satisfies (1.24), provided that $\mathbf{u} \not\perp \mathbf{r}$. Note that this Δ_2 is Hermitian. Show that,

$$\|\Delta_1\|_2 = \cos\angle(\mathbf{r},\mathbf{u})\,\frac{\|\mathbf{r}\|_2}{\|\mathbf{u}\|_2} \leq \frac{\|\mathbf{r}\|_2}{\|\mathbf{u}\|_2} \leq \frac{1}{\cos\angle(\mathbf{r},\mathbf{u})}\,\frac{\|\mathbf{r}\|_2}{\|\mathbf{u}\|_2} = \|\Delta_2\|_2.$$

(c) Consider

$$\Delta_3 \equiv \frac{\mathbf{b}\,\mathbf{b}^*}{\mathbf{b}^*\mathbf{u}} - \frac{\mathbf{A}\mathbf{u}\,(\mathbf{A}\mathbf{u})^*}{\mathbf{u}^*\mathbf{A}^*\mathbf{A}\mathbf{u}}.$$

Show that $\Delta = \Delta_3$ satisfies (1.24). Discuss rank, symmetry and size of Δ_3 (in terms of $\|\mathbf{r}\|_2$).

(d) Assume that $\mathbf{A}$ is positive definite. Prove that $\mathbf{A} + \Delta_3$ is positive definite as well as soon as $\mathbf{b}^*\mathbf{u} > 0$. Show that $\mathbf{b}^*\mathbf{u} > 0$ if $\mathbf{u}$ is sufficiently close to the solution $\mathbf{x}$ of the system $\mathbf{A}\mathbf{x} = \mathbf{b}$.

D Perturbed problems

Exercise 1.15. Let $\mathbf{A}$ be an $n \times n$ matrix.

(a) Show that

$$(\mathbf{I} - \mathbf{A})^{-1} = \mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \mathbf{A}^3 + \ldots. \tag{1.27}$$

holds if $\|\mathbf{A}\| < 1$ for some multiplicative norm. You can do this by combining Theorem 0.12 and (1.13), but give also an elementary proof.
In particular, $\|\mathbf{A}\| < 1$ implies that $\mathbf{I} - \mathbf{A}$ is non-singular.

Theorem 1.11 *Let* $\mathbf{A}$ *and* Δ *be* $n \times n$ *matrices,* $\mathbf{A}$ *is non-singular,* $\|\cdot\|$ *a multiplicative norm. Put* $\delta \equiv \|\mathbf{A}^{-1}\Delta\|$. *Then* $\delta \leq \|\mathbf{A}^{-1}\|\,\|\Delta\|$. *If* $\delta < 1$, *then* $\mathbf{A} + \Delta$ *is non-singular,*

$$\|(\mathbf{A}+\Delta)^{-1}\| \leq \|\mathbf{A}^{-1}\|\frac{1}{1-\delta} \quad \textit{and} \quad \|\mathbf{A}^{-1} - (\mathbf{A}+\Delta)^{-1}\| \leq \|\mathbf{A}^{-1}\|\,\frac{\delta}{1-\delta}. \tag{1.28}$$

Exercise 1.16. ***Proof of Theorem*** 1.11**.**

(a) Assume $\delta < 1$. Prove that $\mathbf{A}+\Delta$ is non-singular and that first estimate in (1.28) is correct.

(b) Show that $\mathbf{A}^{-1} - (\mathbf{A}+\Delta)^{-1} = \mathbf{A}^{-1}\Delta(\mathbf{A}+\Delta)^{-1}$. Derive the second estimate of (1.28).

The following theorem can be viewed as a perturbation theorem, where the off-diagonal entries (the matrix $\mathbf{E}$ in Exercise 1.17) are the perturbations of a diagonal matrix ($\mathbf{D}$ in Exercise 1.17). However, it is also of interest without this interpretation: the theorem gives a simple way of estimating eigenvalues. Moreover, the proof is a simple and nice application of the perturbation Theorem 1.11.

Theorem 1.12 (Gershgorin's theorem) *Let* $\mathbf{A} = (A_{ij})$ *be an* $n \times n$ *matrix.*
A **Gershgorin disk** *is a disk* $\mathcal{D}_i$ *in* $\mathbb{C}$ *with centre* A_{ii} *and radius* $\sum_{j,j\neq i} |A_{ij}|$*:*

$$\mathcal{D}_i \equiv \{\zeta \in \mathbb{C} \,\big|\, |A_{ii} - \zeta| \leq \rho_i\} \quad \textit{with} \quad \rho_i \equiv \sum_{j,j\neq i} |A_{ij}| \quad (i = 1, \ldots, n).$$

Each eigenvalues of $\mathbf{A}$ *is contained in some Gershgorin disk:* $\Lambda(\mathbf{A}) \subset \bigcup_i \mathcal{D}_i$.

Exercise 1.17. **Gershgorin's theorem.** Let $\mathbf{D} \equiv \mathrm{diag}(\mathbf{A})$ be the diagonal of $\mathbf{A}$ and $\mathbf{E}$ the outer diagonal (i.e., $D_{ii} = A_{ii}$ all i, $D_{ij} = 0$ all i, j, $i \neq j$, $\mathbf{E} \equiv \mathbf{A} - \mathbf{D}$).
Put $\rho_i \equiv \sum_{j,j\neq i} |A_{ij}| = \|\mathbf{E}^*\mathbf{e}_i\|_1 \quad (i = 1, \ldots, n)$.

(a) Show that, for $\lambda \in \mathbb{C}$, $\mathbf{A} - \lambda\mathbf{I}$ is non-singular if $\|(\mathbf{D} - \lambda\mathbf{I})^{-1}\mathbf{E}\| < 1$.

(b) Show that $\|(\mathbf{D} - \lambda\mathbf{I})^{-1}\mathbf{E}\|_p \leq \|(\mathbf{D} - \lambda\mathbf{I})^{-1}\|_p\|\mathbf{E}\|_p < 1$ if $\|\mathbf{E}\|_p < \min_i |A_{ii} - \lambda|$.

(c) Conclude that (**Bauer**–**Fike**'s Theorem holds):

$$|A_{ii} - \lambda| \leq \|\mathbf{E}\|_p \quad \text{for some } i, \text{ if } \lambda \text{ is an eigenvalue of } \mathbf{A}.$$

(d) Show that $\|(\mathbf{D} - \lambda\mathbf{I})^{-1}\mathbf{E}\|_\infty < 1$ if and only if $\rho_i < |A_{ii} - \lambda|$ for all i.

(e) Conclude that Theorem 1.12 holds.

Gershgorin disks indicate how eigenvalues can get perturbed ($\mathbf{D}$ is an diagonal matrix of eigenvalues, $\mathbf{E}$ is a perturbation matrix). In practice it appears that the effect of the perturbation A_{ij} on the eigenvalue A_{ii} is often better described by $\frac{|A_{ij}|}{|A_{ii} - A_{jj}|}$ than by $|A_{ij}|$.

(f) Derive a variant of Gershgorin's theorem using the 1-norm rather than the ∞-norm.

Note that the theorem does not exclude the possibility that all eigenvalues are contained in the same Gershgorin disk. The following theorem states that the eigenvalues depend continuously on a parameter if the matrix depends continuously on that parameter. Bauer–Fike's Theorem can be used to prove this result: however, we will not give further details here.

Theorem 1.13 *Assume* $\mathbf{F}(\tau)$ *is an* $n \times n$ *matrix for all* τ *in some subset* $\mathcal{I}$ *of* $\mathbb{C}$. *If* $\mathbf{F}$ *depends continuously on* τ, *then there are continuous complex-valued functions* $\mu_1, \ldots, \mu_n$ *on* $\mathcal{I}$ *such that* $\mu_1(\tau), \ldots, \mu_n(\tau)$ *are the eigenvalues of* $\mathbf{F}(\tau)$ *counted according to multiplicity* ($\tau \in \mathcal{I}$).

This theorem allows a continuity argument to prove that

Theorem 1.14 (Gershgorin's theorem 2) *If precisely p Gershgorin disks are connected, then the union of these p disks contains exactly p eigenvalues of* $\mathbf{A}$.

Exercise 1.18. ***Proof of Theorem*** 1.14***.*** Consider Theorem 1.12.

From Theorem 1.12 we know that $\Lambda(\mathbf{A}) \subset \bigcup \mathcal{D}_i$.

(a) Give an example that shows that not all Gershgorin disk contains at least one eigenvalue of $\mathbf{A}$ (Hint: Replace the $(n,1)$ entry of $\mathbf{I}+\varepsilon\mathbf{S}$ by 1. Here $\mathbf{S}$ is the shift matrix that assigns $\mathbf{e}_{i-1}$ to $\mathbf{e}_i$).

A subset $\mathcal{G}$ of $\mathbb{C}$ is **connected** if for all ζ_0 and ζ_1 in $\mathbb{C}$ there is a continuous curve in $\mathcal{G}$ that connects ζ_0 and ζ_1 (i.e., for some continuous function $\phi : [0,1] \to \mathbb{C}$ we have that $\phi(0) = \zeta_0,\ \phi(1) = \zeta_1,\ \ \phi(t) \in \mathcal{G}\ \ (t \in [0,1])$).

(b) Suppose there is a subset E of $\{1, 2, \ldots, n\}$ of p numbers such that

$$\mathcal{G} \equiv \bigcup_{i \in E} \mathcal{D}_i \quad \text{is connected, while} \quad \mathcal{G} \cap \mathcal{D}_j = \emptyset \quad (j \notin E).$$

Prove that $\mathcal{G}$ contains exactly p eigenvalues of $\mathbf{A}$.

E Rounding errors

Many results an details on effects of rounding errors (specifically in algorithms for dense matrices) can be found in [1].

Convention 1.15 For ease of notation, we follow the following conventions.

- ξ is a number in $[-\mathbf{u}, \mathbf{u}]$ with $\mathbf{u}$ the **relative machine precision** ($\mathbf{u}$ is `0.5*eps` in Matlab).
- ξs on different locations can have different values.[5]
- Formulae involving a ξ are to read from left to right.
- We neglect order $\mathbf{u}^2$ terms. (As an alternative, replace quantities as $n\xi$ by $\frac{n\xi}{1-n\xi}$.)

Exercise 1.19.

(a) Following the above conventions, show that the statement $\xi = 2\xi$ is correct while $2\xi = \xi$ is wrong.

(b) If α and β are scalars then $\alpha\xi + \beta\xi = (|\alpha| + |\beta|)\xi$. Prove that this formula is sharp, i.e., it is correct and there are $\xi_1, \xi_2 \in [-\mathbf{u}, \mathbf{u}]$ for which $\alpha\xi_1 + \beta\xi_2 = (|\alpha| + |\beta|)\mathbf{u}$.

Notation 1.16 If $\mathbf{B}$ is an $n \times k$ matrix (k and n can be 1) to be obtained by computation with computational rules (an algorithm) that are clear from the context, then $\widehat{\mathbf{B}}$ denotes the quantity as actually computed. We assume that the input values (matrix entries) are **machine numbers** (real or complex numbers that can be represented in the computer). If $\mathbf{B}$ is defined by a (longer) expression, then we will also use the notation $\mathbf{B}\widehat{\ }$ or $fl(\mathbf{B})$ instead of $\widehat{\mathbf{B}}$.

We follow the following rules.

Rule 1.17

– If α and β are machine numbers, then the result $(\alpha \bullet \beta)\widehat{\ }$ obtained in the computer by a *floating point operation* (**flop**) or basic arithmetic operation $\bullet$, i.e., $\bullet$ represents $+$, $-$, $*$, or $/$, is the exact result $\alpha \bullet \beta$ with a *relative* error of at most $\mathbf{u}$:

$$(\alpha \bullet \beta)\widehat{\ } = (\alpha \bullet \beta)(1 + \xi).$$

($\beta \neq 0$ if $\bullet$ is $/$). That is, $(\alpha + \beta)\widehat{\ } = (\alpha + \beta)(1 + \xi), \ldots$.

– Operations with 0 ($\alpha - 0$ or $\beta = 0$) are exact (again $\beta \neq 0$ if $\bullet$ is $/$).

[5] We rather put $x_1\xi + x_2\xi$ than $x_1\xi_1 + x_2\xi_2$.

Exercise 1.20. **Rounding errors in elementary operations.** Below, $\mathbf{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$ and $\mathbf{y}$ are n-vectors, α is a scalar, $\mathbf{A}$ is an $m \times n$ matrix. $(\mathbf{x}, \mathbf{y}) \equiv \mathbf{y}^{\mathrm{T}}\mathbf{x}$ is the standard inner product, $\alpha\mathbf{x} \equiv (\alpha x_1, \ldots, \alpha x_n)^{\mathrm{T}}$ and inequalities between vectors and between matrices are coordinate-wise (entry-wise), absolute values of vectors and of matrices are also coordinate-wise (entry-wise).

(a) Prove that, for some vector δ_x,

$$(\mathbf{x}, \mathbf{y})\widehat{\ } = (\mathbf{x} + \delta_x, \mathbf{y}) = (\mathbf{x}, \mathbf{y}) + n(|\mathbf{x}|, |\mathbf{y}|)\xi \quad \delta_x \text{ such that } |\delta_x| \leq n\mathbf{u}\,|x|. \tag{1.29}$$

The inequality is coordinate wise.
Discuss the sharpness of this statement.
Discuss the consequences for the case where $\mathbf{x}$ and $\mathbf{y}$ are almost perpendicular.

(b) Consider the multiplication $\alpha\mathbf{x}$. Why is the statement $(\alpha\mathbf{x})\widehat{\ } = \alpha\mathbf{x} + \xi\alpha|\mathbf{x}|$ generally not correct? Prove that $(\alpha\mathbf{x})\widehat{\ } = \alpha\mathbf{x} + \delta_x$ with $|\delta_x| \leq \mathbf{u}|\alpha||\mathbf{x}|$.

Let $\mathbf{z} \equiv \alpha\mathbf{x} + \mathbf{y}$ (AXPY). Prove that

$$\widehat{\mathbf{z}} = \mathbf{z} + \delta_z, \quad \text{with} \quad |\delta_z| \leq \mathbf{u}(|\alpha||\mathbf{x}| + |\mathbf{z}|), \quad \|\delta_z\|_2 \leq \mathbf{u}\max(\|\mathbf{y}\|_2 + 2\|\mathbf{z}\|_2).$$

(c) Prove that

$$(\mathbf{A}\mathbf{x})\widehat{\ } = (\mathbf{A} + \Delta_A)\mathbf{x}, \quad \text{where} \quad |\Delta_A| \leq p_A\,\mathbf{u}|\mathbf{A}|,$$

where p_A is the maximal number of non-zero entries in the rows of $\mathbf{A}$. Note that Δ_A depends on both $\mathbf{A}$ and $\mathbf{x}$, while the upper-bound does not depend on $\mathbf{x}$.

(d) Let $\mathbf{A}$ and $\mathbf{B}$ matrices of size $m \times n$ and $n \times \ell$, respectively. Prove that

$$(\mathbf{A}\mathbf{B})\widehat{\ } = \mathbf{A}\mathbf{B} + \Delta, \quad \text{where} \quad |\Delta| \leq p_A\,\mathbf{u}|\mathbf{A}|\,|\mathbf{B}|.$$

Note. For estimates of $\|\,|\mathbf{A}|\,\|_2$ and of $\|\,|\mathbf{A}|\,\|_{\mathrm{F}} = \|\mathbf{A}\|_{\mathrm{F}}$ in terms of $\|\mathbf{A}\|_2$ (and the sparsity of $\mathbf{A}$), see Exercise 1.7.

F Full numerical accuracy

In order to verify that a vector $\mathbf{x}$ is the exact solution of a problem, the only option usually is to check whether the residual for this vector is zero. Unfortunately the computed residual will be spoilt by rounding errors and even for the exact solution, the *computed* residual will be non-zero.

Theorem 1.18 *If $\mathbf{x}$ is the exact solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$, p_A is the maximum number of non-zeros per row of $\mathbf{A}$, then, with $\widehat{\mathbf{r}} \equiv fl(\mathbf{b} - \mathbf{A}\mathbf{x})$, we have the sharp estimates*

$$(1 - \mathbf{u})\,|\widehat{\mathbf{r}}| \leq p_A\,\mathbf{u}\,|\mathbf{A}|\,|\mathbf{x}|, \qquad \frac{\|\widehat{\mathbf{r}}\|}{\|\mathbf{A}\|\,\|\mathbf{x}\|} \leq \frac{p_A\,\mathbf{u}}{1 - \mathbf{u}}\,\frac{\|\,|\mathbf{A}|\,\|}{\|\mathbf{A}\|}.$$

We have a similar result for eigenvalue problems. The **estimates are sharp** in the sense that there are examples for which we have equality. However, if p_A is large (for instance, $p_A = n$ if $\mathbf{A}$ is a full matrix), then the computation involves many numbers and it is unlikely that all numbers lead to rounding errors that point in the same direction. In practice, some rounding errors cancel and inaccuracies in the residual (or in any other quantity computed with n arithmetic operations) that are observed match better a bound with $\sqrt{p_A}$ rather than p_A.

Now, suppose we have an $\mathbf{x}'$ available for which $fl(\mathbf{b}-\mathbf{A}\mathbf{x}') \leq p_A\,\mathbf{u}\,|\mathbf{A}|\,|\mathbf{x}'|$ (or $\|fl(\mathbf{b}-\mathbf{A}\mathbf{x}')\| \leq p_A\,\mathbf{u}\,\|\,|\mathbf{A}|\,\|\,\|\mathbf{x}'\|$). The above result implies that we are not able do decide whether $\mathbf{x}'$ is not the exact solution. We will say that $\mathbf{x}'$ has **full numerical accuracy**, or is **numerically exact**.

Note that the quantity

$$p_A\,\|\mathbf{A}^{-1}\|\,\|\,|\mathbf{A}|\,\| = p_A\,\mathcal{C}(\mathbf{A})\,\frac{\|\,|\mathbf{A}|\,\|}{\|\mathbf{A}\|}$$

can viewed as the conditioning of the problem "Solve $\mathbf{Ax} = \mathbf{b}$ for $\mathbf{x}$": rounding errors of size at most $\mathbf{u}$ can lead to errors of size $(p_A \, \|\mathbf{A}^{-1}\| \, \| \, |\mathbf{A}| \, \|)\mathbf{u}$ in the "best" solution.

Exercise 1.21. ***Proof of Theorem*** 1.18***.***

(a) Prove the theorem.

Consider an $n \times k$ orthonormal matrix $\mathbf{V}$. Suppose $\mathbf{v}$ is an n-vector in the span of $\mathbf{V}$. Then $\mathbf{v} - \mathbf{V}(\mathbf{V}^*\mathbf{v}) = \mathbf{0}$. However, we may not expect the computed quantity $\mathit{fl}(\mathbf{v} - \mathbf{V}(\mathbf{V}^*\mathbf{v}))$ to be $\mathbf{0}$. Actually,

$$\delta \equiv \mathit{fl}(\mathbf{v} - \mathbf{V}(\mathbf{V}^*\mathbf{v})) = \Delta_1(\mathbf{V}^*\mathbf{v}) + \mathbf{V}(\Delta_2^*\mathbf{v}), \quad \text{where } |\Delta_1| \le k\,\mathbf{u}\,|\mathbf{V}|, \quad |\Delta_2| \le n\,\mathbf{u}\,|\mathbf{V}|. \tag{1.30}$$

Since $\| \, |\mathbf{V}| \, \|_2 \le \sqrt{k}$, we have the upper bound $\|\delta\|_2 \le \mathbf{u}(k+n)\sqrt{k}\,\|\mathbf{v}\|_2$. This upper bound is sharp. Therefore, if $\mathbf{w}$ is an n-vector for which $\|\mathit{fl}(\mathbf{w} - \mathbf{V}(\mathbf{V}^*\mathbf{w}))\|_2 \le \mathbf{u}(k+n)\sqrt{k}\,\|\mathbf{w}\|_2$, then we will consider $\mathbf{w}$ to be numerically in the span of $\mathbf{V}$: the columns of the matrix $[\mathbf{V}, \mathbf{w}]$ are **numerically linearly dependent**.

Exercise 1.22.

(a) Prove (1.30).

References

[1] Nicholas J. Higham, *Accuracy and stability of numerical algorithms*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996. MR 97a:65047