A Solutions of exercises

In this section solutions are collected for the exercises at the end of the individual chapters. These are not to be understood as "'blue-print"' solutions but rather as suggestions in sketchy form for stimulating the reader's own work.

A.1 Chapter 1

Solution A.1.1: a) Follows directly from

$$0 \leq \Big(\sqrt{\varepsilon}\, a \pm \frac{1}{2\sqrt{\varepsilon}}\, b\Big)^2, \quad a, b \in \mathbb{R},\ \varepsilon \in \mathbb{R}_+.$$

b) The function $f(x) = x^{-1}$ is for $x > 0$ convex and $\sum_{i=1}^n x_i \lambda_i$ is a convex linear combination. Hence, by a geometric argument, one may conclude the asserted estimate.

c) For $n = 0$ the statement is obvious. For $n \in \mathbb{N}$ observe that there are exactly three local extrema, for $x = 0$ (maximum), $x = 1$ (maximum), and $x_{\min} = \frac{1}{1+n}$ (minimum). Furthermore,

$$x_{\min}^2 (1 - x_{\min})^{2n} = \frac{1}{(1+n)^2}\Big(1 - \frac{1}{1+n}\Big)^{2n} = \frac{1}{(1+n)^2}\frac{n^{2n}}{(1+n)^{2n}} \leq \frac{1}{(1+n)^2}.$$

Solution A.1.2: a) Multiplying out yields

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

b) Let $x, y \in \mathbb{R}^n$ be arbitrary and (without loss of generality) $\|x\| = \|y\| = 1$.

$$0 \leq (x-y, x-y) = \|x\|^2 + \|y\|^2 - 2\,(x,y) = 2 - 2\,(x,y).$$

Similarly

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

i. e., $|(x,y)| \leq 1$.

c) The properties of a scalar product follow immediately from those of the Euclidean scalar product and those assumed for the matrix A.

i) Yes. Let $\langle \cdot, \cdot \rangle$ be an arbitrary scalar product and let $\{e_i\}_{1 \leq i \leq n}$ be a Cartesian basis of $\mathbb{R}^n$, such that any $x, y \in \mathbb{R}^n$ have the respresentations $x = \sum_i x_i e_i$, $y = \sum_i x_i e_i \in \mathbb{R}^n$. Define a matrix $A \in \mathbb{R}^{n \times n}$ by $a_{ij} := \langle e_j, e_i \rangle$. Then, there holds $(Ax, y) = \sum_{ij} a_{ij} x_j y_i = \sum_{ij} x_j y_i \langle e_j, e_i \rangle = \langle x, y \rangle$. Furthermore, A is obviously symmetric and positive definite due to the same properties of the scalar product $< \cdot, \cdot >$.

ii) The following statements are equivalent:

1. $< \cdot, \cdot > : \ \mathbb{C} \times \mathbb{C} \to \mathbb{C}$ is a (Hermitian) positive definite sesquilinear form (i.e. a scalar product).
2. There exists a (Hermitian) positive definite matrix $A \in \mathbb{C}^{n \times n}$ such that $\langle x, y \rangle = (Ax, y),\ x, y \in \mathbb{C}^n$.

Solution A.1.3: a) The identity becomes obvious by replacing x by $\tilde{x} := x\|x\|^{-1}$.

b) There holds, for $x \in \mathbb{R}^n \setminus 0$:

$$\|Ax\|_2 = \frac{\|Ax\|_2}{\|x\|}\,\|x\|_2 \leq \sup_{y \in \mathbb{R}^n} \frac{\|Ay\|_2}{\|y\|_2}\,\|x\| = \|A\|_2 \|x\|_2.$$

c) There holds

$$\|AB\|_2 = \sup_{x \in \mathbb{R}^n} \frac{\|ABx\|_2}{\|x\|_2} \leq \sup_{x \in \mathbb{R}^n} \frac{\|A\|_2 \|Bx\|_2}{\|x\|_2} \leq \sup_{x \in \mathbb{R}^n} \frac{\|A\|_2 \|B\|_2 \|x\|_2}{\|x\|_2} = \|A\|_2 \|B\|_2.$$

This relation is not true for any matrix norm. As a counter example, employ the elementwise maximum norm $\|A\|_{\max} := \max_{i,j=1,\cdots,n} |a_{ij}|$ to

$$\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}.$$

d) Let λ be any eigenvalue of A and x a corresponding eigenvector. Then,

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

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,

$$\|Ax\|_2 = \big\|A\big(\sum_i x_i a^i\big)\big\|_2 = \big\|\sum_i \lambda_i x_i a^i\big\|_2 \leq \max_i |\lambda_i|\,\big\|\sum_i x_i a^i\big\|_2| = \max_i |\lambda_i|\,\|x\|_2,$$

and consequently,

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

e) There holds

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

and $\|\bar{A}^T A\|_2 \leq \|\bar{A}^T\|_2 \|A\|_2 = \|A\|_2^2$ (observe that $\|A\|_2 = \|\bar{A}^T\|_2$ due to $\|Ax\|^2 = \|\bar{A}^T \bar{x}\|^2,\ x \in \mathbb{C}^n$).

Solution A.1.4: a) See the description in the text.

b) Let $A \in \mathbb{R}^{n\times n}$ be symmetric and positive definite. Then, there exists an ONB $\{a^1,\dots,a^n\}$ of eigenvectors of A such that with the (regular) matrix $B = [a^1 \dots a^n]$ there holds $A = B\,D\,B^{-1}$ with $D = \mathrm{diag}_i(\lambda_i)$ ($\lambda_i > 0$ the eigenvalues of A). Now define

$$A^{1/2} := B\,\mathrm{diag}_i\big(\lambda_i^{1/2}\big)B^{-1}.$$

This is well defined and independent of the concrete choice of B.

c) Let $A \in \mathbb{C}^{n\times n}$ be positive definite, i. e., $\bar{x}^T Ax \in \mathbb{R}_+$, $x \in \mathbb{C}^n$. Then, A is necessarily Hermitian since for $x, y \in \mathbb{C}$ arbitrary there holds:

$$\left\{\begin{array}{r} (\overline{x+y})^T A(x+y) \in \mathbb{R}, \\ (\overline{x+iy})^T A(x+iy) \in \mathbb{R} \end{array}\right.$$
$$\Longrightarrow \quad \left\{\begin{array}{l} \bar{x}^T Ax + \bar{y}^T Ay + (\bar{x}^T Ay + \bar{y}^T Ax) \in \mathbb{R}, \\ \bar{x}^T Ax + \bar{y}^T Ay + i\,(\bar{x}^T Ay - \bar{y}^T Ax) \in \mathbb{R}. \end{array}\right.$$

Setting $x = e_i$ and $y = e_j$, we see that $a_{ij} + \bar{a}_{ji} \in \mathbb{R}$ and $i(a_{ij} - \bar{a}_{ji}) \in \mathbb{R}$, i.e.,

$$\begin{aligned} &\mathrm{Re}(a_{ji} + a_{ij}) + i\mathrm{Im}(a_{ji} + a_{ij}) \in \mathbb{R}, \\ &i\mathrm{Re}(a_{ji} - a_{ij}) + \mathrm{Im}(a_{ji} - a_{ij}) \in \mathbb{R}. \end{aligned}$$

Hence, $a_{ij} = \mathrm{Re}\,a_{ij} + i\mathrm{Im}\,a_{ij} = \mathrm{Re}\,a_{ji} - i\mathrm{Im}\,a_{ji} = \mathrm{Re}\,\bar{a}_{ji} + i\mathrm{Im}\,\bar{a}_{ji} = \bar{a}_{ji}$.
Remark: The above argument only uses that $\bar{x}^T Ax \in \mathbb{R}$, $x \in \mathbb{C}^n$.

Solution A.1.5: a) $\|v\|_\infty = \max_{1\le i\le n}|v_i|$ and $\|v\|_1 = \sum_{i=1}^n |v_i|$.

b) The "spectrum" $\Sigma(A)$ is defined as $\Sigma(A) := \{\lambda \in \mathbb{C},\ \lambda \text{ eigenvalue of } A\}$.

c) The "Gerschgorin circles" $K_i \subset \mathbb{C}$, $i=1,\dots,n$, are the closed discs
$K_i := \big\{x \in \mathbb{C},\ |x - a_{ii}| \le \sum_{j\ne i} a_{ij}\big\}$.

d) $\rho(A) = \max_{1\le i\le n}\{|\lambda_i|,\ \lambda_i \text{ eigenvalue of } A\}$.

e) $\kappa_2(A) = \|A\|_2\|A^{-1}\|_2 = \max_{1\le i\le n}\sigma_i / \min_{1\le i\le n}\sigma_i$, where σ_i are the "singular values" of A, i. e., the square roots of the (nonnegative) eigenvalues of $\bar{A}^T A$).

Solution A.1.6: a) $a_{ii} \in \mathbb{R}$ follows directly from the property $a_{ii} = \bar{a}_{ii}$ of a Hermitian matrix. Positiveness follows via testing by e_i, which yields $a_{ii} = \bar{e}_i^T Ae_i > 0$.

b) The trace of a matrix is invariant under coordinate transformation, i. e. similarity transformation (may be seen by direct calculation $\sum_{ijk} b_{ij}a_{jk}b_{ki} = \sum_i a_{ii}$ or by applying the product formula for determinants to the characteristic polynomial. Observing that a Hermitian matrix is similar to a diagonal matrix with its eigenvalues on the main diagonal implies the asserted identity.

c) Assume that A is singular. Then $\ker(A) \ne \emptyset$ and there exists $x \ne 0$ such that $Ax = 0$, i. e., zero is an eigenvalue of A. But this contradicts the statement of Gerschgorin's Lemma which bounds all eigenvalues away from zero due to the strict diagonal dominance. If

all diagonal entries $a_{ii} > 0$, then also by Gerschgorin's lemma all Gerschgorin circles (and consequently all eigenvalues) are contained in the right complex half-plane. If A is additionally Hermitian, all these eigenvalues are real and positive and A consequently positive definite.

Solution A.1.7: Define

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

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

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

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

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

Solution A.1.8: To proof the statement, we use a so-called "deformation argument". For $t \in [0,1]$ define the matrix

$$A(t) = (1-t)\mathrm{diag}_i(a_{ii}) + t\,A.$$

Obviously $A(0)$ is a diagonal matrix with eigenvalues $\lambda_i(0) = a_{ii}$. Now observe that the "evolution" of the ith eigenvalue $\lambda_i(t)$ is a continuous function in t (This follows from the fact that a root t_0 of a polynomial p_α is locally arbitrarily differentiable with respect to its coefficients – a direct consequence of the implicit function theorem employed to $p(\alpha, t) = p_\alpha(t)$ and a special treatment of multiple roots).

Furthermore, the Gerschgorin circles of $A(t)$, $0 \leq t \leq 1$ have all the same origin, only the radii are strictly increasing. So, Gerschgorin's Lemma implies that the image of the function $t \to \lambda_i(t)$ lies entirely in the union of all Gerschgorin circles of $A(1)$. And due to the fact that it is continous obviously in the connected component containing a_{ii}.

Solution A.1.9: (i)⇒(ii): Suppose that A and B commute. First observe that for an arbitrary eigenvector z of B with eigenvalue λ there holds:

$$ABz - BAz = 0 \quad \Rightarrow \quad BAz = \lambda Az.$$

So, Az is either 0 or also an eigenvector of B with eigenvalue λ. Due to the fact that B is Hermitian there exists an orthonormal basis $\{v_i\}_{i=1}^n$ of eigenvectors of B. So we

can transfrom B by a change of basis to a diagonal matrix. Furthermore, by virtue of the observation above, A has block diagonal structure with respect to this basis, where a single block solely acts on an eigenspace $E_\lambda(B)$ for a specific eigenvalue λ of B. Due to the fact that A is also hermitian, we can diagonalize $A\big|_{E_\lambda(B)}$ with respect to this subspace by another change of basis. Now observe that the diagonal character of $B\big|_{E_\lambda(B)} = \lambda I$ will be preserved.

(ii)⇒(i): Let $O = \{v_i\}_{i=1}^n$ be the common basis of eigenvectors of A and B, one checks that

$$ABv_i = \lambda_i^A \lambda_i^B v_i = \lambda_i^B \lambda_i^A v_i = BAv_i, \quad i = 1, \ldots, n.$$

Consequently, $ABx = BAx$, $x \in \mathbb{K}^n$, and therefore $AB = BA$.

(i)⇔(iii): For any two Hermitian matrices A and B there holds $BA = \bar{B}^T \bar{A}^T = \overline{AB}^T$ and the asserted equivalence follows immediately.

Solution A.1.10: i) Let $A \in \mathbb{K}^{n\times n}$ be an arbitrary, regular matrix and define $\varphi(x, y) := (Ax, Ay)_2$. It is clear that φ is a sesquilinear form. Furthermore symmetry and positivity follow directly from the corresponding property of $(\,.\,)$. For definiteness observe that $(Ax, Ax) = 0 \Rightarrow Ax = 0 \Rightarrow x = 0$ due to the regularity of A.

ii) The earlier result does not contradict (i) because there holds

$$(Ax, Ay)_2 = (x, \bar{A}^T Ay)_2,$$

and $\bar{A}^T A$ is a hermitian matrix.

Solution A.1.11: i) Let λ_1 and λ_2 be two pairs of eigenvalues with eigenvectors v^1 and v^2. It holds:

$$0 = (v^1, Av^2) - (v^1, Av^2) = (v^1, Av_2) - (Av^1, v^2) = (\lambda_2 - \lambda_1)(v^1, v^2).$$

So, if $\lambda_1 \neq \lambda_2$ it must hold that $(v^1, v^2) = 0$. Yes, this result is true in general for normal matrices (over $\mathbb{C}$) and—more generally—known as the "spectral theorem for normal operators" (see [Bosch, Lineare Algebra, p. 266, Satz 7.5/8], for details).

ii) Let v be an eigenvector for the eigenvalue $\lambda_{\max}$. There holds:

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

Conversely, for arbitrary $x \in \mathbb{K}^n \setminus \{0\}$ there exists a representation $x = \sum_i x_i\, v_i$ with respect to an orthonormal basis $\{v_i\}$, so that

$$\frac{(Ax, x)_2}{\|x\|_2^2} = \frac{(A\sum_i x_i v_i, \sum_i x_i v_i)_2}{\|x\|_2^2} = \frac{\sum_i \lambda_i x_i^2}{\sum_i x_i^2} \leq \lambda_{\max}(A)$$

The corresponding equality for $\lambda_{\min}(A)$ follows by a similar argument. $\lambda_{\min}(A) \leq \lambda_{\max}(A)$ is obvious.

Solution A.1.12: i) We use the definition (c) from the text for the ε-pseudo-spectrum. Let $z \in \sigma_\varepsilon(A)$ and accordingly $v \in \mathbb{K}^n$, $\|v\| = 1$, satisfying $\|(A - zI)v\| \leq \varepsilon$. Then,

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

This proves the asserted relation.

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

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

Hence, setting $w := \|A^{-1}v\|^{-1}A^{-1}v \in \mathbb{K}^n$ with $\|w\| = 1$, we obtain

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

Hence, observing that

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

we conclude that

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

This completes the proof.

A.2 Chapter 2

Solution A.2.1: a) An example for a symmetric, diagonally dominant matrix that is indefinite is

$$A = \begin{pmatrix} -2 & 1 \\ 1 & 2 \end{pmatrix}.$$

On the other hand, a symmetric, positive definite but not (strictly) diagonally dominant matrix is given by

$$B = \begin{pmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{pmatrix},$$

or typically system matrices arising from higher order difference approximations, e. g. the "stretched" 5-point stencil for the Laplace problem in 1D:

$$B^n = \frac{1}{12h}\begin{pmatrix} 30 & -16 & 1 & & & & \\ -16 & 30 & -16 & 1 & & & \\ 1 & -16 & 30 & -16 & 1 & & \\ & & & \ddots & & & \\ & & 1 & -16 & 30 & -16 & 1 \\ & & & 1 & -16 & 30 & -16 \\ & & & & 1 & -16 & 30 \end{pmatrix} \in \mathbb{R}^{n\times n}.$$

Note: To prove that the above $B^n \in \mathbb{R}^{n\times n}$ is positive definite, compute $\det(B^k) > 0$ for $k = 1, \cdots, 3$ and derive a recursion formula of the form

$$\det(B^{k+1}) = 30\det(B^k) \pm \cdots \det(B^{k-1}) \pm \cdots \det(B^{k-2})$$

so that $\det(B^{k+1}) > 0$ follows by induction.

b) Apply the Gerschgorin lemma to the adjoint transpose $\bar{A}^T$. This yields that $0 \notin \sigma(\bar{A}^T)$, i. e., $\bar{A}^T$ is regular. Then, also A is regular.

c) All eigenvalues of the symmetric matrix A are real. Further, all Gerschgorin circles have their centers on the positive real half axis. Hence the strict diagonal dominance implies that all eigenvalues must be positive.

Solution A.2.2: The result of the first $k-1$ elimination steps is a block matrix $A^{(k-1)}$ of the form

$$A^{(k-1)} = \left[\begin{array}{c|c} \overline{R}^{k-1} & * \\ \hline 0 & \overline{A}^{k-1} \end{array}\right], \quad \text{with } \overline{A}^{k-1} \in \mathbb{R}^{(n-k)\times(n-k)} \text{ pos. def. (by induction).}$$

The k-th elimination step reads:

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - \frac{a_{ik}^{(k-1)} a_{kj}^{(k-1)}}{a_{kk}^{(k-1)}}, \quad i,j = k,\ldots,n.$$

i) The main diagonal elements of positive definite matrices are positive, $a_{jj}^{(k-1)} > 0$. For the diagonal elements it follows by symmetry:

$$a_{ii}^{(k)} = a_{ii}^{(k-1)} - \frac{a_{ik}^{(k-1)} a_{ki}^{(k-1)}}{a_{kk}^{(k-1)}} = a_{ii}^{(k-1)} - \frac{|a_{ik}^{(k-1)}|^2}{a_{kk}^{(k-1)}} \le a_{ii}^{(k-1)}, \quad i = k,\ldots,n.$$

ii) The element with maximal modulus of a positive definite matrix $\overline{A}^{(k-1)}$ lies on the main diagonal,

$$\max_{k\le i,j\le n} |a_{ij}^{(k-1)}| \le \max_{k\le i\le n} |a_{ii}^{(k-1)}|.$$

The submatrix $\overline{A}^{(k)}$ obtained in the k-th step is again positive definite. Hence the result (i) implies

$$\max_{k\le i,j\le n} |a_{ij}^{(k)}| \le \max_{k\le i\le n} |a_{ii}^{(k)}| \le \max_{k\le i\le n} |a_{ii}^{(k-1)}| \le \max_{k\le i,j\le n} |a_{ij}^{(k-1)}|.$$

Since in the k-th elimination step the first $k-1$ rows are not changed anymore induction with respect to $k = 1,\ldots,n$ yields:

$$\max_{1\le i,j\le n} |r_{ij}| = \max_{1\le i,j\le n} |a_{ij}^{(n-1)}| \le \max_{1\le i,j\le n} |a_{ij}^{(0)}| \le \max_{1\le i,j\le n} |a_{ij}|.$$

Solution A.2.3: i) Let

$$\mathcal{L} := \{L \in \mathbb{R}^{n\times n},\ L \text{ regular, lower-left triangular matrix mit } l_{ii} = 1\},$$
$$\mathcal{R} := \{R \in \mathbb{R}^{n\times n},\ L \text{ regular upper-right triangular matrix}\}.$$

We have to show the following group properties for the matrix multiplication $\circ$:
(G1) Closedness: $L_1, L_2 \in \mathcal{L} \Rightarrow L_1 \circ L_2 \in \mathcal{L}$.
(G2) Associative law: $L_1, L_2, L_3 \in \mathcal{L} \Rightarrow L_1 \circ (L_2 \circ L_3) = (L_1 \circ L_2) \circ L_3$.
(G3) Neutral element I: $L \in \mathcal{L} \Rightarrow L \circ I = L$.
(G4) Inverse: $L \in \mathcal{L} \Rightarrow \exists L^{-1} \in \mathcal{L} : L \circ L^{-1} = I$.

(G1) follows by computation. (G2) and (G3) follow from the properties of matrix multiplication. (G4) is seen through determination of the inverse by simultaneous elimination:

$$\left[\begin{array}{ccc|ccc} 1 & & 0 & 1 & & 0 \\ & \ddots & & & \ddots & \\ 0 & & 1 & * & & 1 \end{array}\right] \quad \Rightarrow \quad L^{-1} = \left[\begin{array}{ccc} 1 & & 0 \\ & \ddots & \\ * & & 1 \end{array}\right] \in \mathcal{L}.$$

The group $\mathcal{L}$ is *not* abelian as the following 3×3 example shows:

$$\left[\begin{array}{ccc} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right]\left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{array}\right] = \left[\begin{array}{ccc} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & -1 & 1 \end{array}\right] \neq \left[\begin{array}{ccc} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & -1 & 1 \end{array}\right] = \left[\begin{array}{ccc} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right]\left[\begin{array}{ccc} 1 & 0 & 0 \\ 1 & 1 & 0 \\ -1 & -1 & 1 \end{array}\right].$$

The argument for $\mathcal{R}$ is analogous. The group $\mathcal{R}$ is also *not* abelian as the following 2×2 example shows:

$$\left[\begin{array}{cc} 1 & 1 \\ 0 & 1 \end{array}\right]\left[\begin{array}{cc} -1 & 1 \\ 0 & 1 \end{array}\right] = \left[\begin{array}{cc} -1 & 2 \\ 0 & 1 \end{array}\right] \neq \left[\begin{array}{cc} -1 & 0 \\ 0 & 1 \end{array}\right] = \left[\begin{array}{cc} -1 & 1 \\ 0 & 1 \end{array}\right]\left[\begin{array}{cc} 1 & 1 \\ 0 & 1 \end{array}\right].$$

ii) For proving the uniqueness of the LR-decompositiong let for a regular matrix $A \in \mathbb{R}^{n\times n}$ two LR-decompositiongs $A = L_1R_1 = L_2R_2$ be given. Then, by (i) $L_1, L_2 \in \mathcal{L}$ as well as $R_1, R_2 \in \mathcal{R}$ and consequently

$$\underbrace{R_1R_2^{-1}}_{\in\mathcal{R}} = \underbrace{L_1^{-1}L_2}_{\in\mathcal{L}} = \operatorname{diag}(d_{ii}).$$

With L_1 (and L_2) also the inverse L_1^{-1} has ones on the main diagonal. Hence $d_{ii} = 1$ which finally implies $R_1 = R_2$ and $L_2 = L_1$.

Solution A.2.4: Let A be a band matrix with $m_l = m_r =: m$ (Make a sketch of this situation.)

i) The k-th elimination step

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} a_{kj}^{(k-1)}, \quad b_i^{(k)} = b_i^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} b_k^{(k-1)}, \quad i,j = k+1,\ldots,k+m,$$

requires essentially m divisions and m^2 multiplications and additionas. Hence alltogether

$$N_{\text{band}} = nm^2 + O(nm) \quad \text{a. op.},$$

for the $n-1$ steps of the forward elimination for computing the matrix R and simultanously of the matrix L. For the sparse model matrix, we have $N_{\text{band}} = 10^8 + O(10^6)$ a. op. in contrast to $N = \frac{1}{3}10^{12} + O(10^8)$ a. op. for a full matrix.

ii) If A is additionally symmetric (and positive definite) one obtaines the Cholesky decomposition from the LR decomposition by

$$A = \tilde{L}\tilde{L}^T, \qquad \tilde{L} = LD^{1/2},\ D = \text{diag}(r_{ii}).$$

Because of the symmetry of all resulting rduced submatrices only the elements on the main diagonal and the upper diagonals need to be computed. This reduces the work to $N_{\text{band}} = \frac{1}{2}nm^2 + O(nm)$ a. Op., i. e., for the model matrix to $N_{\text{band}} = \frac{1}{2}10^8 + O(10^6)$ a. op., and $N_{\text{band}} = \frac{1}{2}10^{16} + O(10^{12})$ a. op., respectively.

Solution A.2.5: a) The first step of the Gaussian elimination applied on the extended matrix $[A|b]$ produces:

$$\left[\begin{array}{ccc|c} 1 & 3 & -4 & 1 \\ 3 & 9 & -2 & 1 \\ 4 & 12 & -6 & 1 \\ 2 & 6 & 2 & 1 \end{array}\right] \quad \rightarrow \quad \left[\begin{array}{ccc|c} 1 & 3 & -4 & 1 \\ 0 & 0 & 10 & -2 \\ 0 & 0 & 10 & -3 \\ 0 & 0 & 10 & -1 \end{array}\right].$$

The linear system is not solvable because of $\text{rank}\, A = 2 \neq 3 = \text{rank}\,[A|b]$. Observe in particular, that A does not have full rank.

b) A straightforward calculation leads to the following normal equation:

$$\left[\begin{array}{cccc} 1 & 3 & 4 & 2 \\ 3 & 9 & 12 & 6 \\ -4 & -2 & -6 & 2 \end{array}\right] \left[\begin{array}{ccc} 1 & 3 & -4 \\ 3 & 9 & -2 \\ 4 & 12 & -6 \\ 2 & 6 & 2 \end{array}\right] \left[\begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array}\right] = \left[\begin{array}{cccc} 1 & 3 & 4 & 2 \\ 3 & 9 & 12 & 6 \\ -4 & -2 & -6 & 2 \end{array}\right] \left[\begin{array}{c} 1 \\ 1 \\ 1 \\ 1 \end{array}\right].$$

Hence,

$$\begin{bmatrix} 30 & 90 & -30 \\ 90 & 270 & -90 \\ -30 & -90 & 60 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 10 \\ 30 \\ -10 \end{bmatrix}.$$

Because of $\operatorname{Rank} A = 2 < 3$, the kernel of the matrix $A^T A \in \mathbb{R}^{3\times 3}$ is one dimensional. Gaussian elimination:

$$\left[\begin{array}{ccc|c} 30 & 90 & -30 & 10 \\ 90 & 270 & -90 & 30 \\ -30 & -90 & 60 & -10 \end{array}\right] \rightarrow \left[\begin{array}{ccc|c} 30 & 90 & -30 & 10 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 30 & 0 \end{array}\right] \rightarrow \left[\begin{array}{ccc|c} 3 & 9 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0, \end{array}\right]$$

and the solution can be characetized by $x = (\frac{1}{3} - 3t, t, 0)^T$, $t \in \mathbb{R}$.

c) The system of normal equations is solvable but the solution is not unique.

d) No. Due to the fact that A does not have full rank, the matrix $A^T A \in \mathbb{R}^{3\times 3}$ cannot be one-to-one, and can consequently be only semi-definite. (Counter example: $x = (-3, 1, 0)^T$ is a non trivial element of the kernel of $A^T A$.)

A.3 Chapter 3

Solution A.3.1: i) For the maximal absolute column sum it holds

$$\|B\|_1 = \max_{j=1,2,3} \sum_{i=1}^{3} |a_{ij}| = 0.9 < 1.$$

This implies convergence due to the fact that $\operatorname{spr}(B) \leq \|B\|_1 < 1$ and, hence, the iteration is contractive. (Observe that the maximal absolute row sum does not imply convergence because $\|B\|_\infty = 1.4 > 1$.) The limit $z = \lim_{t\to\infty} x^t$ fullfills $z = B\,z + c$. Hence,

$$z = (I - B)^{-1} c.$$

ii) Let λ_i be the eigenvalues of B. It holds

$$\prod_{i=1}^{3} \lambda_i = \det(B) = -1.$$

This implies that at least for one of the eigenvalues there must hold $|\lambda| \geq 1$. So, for the choice (ii) the fixed point iteration cannot be convergent in general: In particular, if $x^0 - x$ happens to be an eigenvector corresponding to the above eigenvalue λ it holds

$$\|x^t - x\| = \|B^t(x^0 - x)\| = \|\lambda^t(x^0 - x)\| = |\lambda|^t \|x^0 - x\| \not\to 0 \quad (t \to \infty).$$

Solution A.3.2: For a general fixed point iteration $x^{t+1} = Bx^t + c$ the following error

estimate holds true in case of convergence to a limit z:

$$\|x^t - z\| \leq \mathrm{spr}(B)^t \|x^0 - z\|.$$

It follows by induction that in order to reduce the initial error by at least a factor of ε it is necessary to perform the following number of iterations:

$$t = \left\lceil \frac{\log_{10}(\varepsilon)}{\log_{10}(\mathrm{spr}(B))} \right\rceil.$$

For the Jacobi- and Gauss-Seidel-Matrix it holds

$$J = \begin{bmatrix} 0 & 1/3 \\ 1/3 & 0 \end{bmatrix}, \qquad H_1 = \begin{bmatrix} 0 & -1/3 \\ 0 & 1/9 \end{bmatrix},$$

hence, $\mathrm{spr}(J) = 1/3$ and $\mathrm{spr}(H_1) = 1/9$. Therefore, the necessary number of iterations is

$$t_J = \left\lceil \frac{6}{\log_{10}(3)} + 1 \right\rceil = 13, \quad \text{and} \quad t_{H_1} = \left\lceil \frac{6}{\log_{10}(9)} \right\rceil = 7, \text{ respectively.}$$

Solution A.3.3: We restate the two definitions of "irreducibility":

a) (With the help of the hint): A matrix $A \in \mathbb{R}^{n\times n}$ is called "irreducible" if for every partition J, K of $\{1, \ldots, n\} =: \mathbb{N}_n$ with $J \cup K = \mathbb{N}_n$ and $J \cap K = \emptyset$, so that $a_{jk} = 0$ for all $j \in J$ and all $k \in K$, it holds that either $K = \emptyset$ or $J = \emptyset$.

b) A matrix $A \in \mathbb{R}^{n\times n}$ is called "irreducible" if for every pair of indices $j, k \in \mathbb{N}_n$ there exists a set of indices $\{i_1, \ldots, i_m\} \in \mathbb{N}_n$ such that $a_{j,i_1} \neq 0$, $a_{i_1,i_2} \neq 0$, $\ldots$, $a_{i_{m-1},i_m} \neq 0$, $a_{i_m,k} \neq 0$.

i) (a) $\Rightarrow$ (b): Let A be irreducible in the sense of (a). Furthermore, let $i \in \mathbb{N}_n$ be an arbitrary index. Let J be the set of all indices $l \in \mathbb{N}_n$, with the property that there exists a sequence of indicies $\{i_1, \ldots, i_m\} \in \mathbb{N}_n$ such that all $a_{i,i_1}, \ldots, a_{i_m,l} \neq 0$. Define its complement $K := \mathbb{N}_n \setminus J$. In order to prove (b) we have to show that $J = \mathbb{N}_n$, or that $K = \emptyset$, respectively.

First of all, it holds that $i \in J$, so J is not empty. Furthermore, observe that for all $p \in K$ it must hold that $a_{l,p} = 0$ for all $l \in J$, otherwise there would exist a sequence from i to p by expanding an arbitrary sequence $a_{i,i_1}, \ldots, a_{i_m,l}$ from i to l (which exists by virtue of $l \in J$) by $a_{l,p}$. So by irreducibility in the sense of (a) it must hold that $K = \emptyset$.

ii) (b) $\Rightarrow$ (a): Let A be irreducible in the sense of (b). Let $\{J, K\}$ be an arbitrary partition in the sense of (a). Then for arbitrary index pairs $\{j, k\} \in J \times K$ there exists a sequence $\{i_1, \ldots, i_m\} \in \mathbb{N}_n$ with $a_{j,i_1} \neq 0$, $\ldots$, $a_{i_m,k} \neq 0$. Inductively, because of $a_{i_\mu,i_\nu} \neq 0$, it follows that $i_\mu \in J$ for $\mu = 1, \ldots, m$, and finally (because of $a_{i_m,k} \neq 0$) also $k \in J$ in contradiction to the original choice $\{j, k\} \in J \times K$. So it must hold that either $J = \emptyset$ or $K = \emptyset$.

Solution A.3.4: a) In case of the matrix A_1, the iteration matrices for the Jacobi and

Gauss-Seidel methods are

$$J = -D^{-1}(L+R) = \begin{bmatrix} 0 & 0.5 & -1 \\ -0.5 & 0 & 1 \\ -1 & -1 & 0 \end{bmatrix}, \quad H_1 = -(D+L)^{-1}R = \begin{bmatrix} 0 & 0.5 & -1 \\ 0 & -0.25 & 1.5 \\ 0 & -0.25 & -0.5 \end{bmatrix}$$

The eigenvalues λ_i of J fulfill $\lambda_1\lambda_2\lambda_3 = \det(J) = -1$, hence $\mathrm{spr}(J) \geq 1$. Therefore, the Jacobi iteration cannot be convergent in general. The matrix H_1 has the characteristic polynomial $\chi(\lambda) = -\lambda(\lambda^2 + \frac{3}{4}\lambda + \frac{1}{2})$ and the eigenvalues $\lambda_1 = 0$, $\lambda_{2/3} = \pm 1/\sqrt{2}$. Consequently, $\mathrm{spr}(H_1) < 1$ and the Gauss-Seidel method is convergent.

b) The matrix A_2 fulfills the weak row sum criterion and is irreducible. Hence, the Jacobi- and Gauss-Seidel methods converge

Solution A.3.5: First, we determine the iteration matrix: It holds

$$\begin{bmatrix} 1 & 0 \\ -\omega a & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 \\ \omega a & 1 \end{bmatrix}.$$

Hence,

$$x^t = \begin{bmatrix} 1 & 0 \\ \omega a & 1 \end{bmatrix} \begin{bmatrix} 1-\omega & \omega a \\ 0 & 1-\omega \end{bmatrix} x^{t-1} + \omega \begin{bmatrix} 1 & 0 \\ \omega a & 1 \end{bmatrix} b,$$

and therefore:

$$B_\omega = \begin{bmatrix} 1-\omega & \omega a \\ \omega a(1-\omega) & \omega^2 a^2 + 1 - \omega \end{bmatrix}.$$

Consequently, it is $\det(B_\omega - \lambda I) = -\lambda\omega^2 a^2 + (1-\omega-\lambda)^2$.

a) With $\omega = 1$ it is

$$\det(B_1 - \lambda I) = -\lambda a^2 + \lambda^2 \quad \Rightarrow \quad \mathrm{spr}(B_1) = a^2,$$

thereby, for $|a| < 1$ the system is convergent.

b) In case of $a = \frac{1}{2}$ it holds

$$\lambda_{1,2} = 1 - \omega + \tfrac{1}{8}\omega^2 \pm \tfrac{1}{2}\omega\sqrt{1 - \omega + \tfrac{1}{16}\omega^2}.$$

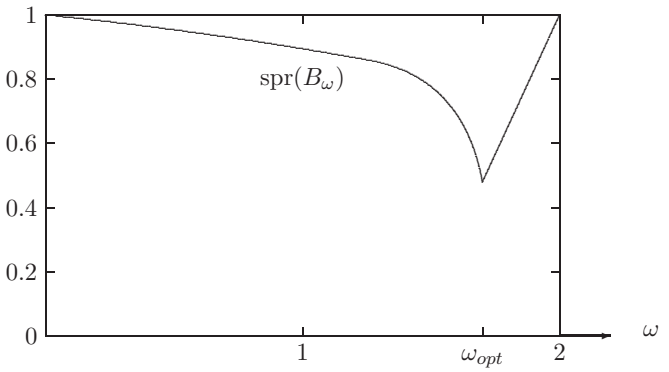
In case of $1 - \omega + \frac{1}{16}\omega^2 \geq 0$, and $\omega \leq 8 - 4\sqrt{3} = 1.07179677\ldots$, respectively, both roots are real valued. For any other choice of ω they are complex. Therefore:

$$\mathrm{spr}(B_\omega) = \begin{cases} 1 - \omega + \frac{1}{8}\omega^2 + \frac{1}{2}\omega\sqrt{1 - \omega + \frac{1}{16}\omega^2}, & 0 \leq \omega \leq 8 - 4\sqrt{3}, \\ \omega - 1, & 8 - 4\sqrt{3} < \omega \leq 2. \end{cases}$$

Evaluating the formula for the stated values:

ω	0.8	0.9	1.0	1.1	1.2	1.3	1.4
$\mathrm{spr}(B_\omega)$	0.476	0.376	0.25	0.1	0.2	0.3	0.4

The graph of the function $\rho(\omega) := \mathrm{spr}(B_\omega)$, $0 \le \omega \le 2$, starts with $\rho(0) = 1$; it has a minimum at $\omega_{\mathrm{opt}} := 8 - 4\sqrt{3}$ with a sharp, down-pointing cusp, behind that it increases linear to $\rho(2) = 1$.



Graph of the spectral radius $\mathrm{spr}(H_\omega)$ *plotted over* $\omega \in [0, 2]$

Solution A.3.6: Let T be a matrix consisting of a (row wise) ONB of A. It holds

$$T^{-1}A^0T = I, \qquad T^{-1}A^1T = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d, \end{pmatrix}$$

and

$$T^{-1}A^kT = T^{-1}AT \cdot T^{-1}AT \cdots T^{-1}AT = \begin{pmatrix} \lambda_1^k & & \\ & \ddots & \\ & & \lambda_d^k \end{pmatrix} \quad \forall k \in \mathbb{N}.$$

So, by virtue of linearity,

$$T^{-1}p(A)T = \begin{pmatrix} p(\lambda_1) & & \\ & \ddots & \\ & & p(\lambda_d) \end{pmatrix}$$

for an arbitrary polynomial p. So, the spectral radius of $p(A)$ is exactly $\max_{i=1,\ldots,n} |p(\lambda_i)|$.

Solution A.3.7: a) It holds

$$X_t = g(X_{t-1}), \qquad g(X) := X(I - AC) + C.$$

Hence,

$$\|g(X) - g(Y)\| = \|(X - Y)(I - AC)\| \leq \|X - Y\| \, \|I - AC\|.$$

Therefore, if $\|I - AC\| =: q < 1$, then g is a contraction. The corresponding fixed-point iteration converges for every initial value X_0. The limit Z fulfills the equation $Z = Z(I - AC) + C$ or $ZAC = C$. This is equivalent to $Z = A^{-1}$.

So, if $q < 1$ the fixed point iteration converges for every initial value $X_0 \in \mathbb{R}^{n \times n}$ to the limit A with the a priori error estimate

$$\|X_t - A^{-1}\| \leq q^t \|X_0 - A^{-1}\|, \quad t \in \mathbb{N}.$$

b) We have

$$X_t = g(X_{t-1}), \qquad g(X) := X(2I - AX).$$

Let Z be an arbitrary fixed point of g. It necessarily fulfills the equation $Z = Z(2I - AZ)$ or $Z = ZAZ$. Suppose that Z is regular, then $Z = A^{-1}$. Note that this assumption is essential because the singular matrix $Z = 0$ is always a valid fixed point of g. To prove convergence (under a yet to be stated assumption) we observe that:

$$\begin{aligned} X_t - Z &= 2X_{t-1} - X_{t-1}AX_{t-1} - Z \\ &= -X_{t-1}AX_{t-.1} + \underbrace{ZA}_{=I} Y_{t-1} + X_{t-1} \underbrace{AZ}_{=I} - \underbrace{ZA}_{=I} Z \\ &= -(X_{t-1} - Z)A(X_{t-1} - Z). \end{aligned}$$

This implies

$$\|X_t - Z\| \leq \|A\| \, \|X_{t-1} - Z\|^2.$$

So, for $Z = A^{-1}$ and under the condition that

$$\|X_0 - Z\| < \frac{1}{\|A\|}$$

the iteration converges quadratically to Z:

$$\|A\| \, \|X_t - Z\| \leq \big(\|A\| \, \|X_{t-1} - Z\|\big)^2 \leq \ \ldots \ \leq \big(\|A\| \, \|X_0 - Z\|\big)^{2^t} \to 0 \quad (t \to \infty).$$

This iteration is exactly Newton's method for calculating the inverse of a matrix.

Remark: It is sufficient to choose a starting value X_0 that fulfills the convergence criterion (for the preconditioner C) in (a):

$$1 > \|I - AX_0\| = \|A\| \, \|A^{-1} - X_0\| \quad \Longleftrightarrow \quad \|A^{-1} - X_0\| < \frac{1}{\|A\|},$$

so, also the criterion of (b) is fulfilled.

Solution A.3.8: Let J be the Jordan normal form of B and T a corresponding transformation matrix such that

$$T^{-1}BT = J.$$

Let $p(X) = \alpha_0 + \alpha_1 X + \alpha_2 X^2 + \cdots + \alpha_k X^k$ be an arbitrary polynomial. Then,

$$T^{-1}p(B)T = \alpha_0 + \alpha_1 T^{-1}BT + \alpha_2 (T^{-1}BT)^2 + \cdots + \alpha_k (T^{-1}BT)^k = p(J).$$

Furthermore, observe that multiplication (or addition) of upper triangular matrices yields another upper triangular matrix, whose diagonal elements are formed by elementwise multiplication (or addition) of the corresponding diagonal elements of the multiplicands. Consequently, $p(J)$ is an upper triangular matrix of the form

$$p(J) = \begin{pmatrix} p(\lambda_1) & & & \\ & p(\lambda_2) & * & \\ & 0 & \ddots & \\ & & & p(\lambda_n) \end{pmatrix},$$

where λ_i are the eigenvalues of B. Hence,

$$\chi_{p(J)}(\lambda) = \det\big(p(J) - \lambda I\big) = \prod_{1 \le i \le n} (p(\lambda_i) - \lambda),$$

which proves the assertion.

Solution A.3.9: In case of a symmetric matrix A, the Jacobi methods reads

$$x^t = -D^{-1}(L + L^T)x^{t-1} + D^{-1}b,$$

with the iteration matrix

$$B = -D^{-1}(L + L^T).$$

The idea of the Chebyshev acceleration is now to construct a sequence of improved approximations $y^t - x = p_t(B)(x^0 - x)$ (instead of the ordinary fixed point iteration $x^t = B^t x^0$) by a smart choice of polynomials

$$p_t(z) = \sum_{s=0}^{t} \gamma_s^t z^s, \quad p_t(1) = 1.$$

It holds

$$\|y^t - x\|_2 \le \|p_t(B)\|_2 \, \|x^0 - x\|_2,$$

with $\|p_t(B)\|_2 = \max_{\lambda \in \sigma(B)} |p(\lambda)|$. So, the optimal choice for the polynomial $p_t(z)$ would be the solution of the minimzation problem

$$\min_{p \in P_t, p(1)=1} \max_{\lambda \in \sigma(B)} |p(\lambda)|.$$

Unfortunately, this is practically impossible because $\sigma(B)$ is usually unknown. But, under the assumption that the Jacobi method is already convergent it holds that

$$\max_{\lambda\in\sigma(B)} |p(\lambda)| \in [-1+\delta, 1-\delta],$$

due to the fact that the resulting iteration matrix B is similar to a symmetric matrix

$$D^{-1/2}(L+L^T)D^{-1/2}.$$

This motivates the modified optimization problem

$$\min_{p\in P_t, p(1)=1} \max_{|x|\leq 1-\delta} |p(x)|.$$

This optimization problem can be solved analytically. The solutions are given by rescaled Chebyshev polynomials:

$$p_t(x) := C_t(x) = \frac{T_t\left(\frac{x}{1-\delta}\right)}{T_t\left(1+\frac{\delta}{2-2\delta}\right)}.$$

Solution A.3.10: a) No, the damped Richardson equation cannot be made convergent in general. A necessary (and sufficient) condition for convergence of the damped Richardson equation (applied to a symmetric coefficient matrix) for arbitrary starting values is that

$$\operatorname{spr}\left(\begin{bmatrix} I & O \\ O & I \end{bmatrix} - \theta \begin{bmatrix} A & B \\ B^T & O \end{bmatrix}\right) < 1.$$

For this to hold true, it is necessary that the eigenvalues of the coefficient matrix are sufficiently small – this can be controlled by θ and is therefore not a problem – and that all eigenvalues are positive. But this does not need to be the case, consider, e. g.,

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} -1 \\ 0 \end{pmatrix}.$$

In this case the coefficient matrix

$$\begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \end{pmatrix}$$

has two positive and one negative eigenvalue $\lambda_1 = 1$, $\lambda_{2,3} = \frac{1}{2} \pm \frac{\sqrt{5}}{2}$.

b) With $A = (a_{ij})$, $B = (b_{ij})$ and employing the fact that A is symmetric and positive definite it holds

$$\left(B^T AB\right)_{il} = \sum_{jk} b_{ji} a_{jk} b_{kl} = \sum_{jk} b_{kl} a_{kj} b_{kl} = \left(B^T AB\right)_{li} \quad 1 \le i, j \le m,$$

$$x^T B^T ABx \ge \|Bx\|^2 \ge 0 \quad \forall x \in \mathbb{R}^m.$$

So, $B^T AB$ is symmetric, positive semidefinite. If $B : \mathbb{R}^m \to \mathbb{R}^n$ is a one to one mapping (because $m \le n$), then, $x^T B^T ABx = 0$ implies $\|Bx\| = 0$. This in turn implies $x = 0$. Therefore, if B has full rank, then $B^T AB$ is positive definite.

The Chebyshev acceleration is most efficiently realized by using the two step recursion formula:

$$\begin{aligned} \xi^t &= 2\frac{\mu_t}{\rho\,\mu_{t+1}} H_1^{\text{sym}} \xi^{t-1} - \frac{\mu_{t-1}}{\mu_{t+1}} \xi^{t-1} + 2\frac{\mu_t}{\rho\,\mu_{t+1}}\,\zeta \\ \mu_{t+1} &= \frac{2}{\rho}\mu_t - \mu_{t-1} \end{aligned}$$

starting from the initial values $\xi^0 = y^0$, $y^1 = H_1^{\text{sym}} y^0 + \zeta$, $\mu_0 = 1$ and $\mu_1 = 1/\rho$.

Hereby, the symmetrized Gauß-Seidel iteration matrix reads

$$H_1^{\text{sym}} = (D + L^T)^{-1} L\,(D + L)^{-1} L^T$$

and the corresponding right hand side of the iterative procedure is

$$\zeta = B^T A^{-1} b - c.$$

We assume that we have an efficient method in estimating the additive splitting

$$B^T A^{-1} B = L + D + L^T$$

and that an estimate $\rho \in (0, 1)$ with $\sigma(H_1^{\text{sym}}) \in (-\rho, \rho)$ is readily available.

Solution A.3.11: One step of the Gauß-Seidel method reads

$$\hat{x}_j^{(1)} = \frac{1}{a_{jj}} \left(b_j - \sum_{k<j} a_{jk} \hat{x}_k^{(1)} + \sum_{k>j} a_{jk} x_k^{(0)} \right).$$

Due to the specific choice of decent directions $r^{(t)} = e_{t+1}$ in the coordinate relaxation, there holds $x_j^{(t+1)} = x_j^{(t)}$ for $j \ne t+1$. Consequently, it suffices to show that in the step $t \to t+1$ the $(t+1)$-th component is set to the correct value. Inserting the step length

$$\alpha_{t+1} = \frac{g_{t+1}^{(t)}}{a_{t+1,t+1}} = \frac{1}{a_{t+1,t+1}} \left(b_{t+1} - \sum_k a_{t+1,k} x_k^{(t)} \right)$$

into the iteration procedure gives

$$\begin{aligned} x_{t+1}^{(t+1)} &= x_{t+1}^{(t)} + \frac{b_{t+1}}{a_{t+1,t+1}} - \frac{1}{a_{t+1,t+1}} \sum_k a_{t+1,k} x_k^{(t)} \\ &= \frac{1}{a_{t+1,t+1}} \left(b_{t+1} - \sum_{k<t+1} a_{t+1,k} x_k^{(t)} - \sum_{k>t+1} a_{t+1,k} x_k^{(t)} \right). \end{aligned} \tag{1.3.1}$$

By induction it follows that $x_k^{(t)} = \hat{x}_k^{(1)}$ for $k < t+1$. Furthermore, $x_k^{(t)} = x_k^{(0)}$ for $k > t+1$, so that:

$$x_{t+1}^{(t+1)} = \frac{1}{a_{t+1,t+1}} \left(b_{t+1} - \sum_{k<t+1} a_{t+1,k} \hat{x}_k^{(1)} - \sum_{k>t+1} a_{t+1,k} x_k^{(0)} \right). \tag{1.3.2}$$

Solution A.3.12: The CG method applied to the normal equation reads: Given an initial value x_0 and an initial decent direction

$$d^{(0)} = A^T(b - Ax^0) = -g^{(0)}$$

iterate by the prescription

$$\alpha_t = \frac{(g^{(t)}, g^{(t)})}{(Ad^{(t)}, Ad^{(t)})}, \quad y^{(t+1)} = y^{(t)} + \alpha_t d^{(t)}, \quad g^{(t+1)} = g^{(t)} + \alpha_t A^T A d^{(t)},$$
$$\beta_t = \frac{(g^{(t+1)}, g^{(t+1)})}{(g^{(t)}, g^{(t)})}, \quad d^{(t+1)} = -g^{(t+1)} + \beta_t d^{(t)}.$$

Remarkably, by efficiently storing and reusing intermediate computational results, there is only one additional matrix-vector multiplication involved in contrast to the original CG method – the term $A^T A d^{(t)}$ has to be computed instead of $Ad^{(t)}$.

The convergence speed, however, is linked to the eigenvalues of $A^T A$ by the result (given in the text) that in order to reduce the error by a factor of ε about

$$t(\varepsilon) \approx \frac{1}{2}\sqrt{\kappa}\, \ln\left(\frac{2}{\varepsilon}\right)$$

steps are required. Now,

$$\kappa = \mathrm{cond}_2(A^T A) = \frac{\max_{\lambda \in \sigma(A^T A)} |\lambda|}{\min_{\lambda \in \sigma(A^T A)} |\lambda|} = \frac{\max_{s \in S(A)} |s|^2}{\min_{s \in S(A)} |s|^2},$$

with the set of singular values $S(A)$ of A. This implies that for symmetric A the relation $\kappa(A^T A) = \kappa(A)^2$ holds and therefore a much slower convergence speed has to be expected.

Solution A.3.13: The asymptotic convergence speed

$$\limsup_{t \to \infty} \left(\frac{\|x^t - x\|}{\|x^0 - x\|} \right)^{1/t}$$

for the different methods in terms of $\kappa = \mathrm{cond}_2(A) = \Lambda/\lambda$ (with Λ maximal absolute eigenvalue and λ minimal absolute eigenvalue) are as follows:

$$\begin{aligned}
\text{Gauß-Seidel:} \quad & \operatorname{spr}(\mathrm{H}_1) = \operatorname{spr}(\mathrm{J})^2 = \Big(1 - \frac{1}{\kappa}\Big)^2 = 1 - 2\frac{1}{\kappa} + \mathcal{O}\big(\frac{1}{\kappa}^2\big), \\
\text{Optimal SOR:} \quad & \operatorname{spr}(\mathrm{H}_{\mathrm{opt}}) = \frac{1 - \sqrt{1 - \operatorname{spr}(\mathrm{J})^2}}{1 + \sqrt{1 - \operatorname{spr}(\mathrm{J})^2}} = 1 - \sqrt{8}\frac{1}{\sqrt{\kappa}} + \mathcal{O}\big(\frac{1}{\kappa}\big), \\
\text{Gradient method:} \quad & \Big(\frac{1 - 1/\kappa}{1 + 1/\kappa}\Big) = 1 - 2\,\frac{1}{\kappa} + \mathcal{O}\big(\frac{1}{\kappa}^2\big), \\
\text{CG method:} \quad & \Big(\frac{1 - 1/\sqrt{\kappa}}{1 + 1/\sqrt{\kappa}}\Big) = 1 - 2\,\frac{1}{\sqrt{\kappa}} + \mathcal{O}\big(\frac{1}{\kappa}\big).
\end{aligned}$$

Solution A.3.14: The CG method applied to the Schur complement

$$B^T A^{-1} B y = B^T A^{-1} b - c$$

reads: Given an initial value y_0 and an initial decent direction

$$d^{(0)} = B^T A^{-1}(b - By^0) - c = -g^{(0)}$$

iterate by the prescription

$$\begin{aligned}
\alpha_t &= \frac{(g^{(t)}, g^{(t)})}{(A^{-1}Bd^{(t)}, Bd^{(t)})}, \quad y^{(t+1)} = y^{(t)} + \alpha_t d^{(t)}, \quad g^{(t+1)} = g^{(t)} + \alpha_t B^T A^{-1} B d^{(t)}, \\
\beta_t &= \frac{(g^{(t+1)}, g^{(t+1)})}{(g^{(t)}, g^{(t)})}, \quad d^{(t+1)} = -g^{(t+1)} + \beta_t d^{(t)}.
\end{aligned}$$

Observe that in each step it is only necessary to compute two matrix vector products (one with B and one with B^T) and one matrix vector product with A^{-1} when evaluating $A^{-1}Bd^{(t)}$. This can be done with the help of an iterative method, e. g. with a preconditioned Richardson method (as introduced in the text)

$$\xi^t = \xi^{t-1} + C^{-1}(b - A\xi^{t-1}).$$

Different choices for the preconditioner C^{-1} are now possible, e. g. by choosing $C = \frac{1}{\omega}(D + \omega L)$ with $A = L + D + R$, one ends up with the SOR method. In practice, it is crucial to have a preconditioner that has good orthogonality preserving features, so one might use another Krylov space method as a preconditioner instead.

Solution A.3.15: There holds

$$\left(\frac{1 - 1/\kappa}{1 + 1/\kappa}\right)^{t(\varepsilon)} \le \varepsilon \quad \Longleftrightarrow \quad \left(\frac{\kappa - 1}{\kappa + 1}\right)^{t(\varepsilon)} \le \varepsilon \quad \Longleftrightarrow \quad \left(\frac{\kappa + 1}{\kappa - 1}\right)^{t(\varepsilon)} \ge \frac{1}{\varepsilon}.$$

Now, without loss of generality, both bases are greater than 1, so that

$$\Longleftrightarrow \quad \left(\frac{\kappa + 1}{\kappa + 1}\right) t(\varepsilon) \ge \log\big(\frac{1}{\varepsilon}\big).$$

Finally, observe that $\log\left(\frac{\kappa+1}{\kappa-1}\right) = 2\left\{\frac{1}{\kappa} + \frac{1}{3}\frac{1}{\kappa^3} + \cdots\right\} \geq 2\,\frac{1}{\kappa}$. Hence,

$$\Longleftarrow \quad 2\frac{1}{\kappa}\,t(\varepsilon) \geq \log\left(\frac{1}{\varepsilon}\right).$$

The corresponding result for the CG method follows by replacing κ with $\sqrt{\kappa}$.

Solution A.3.16: The matrix C can be written in the form $C = KK^T$ with the help of

$$K = \frac{1}{\sqrt{(2-\omega)\,\omega}}\Big(\frac{1}{\omega}D + L\Big)\,D^{-1/2}.$$

A close look reveals that the iteration matrix H_ω^{SSOR} of the SSOR method can be expressed in terms of C and A:

$$H_\omega^{\mathrm{SSOR}} = I - C^{-1}A.$$

In view of $\mathrm{spr}(H_\omega^{\mathrm{SSOR}}) < 1$, the inverse C^{-1} can be viewed as an approximation to A^{-1} that is suitable for preconditioning.

Solution A.3.17: For the model problem matrix A it holds that $\mathrm{spr(A)} < 1$. Hence, the inverse $(I-J)^{-1}$ is well defined and the Neumann series converges:

$$(I-J)^{-1} = \sum_{k=0}^{\infty} J^k.$$

Furthermore, with $J = I - D^{-1}A$ it follows that $(I-J)^{-1} = D\,A^{-1}$. Then,

$$A^{-1} = D^{-1}\sum_{k=0}^{\infty} J^k.$$

Finally, observe that the multiplication of two arbitrary matrices with non-negative entries yields another matrix with non-negative entries. Therefore the matrices $J^k = D^{-k}(-L-R)^k$ are elementwise non-negative. So, A^{-1} viewed as the sum of elementwise non-negative matrices has the same property.

Solution A.3.18: i) The stated inequality is solely a result of the special choice of $x_0 + K_t(d^0; A)$ as affine subspace for the optimization problem – it holds:

$$x^0 + K_t(d^0, A) = x^0 + \mathrm{span}\big\{A^0d^0, \cdots, A^{t-1}d^0\big\} = \big\{x^0 + p(A)d^0 \;:\; p \in P_{t-1}\big\}.$$

Furthermore, $d^0 = g^0 = Ax^0 - b = A(x^0 - x)$, so

$$x^0 + K_t(d^0, A) = \big\{x^0 + Ap(A)(x^0 - x) \;:\; p \in P_{t-1}\big\}.$$

Hence, it follows that

$$\begin{aligned}\|Ax^t_{\text{gmres}} - b\|_2 &= \min_{p\in P_{t-1}} \|A[I + Ap(A)](x^0 - x)\|_2 = \min_{p\in P_t, p(0)=1} \|p(A)A(x^0 - x)\|_2 \\ &\leq \min_{p\in P_t, p(0)=1} \|p(A)\|_2 \, \|A(x^0 - x)\|_2 \, .\end{aligned}$$

ii) Due to the fact that A is symmetric and positive definite there exists an orthonormal basis $\{o_i\}$ of eigenvectors of A with corresponding eigenvalues $\{\lambda_i\}$. Let $y \in \mathbb{R}^n$ be an arbitrary vector with $y = \sum_i y_i o_i$ for suitable coefficients y_i. It holds

$$\|p(A)y\| = \|p(A)\sum_i y_i o_i\|_2 = \|\sum_i p(\lambda_i) y_i o_i\|_2 \leq \sup_i |p(\lambda_i)| \, \|\sum_i y_i o_i\|_2 = \sup_i |p(\lambda_i)| \, \|y\|_2.$$

We conclude that $\|p(A)\|_2 \leq \sup_i |p(\lambda_i)|$ and consequently (Let λ be the smallest and Λ be the biggest eigenvalue of A):

$$\begin{aligned}\|Ax^t_{\text{gmres}} - b\|_2 &\leq \min_{p\in P_t, p(0)=1} \max_i |p(\lambda_i)| \, \|A(x^0 - x)\|_2 \\ &\leq \min_{p\in P_t, p(0)=1} \max_{\lambda\leq\mu\leq\Lambda} |p(\mu)| \, \|A(x^0 - x)\|_2.\end{aligned}$$

But this is (up to the different norms) the very same inequality that was derived for the CG method. So, with the same line of reasoning one derives

$$\|Ax^t_{\text{gmres}} - b\|_2 \leq \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^t \|A(x^0 - x)\|_2.$$

iii) Similarly to (ii):

$$\|p(A)y\|_2 = \|T^{-1}Tp(A)T^{-1}Ty\|_2 = \|T^{-1}p(D)Ty\|_2 \leq \|T^{-1}\|_2 \, \|p(D)\|_2 \, \|T\|_2 \, \|y\|_2.$$

Furthermore, $\|p(D)\|_2 = \max_i |\lambda_i|$, so one concludes that

$$\|p(A)\|_2 \leq \kappa_2(T) \max_i |\lambda_i|.$$

The difficulty of this result lies in the fact that the λ_i are generally complex valued, so some a priori asumption has to be made in order to control $\max_i |\lambda_i|$.

Solution A.3.19: a) It holds

$$\begin{aligned}\lambda_{\max} &= 6 + 2 \times 3\cos\big((1-h)\pi\big) \approx 12, \\ \lambda_{\min} &= 6 - 2 \times 3\cos(h\pi),\end{aligned}$$

and hence,

$$\text{cond}_2(A) \approx \frac{4}{\pi^2 h^2}.$$

In analogy to the text, it holds that the eigenvalues of the Jacobi iteration matrix $J = I - D^{-1}A$ are given by

$$\mu_{ijk} = \frac{1}{2}\big(\cos[ih\pi] + \cos[jh\pi] + \cos[kh\pi]\big), \quad i,j,k = 0,\ldots,m.$$

Consequently,

$$\mathrm{spr(J)} = 1 - \frac{\pi^2}{2}\mathrm{h}^2 + \mathcal{O}(\mathrm{h}^4).$$

b) Due to the fact that the matrix A is consistently ordered, it holds

$$\begin{aligned}\mathrm{spr}(\mathrm{H}_1) &= \rho^2 = 1 - \pi^2 h^2 + \mathcal{O}(h^4),\\ \mathrm{spr}(\mathrm{H}_{\omega_{\mathrm{opt}}}) &= \frac{1-\sqrt{1-\rho^2}}{1+\sqrt{1-\rho^2}} = \frac{1-\pi h + \mathcal{O}(h^2)}{1+\pi h + \mathcal{O}(h^2)} = 1 - 2\pi h + \mathcal{O}(h^2).\end{aligned}$$

The number of required iterations $T_*(\varepsilon) \approx \ln(\varepsilon)/\ln(\mathrm{spr}(*))$ is thus

$$\begin{aligned}T_J(\varepsilon) \approx -\frac{2}{\pi^2 h^2}\ln(\varepsilon) \approx 18\,665, \quad T_{H_1}(\varepsilon) \approx -\frac{1}{\pi^2 h^2}\ln(\varepsilon) \approx 9\,333,\\ T_{H_{\omega_{\mathrm{opt}}}}(\varepsilon) \approx -\frac{1}{2\pi h}\ln(\varepsilon) \approx 147,\end{aligned}$$

and for the gradient and CG method:

$$\begin{aligned}T_G(\varepsilon) = -\frac{1}{2}\kappa\ln(\varepsilon) \approx -\frac{2}{\pi^2 h^2}\ln(\varepsilon) \approx 18\,665,\\ T_{CG}(\varepsilon) = -\frac{1}{2}\sqrt{\kappa}\ln(\varepsilon/2) \approx -\frac{1}{\pi h}\ln(\varepsilon/2) \approx 316.\end{aligned}$$

c) A matrix vector multiplication with A needs roughly $7h^{-3}$ a. op.. With that one concludes that the number of required a. op. for Jacobi, Gauß-Seidel and SOR method is approximately $8h^{-3}$. Similarly the workload for the gradient method is $11h^{-3}$ a. op., whereas the CG method needs $12h^{-3}$ a. op.

A.4 Chapter 4

Solution A.4.1: There holds $z^0 = \sum_{i=1}^n \alpha_i w^i$ and $z^t = \|A^t z^0\|_2^{-1} A^t z^0$ and therefore

$$\begin{aligned}\lambda^t = (Az^t, z^t)_2 &= \frac{(A^{t+1}z^0, A^t z^0)_2}{\|A^t z^0\|_2^2} = \frac{\sum_{i=1}^n |\alpha_i|^2 \lambda_i^{2t+1}}{\sum_{i=1}^n |\alpha_i|^2 \lambda_i^{2t}}\\ &= \frac{(\lambda_n)^{2t+1}\left\{|\alpha_n|^2 + \sum_{i=1}^{n-1}|\alpha_i|^2\left(\frac{\lambda_i}{\lambda_n}\right)^{2t+1}\right\}}{(\lambda_n)^{2t}\left\{|\alpha_n|^2 + \sum_{i=1}^{n-1}|\alpha_i|^2\left(\frac{\lambda_i}{\lambda_n}\right)^{2t}\right\}}\\ &= \lambda_n \frac{|\alpha_n|^2 + \sum_{i=1}^{n-1}|\alpha_i|^2\left(\frac{\lambda_i}{\lambda_n}\right)^{2t} + \sum_{i=1}^{n-1}|\alpha_i|^2\left(\frac{\lambda_i}{\lambda_n}\right)^{2t}\left(\frac{\lambda_i}{\lambda_n}-1\right)}{|\alpha_n|^2 + \sum_{i=1}^{n-1}|\alpha_i|^2\left(\frac{\lambda_i}{\lambda_n}\right)^{2t}}\\ &= \lambda_n + \lambda_n \frac{\sum_{i=1}^{n-1}|\alpha_i|^2\left(\frac{\lambda_i}{\lambda_n}\right)^{2t}\left(\frac{\lambda_i}{\lambda_n}-1\right)}{|\alpha_n|^2 + \sum_{i=1}^{n-1}|\alpha_i|^2\left(\frac{\lambda_i}{\lambda_n}\right)^{2t}} =: \lambda_n + \lambda_n E_t.\end{aligned}$$

The error term on the right can be estimated as follows:

$$|E_t| \le \Big(\frac{\lambda_{n-1}}{\lambda_n}\Big)^{2t} \frac{\sum_{i=1}^{n-1} |\alpha_i|^2}{|\alpha_n|^2} = \Big(\frac{\lambda_{n-1}}{\lambda_n}\Big)^{2t} \frac{\|z^0\|_2^2}{|\alpha_n|^2}.$$

Hence,

$$|\lambda^t - \lambda_n| \le |\lambda_n| \frac{\|z^0\|_2^2}{|\alpha_n|^2} \Big(\frac{\lambda_{n-1}}{\lambda_n}\Big)^{2t}.$$

Solution A.4.2: Let $\mu_i := (\lambda_i - \lambda)^{-1}$ be the eigenvalues of the matrix $(A - \lambda I)^{-1}$. Further, we note that $\mu_{\max} = (\lambda_{\min} - \lambda)^{-1}$. The corresponding iterates generated by the inverse iteration are $\mu^t = (\lambda^t - \lambda)^{-1}$ with $\lambda^t := 1/\mu^t + \lambda$. We begin with the identity

$$\begin{aligned} z^t = \frac{\tilde{z}^t}{\|\tilde{z}^t\|_2} &= \frac{(A - \lambda^{t-1} I)^{-1} z^{t-1}}{\|(A - \lambda^{t-1} I)^{-1} z^{t-1}\|_2} = \frac{(A - \lambda^{t-1} I)^{-1} (A - \lambda^{t-2} I)^{-1} z^{t-2}}{\|(A - \lambda^{t-1} I)^{-1} (A - \lambda^{t-2} I)^{-1} z^{t-2}\|_2} \\ &= \quad \cdots = \frac{\prod_{j=0}^{t-1} (A - \lambda^j I)^{-1} z^0}{\| \prod_{j=0}^{t-1} (A - \lambda^j I)^{-1} z^0 \|_2}, \end{aligned}$$

from which we conclude

$$\begin{aligned} \mu^t &= \big((A - \lambda^{t-1} I)^{-1} z^t, z^t\big)_2 \\ &= \frac{\big((A - \lambda^{t-1} I)^{-1} \prod_{j=0}^{t-1} (A - \lambda^j I)^{-1} z^0, \prod_{j=0}^{t-1} (A - \lambda^j I)^{-1} z^0\big)_2}{\| \prod_{j=0}^{t-1} (A - \lambda^j I)^{-1} z^0 \|_2^2} \\ &= \frac{\sum_{i=1}^n |\alpha_i|^2 (\lambda_i - \lambda^{t-1})^{-1} \prod_{j=0}^{t-1} (\lambda_i - \lambda^j)^{-2}}{\sum_{i=1}^n |\alpha_i|^2 \prod_{j=0}^{t-1} (\lambda_i - \lambda^j)^{-2}}. \end{aligned}$$

Next,

$$\begin{aligned} \mu^t &= \frac{|\alpha_1|^2 (\lambda_1 - \lambda^{t-1})^{-1} \prod_{j=0}^{t-1} (\lambda_1 - \lambda^j)^{-2} + \sum_{i=2}^n |\alpha_i|^2 (\lambda_i - \lambda^{t-1})^{-1} \prod_{j=0}^{t-1} (\lambda_i - \lambda^j)^{-2}}{|\alpha_1|^2 \prod_{j=0}^{t-1} (\lambda_1 - \lambda^j)^{-2} + \sum_{i=2}^n |\alpha_i|^2 \prod_{j=0}^{t-1} (\lambda_i - \lambda^j)^{-2}} \\ &= \frac{(\lambda_1 - \lambda^{t-1})^{-1} \prod_{j=0}^{t-1} (\lambda_1 - \lambda^j)^{-2}}{\prod_{j=0}^{t-1} (\lambda_1 - \lambda^j)^{-2}} \, \frac{|\alpha_1|^2 + \sum_{i=2}^n |\alpha_i|^2 \big(\frac{\lambda_1 - \lambda^{t-1}}{\lambda_i - \lambda^{t-1}}\big) \prod_{j=0}^{t-1} \big(\frac{\lambda_1 - \lambda^j}{\lambda_i - \lambda^j}\big)^2}{|\alpha_1|^2 + \sum_{i=2}^n |\alpha_i|^2 \prod_{j=0}^{t-1} \big(\frac{\lambda_1 - \lambda^j}{\lambda_i - \lambda^j}\big)^2} \\ &= \frac{1}{\lambda_1 - \lambda^{t-1}} \, \frac{|\alpha_1|^2 + \sum_{i=2}^n |\alpha_i|^2 \big(\frac{\lambda_1 - \lambda^j}{\lambda_i - \lambda^j}\big)^2 + \sum_{i=2}^n |\alpha_i|^2 \prod_{j=0}^{t-1} \big(\frac{\lambda_1 - \lambda^j}{\lambda_i - \lambda^j}\big)^2 \big(1 - \big(\frac{\lambda_1 - \lambda^{t-1}}{\lambda_i - \lambda^{t-1}}\big)\big)}{|\alpha_1|^2 + \sum_{i=2}^n |\alpha_i|^2 \prod_{j=0}^{t-1} \big(\frac{\lambda_1 - \lambda^j}{\lambda_i - \lambda^j}\big)^2} \\ &= \frac{1}{\lambda_1 - \lambda^{t-1}} + \frac{1}{\lambda_1 - \lambda^{t-1}} \, \frac{\sum_{i=2}^n |\alpha_i|^2 \prod_{j=0}^{t-1} \big(\frac{\lambda_1 - \lambda^j}{\lambda_i - \lambda^j}\big)^2 \big(1 - \frac{\lambda_1 - \lambda^{t-1}}{\lambda_i - \lambda^{t-1}}\big)}{|\alpha_1|^2 + \sum_{i=2}^n |\alpha_i|^2 \prod_{j=0}^{t-1} \big(\frac{\lambda_1 - \lambda^j}{\lambda_i - \lambda^j}\big)^2} \\ &=: \frac{1}{\lambda_1 - \lambda^{t-1}} + \frac{1}{\lambda_1 - \lambda^{t-1}} E_t. \end{aligned}$$

The error term on the right-hand side can be estimates as follows:

$$|E_t| \leq \frac{\sum_{i=2}^n |\alpha_i|^2 \prod_{j=0}^{t-1} \left|\frac{\lambda_1-\lambda^j}{\lambda_i-\lambda^j}\right|^2 \left|1-\frac{\lambda_1-\lambda^{t-1}}{\lambda_i-\lambda^{t-1}}\right|}{|\alpha_1|^2 + \sum_{i=2}^n |\alpha_i|^2 \prod_{j=0}^{t-1} \left(\frac{\lambda_1-\lambda^j}{\lambda_i-\lambda^j}\right)^2}$$
$$\leq \prod_{j=0}^{t-1} \left|\frac{\lambda_1-\lambda^j}{\lambda_2-\lambda^j}\right|^2 \frac{\sum_{i=2}^n |\alpha_i|^2}{|\alpha_1|^2} = \prod_{j=0}^{t-1} \left|\frac{\lambda_1-\lambda^j}{\lambda_2-\lambda^j}\right|^2 \frac{\|z^0\|_2^2}{|\alpha_1|^2}.$$

This yields

$$\left|\mu^t - \frac{1}{\lambda_1-\lambda^{t-1}}\right| \leq \frac{1}{\lambda_1-\lambda^{t-1}} \prod_{j=0}^{t-1} \left|\frac{\lambda_1-\lambda^j}{\lambda_2-\lambda^j}\right|^2 \frac{\|z^0\|_2^2}{|\alpha_1|^2}.$$

Observing $\mu^t = (\lambda^t - \lambda^{t-1})^{-1}$ or $\lambda^t = 1/\mu^t + \lambda^{t-1}$,

$$\left|\frac{1}{\lambda^t-\lambda^{t-1}} - \frac{1}{\lambda_1-\lambda^{t-1}}\right| = \left|\frac{\lambda_1-\lambda^{t-1}-\lambda^t+\lambda^{t-1}}{(\lambda^t-\lambda^{t-1})(\lambda_1-\lambda^{t-1})}\right| = \left|\frac{\lambda_1-\lambda^t}{(\lambda^t-\lambda^{t-1})(\lambda_1-\lambda^{t-1})}\right|,$$

we obtain the desired estimate

$$|\lambda_1-\lambda^t| \leq |\lambda^t-\lambda^{t-1}| \prod_{j=0}^{t-1} \left|\frac{\lambda_1-\lambda^j}{\lambda_2-\lambda^j}\right|^2 \frac{\|z^0\|_2^2}{|\alpha_1|^2}.$$

Solution A.4.3: It suffices to prove the following two statements about the QR-iteration. The assertion then follows by induction.

1. Let A be a Hessenberg matrix and $A = QR$ its QR-decomposition. Then, $\tilde{A} = RQ$ is also a Hessenberg matrix.

2. Let A be a symmetric matrix and $A = QR$ its QR-decomposition. Then, $\tilde{A} = RQ$ is also a symmetric matrix.

The QR decomposition of a Hessenberg matrix A can be expressed as

$$G_{n-1}\, G_{n-2} \cdots G_1 A = R$$

with

$$G_i = \begin{pmatrix} I_{i-1} & & 0 \\ & \tilde{G}_i & \\ 0 & & I_{n-i-1} \end{pmatrix},$$

and an orthogonal component $\tilde{G}_i \in \mathbb{R}^{2\times 2}$ that eliminates the lower left off diagonal entry of the block

$$\tilde{G}_i \begin{pmatrix} *_{i,i} & *_{i,i+1} \\ a_{i+1,i} & a_{i+1,i+1} \end{pmatrix} = \begin{pmatrix} * & * \\ 0 & * \end{pmatrix}.$$

Apart from eliminating the entry $a_{i+1,i}$, the orthogonal matrix G_i only acts on the upper right part of the (intermediate) matrix. Consequently, R is an upper triangular matrix and it holds

$$\tilde{A} = RQ = RG_1^T G_2^T \cdots G_{n-1}^T.$$

Similarly, it follows by induction that multiplication with G_i^T from the right only introduces at most one (lower-left) off-diagonal element at position $*_{i+1,i}$, so $\tilde{A}$ is indeed a Hessenberg matrix.

Now, let A be symmetric. It holds $QR = A = A^T = R^T Q^T$ and consequently $R = Q^T R^T Q^T$. We conclude that

$$\tilde{A} = RQ = Q^T R^T Q^T Q = (RQ)^T = \tilde{A}^T.$$

Solution A.4.4: Let $A = \tilde{Q}\tilde{R}$ be an arbitrary QR-decomposition of A. Define a unitary matrix $H = \mathrm{diag}(h_i) \in \mathbb{C}^{n\times n}$ by $h_i = \frac{\bar{r}_{ii}}{|r_{ii}|}$ and set $R = H\tilde{R}$, $Q = \tilde{Q}\bar{H}$.

Now, observe that $\bar{A}^T A = \bar{R}^T \bar{Q}^T QR = \bar{R}^T R$ is the Cholesky decomposition of the real valued, symmetric and positive definite matrix $\bar{A}^T A$. Since the Cholesky decomposition (with positive diagonal) is uniquely determined it follows that R is unique and hence also $Q = AR^{-1}$.

Solution A.4.5: i) From the definition of K_m it follows

$$K_{m+1} = \mathrm{span}\{q, A\,K_m\}.$$

Now, if $K_m = K_{m+1} = \mathrm{span}\{q, A\,K_m\}$ one sees by induction that repeated applications of this procedure yield the same space again, hence $K_n = K_m \;\forall\, n \geq m$. On the other hand, given the fact that $K_{m-1} \neq K_m$ it must hold $K_i \neq K_{i+1}$ for $i = 1, \ldots, m-1$. Otherwise, this would already imply $K_{m-1} = K_m$ which is a contradiction.

It holds $\dim K_1 = 1$ because $q \neq 0$. Furthermore, K_m is generated by m vectors. Therefore, one sees by induction that $\dim K_i = i$ as long as $K_i \neq K_{i-1}$, i. e. for $2 \leq i \leq m$.

ii) Let $\lambda \in \sigma(Q^{mT} A Q^m)$ be arbitrary. Then, there exists an eigenvector $v \in \mathbb{C}^m \setminus \{0\}$ with $Q^{mT} A Q^m v = \lambda v$. Multiplication of Q^m from the left and utilizing

$$Q^m Q^{mT}\,. = \sum_{i=1}^{m} q^i\,(q^i, .) = \mathrm{proj}_{K_m}\,.$$

yields

$$\mathrm{proj}_{K_m} A\,Q^m v = \lambda\, Q^m v.$$

But by definition of m it holds that K_m is A-invariant, i. e. $AK_m \subset K_m$, hence $\mathrm{proj}_{K_m}(A\,Q^m v) = A\,Q^m v$ and therefore $\lambda \in \sigma(A)$.

In case of $m = n$ there is $K_m = \mathbb{C}^n$. Consequently, $Q^m \in \mathbb{C}^{n\times n}$ is a regular matrix and

the matrices A and $Q^{m\,T}AQ^m$ are similar; specifically

$$\sigma(Q^{m\,T}AQ^m) = \sigma(A).$$

Solution A.4.6: Let $\{v^1, \ldots, v^m\} \in \mathbb{R}^n$ be a linearly independent set of vectors. The classical Gram-Schmidt orthogonalization procedure reads: For $i = 1, \ldots, m$:

$$\begin{aligned} &\alpha) \quad \tilde{u}^i := v^i - \sum_{j=1}^{i-1} (u^j, v^i) u^j, \\ &\beta) \quad u^i := \tilde{u}^i / \|\tilde{u}^i\|. \end{aligned}$$

The modified Gram-Schmidt orthogonalization procedure takes the form: For $i = 1, \ldots, m$:

$$\begin{aligned} &\alpha) \quad \tilde{u}^{i,1} := v^i, \\ &\qquad \tilde{u}^{i,k} := \tilde{u}^{i,k-1} - (u^{k-1}, \tilde{u}^{i,k-1}) u^{k-1}, \text{ for } k = 2, \ldots, i, \\ &\beta) \quad u^i := \tilde{u}^{i,i} / \|\tilde{u}^{i,i}\|. \end{aligned}$$

i) For the modified Gram-Schmidt algorithm we can assume by induction that

$$\tilde{u}^{i,k-1} = v^i - \mathrm{proj}_{\langle u^1, \ldots, u^{k-2} \rangle}(v^i),$$

hence

$$\begin{aligned} \tilde{u}^{i,k} &= \tilde{u}^{i,k-1} - (u^{k-1}, \tilde{u}^{i,k-1}) u^{k-1} \\ &= v^i - \mathrm{proj}_{\langle u^1, \ldots, u^{k-2} \rangle}(v^i) - \mathrm{proj}_{u^{k-1}}(\tilde{u}^{i,k-1}) \\ &= v^i - \mathrm{proj}_{\langle u^1, \ldots, u^{k-2} \rangle}(v^i) - \mathrm{proj}_{u^{k-1}}(v^i) \\ &= v^i - \sum_{j=1}^{k-1} (u^j, v^i) u^j. \end{aligned}$$

ii) By rewriting step (α) of the classical algorithm in the form

$$\begin{aligned} &\alpha) \quad \tilde{u}^{i,1} := v^i, \\ &\qquad \tilde{u}^{i,k} := \tilde{u}^{i,k-1} - (u^{k-1}, v^i) u^{k-1}, \text{ for } k = 2, \ldots, i, \end{aligned}$$

one observes that the algorithmic complexity of both variants are exactly the same. Both consist of $i - 1$ scalar-products (with n a. op.) with vector scaling and vector addition (with n a. op.) in step (α) which sums up to

$$\sum_{i=1}^{m} (i-1)\,(n+n) = n\,m(m-1) \text{ a. op.}$$

as well as m normalization steps with roughly $2n$ a. op. in (β). In total $n\,m(m+1)$ a. op..

Solution A.4.7: The result by the *classical* Gram-Schmidt algorithm is:

$$\tilde{Q} = \begin{bmatrix} 1 & 0 & 0 \\ \varepsilon & 0 & \frac{\sqrt{2}}{2} \\ \varepsilon & -1 & \frac{\sqrt{2}}{2} \end{bmatrix},$$

with $\|\tilde{Q}^T\tilde{Q} - I\|_\infty = \sqrt{2}\,(\frac{1}{2} + \varepsilon)$. The result by the *modified* Gram-Schmidt algorithm is:

$$\tilde{Q} = \begin{bmatrix} 1 & 0 & 0 \\ \varepsilon & 0 & -1 \\ \varepsilon & -1 & 0 \end{bmatrix},$$

with $\|\tilde{Q}^T\tilde{Q} - I\|_\infty \approx 2\,\varepsilon$.

Solution A.4.8: i) With the help of the Taylor expansion of the cosine:

$$\begin{aligned} &|\lambda_{ijkk} - \lambda_{ijk}^h| = \\ &\left|(i^2+j^2+k^2)\pi^2 - h^{-2}\left\{6 - 6 - \frac{(i^2+j^2+k^2)\pi^2 h^2}{2!} - \frac{(i^4+j^4+k^4)\pi^4 h^4}{4!} - \mathcal{O}(h^6)\right\}\right| \\ &= \frac{(i^4+j^4+k^4)\pi^4 h^2}{4!} + \mathcal{O}(h^4) \le \frac{1}{4!}\lambda_{ijk}^2 h^2 + \mathcal{O}(h^4) \le \frac{1}{12}\lambda_{ijk}^2 h^2 \text{ (for } h \text{ sufficiently small).} \end{aligned}$$

ii) The maximal eigenvalue $\lambda_{\max}$ that can be reliably computed with a relative tolerance TOL fulfills the relation

$$\frac{1}{12}\lambda_{\max} h^2 \approx \text{TOL} \implies \lambda_{\max} \approx \frac{12\,\text{TOL}}{h^2}.$$

The number of reliably approximateable eigenvalues (not counting multiplicities) is the cardinality of the set

$$\big\{i^2+j^2+k^2 :\ (i^2+j^2+k^2) \le \frac{\lambda_{\max}}{\pi^2}, i,j,k \in \mathbb{N},\ 1 \le i,j,k \le m\big\}.$$

For the concrete choice of numbers this leads to:

$$\#\big\{i^2+j^2+k^2 :\ (i^2+j^2+k^2) \le 19,\ i,j,k \in \mathbb{N}\big\},$$

whose cardinality can be counted by hand:

$$\#\big\{(1,1,1),\ (2,1,1),\ (2,2,1),\ (2,2,2),\ (3,1,1),\ (3,2,1),\ (3,2,2),\ (3,3,1),\ (4,1,1)\big\} = 9.$$

iii) The number of reliably approximateable eigenvalues (counting multiplicities) is the

cardinality of the set

$$\Big\{(i,j,k)\in\mathbb{N}^3:\ (i^2+j^2+k^2)\le\frac{\lambda_{\max}}{\pi^2},\ 1\le i,j,k\le m\Big\}.$$

For large numbers a reasonably large subset is given by

$$\Big\{(i,j,k)\in\mathbb{N}^3:\ 1\le i,j,k\le\frac{\sqrt{\lambda_{\max}}}{\sqrt{3}\,\pi}\Big\},$$

which has the cardinality

$$\Big\lfloor\frac{\sqrt{\lambda_{\max}}}{\sqrt{3}\,\pi}\Big\rfloor^3=\Big\lfloor\frac{4\sqrt{\mathrm{TOL}}}{\pi\,h}\Big\rfloor^3.$$

Therefore, h must be chosen such that

$$\Big\lfloor\frac{4\sqrt{\mathrm{TOL}}}{\pi\,h}\Big\rfloor^3\ge 1.000\quad\Longleftrightarrow\quad h\le\frac{6\sqrt{10^{-3}}}{10\,\pi}\approx 6.0\times10^{-3}.$$

Approximately 7-times uniform refinement in 3D, i. e., $n_h\approx h^{-3}\approx 4.6\times10^6$.

Solution A.4.9: i) The inverse iteration for determining the smallest eigenvalue (with shift $\lambda=0$) reads

$$A\tilde z^t=z^{t-1},\quad z^t=\|\tilde z^t\|^{-1}\tilde z^t,\quad t=1,2,\dots$$

with intermediate guesses $\mu^t=(A^{-1}z^t,z^t)$ for the smallest eigenvalue. One iteration of the inverse iteration consists of 1 solving step consisting of cn a. op and a normalization step of roughly $2n$ a. op. Determining the final guess for the eigenvalue needs another solving step and a scalar product, in total $(c+1)\,n$ a. op. So, for 100 iteration steps we end up with

$$(101\,c+201)n\text{ a. op.}$$

The Lanczos algorithm reads: Given initial $q^0=0$, $q^1=\|q\|^{-1}q$, $\beta_1=0$ compute for $1\le t\le m-1$:

$$\begin{aligned}r^t&=A^{-1}q^t,\quad \alpha_t=(r^t,q^t),\quad s^t=r^t-\alpha_tq^t-\beta_tq^{t-1}\\ \beta^{t+1}&=\|s^t\|,\quad q^{t+1}=s^t/\beta^{t+1},\end{aligned}$$

and a final step $r^m=A^{-1}q^m$, $\alpha_m=(r^m,q^m)$. This procedure takes cn a. op. for the matrix vector product with additional $5n$ a. op. per round. In total (respecting initial and final computations):

$$(101\,c+501)\,n\text{ a. op.}$$

The Lanczos algorithm will construct a tridiagonal matrix T^m (with $m=100$ in our case) of which we still have to compute the eigenvalues with the help of the QR method:

$$\begin{aligned}&B^{(0)}=T^m,\\ &B^{(i)}=Q^{(i)}R^{(i)},\quad B^{(i+1)}=R^{(i)}Q^{(i)}.\end{aligned}$$

From a previous exercise we already know that the intermediate $B^{(i)}$ will retain the tridiagonal matrix property, so that a total workload of $\mathcal{O}(m)$ a. op. per round of the QR method can be assumed. For simplicity, we assume that the number of required QR iterations (to achieve good accuracy) also scales with $\mathcal{O}(m)$. Then, the total workload of QR method is $\mathcal{O}(m^2)$ a. op.

ii) Assume that it is possible to start the inverse iteration with a suitable guess for each of the 10 desired eigenvalues. Still, it is necessary to do the full 100 iterations for each eigenvalue independently, resulting in

$$10\,(101\,c + 201)\,n \text{ a. op.}$$

The Lanczos algorithm, in contrast, already approximates the first 10 eigenvalues simultaneoulsy for the choice $m = 100$ (see results of the preceding exercise). Hence, we end up with the same number of a. op.:

$$(101\,c + 501)\,n \text{ a. op.}$$

(except for some possibly higher workload in the QR iteration). Given the fact that c is usually o moderate size somewhere around 5, the Lanczos algorithm clearly wins.

Solution A.4.10: i) Let $A \in \mathbb{C}^{n\times n}$, $x, b \in \mathbb{C}^n$. It is equivalent:

$$\begin{aligned}
& Ax = b \\
\iff\quad & (\operatorname{Re} A + i \operatorname{Im} A)(\operatorname{Re} x + i \operatorname{Im} x) = \operatorname{Re} b + i \operatorname{Im} b \\
\iff\quad & \begin{cases} \operatorname{Re} A \operatorname{Re} x - \operatorname{Im} A \operatorname{Im} x = \operatorname{Re} b \\ -\operatorname{Re} A \operatorname{Im} x - \operatorname{Im} A \operatorname{Re} x = -\operatorname{Im} b \end{cases} \\
\iff\quad & \begin{pmatrix} \operatorname{Re} A & \operatorname{Im} A \\ -\operatorname{Im} A & \operatorname{Re} A \end{pmatrix} \begin{pmatrix} \operatorname{Re} x \\ -\operatorname{Im} x \end{pmatrix} = \begin{pmatrix} \operatorname{Re} b \\ -\operatorname{Im} b \end{pmatrix}.
\end{aligned}$$

ii) For all three properties it holds that they are fullfilled by the block-matrix $\tilde{A}$ if and only if the correspondig complex valued matrix A has the analogous property (in the complex sense):

a) From the above identity we deduce that the complex valued linear system of equations (in the first line) is uniquely solvable for arbitrary $b \in \mathbb{C}^n$ if and only if the same holds true for the real valued linear equation (in the last line) for arbitrary $(\operatorname{Re} b, \operatorname{Im} b) \in \mathbb{R}^{2n}$. Thus $\tilde{A}$ is regular iff A is regular.

b) Observe that

$$\begin{aligned}
& \tilde{A} \text{ symmetric} \\
\iff\quad & \operatorname{Im} A = -\operatorname{Im} A^T \text{ and } \operatorname{Re} A = \operatorname{Re} A^T \\
\iff\quad & \operatorname{Re} A + \operatorname{Im} A = \operatorname{Re} A - \operatorname{Im} A^T \\
\iff\quad & A = \bar{A}^T.
\end{aligned}$$

c) For arbitrary $x \in \mathbb{C}^n$ it holds

$$\begin{aligned} & \operatorname{Re}\bigl(\bar{x}^T A\, x\bigr) > 0 \\ \iff \quad & \operatorname{Re} x^T \operatorname{Re} A \operatorname{Re} x + \operatorname{Im} x^T \operatorname{Re} A \operatorname{Im} x - \operatorname{Im} x^T \operatorname{Im} A \operatorname{Re} x - \operatorname{Re} x^T \operatorname{Im} A \operatorname{Im} x > 0 \\ \iff \quad & \begin{pmatrix} \operatorname{Re} x \\ -\operatorname{Im} x \end{pmatrix}^T \begin{pmatrix} \operatorname{Re} A & \operatorname{Im} A \\ -\operatorname{Im} A & \operatorname{Re} A \end{pmatrix} \begin{pmatrix} \operatorname{Re} x \\ -\operatorname{Im} x \end{pmatrix} > 0. \end{aligned}$$

Solution A.4.11: The statement follows immediately from the equivalent definition

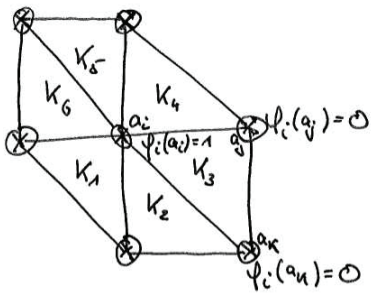
$$\begin{aligned} \sigma_\varepsilon(T) &= \bigl\{ z \in \mathbb{C} \ : \ \sigma_{\min}(zI - T) \le \varepsilon \bigr\}, \text{ with} \\ \sigma_{\min}(T) &:= \min \bigl\{ \lambda^{1/2} \ : \ \lambda \in \sigma(\bar{T}^T T) \bigr\} \end{aligned}$$

and by the observation that similar matrices yield the same set of eigenvalues:

$$\begin{aligned} \sigma_{\min}(T) &= \min \bigl\{ \lambda^{1/2} \ : \ \lambda \in \sigma(\bar{T}^T T) \bigr\} \\ &= \min \bigl\{ \lambda^{1/2} \ : \ \lambda \in \sigma(\bar{Q}^T \bar{T}^T Q \bar{Q}^T T Q) \bigr\} \\ &= \min \Bigl\{ \lambda^{1/2} \ : \ \lambda \in \sigma\Bigl(\overline{(\bar{Q}^T T Q)}^T (\bar{Q}^T T Q) \Bigr) \Bigr\} \\ &= \sigma_{\min}(Q^{-1} T Q). \end{aligned}$$

A.5 Chapter 5

Solution A.5.1: Let a_i be an arbitrary nodal point and φ_h^i be the corresponding nodal basis function. Its support consists of 6 triangles $T_1, \cdots, T_6$:



Outside of $\cup_{i=1}^6 \overline{T_i}$ the function φ_h^i is zero. Due to the fact that φ_h^i is continuous and cellwise linear, its gradient is cellwise defined and constant with values

$$\begin{aligned} \nabla\varphi_h^i\Big|_{K_1} &= \frac{1}{h}\begin{pmatrix} 1 \\ 1 \end{pmatrix}, & \nabla\varphi_h^i\Big|_{K_2} &= \frac{1}{h}\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \\ \nabla\varphi_h^i\Big|_{K_3} &= \frac{1}{h}\begin{pmatrix} -1 \\ 0 \end{pmatrix}, & \nabla\varphi_h^i\Big|_{K_4} &= \frac{1}{h}\begin{pmatrix} -1 \\ -1 \end{pmatrix}, \\ \nabla\varphi_h^i\Big|_{K_5} &= \frac{1}{h}\begin{pmatrix} 0 \\ -1 \end{pmatrix}, & \nabla\varphi_h^i\Big|_{K_6} &= \frac{1}{h}\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \end{aligned}$$

where h denotes the length of the catheti of the triangles. With these preliminaries it follows immediately that

$$b_i = \sum_{\mu=1}^{6} \frac{|K_\mu|}{3} \sum_{j=1}^{3} f(a_j)\varphi_h^i(a_j) = 6\,\frac{1}{6}\,h^2 f(a_i) = h^2 f(a_i).$$

For the stiffness matrix $a_{ij} = (\nabla\varphi_h^i, \nabla\varphi_h^j)$, we have to consider three distinct cases: a) where $a_i = a_j$, b) where a_i and a_j are endpoints of a cathetus, and c) where they are endpoints of a hypotenuse:

a) $$a_{ii} = \sum_{\mu=1}^{6} \frac{|K_\mu|}{3} \sum_{\nu=1}^{3} \left(\nabla\varphi_h^i(a_\nu), \nabla\varphi_h^i(a_\nu)\right) = \frac{1}{6}\,h^2\,3\,(2+1+1+2+1+1)\,h^{-2} = 4.$$

b) $$a_{ij} = \sum_{\mu=1}^{6} \frac{|K_\mu|}{3} \sum_{\nu=1}^{3} \left(\nabla\varphi_h^i(a_\nu), \nabla\varphi_h^j(a_\nu)\right) = \frac{1}{6}\,h^2\,3\,(-1-1)\,h^{-2} = -1.$$

c) $$a_{ij} = \sum_{\mu=1}^{6} \frac{|K_\mu|}{3} \sum_{\nu=1}^{3} \left(\nabla\varphi_h^i(a_\nu), \nabla\varphi_h^j(a_\nu)\right) = \frac{1}{6}\,h^2\,3\,0 = 0.$$

In summary, the *stencil* has the form

$$\begin{pmatrix} 0 & -1 & \\ -1 & 4 & -1 \\ & -1 & 0 \end{pmatrix}.$$

This is, up to a factor of h^{-2} exactly the stencil of the finite different discretization described in the text.

Solution A.5.2: The principal idea for the convergence proof of the two-grid algorithm was to prove a contraction property for

$$e_L^{(t+1)} = \mathrm{ZG}_L(\nu)e_L^{(t)}, \quad \mathrm{ZG}_L(\nu) = \left(A_L^{-1} - p_{L-1}^L A_{L-1}^{-1} r_L^{L-1}\right) A_L S_L^\nu$$

This was done with the help of a so called *smoothing property*,

$$\|A_L S_L^\nu\| \le c_s \nu^{-1} h_L^{-2},$$

and an *approximation property*,

$$\left\|A_L^{-1} - p_{L-1}^L A_{L-1}^{-1} r_L^{L-1}\right\| \le c_a h_L^2.$$

The first property is completely independent of the choice of restriction that is used. The second, however, poses major difficulties for our choice of restriction: In analogy to the proof given in the text let $\psi_L \in V_L$ be arbitrary. Now, $v_L := A_L^{-1}\psi_L$ is the solution of the variational problem

$$a(v_L, \varphi_L) = (\psi_L, \varphi_L) \quad \forall \varphi_L \in V_L,$$

and similarly $v_{L-1} := p_{L-1}^L A_{L-1}^{-1} r_L^{L-1} \psi_L$ is the solution of

$$a\left(v_{L-1}, \varphi_{L-1}\right) = (r_L^{L-1}\psi_L, \varphi_{L-1}) \quad \forall \varphi_{L-1} \in V_{L-1}.$$

Let v and $\tilde{v}$ be the solutions of the corresponding continuous problems:

$$\begin{aligned} a(v, \varphi) &= (\psi_L, \varphi) \quad \forall \varphi \in V, \\ a\left(\tilde{v}, \varphi\right) &= (r_L^{L-1}\psi_L, \varphi) \quad \forall \varphi \in V. \end{aligned}$$

We can employ the usual a priori error estimate (for the Ritz-projection):

$$\begin{aligned} \|v_L - v_{L-1}\| &\leq \|v_L - v\| + \|v_{L-1} - \tilde{v}\| + \|v - \tilde{v}\| \\ &\leq ch^2\big(\|\psi_L\| + \|r_L^{L-1}\psi_L\|\big) + \|v - \tilde{v}\|. \end{aligned}$$

Furthermore, exploiting the finit dimensionality of the spaces involved it is possible to bound $\|r_L^{L-1}\psi_L\|$ in terms of $\|\psi_L\|$, i. e.,

$$\|r_L^{L-1}\psi_L\| \leq c\|\psi_L\|.$$

But, now, r_L^{L-1} is not the L^2-projection. So we have to assume that in general

$$(r_L^{L-1}\psi_L, \varphi_{L-1}) \neq (\psi_L, \varphi_{L-1}),$$

and hence $v \neq \tilde{v}$. This is a problem because a necessary bound of the form

$$\|v - \tilde{v}\| \leq ch_L^2 \|\psi_L\|.$$

does not hold in general.

Solution A.5.3: This time, the problem when trying to convert the proof to the given problem arises in the smoothing property. The proof of the approximation property does not need symmetry. We still have an inverse property of the form $\|A_L\| \leq ch^{-2}$. So, it remains to show that

$$\|S_L\| \leq c < 1$$

for $S_L = I_L - \theta A_L$ with a constant c that is independent of L. Because A_L is not symmetric, it is not possible to copy the arguments (that utilize spectral theory) from the text. We proceed differently: First of all observe that for all $u_L \in V_L$ it holds that

$$\begin{aligned} (A_L u_L, u_L) = a(u_L, u_L) &= \|\nabla u\|^2 + (\partial_1 u, u) \\ &= \|\nabla u\|^2 + \frac{1}{2}\int_\Omega \partial_1(u^2)\, dx \\ &= \|\nabla u\|^2 + \frac{1}{2}\int_{\partial\Omega} n_1 u^2\, ds \\ &= \|\nabla u\|^2. \end{aligned}$$

Hence, A_L is positive definite – or, equivalently, for all (complex valued) eigenvalues λ_i, $i = 1, ..., N_L$ of A_L it holds:

$$\mathrm{Re}\lambda_i > 0 \quad i = 1, ..., N_L.$$

The eigenvalues of $S_L = I_L - \theta A_L$ are $1 - \theta\lambda_i$, $i = 1, ..., N_L$. Furthermore,

$$\begin{aligned} |1 - \theta\lambda| = |1 - \theta\mathrm{Re}\lambda - \theta\mathrm{Im}\lambda| &= \left\{(1 - \theta\mathrm{Re}\lambda)^2 + \theta^2(\mathrm{Im}\lambda)^2\right\}^{1/2} \\ &= \left\{1 - 2\theta\mathrm{Re}\lambda + \theta^2(\mathrm{Im}\lambda)^2 + (\mathrm{Re}\lambda)^2)\right\}^{\frac{1}{2}} \end{aligned}$$

So finally, the choice

$$\theta < \min_{i=1,...,N_L} \frac{2\mathrm{Re}\lambda_i}{|\lambda_i|^2},$$

leads to

$$\mathrm{spr(S_L)} = \max_{\mathrm{i=1...N_L}} |1 - \theta\lambda_\mathrm{i}| < \mathrm{c} < 1.$$

with a constant c independent of L. The smoothing property now follows with the general observation that for every $\varepsilon > 0$ there exists an (operator, or induced matrix) norm $\|\cdot\|_*$ with

$$\|S_L\|_* \leq c + \varepsilon.$$

The question remains whether this extends to an L independent convergence rate in the norm $\|\cdot\|$.

Solution A.5.4: Applying one step of the Richardson iteration $\bar{x}^{n+1} = \bar{x}^n + \theta(b - A_L\bar{x}^n)$ needs essentially one matrix vector multiplication with a complexity of $9N_L$ a. op. (due to the fact that at most 9 matrix entries per row are non-zero). Together with the necessary addition processes S_L^ν needs $11\,\nu N_L$ a. op.
Calculating the defect $d_l = f_l - A_l x^l$ needs another $10N_L$ a. op. For the L^2 projektion onto the coarser grid, we need to calculate

$$\tilde{d}^{l-1} := r_l^{l-1} d_l.$$

This can be done very efficiently: Let $\{\varphi_i^l\}$ be the nodal basis on level l. The i-th component of the L^2 projection of $\tilde{d}^{l-1}$ is given by

$$\tilde{d}_i^{l-1} = (r_l^{l-1} d^l, \varphi_i^{l-1}) = (d^l, \varphi_i^{l-1}).$$

Due to the fact that $V_{l-1} \subset V_l$, it is possible to express φ_i^{l-1} as

$$\varphi_i^{l-1} = \sum_{j=1}^{N_l} \mu_{ij} \varphi_j^l,$$

where at most 9 values μ_{ij} are non trivial. This reduces the computation of the L^2 projection to

$$\tilde{d}_i^{l-1} = \sum_{j=1}^{N_l} \mu_{ij}(d^l, \varphi_i^l) = \sum_{j=1}^{N_l} \mu_{ij} d_i^l$$

and needs $9N_l$ a. op.. Contrary to this, the prolongation is relatively cheap with roughly $2N_l$ a. op. (interpolating intermediate values, neglecting the one in the middle and the boundary, ...). Additionally, we account another N_l a. op. for adding the correction. In total:

$$(2 \cdot 11 + 10 + 9 + 2 + 1)N_l = 44N_l \text{ a. op. on level } l.$$

The dimension of the subspaces behaves roughly like

$$N_{l-k} \approx 2^{-2k} N_l.$$

Within a V-cycle all operations have to be done exactly once on every level, hence (neglecting the cost for solving on the coarsest level) we end up with:

$$\sum_{k=0}^{l} 44N_{l-k} = \sum_{k=0}^{l} \frac{44}{2^{2k}} N_l = \frac{4}{3} 44N_l \left(1 - 2^{-(2k+2)}\right) \leq \frac{4}{3} 44N_l \text{ a. op..}$$

Within a W-cycle, we have to do 2^k steps on level $l-k$. This leads to

$$\sum_{k=0}^{l} 2^k 44N_{l-k} = \sum_{k=0}^{l} \frac{44}{2^k} N_l = 2 \cdot 44N_l \left(1 - 2^{-k-1}\right) \leq 2 \cdot 44N_l \text{ a. op.}$$

A.5.1 Solutions for the general exercises

Solution A.5.5: a) If there exists a regular $T \in \mathbb{R}^{n \times n}$ and a diagonal matrix $D \in \mathbb{R}^{n \times n}$ such that

$$T^{-1}AT = D.$$

b) A matrix $A = (a_{ij})$ is diagonally dominant if there holds

$$\sum_{j=1, j \neq i}^{n} |a_{ij}| \leq |a_{ii}|, \quad i = 1, \ldots, n.$$

c) A matrix $A \in \mathbb{C}^{n \times n}$ is called normal if $\bar{A}^T A = A\bar{A}^T$. Yes, if A is Hermitian, it is automatically normal.

d) The Rayleigh quotient is defined as $(Av, v)_2 / \|v\|_2^2$ for a given vector $v \neq 0$. It can be used to calculate an eigenvalue approximation from a given eigenvector approximation.

e) $\text{cond}_2 = \|A\|_2 \|A^{-1}\|_2 = |\sigma_{\max}| / |\sigma_{\min}|$, where $\|\cdot\|_2$ is the matrix norm induced by $\|\cdot\|_2 : \mathbb{C}^n \to \mathbb{R}_0^+$, and $\sigma_{\min}$ and $\sigma_{\max}$ are the smallest and largest singular value of A.

g) A Gerschgorin circle is a closed disc, denoted by $\bar{K}_\rho(a_{ii})$, and associated with a row (or column) of a matrix by the diagonal value a_{ii} and the absolute sum of the off-diagonal elements $\rho = \sum_{j \neq i} |a_{ij}|$ (or $\rho = \sum_{j \neq i} |a_{ji}|$, respectively). The union of all Gerschgorin

circles of a matrix has the property that it contains all eigenvalues of the matrix.

h) The restriction $r_l^{l-1} : V_l \to V_{l-1}$ is used to transfer an intermediate value $v_l \in V_l$ to the next coarser level V_{l-1}, typically a given finite element function to the next coarser mesh. The prolongation operator $p_{l-1}^l : V_{l-1} \to V_l$ does the exact opposite. It transfers an intermediate result from V_{l-1} to the next finer level.

i) Given an arbitrary $b \in \mathbb{C}^n$ it is defined as

$$K_m(b; A) = \text{span}\big\{b, Ab, \ldots, A^{m-1}b\big\}.$$

j) It refers to the damping parameter $\theta \in (0, 1]$ in the Richardson iteration:

$$x^{(k+1)} = x^{(k)} + \omega\big(b - Ax^{(k)}\big).$$

k) The difference lies in the evaluation of the term

$$\tilde{u}^k = v^k - \sum_{i=1}^{k-1} (v^k, u^i)\, u^i.$$

In the classical Gram-Schmidt method this is done in a straight forward manner, in the modified version a slightly different algorithm is used:

$$\tilde{u}^{k,1} := v^k, \quad \tilde{u}^{k,i} := \tilde{u}^{k,i-1} - (u^{i-1}, \tilde{u}^{k,i-1})u^{i-1}, \text{ for } i = 2, \ldots, k,$$

with $u^k = \tilde{u}^{k,k}/|\tilde{u}^{k,k}|$. Both algorithms are equivalent in exact arithmetic, but the latter is much more stable in floating point arithmetic.

Solution A.5.6: i) The matrix A_1 fulfils the weak row-sum criterion. Therefore the Jacobi and Gauß-Seidel methods converge. Furthermore, A_1 is symmetric and positive definite (because it is regular and diagonally dominant), hence the CG method is applicable.

ii) For A_2 the Jacobi matrix reads

$$J = \begin{pmatrix} 0 & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

with eigenvalues $fulfils \lambda_1 = -1$, $fulfils \lambda_{2,3} = \pm\sqrt{2}/2$. Hence, no convergence in general. The Gauß-Seidel matrix is

$$H_1 = \frac{1}{8} \begin{pmatrix} 0 & 4 & -4 \\ 0 & 2 & 2 \\ 0 & -1 & 3 \end{pmatrix}$$

with eigenvalues $fulfils \lambda_1 = 0$, $fulfils \lambda_{2,3} = -\frac{5}{16} \pm \frac{i\sqrt{7}}{16}$. Hence, the Gauß-Seidel iteration does converge. A_2 is symmetric and positive definite (because it is regular and diagonally dominant).

iii) The matrix A_3 is not symmetric, so the CG method is not directly applicable. For the Jacobi method:

$$J = \begin{pmatrix} 0 & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix},$$

with corresponding eigenvalues $fulfils \lambda_1 = 0$, $fulfils \lambda_{2,3} = \pm\frac{1}{2}$. Hence, the method does converge. Similarly for the Gauß-Seidel method:

$$H_1 = \frac{1}{8} \begin{pmatrix} 0 & 4 & -4 \\ 0 & 2 & 2 \\ 0 & 3 & -1 \end{pmatrix},$$

with eigenvalues $fulfils \lambda_1 = 0$, $fulfils \lambda_{2,3} = -\frac{1}{16} \pm \frac{\sqrt{33}}{16}$. The method does converge.

Solution A.5.7: Given a diagonal matrix $D = \mathrm{diag}(d, 1, 1)$ it holds

$$D^{-1}AD = \begin{pmatrix} 1 & 10^{-3}d^{-1} & 10^{-4}d^{-1} \\ 10^{-3}d & 2 & 10^{-3} \\ 10^{-4}d & 10^{-3} & 3 \end{pmatrix}.$$

Now, we choose $d \in \mathbb{R}$ in such a way that the Gerschgorin circle defined by the first column has minimal radius but is still disjunct from the other two Gerschgorin circles. Therefore, a suitable choice of d must fulfil (the first two Gerschgorin circles must not touch):

$$1 + 1.1 \times 10^{-3}\, d < 2 - 10^{-3} - 10^{-3}\, d^{-1}.$$

Solving this quadratic inequality leads to a necessary condition $d > 0.001001$ (and $\ldots$), hence $d = 0.0011$ is a suitable choice. This improves the radius of the first Gerschgorin circle to

$$\rho_1 = (1.1 \times 10^{-3})^2 = 1.21 \times 10^{-6} : \quad \overline{K}_{1.21\times 10^{-6}}(1).$$

Similarly, for the third Gerschgorin circle and with the choice $D = \mathrm{diag}(1, 1, d)$:

$$3 - 1.1 \times 10^{-3}\, d > 2 + 10^{-3} + 10^{-3}\, d^{-1}.$$

This is the same inequality as already discussed. Therefore:

$$\rho_3 = (1.1 \times 10^{-3})^2 = 1.21 \times 10^{-6} : \quad \overline{K}_{1.21\times 10^{-6}}(3).$$

For the second eigenvalue and with the choice $D = \mathrm{diag}(1, d, 1)$:

$$2 - 2 \times 10^{-3}\, d > 1 + 10^{-4} + 10^{-3}\, d^{-1}, \text{ and}$$
$$2 + 2 \times 10^{-3}\, d < 3 - 10^{-4} - 10^{-3}\, d^{-1},$$

with an inequality of the form

$$d > \frac{22}{9999 + \sqrt{99979201}} \approx 0.0011\ldots$$

and an (obviously) appropriate choice of $d = 0.002$. Hence:

$$\rho_2 = (2 \times 10^{-3})^2 = 4 \times 10^{-6} : \quad \overline{K}_{4\times 10^{-6}}(2).$$

Solution A.5.8: Let $z^0 \in \mathbb{C}^n$ with $\|z^0\| = 1$ be an arbitrary starting point. Then, construct a sequence $z^t \in \mathbb{C}^n$, $t = 1, 2, \ldots$ by

$$\tilde{z}^t := Az^{t-1}, \quad z^t = \tilde{z}^t / \|\tilde{z}^t\|.$$

In case of a general matrix the corrsponding eigenvalue approximation is given by

$$\lambda^t := \frac{(Az^t)_r}{z_r^t},$$

where r is an index such that $|z_r^t| = \max_{j=1,\ldots,n} |z_j^t|$. In case of a Hermitian matrix A, the eigenvalue approximation can be determined with the help of the Rayleigh quotient:

$$\lambda^t := \frac{(Az^t, z^t)}{\|z^t\|^2}.$$

i) The power method converges if A is diagonalizable and the eigenvalue with largest modulus is separated from the other eigenvalues, i. e. $|\lambda_n| > |\lambda_i|$ for $i < n$. Furthermore the starting vector z^0 must have a non-trivial component in the direction of the eigenvector w_n corresponding to λ_n.

ii) The separation of the biggest eigenvalue from the others is the most crucial restriction because the convergence rate is directly connected to this property (see iii)), and the other two conditions are usually fulfilled (due to round-off errors).

iii) The power method has the following a priori error estimate (for a general matrix):

$$\lambda^t = \lambda_{\max} + \mathcal{O}\Big(\left| \frac{\lambda_{n-1}}{\lambda_n} \right|^t \Big), \quad t \to \infty.$$