3 Iterative Methods for Linear Algebraic Systems

In this chapter, we discuss *iterative* methods for solving linear systems. The underlying problem has the form

$$Ax = b, \tag{3.0.1}$$

with a *real* square matrix $A = (a_{ij})_{i,j=1}^n \in \mathbb{R}^{n\times n}$ and a vector $b = (b_j)_{j=1}^n \in \mathbb{R}^n$. Here, we concentrate on the higher-dimensional case $n \gg 10^3$, such that, besides arithmetical complexity, also storage requirement becomes an important issue. In practice, high-dimensional matrices usually have very special structure, e. g., band structure and extreme sparsity, which needs to be exploited by the solution algorithms. The most cost-intensive parts of the considered algorithms are simple matrix-vector multiplications $x \to Ax$. Most of the considered methods and results are also applicable in the case of matrices and right-hand sides with *complex* entries.

3.1 Fixed-point iteration and defect correction

For the construction of cheap iterative methods for solving problem (3.0.1), one rewrites it in form of an equivalent fixed-point problem,

$$Ax = b \quad \Leftrightarrow \quad Cx = Cx - Ax + b \quad \Leftrightarrow \quad x = (I - C^{-1}A)x + C^{-1}b,$$

with a suitable regular matrix $C \in \mathbb{R}^{n\times n}$, the so-called "preconditioner". Then, starting from some initial value x^0, one uses a simple fixed-point iteration,

$$x^t = \underbrace{(I - C^{-1}A)}_{=:\, B}\, x^{t-1} + \underbrace{C^{-1}b}_{=:\, c}, \quad t = 1, 2, \dots . \tag{3.1.2}$$

Here, the matrix $B = I - C^{-1}A$ is called the "iteration matrix" of the fixed-point iteration. Its properties are decisive for the convergence of the method. In practice, such a fixed-point iteration is organized in form of a "defect correction" iteration, which essentially requires in each step only a matrix-vector multiplication and the solution of a linear system with the matrix C as coefficient matrix:

$$d^{t-1} = b - Ax^{t-1} \text{ (residual)}, \quad C\delta x^t = d^{t-1} \text{ (correction)}, \quad x^t = x^{t-1} + \delta x^t \text{ (update)}.$$

Example 3.1: The simplest method of this type is the *(damped) Richardson*[1] *method*, which for a suitable parameter $\theta \in (0, 2\lambda_{max}(A)^{-1}]$ uses the matrices

$$C = \theta^{-1}I, \qquad B = I - \theta A. \tag{3.1.3}$$

[1] Lewis Fry Richardson (1881–1953): English mathematician and physicist; worked at several institutions in in England and Scotland; a typical "applied mathematician"; pioneered modeling and numerics in weather prediction.

Starting from some initial value x^0 the iteration looks like

$$x^t = x^{t-1} + \theta(b - Ax^{t-1}), \quad t = 1, 2, \dots . \tag{3.1.4}$$

In view of the Banach fixed-point theorem a sufficient criterion for the convergence of the fixed-point iteration (3.1.2) is the contraction property of the corresponding fixed-point mapping $g(x) := Bx + c$,

$$\|g(x) - g(y)\| = \|B(x - y)\| \le \|B\|\|x - y\|, \qquad \|B\| < 1,$$

in some vector norm $\|\cdot\|$. For a given iteration matrix B the property $\|B\| < 1$ may depend on the particular choice of the norm. Hence, it is desirable to characterize the convergence of this iteration in terms of norm-independent properties of B. For this, the appropriate quantity is the "spectral radius"

$$\mathrm{spr(B)} := \max\{\,|\lambda| : \lambda \in \sigma(\mathrm{B})\,\}.$$

Obviously, $\mathrm{spr(B)}$ is the radius of the smallest circle in $\mathbb{C}$ around the origin, which contains all eigenvalues of B. For any natural matrix norm $\|\cdot\|$, there holds

$$\mathrm{spr(B)} \le \|\mathrm{B}\|. \tag{3.1.5}$$

For symmetric B, we even have

$$\mathrm{spr(B)} = \|\mathrm{B}\|_2 = \sup_{\mathrm{x}\in\mathbb{R}^{\mathrm{n}}\setminus\{0\}} \frac{\|\mathrm{Bx}\|_2}{\|\mathrm{x}\|_2}. \tag{3.1.6}$$

However, we note that $\mathrm{spr}(\cdot)$ does not define a norm on $\mathbb{R}^{n\times n}$ since the triangle inequality does not hold in general.

Theorem 3.1 (Fixed-point iteration): *The fixed-point iteration (3.1.2) converges for any starting value x^0 if and only if*

$$\rho := \mathrm{spr}(B) < 1. \tag{3.1.7}$$

In case of convergence the limit is the uniquely determined fixed point x. The asymptotic convergence behavior with respect to any vector norm $\|\cdot\|$ is characterized by

$$\sup_{x^0\in\mathbb{R}^n} \limsup_{t\to\infty} \Big(\frac{\|x^t - x\|}{\|x^0 - x\|}\Big)^{1/t} = \rho. \tag{3.1.8}$$

Hence, the number of iteration steps necessary for an asymptotic error reduction by a small factor $\mathrm{TOL} > 0$ is approximately given by

$$t(\mathrm{TOL}) \approx \frac{\ln(1/\mathrm{TOL})}{\ln(1/\rho)}. \tag{3.1.9}$$

Proof. Assuming the existence of a fixed point x, we introduce the notation $e^t := x^t - x$. Recalling that $x = Bx + c$, we find

$$e^t = x^t - x = Bx^{t-1} + c - (Bx + c) = Be^{t-1} = \cdots = B^t e^0.$$

i) In case that spr(B) < 1, in view of Lemma 3.1 below, there exists a vector norm $\|\cdot\|_{B,\varepsilon}$ depending on B and some $\varepsilon > 0$ chosen sufficiently small, such that the corresponding natural matrix norm $\|\cdot\|_{B,\varepsilon}$ satisfies

$$\|B\|_{B,\varepsilon} \le \mathrm{spr(B)} + \varepsilon = \rho + \varepsilon < 1.$$

Consequently, by the Banach fixed-point theorem, there exists a unique fixed-point x and the fixed-point iteration converges for any starting value x^0:

$$\|e^t\|_{B,\varepsilon} = \|B^t e^0\|_{B,\varepsilon} \le \|B^t\|_{B,\varepsilon}\, \|e^0\|_{B,\varepsilon} \le \|B\|_{B,\varepsilon}^t\, \|e^0\|_{B,\varepsilon} \ \to\ 0.$$

In view of the norm equivalence in $\mathbb{R}^n$ this means convergence $x^t \to x\ (t \to \infty)$.

ii) Now, we assume convergence for any starting value x^0. Let λ be an eigenvalue of B such that $|\lambda| = \rho$ and $w \neq 0$ a corresponding eigenvector. Then, for the particular starting value $x^0 := x + w$, we obtain

$$\lambda^t e^0 = \lambda^t w = B^t w = B^t e^0 = e^t \to 0 \quad (t \to \infty).$$

This necessarily requires spr(B) $= |\lambda| < 1$. As byproduct of this argument, we see that in this particular case

$$\Big(\frac{\|e^t\|}{\|e^0\|}\Big)^{1/t} = \rho, \quad t \in \mathbb{N}.$$

iii) For an arbitrary small $\varepsilon > 0$ let $\|\cdot\|_{B,\varepsilon}$ again be the above special norm for which $\|B\|_{B,\varepsilon} \le \rho + \varepsilon$. Then, by the norm equivalence for any other vector norm $\|\cdot\|$ there exist positive numbers $m = m(B,\varepsilon)$, $M = M(B,\varepsilon)$ such that

$$m\|x\| \le \|x\|_{B,\varepsilon} \le M\|x\|, \quad x \in \mathbb{R}^n.$$

Using this notation, we obtain

$$\|e^t\| \le \frac{1}{m}\|e^t\|_{B,\varepsilon} = \frac{1}{m}\|B^t e^0\|_{B,\varepsilon} \le \frac{1}{m}\|B\|_{B,\varepsilon}^t \|e^0\|_{B,\varepsilon} \le \frac{M}{m}(\rho+\varepsilon)^t\|e^0\|,$$

and, consequently, observing that $\big(\frac{M}{m}\big)^{1/t} \to 1\ (t \to \infty)$:

$$\limsup_{t\to\infty} \Big(\frac{\|e^t\|}{\|e^0\|}\Big)^{1/t} \le \rho + \varepsilon.$$

Since $\varepsilon > 0$ can be chosen arbitrarily small and recalling the last identity in (ii), we obtain the asserted identity (3.1.8).

iv) Finally requiring an error reduction by $\mathrm{TOL} > 0$, we have to set

$$\frac{\|x^t - x\|}{\|x^0 - x\|} \le (\rho + \varepsilon)^t \approx \mathrm{TOL}, \quad t \ge t(\mathrm{TOL}),$$

from which we obtain

$$t(\mathrm{TOL}) \approx \frac{\ln(1/\mathrm{TOL})}{\ln(1/\rho)}.$$

This completes the proof. Q.E.D.

The spectral radius of the iteration matrix determines the general asymptotic convergence behavior of the fixed-point iteration. The relation (3.1.9) can be interpreted as follows: In case that $\rho = \mathrm{spr(B)} < 1$ the error obtained in the t-th step (t sufficiently large) can be further reduced by a factor 10^{-1}, i. e., gaining one additional decimal in accuracy, by

$$t(10^{-1}) \approx \frac{\ln(1/10)}{\ln(1/\rho)}$$

more iterations. For example, for $\rho \sim 0.99$, which is not at all unrealistic, we have $t_1 \sim 230$. For large systems with $n \gg 10^6$ this means substantial work even if each iteration step only requires $\mathcal{O}(n)$ arithmetic operations.

We have to provide the auxiliary lemma used in the proof of Theorem 3.1.

Lemma 3.1 (Spectral radius): *For any matrix $B \in \mathbb{R}^{n\times n}$ and any small $\varepsilon > 0$ there exists a natural matrix norm $\|\cdot\|_{B,\varepsilon}$, such that*

$$\mathrm{spr}(B) \le \|B\|_{B,\varepsilon} \le \mathrm{spr}(B) + \varepsilon. \tag{3.1.10}$$

Proof. The matrix B is similar to an upper triangular matrix (e. g., its Jordan normal form),

$$B = T^{-1}RT\,, \quad R = \begin{bmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ 0 & & r_{nn} \end{bmatrix},$$

with the eigenvalues of B on its main diagonal. Hence,

$$\mathrm{spr(B)} = \max_{1\le \mathrm{i} \le \mathrm{n}} |\mathrm{r_{ii}}|.$$

For an arbitrary $\delta \in (0,1]$, we set

$$S_\delta = \begin{bmatrix} 1 & & & 0 \\ & \delta & & \\ & & \ddots & \\ 0 & & & \delta^{n-1} \end{bmatrix} \quad R_0 = \begin{bmatrix} r_{11} & & 0 \\ & \ddots & \\ 0 & & r_{nn} \end{bmatrix} \quad Q_\delta = \begin{bmatrix} 0 & r_{12} & \delta r_{13} & \cdots & \delta^{n-2} r_{1n} \\ & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & \delta r_{n-2,n} \\ & & & \ddots & r_{n-1,n} \\ & & & & 0 \end{bmatrix},$$

and, with this notation, have

$$R_\delta := S_\delta^{-1} R S_\delta = \begin{bmatrix} r_{11} & \delta r_{12} & \cdots & \delta^{n-1} r_{1n} \\ & \ddots & \ddots & \vdots \\ & & \ddots & \delta r_{n-1,n} \\ 0 & & & r_{nn} \end{bmatrix} = R_0 + \delta Q_\delta.$$

In view of the regularity of $S_\delta^{-1}T$, a vector norm is defined by

$$\|x\|_\delta := \|S_\delta^{-1} T x\|_2 , \quad x \in \mathbb{R}^n.$$

Then, observing $R = S_\delta R_\delta S_\delta^{-1}$, there holds

$$B = T^{-1} R T = T^{-1} S_\delta R_\delta S_\delta^{-1} T.$$

Hence for all $x \in \mathbb{R}^n$ and $y = S_\delta^{-1} T x$:

$$\begin{aligned} \|Bx\|_\delta &= \|T^{-1} S_\delta R_\delta S_\delta^{-1} T x\|_\delta = \|R_\delta y\|_2 \\ &\le \|R_0 y\|_2 + \delta \|Q_\delta y\|_2 \le \{\max_{1\le i\le n} |r_{ii}| + \delta\mu\} \, \|y\|_2 \\ &\le \{\mathrm{spr(B)} + \delta\mu\} \|\mathrm{x}\|_\delta \end{aligned}$$

with the constant

$$\mu = \Big(\sum_{i,j=1}^{n} |r_{ij}|^2 \Big)^{1/2}.$$

This implies

$$\|B\|_\delta = \sup_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\|Bx\|_\delta}{\|x\|_\delta} \le \mathrm{spr(B)} + \mu\delta,$$

and setting $\delta := \varepsilon/\mu$ the desired vector norm is given by $\|\cdot\|_{B,\varepsilon} := \|\cdot\|_\delta$. Q.E.D.

3.1.1 Stopping criteria

In using an iterative method, one needs "stopping criteria", which for some prescribed accuracy TOL terminates the iteration, in the ideal case, once this required accuracy is reached.

i) **Strategy 1.** From the Banach fixed-point theorem, we have the general error estimate

$$\|x^t - x\| \le \frac{q}{1-q}\,\|x^t - x^{t-1}\|, \tag{3.1.11}$$

with the "contraction constant" $q = \|B\| < 1$. For a given error tolerance $\mathrm{TOL} > 0$ the iteration could be stopped when

$$\frac{\|B\|}{1-\|B\|}\frac{\|x^t - x^{t-1}\|}{\|x^t\|} \le \mathrm{TOL}. \tag{3.1.12}$$

The realization of this strategy requires an quantitatively correct estimate of the norm $\|B\|$ or of spr(B). That has to be generated from the computed iterates x^t, i. e., *a posteriori* in the course of the computation. In general the iteration matrix $B = I - C^{-1}A$ cannot be computed explicitly with acceptable work. Methods for estimating spr(B) will be considered in the chapter about the iterative solution of eigenvalue problems, below.

ii) **Strategy 2.** Alternatively, one can evaluate the "residual" $\|Ax^t - b\|$. Observing that $e^t = x^t - x = A^{-1}(Ax^t - b)$ and $x = A^{-1}b$, it follows that

$$\|e^t\| \le \|A^{-1}\|\,\|Ax^t - b\|\,, \qquad \frac{1}{\|b\|} \ge \frac{1}{\|A\|\,\|x\|},$$

and further

$$\frac{\|e^t\|}{\|x\|} \le \|A^{-1}\|\|A\|\frac{\|Ax^t - b\|}{\|b\|} = \mathrm{cond(A)}\,\frac{\|\mathrm{Ax^t - b}\|}{\|\mathrm{b}\|}.$$

This leads us to the stopping criterion

$$\mathrm{cond(A)}\,\frac{\|\mathrm{Ax^t - b}\|}{\|\mathrm{b}\|} \le \mathrm{TOL}. \tag{3.1.13}$$

The evaluation of this criterion requires an estimate of cond(A), which may be as costly as the solution of the equation $Ax = b$ itself. Using the spectral norm $\|\cdot\|_2$ the condition number is related to the singular values of A (square roots of the eigenvalues of A^TA),

$$\mathrm{cond_2(A)} = \frac{\sigma_{\max}}{\sigma_{\min}}.$$

Again generating accurate estimates of these eigenvalues may require more work than the solution of $Ax = b$. This short discussion shows that designing useful stopping criteria for iterative methods is an not at all an easy task. However, in the context of linear systems originating from the "finite element discretization" ("FEM") of partial differential equations there are approaches based on the concept of "Galerkin orthogonality", which allow for a systematic balancing of iteration and discretization errors. In this way, practical stopping criteria can be designed, by which the iteration may be terminated once the level of the discretization error is reached. Here, the criterion is essentially the approximate solution's "violation of Galerkin orthogonality" (s. Meidner et al. [43] and Rannacher et al. [45] for more details).

3.1.2 Construction of iterative methods

The construction of concrete iterative methods for solving the linear system $Ax = b$ by defect correction requires the specification of the preconditioner C. For this task two particular goals have to be observed:

- $\mathrm{spr(I - C^{-1}A)}$ should be as small as possible.
- The correction equation $C\delta x^t = b - Ax^{t-1}$ should be solvable with $\mathcal{O}(n)$ a. op., requiring storage space not much exceeding that for storing the matrix A itself.

Unfortunately, these requirements contradict each other. The two extreme cases are:

$$\begin{aligned} C &= A & &\Rightarrow & &\mathrm{spr(I{-}C^{-1}A)} = 0 \\ C &= \theta^{-1}I & &\Rightarrow & &\mathrm{spr(I{-}C^{-1}A)} \approx 1. \end{aligned}$$

The simplest preconditioners are defined using the natural additive decomposition of the matrix, $A = L + D + R$, where

$$D = \begin{bmatrix} a_{11} & & \cdots & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & \cdots & & a_{nn} \end{bmatrix} \quad L = \begin{bmatrix} 0 & & \cdots & 0 \\ a_{21} & \ddots & & \\ \vdots & \ddots & \ddots & \\ a_{n1} & \cdots & a_{n,n-1} & 0 \end{bmatrix} \quad R = \begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ & \ddots & \ddots & \vdots \\ & & \ddots & a_{n-1,n} \\ 0 & \cdots & & 0 \end{bmatrix}.$$

Further, we assume that the main diagonal elements of A are nonzero, $a_{ii} \neq 0$.

1. *Jacobi*[2] *method ("Gesamtschrittverfahren" in German):*

$$C = D, \qquad B = -D^{-1}(L + R) =: J \quad \text{(iteration matrix).} \tag{3.1.14}$$

The iteration of the Jacobi method reads

$$Dx^t = b - (L{+}R)x^{t-1}, \quad t = 1, 2, \dots\,, \tag{3.1.15}$$

or written component-wise:

$$a_{ii}x_i^t = b_i - \sum_{j=1}^{n} a_{ij}x_j^t, \quad i = 1, \dots, n,$$

2. *Gauß-Seidel method ("Einzelschrittverfahren" in German):*

$$C = D + L, \qquad B = -(D + L)^{-1}R =: H_1 \quad \text{(iteration matrix).} \tag{3.1.16}$$

[2] Carl Gustav Jakob Jacobi (1804–1851): German mathematician; already as child highly gifted; worked in Königsberg and Berlin; contributions to many parts of mathematics: Number Theory, elliptic functions, partial differential equations, functional determinants, and Theoretical Mechanics.

The iteration of the Gauß-Seidel method reads as follow:

$$(D+L)x^t = b - Rx^{t-1}, \quad t = 1, 2, \dots .$$

Writing this iteration componentwise,

$$a_{ii}x_i^t = b_i - \sum_{j<i} a_{ij}x_j^t - \sum_{j>i} a_{ij}x_j^{t-1}, \quad i = 1, \dots, n,$$

one sees that Jacobi and Gauß-Seidel method have exactly the same arithmetic complexity per iteration step and require the same amount of storage. However, since the latter method uses a better approximation of the matrix A as preconditioner it is expected to have an iteration matrix with smaller spectral radius, i. e., converges faster. It will be shown below that this is actually the case for certain classes of matrices A.

3. *SOR method ("Successive Over-Relaxation"):* $\omega \in (0,2)$

$$C = \frac{1}{\omega}(D+\omega L), \qquad B = -(D+\omega L)^{-1}[(\omega-1)D + \omega R]. \tag{3.1.17}$$

The SOR method is designed to accelerate the Gauß-Seidel method by introducing a "relaxation parameter" $\omega \in \mathbb{R}$, which can be optimized in order to minimize the spectral radius of the corresponding iteration matrix. Its iteration reads as follows:

$$(D+\omega L)x^t = \omega b - [(\omega-1)D + \omega R]x^{t-1}, \quad t = 1, 2, \dots .$$

The arithmetic complexity is about that of Jacobi and Gauß-Seidel method. But the parameter ω can be optimized for a certain class of matrices resulting in a significantly faster convergence than that of the other two simple methods.

4. *ILU method ("Incomplete LU Decomposition"):*

$$C = \tilde{L}\tilde{R}, \qquad B = I - \tilde{R}^{-1}\tilde{L}^{-1}A. \tag{3.1.18}$$

For a symmetric, positive definite matrix A the ILU method naturally becomes the ILL^T method ("Incomplete Cholesky decomposition"). The ILU decomposition is obtained by the usual recursive process for the direct computation of the LU decomposition from the relation $LU = A$ by setting all matrix elements to zero, which correspond to index pairs $\{i,j\}$ for which $a_{ij} = 0$:

$$\begin{aligned}
i = 1, \dots, n: \quad & \tilde{r}_{il} = a_{il} - \sum_{k=1}^{i-1} \tilde{l}_{ik}\tilde{r}_{kl} \quad (l = 1, \dots, n) \\
& \tilde{l}_{ii} = 1, \quad \tilde{l}_{ki} = \tilde{r}_{ii}^{-1}\Big\{a_{ki} - \sum_{l=1}^{i-1} \tilde{l}_{kl}\tilde{r}_{li}\Big\} \quad (k = i+1, \dots, n) \\
& \tilde{l}_{ij} = 0, \ \tilde{r}_{ij} = 0, \ \text{for } a_{ij} = 0\,.
\end{aligned}$$

If this process stops because some $\tilde r_{ii} = 0$, we set $\tilde r_{ii} := \delta > 0$ and continue. The iteration of the ILU method reads as follows:

$$\tilde L \tilde R x^t = (\tilde L \tilde R - A)x^{t-1} + b, \quad t = 1, 2, \dots .$$

We note that here, L and U stand for "lower" and "upper" triangular matrix, respectively, in contrast to the notion L and R for "left" and "right" triangular matrix as used before in the context of multiplicative matrix decomposition.

Again this preconditioner is cheap, for sparse matrices, $\mathcal{O}(n)$ a. op. per iteration step, but its convergence is difficult to analyze and will not be discussed further. However, in certain situations the ILU method plays an important role as a robust "smoothing iteration" within "multigrid methods" to be discussed below.

5. *ADI method ("Alternating-Direction Implicit Iteration"):*

$$\begin{aligned} C &= (A_x + \omega I)(A_y + \omega I), \\ B &= (A_y + \omega I)^{-1}(\omega I - A_x)(A_x + \omega I)^{-1}(\omega I - A_y). \end{aligned} \tag{3.1.19}$$

The ADI method can be applied to matrices A which originate from the discretization of certain elliptic partial differential equations, in which the contributions from the different spatial directions (x-direction and y-direction in 2D) are separated in the form $A = A_x + A_y$. A typical example is the central difference approximation of the Poisson equation described in Chapter 0.4.2. The iteration of the ADI method reads as follows:

$$(A_x + \omega I)(A_y + \omega I)x^t = \big((A_x + \omega I)(A_y + \omega I) - A\big)x^{t-1} + b, \quad t = 1, 2, \dots .$$

Here, the matrices $A_x + \omega I$ and $A_y + \omega I$ are tri-diagonal, such that the second goal "solution efficiency" is achieved, while the full matrix A is five-diagonal. This method can be shown to converge for any choice of the parameter $\omega > 0$. For certain classes of matrices the optimal choice of ω leads to convergence, which is at least as fast as that of the optimal SOR method. We will not discuss this issue further since the range of applicability of the ADI method is rather limited.

Remark 3.1 (Block-versions of fixed-point iterations): Sometimes the coefficient matrix A has a regular block structure for special numberings of the unknowns (e. g., in the discretization of the Navier-Stokes equations when grouping the velocity and pressure unknowns together at each mesh point):

$$A = \begin{bmatrix} A_{11} & & \cdots & A_{1r} \\ & \ddots & & \\ \vdots & & \ddots & \vdots \\ A_{r1} & \cdots & & A_{rr} \end{bmatrix},$$

where the submatrices A_{ij} are of small dimension, $3-10$, such that the explicit inversion of the diagonal blocks A_{ii} is possible without spoiling the overall complexity of $\mathcal{O}(n)$ a. op. per iteration step.

3.1.3 Jacobi- and Gauß-Seidel methods

In the following, we will give a complete convergence analysis of Jacobi and Gauß-Seidel method. As already stated above, both methods have the same arithmetic cost (per iteration step) and require not much more storage as needed for storing the matrix A. This simplicity suggests that both methods may not be very fast, which will actually be seen below at the model matrix in Example (2.7) of Section 2.2.

Theorem 3.2 (Strong row-sum criterion): *If the row sums or the column sums of the matrix $A \in \mathbb{R}^{n\times n}$ satisfy the condition (strict diagonal dominance)*

$$\sum_{k=1,k\neq j}^{n} |a_{jk}| < |a_{jj}| \quad or \quad \sum_{k=1,k\neq j}^{n} |a_{kj}| < |a_{jj}|, \quad j=1,\dots,n, \tag{3.1.20}$$

then, $\operatorname{spr}(J) < 1$ *and* $\operatorname{spr}(H_1) < 1$ *, i. e., Jacobi and Gauß-Seidel method converge.*

Proof. First, assume that the matrix A is strictly diagonally dominant. Let $\lambda \in \sigma(J)$ and $\mu \in \sigma(H_1)$ with corresponding eigenvectors v and w, respectively. Then, noting that $a_{jj} \neq 0$, we have

$$\lambda v = Jv = -D^{-1}(L+R)v$$

and

$$\mu w = H_1 w = -(D+L)^{-1}Rw \quad \Leftrightarrow \quad \mu w = -D^{-1}(\mu L+R)w.$$

From this it follows that for $\|v\|_\infty = \|w\|_\infty = 1$ and using the strict diagonal dominance of A:

$$|\lambda| \le \|D^{-1}(L+R)\|_\infty = \max_{j=1,\dots,n} \Big\{ \frac{1}{|a_{jj}|} \sum_{k=1,k\neq j}^{n} |a_{jk}| \Big\} < 1.$$

Hence, spr(J) < 1. Further,

$$|\mu| \le \|D^{-1}(\mu L+R)\|_\infty \le \max_{1\le j\le n} \Big\{ \frac{1}{|a_{jj}|} \Big[\sum_{k<j} |\mu|\,|a_{jk}| + \sum_{k>j} |a_{jk}| \Big] \Big\}.$$

For $|\mu| \ge 1$, we would obtain the contradiction

$$|\mu| \le |\mu|\,\|D^{-1}(L+R)\|_\infty < |\mu|,$$

so that also spr(H$_1$) < 1. If instead of A its transpose A^T is strictly diagonally dominant, we can argue analogously since, in view of $\lambda(\bar{A}^T) = \overline{\lambda(A)}$, the spectral radii of these two matrices coincide. Q.E.D.

Remark 3.2: We show an example of a non-symmetric matrix A, which satisfies the strong column- but not the strong row-sum criterion:

$$A = \begin{bmatrix} 4 & 4 & 1 \\ 2 & 5 & 3 \\ 1 & 0 & 5 \end{bmatrix}, \qquad A^T = \begin{bmatrix} 4 & 2 & 1 \\ 4 & 5 & 0 \\ 1 & 3 & 5 \end{bmatrix}.$$

Clearly, for symmetric matrices the two conditions are equivalent.

The strict diagonal dominance of A or A^T required in Theorem 3.2 is a too restrictive condition for the needs of many applications. In most cases only simple "diagonal dominance" is given as in the Example (2.7) of Section 2.2,

$$A = \left.\begin{bmatrix} B & -I_4 & & \\ -I_4 & B & -I_4 & \\ & -I_4 & B & -I_4 \\ & & -I_4 & B \end{bmatrix}\right\} 16\,, \qquad B = \left.\begin{bmatrix} 4 & -1 & & \\ -1 & 4 & -1 & \\ & -1 & 4 & -1 \\ & & -1 & 4 \end{bmatrix}\right\} 4$$

However, this matrix is *strictly* diagonally dominant in some of its rows, which together with an additional structural property of A can be used to guarantee convergence of Jacobi and Gauß-Seidel method.

Definition 3.1: *A matrix $A \in \mathbb{R}^{n\times n}$ is called "reducible", if there exists a permutation matrix P such that*

$$PAP^T = \begin{bmatrix} \tilde A_{11} & 0 \\ \tilde A_{21} & \tilde A_{22} \end{bmatrix},$$

(simultaneous row and column permutation) with matrices $\tilde A_{11} \in \mathbb{R}^{p\times p}$, $\tilde A_{22} \in \mathbb{R}^{q\times q}$, $\tilde A_{21} \in \mathbb{R}^{q\times p}$, $p, q > 0$, $p + q = n$. It is called "irreducible" if it is not reducible.

For a reducible matrix A the linear system $Ax = b$ can be transformed into an equivalent system of the form $PAP^Ty = Pb, x = P^Ty$ which is decoupled into two separate parts such that it could be solved in two successive steps. The following lemma provides a criterion for the irreducibility of the matrix A, which can be used in concrete cases. For example, the above model matrix A is irreducible.

Lemma 3.2 (Irreducibility): *A matrix $A \in \mathbb{R}^{n\times n}$ is irreducible if and only if the associated directed graph*

$$G(A) := \big\{ \textit{knots}\ P_1, ..., P_n\,,\ \textit{edges}\ \overline{P_jP_k} \Leftrightarrow a_{jk} \neq 0,\ j,k = 1, ..., n \big\}$$

is connected, i. e., for each pair of knots $\{P_j, P_k\}$ there exists a directed connection between P_j and P_k.

Proof. The reducibility of A can be formulated as follows: There exists a non-trivial decomposition $N_n = J \cup K$ of the index set $N_n = \{1, ..., n\}$, $J,\, K \neq \emptyset$, $J \cap K = \emptyset$ such that $a_{jk} = 0$ for all pairs $\{j,k\} \in J \times K$. Connectivity of the graph $G(A)$ now means that for any pair of indices $\{j,k\}$ there exists a chain of indices $i_1, \dots, i_m \in \{1, \dots, n\}$ such that

$$a_{ji_1} \neq 0\,,\ a_{i_1 i_2} \neq 0\,, \dots\,, a_{i_{m-1} i_m} \neq 0\,,\ a_{i_m k} \neq 0.$$

From this, we can conclude the asserted characterization (left as exercise). Q.E.D.

For irreducible matrices the condition in the strong row-sum criterion can be relaxed.

Theorem 3.3 (Weak row-sum criterion): *Let the matrix $A \in \mathbb{R}^{n\times n}$ be irreducible and diagonally dominant,*

$$\sum_{k=1,k\neq j}^{n} |a_{jk}| \leq |a_{jj}| \quad j = 1, \dots, n, \tag{3.1.21}$$

and let for at least one index $r \in \{1, \dots, n\}$ the corresponding row sum satisfy

$$\sum_{k=1,k\neq r}^{n} |a_{rk}| < |a_{rr}|. \tag{3.1.22}$$

Then, A is regular and $\mathrm{spr}(J) < 1$ and $\mathrm{spr}(H_1) < 1$, i. e., Jacobi and Gauß-Seidel method converge. An analogous criterion holds in terms of the column sums of A.

Proof. i) Because of the assumed irreducibility of the matrix A there necessarily holds

$$\sum_{k=1}^{n} |a_{jk}| \,>\, 0\,, \quad j = 1, \dots, n\,,$$

and, consequently, by its diagonal dominance, $a_{jj} \neq 0,\, j = 1, \dots, n$. Hence, Jacobi and Gauß-Seidel method are feasible. With the aid of the diagonal dominance, we conclude analogously as in the proof of Theorem 3.2 that

$$\mathrm{spr(J)} \leq 1\,, \quad \mathrm{spr(H_1)} \leq 1.$$

ii) Suppose now that there is an eigenvalue $\lambda \in \sigma(J)$ with modulus $|\lambda| = 1$. Let $v \in \mathbb{C}^n$ be a corresponding eigenvector with a component v_s satisfying $|v_s| = \|v\|_\infty = 1$. There holds

$$|\lambda|\, |v_i| \leq |a_{ii}|^{-1} \sum_{k\neq i} |a_{ik}|\, |v_k|\,, \quad i = 1, \dots, n. \tag{3.1.23}$$

By the assumed irreducibility of A in the sense of Lemma 3.2 there exist a chain of indices $i_1, \dots, i_m$ such that $a_{si_1} \neq 0\,, \dots, a_{i_m r} \neq 0$. Hence, by multiple use of the inequality

(3.1.23), we obtain the following contradiction (observe that $|\lambda| = 1$)

$$\begin{aligned}
|v_r| &= |\lambda v_r| \le |a_{rr}|^{-1} \sum_{k\neq r} |a_{rk}|\, \|v\|_\infty < \|v\|_\infty,\\
|v_{i_m}| &= |\lambda v_{i_m}| \le |a_{i_m i_m}|^{-1} \Big\{ \sum_{k\neq i_m, r} |a_{i_m k}|\, \|v\|_\infty + \underbrace{|a_{i_m r}|}_{\neq 0}\, |v_r| \Big\} < \|v\|_\infty,\\
&\;\vdots\\
|v_{i_1}| &= |\lambda v_{i_1}| \le |a_{i_1 i_1}|^{-1} \Big\{ \sum_{k\neq i_1, i_2} |a_{i_1 k}|\, \|v\|_\infty + \underbrace{|a_{i_1 i_2}|}_{\neq 0}\, |v_{i_2}| \Big\} < \|v\|_\infty,\\
\|v\|_\infty &= |\lambda v_s| \le |a_{ss}|^{-1} \Big\{ \sum_{k\neq s, i_1} |a_{sk}|\, \|v\|_\infty + \underbrace{|a_{s i_1}|}_{\neq 0}\, |v_{i_1}| \Big\} < \|v\|_\infty.
\end{aligned}$$

Consequently, there must hold $\mathrm{spr(J)} < 1$. Analogously, we also conclude $\mathrm{spr(H_1)} < 1$. Finally, in view of $A = D(I-J)$ the matrix A must be regular. Q.E.D.

3.2 Acceleration methods

For practical problems Jacobi and Gauß-Seidel method are usually much too slow. Therefore, one tries to improve their convergence by several strategies, two of which will be discussed below.

3.2.1 SOR method

The SOR method can be interpreted as combining the Gauß-Seidel method with an extra "relaxation step". Starting from a standard Gauß-Seidel step in the t-th iteration,

$$\tilde{x}_j^t = \frac{1}{a_{jj}} \Big\{ b_j - \sum_{k<j} x_k^t - \sum_{k>j} x_k^{t-1} \Big\},$$

the next iterate x_j^t is generated as a convex linear combination ("relaxation") of the form

$$x_j^t = \omega \tilde{x}_j^t + (1-\omega)\, x_j^{t-1},$$

with a parameter $\omega \in (0,2)$. For $\omega = 1$ this is just the Gauß-Seidel iteration. For $\omega < 1$, one speaks of "underrelaxation" and for $\omega > 1$ of "overrelaxation". The iteration matrix of the SOR methods is obtained from the relation

$$x^t = \omega D^{-1}\{b - Lx^t - Rx^{t-1}\} + (1-\omega)x^{t-1}$$

as

$$H_\omega = -(D+\omega L)^{-1}\,[(\omega - 1)\, D + \omega R].$$

Hence, the iteration reads

$$x^t = H_\omega x^{t-1} + \omega(D+\omega L)^{-1}b, \tag{3.2.24}$$

or in componentwise notation:

$$x_i^t = (1-\omega)x_i^{t-1} + \frac{\omega}{a_{ii}}\Big(b_i - \sum_{j<i} a_{ij}x_j^t - \sum_{j>i} a_{ij}x_j^{t-1}\Big), \quad i=1,\dots,n. \tag{3.2.25}$$

The following lemma shows that in the relaxation parameter has to be picked in the range $0<\omega<2$ if one wants to guarantee convergence.

Lemma 3.3 (Relaxation): *For an arbitrary matrix $A\in\mathbb{R}^{n\times n}$ with regular D there holds*

$$\operatorname{spr}(H_\omega) \ge |\omega-1|\,, \quad \omega\in\mathbb{R}. \tag{3.2.26}$$

Proof. We have

$$H_\omega = (D+\omega L)^{-1}[\,(1-\omega)\,D - \omega R\,] = (I+\omega\underbrace{D^{-1}L}_{=:\,L'})^{-1}\,\underbrace{D^{-1}D}_{=\,I}\,[\,(1-\omega)\,I - \omega\underbrace{D^{-1}R}_{=:\,R'}\,].$$

Then,

$$\det(H_\omega) = \underbrace{\det(I+\omega L')}_{=\,1})^{-1}\cdot\underbrace{\det((1-\omega)\,I-\omega R'\,)}_{=\,(1-\omega)^n} = (1-\omega)^n.$$

Since $\det(H_\omega) = \prod_{i=1}^n \lambda_i$ ($\lambda_i\in H_\omega$) it follows that

$$\operatorname{spr}(H_\omega) = \max_{1\le i\le n}|\lambda_i| \ge \Big(\prod_{i=1}^n |\lambda_i|\Big)^{1/n} = |1-\omega|,$$

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

For positive definite matrices the assertion of Lemma 3.3 can be reversed in a certain sense. This is the content of the following theorem of of Ostrowski[3] and Reich[4] .

–

[3] Alexander Markowitsch Ostrowski (1893–1986): Russian-German-Swiss mathematician; studied at Marburg, Göttingen (with D. Hilbert and E. Landau) and Hamburg, since 1927 Prof. in Basel; worked on Dirichlet series, in Valuation Theory and especially in Numerical Analysis: “On the linear iteration procedures for symmetric matrices”, Rend. Mat. Appl. 5, 140–163 (1954).

[4] Edgar Reich (1927–2009): US-American mathematician of German origin; start as Electrical Engineer at MIT (Massachusetts, USA) and Rand Corp. working there on numerical methods and Queuing Theory: “On the convergence of the classical iterative method for solving linear simultaneous equations”, Ann. Math. Statist. 20. 448–451 (1949); PhD at UCLA and 2-year postdoc at Princeton, since 1956 Prof. at Univ. of Minnesota (Minneapolis, USA), work in Complex Analysis especially on quasi-conformal mappings.

Theorem 3.4 (Theorem of Ostrowski-Reich): *For a positive definite matrix $A \in \mathbb{R}^{n\times n}$ there holds*

$$\mathrm{spr}(H_\omega) < 1\,, \quad \textit{for} \quad 0 < \omega < 2. \tag{3.2.27}$$

Hence, especially the Gauß-Seidel method ($\omega = 1$) is convergent. Its asymptotic convergence speed can be estimated by

$$\mathrm{spr}(H_1) \le 1 - \frac{2}{\mu} + \frac{2}{\mu(\mu+1)}, \qquad \mu := \frac{\lambda_{\max}(D)}{\lambda_{\min}(A)}, \tag{3.2.28}$$

assuming the quantity $\mu \approx \mathrm{cond}_2(A)$ to be large.

Proof. i) In view of the symmetry of A, we have $R = L^T$, i. e., $A = L + D + L^T$. Let $\lambda \in \sigma(H_\omega)$ be arbitrary for $0 < \omega < 2$, with some eigenvector $v \in \mathbb{R}^n$, i. e., $H_\omega v = \lambda v$. Thus, there holds

$$\big((1-\omega)\,D - \omega L^T\big)\, v = \lambda\,(D+\omega L)\, v$$

and

$$\omega\,(D+L^T)\, v = (1-\lambda)\, Dv - \lambda\omega Lv.$$

From this, we conclude that

$$\begin{aligned} \omega Av &= \omega\,(D+L^T)\, v + \omega Lv \\ &= (1-\lambda)\, Dv - \lambda\omega Lv + \omega Lv = (1-\lambda)\, Dv + \omega\,(1-\lambda)\, Lv, \end{aligned}$$

and

$$\begin{aligned} \lambda\omega Av &= \lambda\omega\,(D+L^T)\, v + \lambda\omega Lv \\ &= \lambda\omega\,(D + L^T)\, v + (1-\lambda)\, Dv - \omega\,(D+L^T)\, v \\ &= (\lambda-1)\omega(D+L^T)\, v + (1-\lambda)\, Dv = (1-\lambda)(1-\omega)\, Dv - (1-\lambda)\,\omega L^T v. \end{aligned}$$

Observing $v^T Lv = v^T L^T v$ implies

$$\begin{aligned} \omega v^T Av &= (1-\lambda)\, v^T Dv + \omega\,(1-\lambda)\, v^T Lv \\ \lambda\omega v^T Av &= (1-\lambda)(1-\omega)\, v^T Dv - (1-\lambda)\,\omega v^T Lv, \end{aligned}$$

and further by adding the two equations,

$$\omega\,(1+\lambda)\, v^T Av = (1-\lambda)\,(2-\omega)\, v^T Dv.$$

As with A also D is positive definite there holds $v^T Av > 0$, $v^T Dv > 0$. Consequently (observing $0 < \omega < 2$), $\lambda \ne \pm 1$, and it follows that

$$\mu := \frac{1+\lambda}{1-\lambda} = \frac{2-\omega}{\omega}\,\frac{v^T Dv}{v^T Av} > 0.$$

Resolving this for λ, we finally obtain the estimate

$$|\lambda| = \left|\frac{\mu - 1}{\mu + 1}\right| < 1, \tag{3.2.29}$$

what was to be shown.

ii) To derive the quantitative estimate (3.2.28), we rewrite (3.2.29) in the form

$$|\lambda| = \left|\frac{\mu - 1}{\mu + 1}\right| = \left|\frac{1 - 1/\mu}{1 + 1/\mu}\right| \le 1 - \frac{2}{\mu} + \frac{2}{\mu(\mu + 1)},$$

where

$$\mu = \frac{v^T D v}{v^T A v} \le \frac{\max_{\|y\|_2=1} y^T D y}{\min_{\|y\|_2=1} y^T A y} \le \frac{\lambda_{\max}(D)}{\lambda_{\min}(A)}.$$

This completes the proof. Q.E.D.

Remark 3.3: The estimate (3.2.28) for the convergence rate of the Gauß-Seidel method in the case of a symmetric, positive definite matrix A has an analogue for the Jacobi method,

$$\operatorname{spr}(J) \le 1 - \frac{1}{\mu}, \tag{3.2.30}$$

where μ is defined as in (3.2.28). This is easily seen by considering any eigenvalue $\lambda \in \sigma(J)$ with corresponding normalized eigenvector v, $\|v\|_2 = 1$, satisfying

$$\lambda D v = D v - A v.$$

Multiplying by v and observing that A as well as D are positive definite, then yields

$$\lambda = 1 - \frac{v^T A v}{v^T D v} \le 1 - \frac{1}{\mu}.$$

Comparing this estimate with (3.2.28) and observing that

$$\operatorname{spr}(J)^2 = (1 - \mu^{-1})^2 \approx 1 - 2\mu^{-1} \approx \operatorname{spr}(H_1),$$

for $\mu \gg 1$, indicates that the Gauß-Seidel method may be almost twice as fast as the Jacobi method. That this is actually the case will be seen below for a certain class of matrices.

Definition 3.2: *A matrix $A \in \mathbb{R}^{n\times n}$ with the usual additive splitting $A = L + D + R$ is called "consistently ordered" if the eigenvalues of the matrices*

$$J(\alpha) = -D^{-1}\{\alpha L + \alpha^{-1} R\}, \quad \alpha \in \mathbb{C},$$

are independent of the parameter α, i. e., equal to the eigenvalues of the matrix $J = J(1)$.

The importance of this property lies in the fact that in this case there are explicit relations between the eigenvalues of J and those of H_ω.

Example 3.2: Though the condition of "consistent ordering" appears rather strange and restrictive, it is satisfied for a large class of matrices. Consider the model matrix in Subsection 0.4.2 of Chapter 0. Depending on the numbering of the mesh points matrices with different block structures are encountered.

i) If the mesh points are numbered in a checker-board manner a block-tridiagonal matrix

$$A = \begin{bmatrix} D_1 & A_{12} & & \\ A_{21} & D_2 & \ddots & \\ & \ddots & \ddots & A_{r-1,r} \\ & & A_{r,r-1} & D_r \end{bmatrix},$$

occurs where the D_i are diagonal and regular. Such a matrix is consistently ordered, which is seen by applying a suitable similarity transformation,

$$T = \begin{bmatrix} I & & & \\ & \alpha I & & \\ & & \ddots & \\ & & & \alpha^{r-1} I \end{bmatrix}, \qquad \alpha D^{-1}L + \alpha^{-1}D^{-1}R \;=\; T(D^{-1}L + D^{-1}R)T^{-1}.$$

and observing that similar matrices have the same eigenvalues.

ii) If the mesh points are numbered in a row-wise manner a block-tridiagonal matrix

$$A = \begin{bmatrix} A_1 & D_{12} & & \\ D_{21} & A_2 & \ddots & \\ & \ddots & \ddots & D_{r-1,r} \\ & & D_{r,r-1} & A_r \end{bmatrix},$$

occurs where the A_i are tridiagonal and the D_{ij} diagonal. Such a matrix is consistently ordered, which is seen by first applying the same similarity transformation as above,

$$TAT^{-1} = \begin{bmatrix} A_1 & \alpha^{-1}D_{12} & & \\ \alpha D_{21} & A_2 & \ddots & \\ & \ddots & \ddots & \alpha^{-1}D_{r-1,r} \\ & & \alpha D_{r,r-1} & A_r \end{bmatrix},$$

and then a similarity transformation with the diagonal-block matrix

$$S = \operatorname{diag}\{S_1, \dots, S_m\},$$

where $S_i = \operatorname{diag}\{1, \alpha, \alpha^2, \dots, \alpha^{r-1}\}$, $i = 1, \dots, m$, resulting in

$$STAT^{-1}S^{-1} = \begin{bmatrix} S_1A_1S_1^{-1} & \alpha^{-1}D_{12} & & \\ \alpha D_{21} & S_2A_2S_2^{-1} & \ddots & \\ & \ddots & \ddots & \alpha^{-1}D_{r-1,r} \\ & & \alpha D_{r,r-1} & S_rA_rS_r^{-1} \end{bmatrix}.$$

Here, it has been used that the blocks D_{ij} are diagonal. Since the main-diagonal blocks are tri-diagonal, they split like $A_i = D_i + L_i + R_i$ and there holds

$$S_iA_iS_i^{-1} = D_i + \alpha L + \alpha^{-1}R.$$

This implies that the matrix A is consistently ordered.

Theorem 3.5 (Optimal SOR method): *Let the matrix $A \in \mathbb{R}^{n\times n}$ be consistently ordered and $0 \le \omega \le 2$. Then, the eigenvalues $\mu \in \sigma(J)$ and $\lambda \in \sigma(H_\omega)$ are related through the identity*

$$\lambda^{1/2}\omega\mu = \lambda + \omega - 1. \tag{3.2.31}$$

Proof. Let $\lambda,\, \mu \in \mathbb{C}$ two numbers, which satisfy equation (3.2.31). If $0 \ne \lambda \in \sigma(H_\omega)$ the relation $H_\omega v = \lambda v$ is equivalent to

$$\big((1-\omega)I - \omega D^{-1}R\big)\, v = \lambda(I + \omega D^{-1}L)v$$

and

$$(\lambda + \omega - 1)v = -\lambda^{1/2}\omega\left(\lambda^{1/2}D^{-1}L + \lambda^{-1/2}D^{-1}R\right) v = \lambda^{1/2}\omega J(\lambda^{1/2})\, v.$$

Thus, v is eigenvector of $J(\lambda^{1/2})$ corresponding to the eigenvalue

$$\mu = \frac{\lambda + \omega - 1}{\lambda^{1/2}\omega}.$$

Then, by the assumption on A also $\mu \in \sigma(J)$. In turn, for $\mu \in \sigma(J)$, by the same relation we see that $\lambda \in \sigma(H_\omega)$. Q.E.D.

As direct consequence of the above result, we see that for consistently ordered matrices the Gauß-Seidel matrix (case $\omega = 1$) either has spectral radius $\mathrm{spr(H_1)} = 0$ or there holds

$$\mathrm{spr(H_1)} = \mathrm{spr(J)}^2. \tag{3.2.32}$$

In case $\mathrm{spr(J)} < 1$ the Jacobi method converges. For reducing the error by the factor 10^{-1} the Gauß-Seidel method only needs half as many iterations than the Jacobi method and is therefore to be preferred. However, this does not necessarily hold in general since one can construct examples for which one or the other method converges or diverges.

For consistently ordered matrices from the identity (3.2.31), we can derive a formula for the "optimal" relaxation parameter ω_{opt} with $\mathrm{spr}(H_{\omega_{\mathrm{opt}}}) \le \mathrm{spr}(H_\omega)$, $\omega \in (0,2)$. If there holds $\rho := \mathrm{spr}(J) < 1$, then:

$$\mathrm{spr}(H_\omega) = \begin{cases} \omega - 1 & , \quad \omega_{\mathrm{opt}} \le \omega \\ \frac{1}{4}\big(\rho\,\omega + \sqrt{\rho^2\omega^2 - 4(\omega-1)}\,\big)^2 & , \quad \omega \le \omega_{\mathrm{opt}}\,. \end{cases}$$

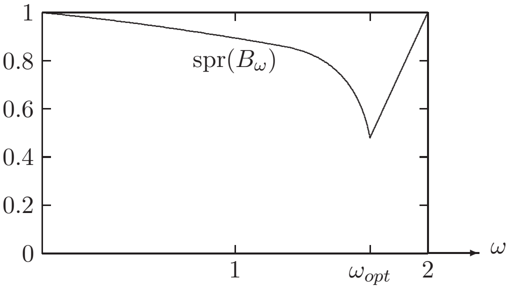


Figure 3.1: *Spectral radius of the SOR matrix H_ω as function of ω*

.

Then, there holds

$$\omega_{\mathrm{opt}} = \frac{2}{1+\sqrt{1-\rho^2}}, \qquad \mathrm{spr}(H_{\omega_{\mathrm{opt}}}) = \omega_{\mathrm{opt}} - 1 = \frac{1-\sqrt{1-\rho^2}}{1+\sqrt{1-\rho^2}} < 1. \tag{3.2.33}$$

In general the exact value for $\mathrm{spr}(J)$ is not known. Since the left-sided derivative of the function $f(\omega) = \mathrm{spr}(H_\omega)$ for $\omega \to \omega_{\mathrm{opt}}$ is singular, in estimating ω_{opt} it is better to take a value slightly larger than the exact one. Using inclusion theorems for eigenvalues or simply the bound $\rho \le \|J\|_\infty$ one obtains estimates $\bar\rho \ge \rho$. In case $\bar\rho < 1$ this yields an upper bound $\bar\omega \ge \omega_{\mathrm{opt}}$

$$\bar\omega := \frac{2}{1+\sqrt{1-\bar\rho^2}} \ge \frac{2}{1+\sqrt{1-\rho^2}} = \omega_{\mathrm{opt}}$$

for which

$$\mathrm{spr}(H_{\bar\omega}) = \bar\omega - 1 = \frac{1-\sqrt{1-\bar\rho^2}}{1+\sqrt{1-\bar\rho^2}} < 1. \tag{3.2.34}$$

However, this consideration requires the formula (3.2.33) to hold true.

Example 3.3: To illustrate the possible improvement of convergence by optimal overrelaxation, we note that

$$\mathrm{spr}(H_1) = \mathrm{spr}\,(J)^2 = \begin{cases} 0.81 \\ 0.99 \end{cases} \quad \Rightarrow \quad \mathrm{spr}(H_{\omega_{\mathrm{opt}}}) = \begin{cases} 0.39 \\ 0.8 \end{cases}$$

This will be further discussed for the model matrix in Section 3.4, below.

3.2.2 Chebyshev acceleration

In the following, we discuss another method of convergence acceleration, termed "Chebyshev acceleration", which can be used in the case of a *symmetric* coefficient matrices A, for fixed-point iterations of the form

$$x^t = Bx^{t-1} + c, \quad t = 1, 2, \dots , \tag{3.2.35}$$

with *diagonalizable* iteration matrix B. First, we describe the general principle of this approach and then apply it to a symmetrized version of the SOR method. Suppose that the above fixed-point iteration converges to the solution $x \in \mathbb{R}^n$ of the linear system

$$Ax = b \quad \Leftrightarrow \quad x = Bx + c, \tag{3.2.36}$$

i. e., that $\mathrm{spr(B)} < 1$. The idea of Chebyshev acceleration is to construct linear combinations

$$y^t := \sum_{s=0}^{t} \gamma_s^t x^s, \quad t \geq 1, \tag{3.2.37}$$

with certain coefficients γ_s^t, such that the new sequence $(y^t)_{t\geq 0}$ converges faster to the fixed point x than the original sequence $(x^t)_{t\geq 0}$. Once the fixed-point has been reached, i. e., $x^t \approx x$, the new iterates should also be close to x. This imposes the consistency condition

$$\sum_{s=0}^{t} \gamma_s^t = 1. \tag{3.2.38}$$

Then, the corresponding error has the form

$$y^t - x = \sum_{s=0}^{t} \gamma_s^t (x^s - x) = \sum_{s=0}^{t} \gamma_s^t B^s (x^0 - x) = p_t(B)(x^0 - x), \tag{3.2.39}$$

with the polynomial $p_t \in P_t$ of degree t given by

$$p_t(z) = \sum_{s=0}^{t} \gamma_s^t z^s, \quad p_t(1) = 1. \tag{3.2.40}$$

This iteration may be viewed as one governed by a sequence of "iteration matrices" $p_t(B)$, $t = 1, 2, \dots$, and therefore, we may try to characterize its convergence by the spectral radius $\mathrm{spr(p_t(B))}$ as in the standard situation of a "stationary fixed-point iteration (i. e., one with a fixed iteration matrix). This requires us to relate the eigenvalues of $p_t(B)$ to those of B,

$$\lambda(p_t(B)) = p_t(\lambda(B)). \tag{3.2.41}$$

This leads us to consider the following optimization problem

$$\mathrm{spr}(\mathrm{p_t(B)}) = \min_{\mathrm{p\in P_t,p(1)=1}} \max_{\mathrm{\lambda\in\sigma(B)}} |\mathrm{p}(\lambda)|. \tag{3.2.42}$$

The eigenvalues $\lambda \in \mathrm{spr(B)}$ are usually not known, but rather the bound $\mathrm{spr}(B) \le 1-\delta$ with some small $\delta > 0$ may be available. Hence, this optimization problem has to be relaxed to

$$\mathrm{spr}(\mathrm{p_t(B)}) \le \min_{\mathrm{p\in P_t,p(1)=1}} \max_{\mathrm{|x|\le 1-\delta}} |\mathrm{p_t(x)}|. \tag{3.2.43}$$

This optimization problem can be explicitly solved in the case $\sigma(B) \in \mathbb{R}$. Therefore, we make the following assumption.

Assumption 3.1: *The coefficient matrix* $A = L + D + L^T$ *is assumed to be symmetric and the iteration matrix* B *of the base iteration (3.2.35) to be similar to a symmetric matrix and, therefore, is diagonalizable with real eigenvalues,*

$$\sigma(B) \subset \mathbb{R}. \tag{3.2.44}$$

Remark 3.4: In general the iteration matrix B cannot be assumed to be symmetric and not even similar to a symmetric matrix (e. g., in the Gauß-Seidel method with $H_1 = -(D+L)^{-1}L^T$). But if this were the case (e. g., in the Richardson method with $B = I - \theta A$ or in the Jacobi method with $J = -D^{-1}(L+L^T)$) the analysis of the new sequence $(y_t)_{t\ge 0}$ may proceed as follows. Taking spectral-norms, we obtain

$$\|y^t - x\|_2 \le \|p_t(B)\|_2 \|x^0 - x\|_2. \tag{3.2.45}$$

Hence, the convergence can be improved by choosing the polynomial p_t such the the norm $\|p_t(B)\|_2$ becomes minimal,

$$\frac{\|y^t - x\|_2}{\|x^0 - x\|_2} \le \min_{p_t\in P_t, p_t(1)=1} \|p_t(B)\|_2 \ll \|B^t\|_2 \le \|B\|_2^t. \tag{3.2.46}$$

Using the representation of the spectral norm, valid for symmetric matrices,

$$\|p_t(B)\|_2 = \max_{\lambda\in\sigma(B)} |p_t(\lambda)|. \tag{3.2.47}$$

and observing $\sigma(B) \in [-1+\delta, 1-\delta]$, for same small $\delta > 0$, the optimization problem takes the form

$$\min_{p_t\in P_t, p_t(1)=1} \max_{|x|\le 1-\delta} |p_t(x)|. \tag{3.2.48}$$

The solution of the optimization problem (3.2.43) is given by the well-known Chebyshev polynomials (of the first kind), which are the orthogonal polynomials obtained by successively orthogonalizing (using the the Gram-Schmidt algorithm with exact arithmetic) the monomial basis $\{1, x, x^2, \dots, x^t\}$ with respect to the scalar product

$$(p, q) := \int_{-1}^{1} p(x)q(x) \, \frac{dx}{\sqrt{1-x^2}}, \quad p, q \in P_t,$$

defined on the function space $C[-1, 1]$. These polynomials, named $T_t \in P_t$, are usually normalized to satisfy $T_t(1) = 1$,

$$\int_{-1}^{1} T_t(x)T_s(x) \frac{dx}{\sqrt{1-x^2}} = \begin{cases} 0, & t \neq s, \\ \pi, & t = s = 0, \\ \pi/2, & t = s \neq 0. \end{cases}$$

They can be written in explicit form as (see, e. g., Stoer & Bulirsch [50] or Rannacher [1]):

$$T_t(x) = \begin{cases} (-1)^t \cosh(t \operatorname{arccosh}(-x)), & x \leq -1, \\ \cos(t \arccos(x)), & -1 \leq x \leq 1, \\ \cosh(t \operatorname{arccosh}(x)), & x \geq 1. \end{cases} \tag{3.2.49}$$

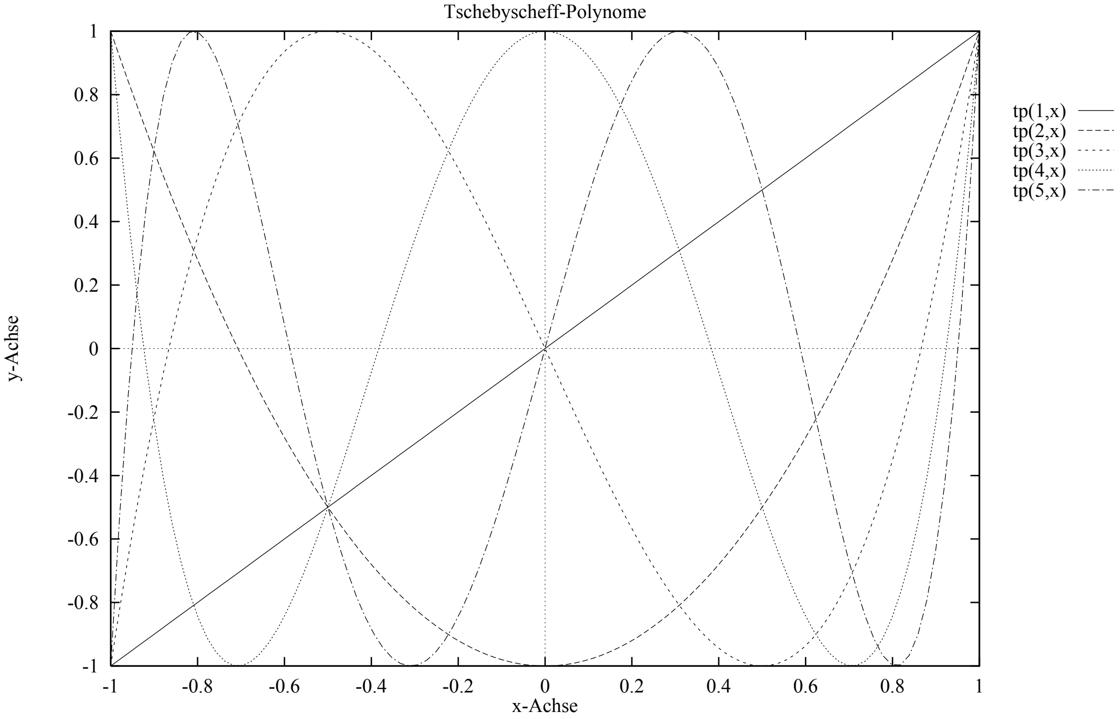


Figure 3.2: *Chebyshev polynomials* T_t, $t = 0, 1, \dots, 5$.

That the so defined functions are actually polynomials can be seen by induction. Further, there holds the three-term recurrence relation

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_{t+1}(x) = 2xT_t(x) - T_{t-1}(x), \quad t \ge 1, \tag{3.2.50}$$

which allows the numerically stable computation and evaluation of the Chebyshev polynomials. Sometimes the following alternative global representation is useful:

$$T_t(x) = \tfrac{1}{2}\big([x + \sqrt{x^2-1}\,]^t + [x - \sqrt{x^2-1}\,]^t\big), \quad x \in \mathbb{R}. \tag{3.2.51}$$

With this notation, we have the following basic result.

Theorem 3.6 (Chebyshev polynomials): *Let $[a,b] \subset \mathbb{R}$ be a non-empty interval and let $c \in \mathbb{R}$ be any point outside this interval. Then, the minimum*

$$\min_{p\in P_t, p(c)=1} \max_{x\in[a,b]} |p(x)| \tag{3.2.52}$$

is attained by the uniquely determined polynomial

$$p(x) := C_t(x) = \frac{T_t(1 + 2\frac{x-b}{b-a})}{T_t(1 + 2\frac{c-b}{b-a})}, \quad x \in [a,b]. \tag{3.2.53}$$

Furthermore, for $a < b < c$ there holds

$$\min_{p\in P_t, p(c)=1} \max_{x\in[a,b]} |p(x)| = \frac{1}{T_t(1 + 2\frac{c-b}{b-a})} = \frac{2\gamma^t}{1+\gamma^{2t}} \le 2\gamma^t, \tag{3.2.54}$$

where

$$\gamma := \frac{1 - 1/\sqrt{\kappa}}{1 + 1/\sqrt{\kappa}}, \qquad \kappa := \frac{c-a}{c-b}.$$

Proof. i) By affine transformation, which does not change the max-norm, we may restrict ourselves to the standard case $[a,b] = [-1,1]$ and $c \in \mathbb{R}\setminus[-1,1]$. Then, $C_t(x) = \tilde{C}T_t(x)$ with constant $\tilde{C} = T_t(c)^{-1}$. The Chebyshev polynomial $T_t(x) = \cos(t \arccos(x))$ attains the values ± 1 at the points $x_i = \cos(i\pi/t), i = 0, \dots, t,$ and it alternates between 1 and -1, i.e., $T_t(x_i)$ and $T_t(x_{i+1})$ have opposite signs. Furthermore, $\max_{[-1,1]} |T_t| = 1$, implying $\max_{[-1,1]} |C_t| = |\tilde{C}|$.

ii) Assume now the existence of $q \in P_t$ such that $\max_{[-1,1]} |q| < \max_{[-1,1]} |C_t| = |\tilde{C}|$ and $q(c) = 1$. Then, the polynomial $r = C_t - q$ changes sign t-times in the interval $[-1,1]$ since $\operatorname{sign} r(x_i) = \operatorname{sign} T_t(x_i), i = 0, \dots, t.$ Thus, r has at least t zeros in $[-1,1]$. Additionally, $r(c) = 0$. Hence, $r \in P_t$ has at least $t+1$ zeros; thus, $r \equiv 0$, which leads to a contradiction.

iii) By definition, there holds $|T_t(x)| \le 1, x \in [-1,1]$. This implies that

$$\max_{x\in[a,b]} |C_t(x)| = \frac{1}{T_t(1 + 2\frac{c-b}{b-a})}.$$

The assertion then follows from the explicit representation of the T_t given above and some manipulations (for details see the proof of Theorem 3.11, below). Q.E.D.

Practical use of Chebyshev acceleration

We now assume $\sigma(B) \subset (-1,1)$, i. e., convergence of the primary iteration. Moreover, we assume that a parameter $\rho \in (-1,1)$ is known such that $\sigma(B) \subset [-\rho,\rho]$. With the parameters $a = -\rho$, $b = \rho$, and $c = 1$, we use the polynomials $p_t = C_t$ given in Theorem 3.6 in defining the secondary iteration (3.2.37). This results in the "Chebyshev-accelerated" iteration scheme. This is a consistent choice since $T_t(1) = 1$.

The naive evaluation of the secondary iterates (3.2.37) would require to store the whole convergence history of the base iteration $(x^t)_{t\geq 0}$, which may not be possible for large problems. Fortunately, the three-term recurrence formula (3.2.50) for the Chebyshev polynomials carries over to the corresponding iterates $(y^t)_{t\geq 0}$, making the whole process feasible at all.

Since the T_t satisfy the three-term recurrence (3.2.50), and so do the polynomials $p = C_t$ from (3.2.53):

$$\mu_{t+1}p_{t+1}(x) = \frac{2x}{\rho}\mu_t p_t(x) - \mu_{t-1}p_{t-1}(x), \quad t \geq 1, \quad \mu_t = T_t(1/\rho), \tag{3.2.55}$$

with initial functions

$$p_0(x) \equiv 1, \quad p_1(x) = \frac{T_1(x/\rho)}{T_1(1/\rho)} = \frac{x/\rho}{1/\rho} = x,$$

i. e., $a_{0,0} = 1$ and $a_{1,0} = 0$, $a_{1,1} = 1$. We also observe the important relation

$$\mu_{t+1} = \frac{2}{\rho}\mu_t - \mu_{t-1}, \quad \mu_0 = 1, \;\; \mu_1 = 1/\rho. \tag{3.2.56}$$

which can be concluded from (3.2.55) observing that $p_t(1) = 1$. With these preparations, we can now implement the Chebyshev acceleration scheme. With the limit $x := \lim_{t\to\infty} x^t$, we obtain for the error $y^t - x = \tilde{e}^t = p_t(B)e^0$:

$$\begin{aligned}
y^{t+1} &= x + \tilde{e}^{t+1} = x + p_{t+1}(B)e^0 = x + 2\frac{\mu_t}{\rho\mu_{t+1}}Bp_t(B)e^0 - \frac{\mu_{t-1}}{\mu_{t+1}}p_{t-1}(B)e^0 \\
&= x + 2\frac{\mu_t}{\rho\mu_{t+1}}B\tilde{e}^t - \frac{\mu_{t-1}}{\mu_{t+1}}\tilde{e}^{t-1} = x + 2\frac{\mu_t}{\rho\mu_{t+1}}B(y^t - x) - \frac{\mu_{t-1}}{\mu_{t+1}}(y^{t-1} - x) \\
&= 2\frac{\mu_t}{\rho\mu_{t+1}}By^t - \frac{\mu_{t-1}}{\mu_{t+1}}y^{t-1} + \frac{1}{\mu_{t+1}}\Big(\mu_{t+1} - \frac{2}{\rho}\mu_t B + \mu_{t-1}\Big)x.
\end{aligned}$$

Now, using the fixed-point relation $x = Bx+c$ and the recurrence (3.2.56), we can remove the appearance of x in the above recurrence obtaining

$$y^{t+1} = 2\frac{\mu_t}{\rho\mu_{t+1}}By^t - \frac{\mu_{t-1}}{\mu_{t+1}}y^{t-1} + 2\frac{\mu_t}{\rho\mu_{t+1}}c, \quad y^0 = x^0, \;\; y^1 = x^1 = Bx^0 + c. \tag{3.2.57}$$

Hence, the use of Chebyshev acceleration for the primary iteration (3.2.35) consists in evaluating the three-term recurrences (3.2.56) and (3.2.57), which is of similar costs as the primary iteration (3.2.35) itself, in which the most costly step is the matrix-vector multiplication By^t.

In order to quantify the acceleration effect of this process, we write the secondary iteration in the form

$$y^t - x = \sum_{s=0}^{t} \gamma_s^t (x^s - x) = p_t(B)(x^0 - x),$$

where γ_s^t are the coefficients of the polynomial p_t. There holds

$$p_t(x) = C_t(x) = \frac{T_t(x/\rho)}{T_t(1/\rho)}.$$

By the estimate (3.2.54) of Theorem 3.6 it follows that

$$\operatorname{spr}(p_t(B)) = \max_{\lambda\in\sigma(B)} |p_t(x)| = \frac{2\gamma^t}{1+\gamma^{2t}} \le 2\gamma^t, \quad \gamma := \frac{1-1/\sqrt{\kappa}}{1+1/\sqrt{\kappa}}, \quad \kappa := \frac{1+\rho}{1-\rho}.$$

Hence, for the primary and the secondary iteration, we find the asymptotic error behavior

$$\limsup \Big(\frac{\|e^t\|}{\|e^0\|}\Big)^{1/t} = \operatorname{spr}(B) \le \rho = 1 - \delta, \tag{3.2.58}$$

$$\limsup \Big(\frac{\|\tilde{e}^t\|}{\|e^0\|}\Big)^{1/t} \le \frac{1-1/\sqrt{\kappa}}{1+1/\sqrt{\kappa}} \le 1 - c'\sqrt{\delta}, \tag{3.2.59}$$

i. e., in the case $0 < \delta \ll 1$ by Chebyshev acceleration a significant improvement can be achieved for the convergence speed.

Application for accelerating the SOR method

We want to apply the concept of Chebyshev acceleration to the SOR method with the iteration matrix (recalling that A is symmetric)

$$H_\omega = (D+\omega L)^{-1}\big((1-\omega)\,D - \omega L^T\big), \quad \omega \in (0,2).$$

However, it is not obvious whether this matrix is diagonalizable. Therefore, one introduces a symmetrized version of the SOR method, which is termed “SSOR method”,

$$\begin{aligned}(D+\omega L)y^t &= [(1{-}\omega)D - \omega L^T]x^{t-1} + b,\\ (D+\omega L^T)x^t &= [(1{-}\omega)D - \omega L]y^t + b,\end{aligned}$$

or equivalently,

$$x^t = (D+\omega L^T)^{-1}[(1-\omega)\,D-\omega L](D+\omega L)^{-1}[(1-\omega)D-\omega L^T]x^{t-1}+b\big)+b\big), \quad (3.2.60)$$

with the iteration matrix

$$H_\omega^{\rm SSOR} := (D+\omega L^T)^{-1}\big((1-\omega)\,D-\omega L\big)(D+\omega L)^{-1}\big((1-\omega)D-\omega L^T\big).$$

The SSOR-iteration matrix is similar to a symmetric matrix, which is seen from the relation

$$\begin{aligned}(D+\omega L^T)H_\omega^{\rm SSOR}(D+\omega L^T)^{-1} &= [(1-\omega)\,D-\omega L](D+\omega L)^{-1}[(1-\omega)D-\omega L^T](D+\omega L^T)^{-1}\\ &= [(1-\omega)\,D-\omega L](D+\omega L)^{-1}(D+\omega L^T)^{-1}[(1-\omega)D-\omega L^T].\end{aligned}$$

The optimal relaxation parameter of the SSOR method is generally different from that of the SOR method.

Remark 3.5: In one step of the SSOR method the SOR loop is successively applied twice, once in the standard "forward" manner based on the splitting $A=(L+D)+L^T$ and then in "backward" form based on $A=L+(D+L^T)$. Hence, it is twice as expensive compared to the standard SOR method. But this higher cost is generally not compensated by faster convergence. Hence, the SSOR method is attractive mainly in connection with the Chebyshev acceleration as described above and not so much as a stand-alone method.

3.3 Descent methods

In the following, we consider a class of iterative methods, which are especially designed for linear systems with symmetric and positive definite coefficient matrices A, but can also be extended to more general situations. In this section, we use the abbreviated notation $(\cdot,\cdot):=(\cdot,\cdot)_2$ and $\|\cdot\|:=\|\cdot\|_2$ for the Euclidian scalar product and norm.

Let $A\in\mathbb{R}^{n\times n}$ be a symmetric positive definite (and hence regular) matrix,

$$(Ax,y)=(x,Ay), \quad x,\,y\in\mathbb{R}^n, \qquad (Ax,x)>0, \quad x\in\mathbb{R}^{n\times n}\setminus\{0\}. \quad (3.3.61)$$

This matrix generates the so-called "A-scalar product" and the corresponding "A-norm",

$$(x,y)_A := (Ax,y), \quad \|x\|_A := (Ax,x)^{1/2}, \quad x,\,y\in\mathbb{R}^n. \quad (3.3.62)$$

Accordingly, vectors with the property $(x,y)_A=0$ are called "A-orthogonal". The positive definite matrix A has important properties. Its eigenvalues are real and positive $0<\lambda:=\lambda_1\le\ldots\le\lambda_n=:\Lambda$ and there exists an ONB of eigenvectors $\{w_1,\ldots,w_n\}$. For its spectral radius and spectral condition number, there holds

$$\mathrm{spr}(A)=\Lambda\,, \quad \mathrm{cond_2(A)}=\frac{\Lambda}{\lambda}. \quad (3.3.63)$$

The basis for the descent methods discussed below is provided by the following theorem, which characterizes the solution of the linear system $Ax = b$ as the minimum of a quadratic functional.

Theorem 3.7 (Minimization property): *The matrix A be symmetric positive definite. The uniquely determined solution of the linear system $Ax = b$ is characterized by the property*

$$Q(x) < Q(y) \quad \forall\, y \in \mathbb{R}^n \setminus \{x\}, \qquad Q(y) := \tfrac{1}{2}(Ay, y)_2 - (b, y)_2. \tag{3.3.64}$$

Proof. Let $Ax = b$. Then, in view of the definiteness of A for $y \neq x$ there holds

$$\begin{aligned} Q(y) - Q(x) &= \tfrac{1}{2}\{\, (Ay, y) - 2(b, y) - (Ax, x) + 2(b, x) \,\} \\ &= \tfrac{1}{2}\{\, (Ay, y) - 2(Ax, y) + (Ax, x) \,\} = \tfrac{1}{2}\,(A[x - y], x - y) > 0. \end{aligned}$$

In turn, if $Q(x) < Q(y)$, for $x \neq y$, i. e., if x is a strict minimum of Q on $\mathbb{R}^n$, there must hold $\mathrm{grad\,Q(x) = 0}$. This means that (observe $a_{jk} = a_{kj}$)

$$\frac{\partial Q}{\partial x_i}(x) = \frac{1}{2}\frac{\partial}{\partial x_i}\sum_{j,k=1}^{n} a_{jk}x_jx_k - \frac{\partial}{\partial x_i}\sum_{k=1}^{n} b_kx_k = \sum_{k=1}^{n} a_{ik}x_k - b_i = 0, \quad i = 1, \dots, n,$$

i. e., $Ax = b$. Q.E.D.

We note that the gradient of Q in a point $y \in \mathbb{R}^n$ is given by

$$\mathrm{grad}\, Q(y) = \tfrac{1}{2}\,(A + A^T)y - b = Ay - b. \tag{3.3.65}$$

This coincides with the "defect" of the point y with respect to the equation $Ax = b$ (negative "residual" $b - Ay$). The so-called "descent methods", starting from some initial point $x^{(0)} \in \mathbb{R}^n$, determine a sequence of iterates x^t, $t \geq 1$, by the prescription

$$x^{t+1} = x^t + \alpha_t r^t, \quad Q(x^{t+1}) = \min_{\alpha \in \mathbb{R}} Q(x^t + \alpha r^t). \tag{3.3.66}$$

Here, the "descent directions" r^t are a priori determined or adaptively chosen in the course of the iteration. The prescription for choosing the "step length" α_t is called "line search". In view of

$$\frac{d}{d\alpha}\, Q(x^t + \alpha r^t) = \mathrm{grad}Q(x^t + \alpha r^t) \cdot r^t = (Ax^t - b, r^t) + \alpha(Ar^t, r^t),$$

we obtain the formula

$$\alpha_t = -\frac{(g^t, r^t)}{(Ar^t, r^t)}, \quad g^t \,:=\, Ax^t - b = \mathrm{gradQ(x^t)}.$$

Definition 3.3: *The general descent method, starting from some initial point $x^0 \in \mathbb{R}^n$, determines a sequence of iterates $x^t \in \mathbb{R}^n$, $t \geq 1$, by the prescription* *i) gradient* $g^t = Ax^t - b$,

ii) descent direction r^t,

iii) step length $\alpha_t = -\dfrac{(g^t, r^t)}{(Ar^t, r^t)}$,

iv) descent step $x^{t+1} = x^t + \alpha_t r^t$.

Each descent step as described in the above definition requires two matrix-vector multiplications. By rewriting the algorithm in a slightly different way, one can save one of these multiplications at the price of additionally storing the vector Ar^t.

General descent algorithm:

Starting values: $x^0 \in \mathbb{R}^n, \quad g^0 := Ax^0 - b.$

Iterate for $t \geq 0$: descent direction r^t

$$\alpha_t = -\frac{(g^t, r^t)}{(Ar^t, r^t)}, \quad x^{t+1} = x^t + \alpha_t r^t, \quad g^{t+1} = g^t + \alpha_t Ar^t.$$

Using the notation $\|y\|_B := (By, y)^{1/2}$ there holds

$$2Q(y) = \|Ay - b\|_{A^{-1}}^2 - \|b\|_{A^{-1}}^2 = \|y - x\|_A^2 - \|x\|_A^2, \tag{3.3.67}$$

i. e., the minimization of the functional $Q(\cdot)$ is equivalent to the minimization of the Defect norm $\|Ay - b\|_{A^{-1}}$ or the error norm $\|y - x\|_A$.

3.3.1 Gradient method

The various descent methods essentially differ by the choice of the descent directions r^t. One of the simplest a priori strategies uses in a cyclic way the Cartesian coordinate direction $\{e^1, \dots, e^n\}$. The resulting method is termed “coordinate relaxation” and is sometimes used in the context of *nonlinear* systems. For solving linear systems it is much too slow as it is in a certain sense equivalent to the Gauß-Seidel method (exercise). A more natural choice are the directions of steepest descent of $Q(\cdot)$ in the points x^t:

$$r^t = -\mathrm{gradQ(x^t)} = -\mathrm{g^t}. \tag{3.3.68}$$

Definition 3.4: *The “gradient method” determines a sequence of iterates* $x^t \in \mathbb{R}^n$, $t \geq 0$, *by the prescription*

Starting values: $x^0 \in \mathbb{R}^n, \quad g^0 := Ax^0 - b.$

Iterate for $t \geq 0$*:* $$\alpha_t = \frac{\|g^t\|^2}{(Ag^t, g^t)}, \quad x^{t+1} = x^t - \alpha_t g^t, \quad g^{t+1} = g^t - \alpha_t Ag^t.$$

In case that $(Ag^t, g^t) = 0$ for some $t \geq 0$ there must hold $g^t = 0$, i. e., the iteration can only terminate with $Ax^t = b$.

Theorem 3.8 (Gradient method): *For a symmetric positive definite matrix $A \in \mathbb{R}^{n\times n}$ the gradient method converges for any starting point $x^0 \in \mathbb{R}^n$ to the solution of the linear system $Ax = b$.*

Proof. We introduce the "error functional"

$$E(y) := \|y-x\|_A^2 = (y-x, A[y-x]), \quad y \in \mathbb{R}^n,$$

and for abbreviation set $e^t := x^t - x$. With this notation there holds

$$\begin{aligned}\frac{E(x^t)-E(x^{t+1})}{E(x^t)} &= \frac{(e^t, Ae^t) - (e^{t+1}, Ae^{t+1})}{(e^t, Ae^t)} \\ &= \frac{(e^t, Ae^t) - (e^t - \alpha_t g^t, A[e^t - \alpha_t g^t])}{(e^t, Ae^t)} \\ &= \frac{2\alpha_t (e^t, Ag^t) - \alpha_t^2 (g^t, Ag^t)}{(e^t, Ae^t)}\end{aligned}$$

and consequently, because of $Ae^t = Ax^t - Ax = Ax^t - b = g^t$,

$$\frac{E(x^t)-E(x^{t+1})}{E(x^t)} = \frac{2\alpha_t \|g^t\|^2 - \alpha_t^2 (g^t, Ag^t)}{(g^t, A^{-1}g^t)} = \frac{\|g^t\|^4}{(g^t, Ag^t)(g^t, A^{-1}g^t)}.$$

For the positive definite matrix A there holds

$$\lambda\|y\|^2 \le (y, Ay) \le \Lambda\|y\|^2, \quad \Lambda^{-1}\|y\|^2 \le (y, A^{-1}y) \le \lambda^{-1}\|y\|^2,$$

with $\lambda = \lambda_{\min}(A)$ and $\Lambda = \lambda_{\max}(A)$. In the case $x^t \neq x$, i. e., $E(x^t) \neq 0$ and $g^t \neq 0$, we conclude that

$$\frac{\|g^t\|^4}{(g^t, Ag^t)(g^t, A^{-1}g^t)} \ge \frac{\|g^t\|^4}{\Lambda\|g^t\|^2\,\lambda^{-1}\|g^t\|^2} = \frac{\lambda}{\Lambda},$$

and, consequently,

$$E(x^{t+1}) \le \{\, 1 - \kappa^{-1} \,\}\, E(x^t), \quad \kappa := \mathrm{cond}_{\mathrm{nat}}(\mathrm{A}).$$

Since $0 < 1 - 1/\kappa < 1$ for any $x^0 \in \mathbb{R}^n$ the error functional $E(x^t) \to 0$ $(t \to \infty)$, i. e., $x^t \to x$ $(t \to \infty)$. Q.E.D.

For the quantitative estimation of the speed of convergence of the gradient method, we need the following result of Kantorovich[5] .

Lemma 3.4 (Lemma of Kantorovich): *For a symmetric and positive definite matrix*

[5] Leonid Vitalyevich Kantorovich (1912–1986): Russian Mathematician; Prof. at the U of Leningrad (1934–1960), at the Academy of Sciences (1961–1971) and at the U Moscow (1971-1976); fundamental contributions to linear optimization in Economy, to Functional Analysis and to Numerics (Theorem of Newton-Kantorovich).

$A\in\mathbb{R}^n$ with smallest and largest eigenvalues λ and Λ, respectively, there holds

$$4\,\frac{\lambda\Lambda}{(\lambda+\Lambda)^2}\le\frac{\|y\|^4}{(y,Ay)(y,A^{-1}y)}\,,\quad y\in\mathbb{R}^n. \tag{3.3.69}$$

Proof. Let $\lambda=\lambda_1\le\ldots\le\lambda_n=\Lambda$ be the eigenvalues of A and $\{w_1,\ldots,w_n\}$ a corresponding ONB of eigenvectors. An arbitrary vector $y\in\mathbb{R}^n$ admits an expansion $y=\sum_{i=1}^n y_iw_i$ with the coefficients $y_i=(y,w_i)$. Then,

$$\frac{\|y\|^4}{(y,Ay)(y,A^{-1}y)}=\frac{(\sum_{i=1}^n y_i^2)^2}{(\sum_{i=1}^n\lambda_iy_i^2)\,(\sum_{i=1}^n\lambda_i^{-1}y_i^2)}=\frac{1}{(\sum_{i=1}^n\lambda_i\zeta_i)\,(\sum_{i=1}^n\lambda_i^{-1}\zeta_i)}=\frac{\varphi(\zeta)}{\psi(\zeta)},$$

with the notation

$$\zeta=(\zeta_i)_{i=1,\ldots,n}\,,\quad \zeta_i=y_i^2(\sum_{i=1}^n y_i^2)^{-1},$$

$$\psi(\zeta)=\sum_{i=1}^n\lambda_i^{-1}\zeta_i\,,\quad \varphi(\zeta)=(\sum_{i=1}^n\lambda_i\zeta_i)^{-1}.$$

Since the function $f(\lambda)=\lambda^{-1}$ is convex it follows from $0\le\zeta_i\le1$ and $\sum_{i=1}^n\zeta_i=1$ that

$$\sum_{i=1}^n\lambda_i^{-1}\zeta_i\ge(\sum_{i=1}^n\lambda_i\zeta_i)^{-1}.$$

We set $g(\lambda):=(\lambda_1+\lambda_n-\lambda)/(\lambda_1\lambda_n)$.

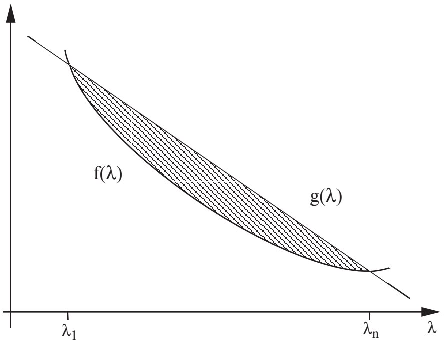


Figure 3.3: *Sketch to the proof of the Lemma of Kantorovich.*

Obviously, the graph of $\varphi(\zeta)$ lies, for all arguments ζ on the curve $f(\lambda)$, and that of $\psi(\zeta)$ between the curves $f(\lambda)$ and $g(\lambda)$ (shaded area). This implies that

$$\frac{\varphi(\zeta)}{\psi(\zeta)}\ge\min_{\lambda_1\le\lambda\le\lambda_n}\frac{f(\lambda)}{g(\lambda)}=\frac{f([\lambda_1+\lambda_n]/2)}{g([\lambda_1+\lambda_n]/2)}=\frac{4\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2},$$

which concludes the proof. Q.E.D.

Theorem 3.9 (Error estimate for gradient method): *Let the matrix* $A\in\mathbb{R}^{n\times n}$ *be symmetric positive definite. Then, for the gradient method the following error estimate holds:*

$$\|x^t-x\|_A\le\Big(\frac{1-1/\kappa}{1+1/\kappa}\Big)^t\,\|x^0-x\|_A\,,\quad t\in\mathbb{N},\tag{3.3.70}$$

with the spectral condition number $\kappa=\mathrm{cond}_2(A)=\Lambda/\lambda$ *of* A*. For reducing the initial error by a factor* TOL *the following number of iterations is required:*

$$t(\mathrm{TOL})\approx\tfrac12\kappa\,\ln(1/\mathrm{TOL}).\tag{3.3.71}$$

Proof. i) In the proof of Theorem 3.8 the following error identity was shown:

$$E(x^{t+1})=\Big\{1-\frac{\|g^t\|^4}{(g^t,Ag^t)\,(g^t,A^{-1}g^t)}\Big\}\,E(x^t).$$

This together with the inequality (3.3.69) in the Lemma of Kantorovich yields

$$E(x^{t+1})\le\Big\{1-4\frac{\lambda\Lambda}{(\lambda+\Lambda)^2}\Big\}\,E(x^t)=\Big(\frac{\lambda-\Lambda}{\lambda+\Lambda}\Big)^2\,E(x^t).$$

From this, we conclude by successive use of the recurrence that

$$\|x^t-x\|_A^2\le\Big(\frac{\lambda-\Lambda}{\lambda+\Lambda}\Big)^{2t}\,\|x^0-x\|_A^2\,,\quad t\in\mathbb{N},$$

Which proves the asserted estimate (3.3.70).

ii) To prove (3.3.71), we take the logarithm on both sides of the relations

$$\Big(\frac{1-1/\kappa}{1+1/\kappa}\Big)^{t(\mathrm{TOL})}=\Big(\frac{\kappa-1}{\kappa+1}\Big)^{t(\mathrm{TOL})}<\mathrm{TOL},\qquad\Big(\frac{\kappa+1}{\kappa-1}\Big)^{t(\mathrm{TOL})}>\frac{1}{\mathrm{TOL}},$$

obtaining

$$t(\mathrm{TOL})>\ln\Big(\frac{1}{\mathrm{TOL}}\Big)\ln\Big(\frac{\kappa+1}{\kappa-1}\Big)^{-1}.$$

Since

$$\ln\frac{x+1}{x-1}=2\Big\{\frac1x+\frac13\frac1{x^3}+\frac15\frac1{x^5}+\dots\Big\}\ge\frac2x$$

this is satisfied for $t(\mathrm{TOL})\ge\frac12\kappa\,\ln(1/\mathrm{TOL})$. Q.E.D.

The relation

$$(g^{t+1},g^t)=(g^{(t)}-\alpha_tAg^t,g^t)=\|g^t\|^2-\alpha_t(Ag^t,g^t)=0\tag{3.3.72}$$

shows that the descent directions $r^t=-g^t$ used in the gradient method in consecutive steps are orthogonal to each other, while g^{t+2} may be far away form being orthogonal to g^t. This may lead to strong oscillations in the convergence behavior of the gradient method especially for matrices A with large condition number, i. e., $\lambda\ll\Lambda$. In the twodimensional case this effect can be illustrated by the contour lines of the functional $Q(\cdot)$, which are eccentric ellipses, leading to a zickzack path of the iteration (see Fig. 3.3.1).

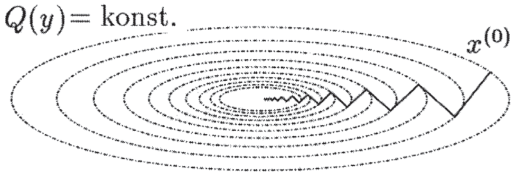


Figure 3.4: *Oscillatory convergence of the gradient method*

3.3.2 Conjugate gradient method (CG method)

The gradient method utilizes the particular structure of the functional $Q(\cdot)$, i. e., the distribution of the eigenvalues of the matrix A, only locally from one iterate x^t to the next one, x^{t+1}. It seems more appropriate to utilize the already obtained information about the global structure of $Q(\cdot)$ in determining the descent directions, e. g., by choosing the descent directions mutually orthogonal. This is the basic idea of the "conjugate gradient method" ("CG method") of Hestenes[6] and Stiefel[7] (1952), which successively generates a sequence of descent directions d^t which are mutually "A-orthogonal", i. e., orthogonal with respect to the scalar product $(\cdot,\cdot)_A$.

For developing the CG method, we start from the ansatz

$$B_t := \operatorname{span}\{d^0, \cdots, d^{t-1}\} \tag{3.3.73}$$

with a set of linearly independent vectors d^i and seek to determine the iterates in the form

$$x^t = x^0 + \sum_{i=0}^{t-1} \alpha_i d^i \;\in\; x^0 + B_t, \tag{3.3.74}$$

such that

$$Q(x^t) = \min_{y\in x^0+B_t} Q(y) \quad\Leftrightarrow\quad \|Ax^t - b\|_{A^{-1}} = \min_{y\in x^0+B_t} \|Ay-b\|_{A^{-1}}. \tag{3.3.75}$$

Setting the derivatives of $Q(\cdot)$ with respect to the α_i to zero, we see that this is equivalent

[6] Magnus R. Hestenes (1906–1991): US-American mathematician; worked at the National Bureau of Standards (NBS) and the University of California at Los Angeles (UCLA); contributions to optimization and control theory and to numerical linear algebra.

[7] Eduard Stiefel (1909–1978): Swiss mathematician; since 1943 Prof. for Applied Mathematics at the ETH Zurich; important contributions to Topology, Groupe Theory, Numerical Linear Algebra (CG method), Approximation Theory and Celestrian Mechanics.

to solving the so-called "Galerkin[8] equations":

$$(Ax^t - b, d^j) = 0\,, \quad j = 0, \dots, t-1, \tag{3.3.76}$$

or in compact form: $Ax^t - b = g^t \perp B_t$. Inserting the above ansatz for x^t into this orthogonality condition, we obtain a regular linear system for the coefficients α_i, $i = 0, \dots, t{-}1$,

$$\sum_{i=1}^{n} \alpha_i (Ad^i, d^j) = (b, d^j) - (Ax^0, d^j), \quad j = 0, \dots, t-1. \tag{3.3.77}$$

Remark 3.6: We note that (3.3.76) does not depend on the symmetry of the matrix A. Starting from this relation one may construct CG-like methods for linear systems with asymmetric and even indefinite coefficient matrices. Such methods are generally termed "projection methods". Methods of this type will be discussed in more detail below.

Recall that the Galerkin equations (3.3.76) are equivalent to minimizing the defect norm $\|Ax^t - b\|_{A^{-1}}$ or the error norm $\|x^t - x\|_A$ on $x^0 + B_t$. Natural choices for the spaces B_t are the so-called Krylov[9] spaces

$$B_t = K_t(d^0; A) := \operatorname{span}\{d^0, Ad^0, \dots, A^{t-1}d^0\}, \tag{3.3.78}$$

with some vector d^0, e. g., the (negative) initial defect $d^0 = b - Ax^0$ of an arbitrary vector x^0. This is motivated by the observation that from $A^t d^0 \in K_t(d^0; A)$, we necessarily obtain

$$-g^t = b - Ax^t = d^0 + A(x^0 - x^t) \in d^0 + AK_t(d^0; A) \in K_t(d^0; A).$$

Because $g^t \perp K_t(d^0; A)$, this implies $g^t = 0$ by construction.

Now the CG method constructs a sequence of descent directions, which form an A-orthogonal basis of the Krylov spaces $K_t(d^0; A)$. We proceed in an inductive way: Starting from an arbitrary point x^0 with (negative) defect $d^0 = b - Ax^0$ let iterates x^i and corresponding descent directions $d^i (i = 0, ..., t-1)$ already been determined such that $\{d^0, ..., d^{t-1}\}$ is an A-orthogonal basis of $K_t(d^0; A)$. For the construction of the next descent direction $d^t \in K_{t+1}(d^0; A)$ with the property $d^t \perp_A K_t(d^0; A)$ we make the ansatz

$$d^t = -g^t + \sum_{j=0}^{t-1} \beta_j^{t-1} d^j \;\in\; K_{t+1}(d^0; A). \tag{3.3.79}$$

[8] Boris Grigorievich Galerkin (1871–1945): Russian civil engineer and mathematician; Prof. in St. Petersburg; contributions to Structural Mechanics especially Plate Bending Theory.

[9] Aleksei Nikolaevich Krylov (1863–1945): Russian mathematician; Prof. at the Sov. Academy of Sciences in St. Petersburg; contributions to Fourier Analysis and differential equations, applications in ship building.

Here, we can assume that $g^t = Ax^t - b \notin K_t(d^0;A)$ as otherwise $g^t = 0$ and, consequently, $x^t = x$. Then, for $i = 0, ..., t-1$ there holds

$$(d^t, Ad^i) = (-g^t, Ad^i) + \sum_{j=0}^{t-1} \beta_j^{t-1}(d^j, Ad^i) = (-g^t + \beta_i^{t-1}\, d^i, Ad^i). \tag{3.3.80}$$

For $i < t-1$, we have $(g^t, Ad^i) = 0$ since $Ad^i \in K_t(d^0;A)$ and, consequently, $\beta_i^{t-1} = 0$. For $i = t-1$, the condition

$$0 = (-g^t, Ad^{t-1}) + \beta_{t-1}^{t-1}(d^{t-1}, Ad^{t-1}) \tag{3.3.81}$$

leads us to the formulas

$$\beta_{t-1} := \beta_{t-1}^{t-1} = \frac{(g^t, Ad^{t-1})}{(d^{t-1}, Ad^{t-1})}, \quad d^t = -g^t + \beta_{t-1} d^{t-1}. \tag{3.3.82}$$

The next iterates x^{t+1} and $g^{t+1} = Ax^{t+1} - b$ are then determined by

$$\alpha_t = -\frac{(g^t, d^t)}{(d^t, Ad^t)}, \quad x^{t+1} = x^t + \alpha_t d^t, \quad g^{t+1} = g^t + \alpha_t Ad^t. \tag{3.3.83}$$

These are the recurrence equations of the CG method. By construction there holds

$$(d^t, Ad^i) = (g^t, d^i) = 0, \quad i \le t-1, \quad (g^t, g^{t-1}) = 0. \tag{3.3.84}$$

From this, we conclude that

$$\|g^t\|^2 = (d^t - \beta_{t-1} d^{t-1}, -g^{t+1} + \alpha_t Ad^t) = \alpha_t (d^t, Ad^t), \tag{3.3.85}$$
$$\|g^{t+1}\|^2 = (g^t + \alpha_t Ad^t, g^{t+1}) = \alpha_t (Ad^t, g^{t+1}). \tag{3.3.86}$$

This allows for the following simplifications in the above formulas:

$$\alpha_t = \frac{\|g^t\|^2}{(d^t, Ad^t)}, \quad \beta_t = \frac{\|g^{t+1}\|^2}{\|g^t\|^2}, \tag{3.3.87}$$

as long as the iteration does not terminate with $g^t = 0$.

Definition 3.5: *The CG method determines a sequence of iterates $x^t \in \mathbb{R}^n$, $t \ge 0$, by the prescription*

$$\begin{aligned} &\textit{Starting values:} \quad && x^0 \in \mathbb{R}^n, \quad d^0 = -g^0 = b - Ax^0, \\ &\textit{Iterate for } t \ge 0: && \alpha_t = \frac{\|g^t\|^2}{(d^t, Ad^t)}, \quad x^{t+1} = x^t + \alpha_t d^t, \quad g^{t+1} = g^t + \alpha_t Ad^t, \\ & && \beta_t = \frac{\|g^{t+1}\|^2}{\|g^t\|^2}, \quad d^{t+1} = -g^{t+1} + \beta_t d^t. \end{aligned}$$

By construction the CG method generates a sequence of descent directions d^t, which are automatically A-orthogonal. This implies that the vectors $d^0, \dots, d^t$ are linearly independent and that therefore $\operatorname{span}\{d^0, \dots, d^{n-1}\} = \mathbb{R}^n$. We formulate the properties of the CG method derived so far in the following theorem.

Theorem 3.10 (CG method): *Let the matrix $A \in \mathbb{R}^{n\times n}$ be symmetric positive definite. Then, (assuming exact arithmetic) the CG method terminates for any starting vector $x^0 \in \mathbb{R}^n$ after at most n steps at $x^n = x$. In each step there holds*

$$Q(x^t) = \min_{y\in x^0+B_t} Q(y), \tag{3.3.88}$$

and, equivalently,

$$\|x^t - x\|_A = \|Ax^t - b\|_{A^{-1}} = \min_{y\in x^0+B_t} \|Ay - b\|_{A^{-1}} = \min_{y\in x^0+B_t} \|y - x\|_A, \tag{3.3.89}$$

where $B_t := \operatorname{span}\{d^0, \dots, d^{t-1}\}$.

In view of the result of Theorem 3.10 the CG method formally belongs to the class of “direct” methods. In practice, however, it is used like an iterative method, since:

1. Because of round-off errors the descent directions d^t are not exactly A-orthogonal such that the iteration does not terminate.
2. For large matrices one obtains accurate approximations already after $t \ll n$ iterations.

As preparation for the main theorem about the convergence of the CG method, we provide the following auxiliary lemma.

Lemma 3.5 (Polynomial norm bounds): *Let A be a symmetric positive definite matrix with spectrum $\sigma(A) \subset [a,b]$. Then, for any polynomial $p \in P_t$, $p(0) = 1$ there holds*

$$\|x^t - x\|_A \le M\,\|x^0 - x\|_A, \quad M := \sup_{\mu\in[a,b]} |p(\mu)|. \tag{3.3.90}$$

Proof. Observing the relation

$$\|x^t - x\|_A = \min_{y\in x^0+B_t} \|y - x\|_A,$$

$$B_t = \operatorname{span}\{d^0, \dots, d^{t-1}\} = \operatorname{span}\{A^0 g^{(0)}, \dots, A^{t-1} g^0\},$$

we find

$$\|x^t - x\|_A = \min_{p\in P_{t-1}} \|x^0 - x + p(A)g^0\|_A.$$

Since $g^0 = Ax^0 - b = A(x^0 - x)$ it follows that

$$\begin{aligned}\|x^t - x\|_A &= \min_{p\in P_{t-1}} \|[I + Ap(A)](x^0 - x)\|_A \\ &\le \min_{p\in P_{t-1}} \|I + Ap(A)\|_A \, \|x^0 - x\|_A \\ &\le \min_{p\in P_t,\, p(0)=1} \|p(A)\|_A \, \|x^0 - x\|_A,\end{aligned}$$

with the natural matrix norm $\|\cdot\|_A$ generated from the A-norm $\|\cdot\|_A$. Let λ_i, $i = 1,\dots,n$, be the eigenvalues and $\{w^1,\dots,w^n\}$ a corresponding ONS of eigenvectors of the symmetric, positive definite matrix A. Then, for arbitrary $y \in \mathbb{R}^n$ there holds

$$y = \sum_{i=1}^n \gamma_i w_i\,, \quad \gamma_i = (y, w_i),$$

and, consequently,

$$\|p(A)y\|_A^2 = \sum_{i=1}^n \lambda_i p(\lambda_i)^2 \gamma_i^2 \le M^2 \sum_{i=1}^n \lambda_i \gamma_i^2 = M^2\, \|y\|_A^2.$$

This implies

$$\|p(A)\|_A = \sup_{y\in\mathbb{R}^n,\, y\ne 0} \frac{\|p(A)y\|_A}{\|y\|_A} \le M,$$

which completes the proof. Q.E.D.

As a consequence of Lemma 3.5, we obtain the following a priori error estimate.

Theorem 3.11 (Error estimate for CG method): *Let A be a symmetric positive definite matrix. Then, for the CG method there holds the error estimate*

$$\|x^t - x\|_A \le 2\Big(\frac{1 - 1/\sqrt{\kappa}}{1 + 1/\sqrt{\kappa}}\Big)^t \|x^0 - x\|_A\,, \quad t \in \mathbb{N}\,, \tag{3.3.91}$$

with the spectral condition number $\kappa = \mathrm{cond}_2(A) = \Lambda/\lambda$ of A. For reducing the initial error by a factor TOL the following number of iteration is required:

$$t(\mathrm{TOL}) \approx \tfrac{1}{2}\sqrt{\kappa}\,\ln(2/\mathrm{TOL}). \tag{3.3.92}$$

Proof. i) Setting $[a,b] := [\lambda,\Lambda]$ in Lemma 3.5, we obtain

$$\|x^t - x\|_A \le \min_{p\in P_t,\, p(0)=1} \Big\{ \sup_{\lambda\le\mu\le\Lambda} |p(\mu)| \Big\} \|x^0 - x\|_A.$$

This yields the assertion if we can show that

$$\min_{p\in P_t,\, p(0)=1} \Big\{ \sup_{\lambda\le\mu\le\Lambda} |p(\mu)| \Big\} \le 2\,\Big(\frac{1 - \sqrt{\lambda/\Lambda}}{1 + \sqrt{\lambda/\Lambda}}\Big)^t.$$

This is again a problem of approximation theory with respect to the max-norm (Chebyshev approximation), which can be solved using the Chebyshev polynomials described above in Subsection 3.2.2. The solution $p_t \in P_t$ is give by

$$p_t(\mu) = T_t\Big(\frac{\Lambda + \lambda - 2\mu}{\Lambda - \lambda}\Big)\, T_t\Big(\frac{\Lambda + \lambda}{\Lambda - \lambda}\Big)^{-1},$$

with the t-th Chebyshev polynomial T_t on $[-1,1]$. There holds

$$\sup_{\lambda \le \mu \le \Lambda} p_t(\mu) = T_t\Big(\frac{\Lambda + \lambda}{\Lambda - \lambda}\Big)^{-1}.$$

From the representation

$$T_t(\mu) = \tfrac{1}{2}\left[\big(\mu + \sqrt{\mu^2 - 1}\big)^t + \big(\mu - \sqrt{\mu^2 - 1}\big)^t\right], \quad \mu \in [-1,1],$$

for the Chebyshev polynomials and the identity

$$\frac{\kappa + 1}{\kappa - 1} + \sqrt{\Big(\frac{\kappa + 1}{\kappa - 1}\Big)^2 - 1} = \frac{\kappa + 1}{\kappa - 1} + \frac{2\sqrt{\kappa}}{\kappa - 1} = \frac{(\sqrt{\kappa} + 1)^2}{\kappa - 1} = \frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1},$$

we obtain the estimate

$$T_t\Big(\frac{\Lambda + \lambda}{\Lambda - \lambda}\Big) = T_t\Big(\frac{\kappa + 1}{\kappa - 1}\Big) = \frac{1}{2}\left[\Big(\frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1}\Big)^t + \Big(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\Big)^t\right] \ge \frac{1}{2}\Big(\frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1}\Big)^t.$$

Hence,

$$\sup_{\lambda \le \mu \le \Lambda} p_t(\mu) \le 2\Big(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\Big)^t,$$

which implies (3.3.91).

ii) For deriving (3.3.92), we require

$$2\Big(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\Big)^{t(\varepsilon)} \le \mathrm{TOL},$$

and, equivalently,

$$t(\mathrm{TOL}) > \ln\Big(\frac{2}{\mathrm{TOL}}\Big) \ln\Big(\frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1}\Big)^{-1}.$$

Since

$$\ln\frac{x + 1}{x - 1} = 2\left\{\frac{1}{x} + \frac{1}{3}\frac{1}{x^3} + \frac{1}{5}\frac{1}{x^5} + \dots\right\} \ge \frac{2}{x},$$

this is satisfied for $t(\mathrm{TOL}) \ge \frac{1}{2}\sqrt{\kappa}\, \ln(2/\mathrm{TOL})$. Q.E.D.

Since $\kappa = \mathrm{cond}_{\mathrm{nat}}(\mathrm{A}) > 1$, we have $\sqrt{\kappa} < \kappa$. Observing that the function $f(\lambda) = (1 - \lambda^{-1})\,(1 + \lambda^{-1})^{-1}$ is strictly monotonically increasing for $\lambda > 0$ $(f'(\lambda) > 0)$, there holds

$$\frac{1-1/\sqrt{\kappa}}{1+1/\sqrt{\kappa}} < \frac{1-1/\kappa}{1+1/\kappa},$$

implying that the CG method should converge faster than the gradient method. This is actually the case in practice. Both methods converge the faster the smaller the condition number is. However, in case $\Lambda \gg \lambda$, which is frequently the case in practice, even the CG method is too slow. An acceleration can be achieved by so-called "preconditioning", which will be described below.

3.3.3 Generalized CG methods and Krylov space methods

For solving a general linear system $Ax = b$, with regular but not necessarily symmetric and positive definite matrix $A \in \mathbb{R}^n$, by the CG method, one may consider the equivalent system

$$A^TAx = A^Tb \tag{3.3.93}$$

with the symmetric, positive definite matrix A^TA. Applied to this system the CG method takes the following form:

$$\begin{aligned}
&\text{Starting values:} \quad && x^0 \in \mathbb{R}^n\,, \quad d^0 = A^T(b - Ax^0) = -g^0,\\
&\text{for } t \geq 0: && \alpha_t = \frac{\|g^t\|^2}{\|Ad^t\|^2}, \quad x^{t+1} = x^t + \alpha_t d^t, \quad g^{t+1} = g^t + \alpha_t A^TAd^t,\\
& && \beta_t = \frac{\|g^{t+1}\|^2}{\|g^t\|^2}, \quad d^{t+1} = -g^{t+1} + \beta_t d^t.
\end{aligned}$$

This approach is referred to as CGS method ("Conjugate Gradient Squared") of P. Sonneveld (1989). The convergence speed is characterized by $\mathrm{cond}_2(A^TA)$. The whole method is equivalent to minimizing the functional

$$Q(y) := \tfrac{1}{2}\,(A^TAy, y) - (A^Tb, y) = \tfrac{1}{2}\|Ay - b\|^2 - \tfrac{1}{2}\,\|b\|^2. \tag{3.3.94}$$

Since $\mathrm{cond}_2(A^TA) \approx \mathrm{cond}_2(A)^2$ the convergence of this variant of the CG method may be rather slow. However, its realization does not require the explicit evaluation of the matrix product A^TA but only the computation of the matrix-vector products $z = Ay$ and A^Tz.

On the basis of the formulation (3.3.75) the standard CG method is limited to linear systems with symmetric, positive definite matrices. But starting from the (in this case equivalent) Galerkin formulation (3.3.76) the method becomes meaningful also for more general matrices. In fact, in this way one can derive effective generalizations of the CG method also for nonsymmetric and even indefinite matrices. These modified CG methods are based on the Galerkin equations (3.3.76) and differ in the choices of "ansatz spaces"

K_t and "test spaces" K_t^*,

$$x^t \in x^0 + K_t : \quad (Ax^t - b, y) = 0 \quad \forall\, y \in K_t^*. \tag{3.3.95}$$

Here, one usually uses the Krylov spaces

$$K_t = \text{span}\{d^0, Ad^0, ..., A^{t-1}d^0\},$$

combined with the test spaces $K_t^* = K_t$, or

$$K_t^* = \text{span}\{d^0, A^T d^0, ..., (A^T)^{t-1}d^0\}.$$

This leads to the general class of "Krylov space methods". Most popular representatives are the following methods, which share one or the other property with the original CG method but generally do not allow for a similarly complete error analysis.

1. GMRES with or without restart ("Generalized Minimal Residual") of Y. Saad and M. H. Schultz (1986): $K_t = \text{span}\{d^0, Ad^0, ..., A^{t-1}d^0\} = K_t^*$,

   $$\|Ax^t - b\| = \min_{y \in x^0 + K_t} \|Ay - b\|. \tag{3.3.96}$$

   Since this method minimizes the residual over spaces of increasing dimension as the CG method also the GMRES methods yields the exact solution after at most n steps. However, for general nonsymmetric matrices the iterates x^t cannot be obtained by a simple tree-term recurrence as in the CG method. It uses a full recurrence, which results in high storage requirements. Therefore, to limit the costs the GMRES method is stopped after a certain number of steps, say k steps, and then restarted with x^k as new starting vector. The latter variant is denoted by "GMRES(k) method".

2. BiCG and BiCGstab ("Biconjugate Gradient Stabilized") of H. A. Van der Vorst, (1992): $K_t = \text{span}\{d^0, Ad^0, ..., A^{t-1}d^0\}$, $K_t^* = \text{span}\{d^0, A^T d^0, ..., (A^T)^{t-1}d^0\}$,

   $$x^t \in x^0 + K_t : \quad (Ax^t - b, y) = 0, \quad \forall y \in K_t^*. \tag{3.3.97}$$

   In the BiCG method the iterates x^t are obtained by a three-term recurrence but for an unsymmetric matrix the residual minimization property gets lost and the method may not even converge. Additional stability is provided in the "BiCGstab" method.

Both methods, GMRES(k) and BiCGstab, are especially designed for unsymmetric but definite matrices. They have there different pros and cons and are both not universally applicable. One can construct matrices for which one or the other of the methods does not work. The methods for the practical computation of the iterates x^t in the Krylov spaces K_t are closely related to the Lanczos and Arnoldi algorithms used for solving the corresponding eigenvalue problems discussed in Chapter 4, below.

3.3.4 Preconditioning (PCG methods)

The error estimate (3.3.91) for the CG method indicates a particularly good convergence if the condition number of the matrix A is close to one. In case of large $\mathrm{cond}_2(A) \gg 1$, one uses "preconditioning", i. e., the system $Ax = b$ is transformed into an equivalent one, $\tilde{A}\tilde{x} = \tilde{b}$ with a better conditioned matrix $\tilde{A}$. To this end, let C be a symmetric, positive definite matrix, which is explicitly given in product form

$$C = KK^T, \tag{3.3.98}$$

with a regular matrix K. The system $Ax = b$ can equivalently be written in the form

$$\underbrace{K^{-1}A\,(K^T)^{-1}}_{\tilde{A}}\,\underbrace{K^Tx}_{\tilde{x}} = \underbrace{K^{-1}b}_{\tilde{b}}. \tag{3.3.99}$$

Then, the CG method is formally applied to the transformed system $\tilde{A}\tilde{x} = \tilde{b}$, while it is hoped that $\mathrm{cond}_2(\tilde{A}) \ll \mathrm{cond}_2(A)$ for an appropriate choice of C. The relation

$$(K^T)^{-1}\tilde{A}K^T = (K^T)^{-1}K^{-1}A(K^T)^{-1}K^T = C^{-1}A \tag{3.3.100}$$

shows that for $C \equiv A$ the matrix $\tilde{A}$ is similar to I, and thus $\mathrm{cond}_2(\tilde{A}) = \mathrm{cond}_2(I) = 1$. Consequently, one chooses $C = KK^T$ such that C^{-1} is a good approximation to A^{-1}.

The CG method for the transformed system $\tilde{A}\tilde{x} = \tilde{b}$ can then be written in terms of the quantities A, b, and x as so-called "PCG method" ("Preconditioned CG" method) as follows:

$$\begin{aligned}
&\text{Starting value:} && x^0 \in \mathbb{R}^N,\;\; d^0 = r^0 = b - Ax^0,\;\; \mathbf{C}\rho^{\mathbf{0}} = \mathbf{r^0},\\
&\text{for } t \ge 0: && \alpha_t = \frac{\langle r^t, \rho^t\rangle}{\langle Ad^t, d^t\rangle},\quad x^{t+1} = x^t + \alpha_t d^t,\;\; r^{t+1} = r^t - \alpha_t Ad^t,\;\; \mathbf{C}\rho^{\mathbf{t+1}} = \mathbf{r^{t+1}},\\
& && \beta_t = \frac{\langle r^{t+1}, \rho^{t+1}\rangle}{\langle r^t, \rho^t\rangle},\quad d^{t+1} = r^{t+1} + \beta_t d^t.
\end{aligned}$$

Compared to the normal CG method the PCG iteration in each step additionally requires the solution of the system $C\rho^{t+1} = r^{t+1}$, which is easily accomplished using the decomposition $C = KK^T$. In order to preserve the work complexity $\mathcal{O}(n)$ a. op. in each step the triangular matrix K should have a sparsity pattern similar to that of the lower triangular part L of A. This condition is satisfied by the following popular preconditioners:

1) Diagonal preconditioning (scaling): $C := D = D^{1/2}D^{1/2}$.
The scaling ensures that the elements of A are brought to approximately the same size, especially with $\tilde{a}_{ii} = 1$. This reduces the condition number since

$$\mathrm{cond}_2(A) \ge \frac{\max_{1\le i\le n} a_{ii}}{\min_{1\le i\le n} a_{ii}}. \tag{3.3.101}$$

Example: The matrix $A = \mathrm{diag}\{\lambda_1 = ... = \lambda_{n-1} = 1, \lambda_n = 10^k\}$ has the condition number

$\mathrm{cond}_2(A) = 10^k$, while the scaled matrix $\tilde{A} = D^{-1/2}AD^{-1/2}$ has the optimal condition number $\mathrm{cond}_2(\tilde{A}) = 1$.

2) SSOR preconditioning: We choose

$$\begin{aligned} C &:= (D+L)D^{-1}(D+L^T) = D + L + L^T + LD^{-1}L^T \\ &= (\underbrace{D^{1/2} + LD^{-1/2}}_{K})(\underbrace{D^{1/2} + D^{-1/2}L^T}_{K^T}), \end{aligned}$$

or, more generally, involving a relaxation parameter $\omega \in (0,2)$,

$$\begin{aligned} C &:= \frac{1}{2-\omega}\Big(\frac{1}{\omega}D + L\Big)\Big(\frac{1}{\omega}D\Big)^{-1}\Big(\frac{1}{\omega}D + L^T\Big) \\ &= \underbrace{\frac{1}{\sqrt{(2-\omega)\omega}}(D^{1/2} + \omega LD^{-1/2})}_{K}\ \underbrace{\frac{1}{\sqrt{(2-\omega)\omega}}(D^{1/2} + \omega D^{-1/2}L^T)}_{K^T}. \end{aligned}$$

Obviously, the triangular matrix K has the same sparsity pattern as L. Each step of the preconditioned iteration costs about twice as much work as the basic CG method. For an optimal choice of the relaxation parameter ω (not easy to determine) there holds

$$\mathrm{cond}_2(\tilde{A}) = \sqrt{\mathrm{cond}_2(A)}.$$

3) ICCG preconditioning (Incomplete Cholesky Conjugate Gradient): The symmetric, positive definite matrix A has a Cholesky decomposition $A = LL^T$ with an lower triangular matrix $L = (l_{ij})_{i,j=1}^n$. The elements of L are successively determined by the recurrence formulas

$$l_{ii} = \Big(a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2\Big)^{1/2}, \quad i = 1,\dots,n, \qquad l_{ji} = \frac{1}{l_{ii}}\Big(a_{ji} - \sum_{k=1}^{i-1} l_{jk}l_{ik}\Big), \quad j = i+1,\dots,n.$$

The matrix L generally has nonzero elements in the whole band of A, which requires much more memory than A itself. This can be avoided by performing (such as in the ILU approach discussed in Subsection 3.1.2) only an "incomplete" Cholesky decomposition where within the elimination process some of the l_{ji} are set to zero, e.g., those for which $a_{ji} = 0$. This results in an incomplete decomposition

$$A \;=\; \tilde{L}\tilde{L}^T + E \tag{3.3.102}$$

with a lower triangular matrix $\tilde{L} = (\tilde{l}_{ij})_{i,j=1}^n$, which has a similar sparsity pattern as A. In this case, one speaks of the "$ICCG(0)$ variant". In case of a band matrix A, one may allow the elements of $\tilde{L}$ to be nonzero in further p off-diagonals resulting in the so-called "*ICCG*(p) variant" of the *ICCG* method, which is hoped to provide a better approximation $C^{-1} \approx A^{-1}$ for increasing p. Then, for preconditioning the matrix

$$C = KK^T := \tilde{L}\tilde{L}^T \tag{3.3.103}$$

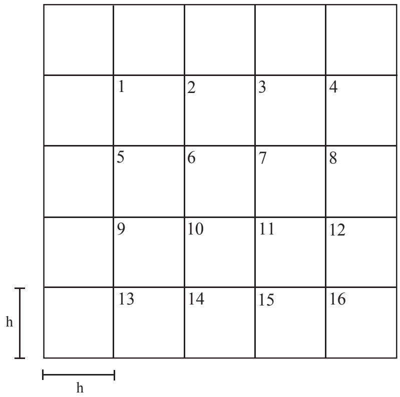
is used. Although, there is no full theoretical justification yet for the success of the *ICCG* preconditioning practical tests show a significant improvement in the convergence behavior. This may be due to the fact that, though the condition number is not necessarily decreased, the eigenvalues of the corresponding transformed matrix $\tilde{A}$ cluster more around $\lambda = 1$.

3.4 A model problem

At the end of the discussion of the classical iterative methods for solving linear systems $Ax = b$, we will determine their convergence rates for the model situation already described in Section 0.4.2 of Chapter 0. We consider the so-called "1-st boundary value problem of the Laplace operator"

$$\begin{aligned} -\frac{\partial^2 u}{\partial x^2}(x,y) - \frac{\partial^2 u}{\partial y^2}(x,y) &= f(x,y) \quad \text{for } (x,y) \in \Omega \\ u(x,y) &= 0 \quad \text{for } (x,y) \in \partial\Omega, \end{aligned} \tag{3.4.104}$$

on the unit square $\Omega = (0,1) \times (0,1) \subset \mathbb{R}^2$. For solving this problem the domain Ω is covered by a uniform mesh as shown in Fig. 3.4.



$h = \frac{1}{m+1}$ mesh size

$n = m^2$ number of unknown mesh values

Figure 3.5: *Mesh for the discretization of the model problem*

The "interior" mesh points are numbered row-wise. On this mesh the second derivatives in the differential equation (3.4.104) are approximated by second-order central difference quotients leading to the following difference equations for the mesh unknowns $U(x,y) \approx u(x,y)$:

$$-h^{-2}\big\{U(x{+}h,y) - 2U(x,y) + U(x{-}h,y) + U(x,y{+}h) - 2U(x,y) + U(x,y{-}h)\big\} = f(x,y).$$

Observing the boundary condition $u(x,y) = 0$ for $(x,y) \in \partial\Omega$ this set of difference equations is equivalent to the linear system

$$Ax = b, \tag{3.4.105}$$

for the vector $x \in \mathbb{R}^n$ of unknown mesh values $x_i \approx u(P_i)$, P_i interior mesh point. The matrix A has the already known form

$$A = \left.\begin{bmatrix} B & -I & & \\ -I & B & -I & \\ & -I & B & \ddots \\ & & \ddots & \ddots \end{bmatrix}\right\} n \qquad B = \left.\begin{bmatrix} 4 & -1 & & \\ -1 & 4 & -1 & \\ & -1 & 4 & \ddots \\ & & \ddots & \ddots \end{bmatrix}\right\} m$$

with the $m{\times}m$-unit matrix I. The right-hand side is given by $b = h^2(f(P_1), \dots, f(P_n))^T$. The matrix A has several special properties:

- "sparse band matrix" with bandwidth $2m+1$;
- "irreducible" and "strongly diagonally dominant";
- "symmetric" and "positive definite";
- "consistently ordered";
- "of nonnegative type" ("M-matrix"): $a_{ii} > 0,\;\; a_{ij} \le 0,\; i \neq j.$
  The importance of this last property will be illustrated in an exercise.

For this matrix eigenvalues and eigenvectors can be explicitly determined $(h = 1/(m{+}1))$:

$$\lambda_{kl} = 4 - 2\,(\cos[kh\pi] + \cos[lh\pi]), \quad w^{kl} = (\sin[ikh\pi]\,\sin[jlh\pi])_{i,j=1,\dots,m}, \quad k,l = 1,\dots,m,$$

i. e., $Aw^{kl} = \lambda_{kl} w^{kl}$. Hence for $h \ll 1$, we have

$$\Lambda := \lambda_{\max} = 4 - 4\,\cos(1-h)\pi \;\approx\; 8,$$
$$\lambda := \lambda_{\min} = 4 - 4\,\cos(h\pi) = 4 - 4\,(1 - \frac{\pi^2}{2}\,h^2 + O(h^4)) \;\approx\; 2\pi^2 h^2,$$

and consequently

$$\kappa \;:=\; \mathrm{cond}_2(\mathrm{A}) \;\approx\; \frac{4}{\pi^2 \mathrm{h}^2}. \tag{3.4.106}$$

Then, the eigenvalues of the Jacobi iteration matrix $J = -D^{-1}(L+R)$ are given by

$$\mu_{kl}(J) = \tfrac{1}{2}\,(\cos[kh\pi] + \cos[lh\pi]), \quad k,l = 1,\dots,m.$$

Hence,

$$\rho := \mathrm{spr(J)} = \mu_{\max}(\mathrm{J}) = \cos[\mathrm{h}\pi] = 1 - \frac{\pi^2}{2}\,\mathrm{h}^2 + \mathrm{O(h^4)}. \tag{3.4.107}$$

For the iteration matrices of the Gauß-Seidel and the optimal SOR iteration matrices,

H_1 and $H_{\omega_{\rm opt}}$, respectively, there holds

$$\mathrm{spr}(\mathrm{H}_1) = \rho^2 = 1 - \pi^2 h^2 + O(h^4), \tag{3.4.108}$$

$$\mathrm{spr}(\mathrm{H}_{\omega_{\rm opt}}) = \frac{1-\sqrt{1-\rho^2}}{1+\sqrt{1-\rho^2}} = \frac{1-\pi h + O(h^2)}{1+\pi h + O(h^2)} = 1 - 2\,\pi h + O(h^2)\,. \tag{3.4.109}$$

Comparison of convergence speed

Now, we make a comparison of the convergence speed of the various iterative methods considered above. The reduction of the initial error $\|x^{(0)} - x\|_2$ in a fixed-point iteration by the factor $\varepsilon \ll 1$ requires about $T(\varepsilon)$ iterations,

$$T(\varepsilon) \approx \frac{\ln(1/\varepsilon)}{\ln(1/\rho)}, \quad \rho = \mathrm{spr}(\mathrm{B}), \;\; \mathrm{B} = \mathrm{I} - \mathrm{C}^{-1}\mathrm{A} \quad \text{iteration matrix.} \tag{3.4.110}$$

Using the above formulas, we obtain:

$$\begin{aligned} T_J(\varepsilon) &\approx -\frac{\ln(1/\varepsilon)}{\ln(1-\frac{\pi^2}{2}h^2)} \approx 2\frac{\ln(1/\varepsilon)}{\pi^2 h^2} = \frac{2}{\pi^2} n \; \ln(1/\varepsilon), \\ T_{GS}(\varepsilon) &\approx -\frac{\ln(1/\varepsilon)}{\ln(1-\pi^2 h^2)} \approx \frac{\ln(1/\varepsilon)}{\pi^2 h^2} = \frac{1}{\pi^2} n \; \ln(1/\varepsilon), \\ T_{SOR}(\varepsilon) &\approx -\frac{\ln(1/\varepsilon)}{\ln(1-2\pi h)} \approx \frac{\ln(1/\varepsilon)}{2\pi h} = \frac{1}{2\pi}\sqrt{n} \; \ln(1/\varepsilon). \end{aligned}$$

The gradient method and the CG method require for the reduction of the initial error $\|x^0 - x\|_2$ by the factor $\varepsilon \ll 1$ the following numbers of iterations:

$$\begin{aligned} T_G(\varepsilon) &= \frac{1}{2}\,\kappa \; \ln(2/\varepsilon) \approx \frac{2}{\pi^2 h^2} \; \ln(1/\varepsilon) \approx \frac{2}{\pi^2} n \; \ln(1/\varepsilon), \\ T_{CG}(\varepsilon) &= \frac{1}{2}\,\sqrt{\kappa} \; \ln(2/\varepsilon) \approx \frac{1}{\pi h} \; \ln(2/\varepsilon) \approx \frac{1}{\pi}\,\sqrt{n} \; \ln(2/\varepsilon). \end{aligned}$$

We see that the Jacobi method and the gradient method converge with about the same speed. The CG method is only half as fast as the (optimal) SOR method, but it does not require the determination of an optimal parameter (while the SOR method does not require the matrix A to be symmetric). The Jacobi method with Chebyshev acceleration is as fast as the "optimal" SOR method but also does not require the determination of an optimal parameter (but a guess for $\mathrm{spr}(J)$).

For the special right-hand side function $f(x,y) = 2\pi^2 \sin(\pi x)\sin(\pi y)$ the exact solution of the boundary value problem is given by

$$u(x,y) = \sin(\pi x)\sin(\pi y). \tag{3.4.111}$$

The error caused by the finite difference discretization considered above can be estimated as follows:

$$\max_{P_i} |u(P_i) - x_i| \le \frac{\pi^4}{12} h^2 + O(h^4). \tag{3.4.112}$$

Hence, for achieving a relative accuracy of $\mathrm{TOL} = 10^{-3}$ (three decimals) a mesh size

$$h \approx \frac{\sqrt{12}}{\pi^2} 10^{-3/2} \approx 10^{-2},$$

is required. This results in $n \approx 10^4$ unknowns. In this case, we obtain for the above spectral radii, conditions numbers and numbers of iterations required for error reduction by $\varepsilon = 10^{-4}$ (including a safety factor of $1/10$) the following values $(\ln(1/\varepsilon) \sim 10)$:

$$\begin{aligned}
\mathrm{spr(J)} &\approx 0,9995 & T_{\mathrm{J}} &\approx 20.000 \\
\mathrm{spr(H_1)} &\approx 0,999 & T_{\mathrm{GS}} &\approx 10.000 \\
\mathrm{spr(H_{\omega *})} &\approx 0,9372 & T_{\mathrm{SOR}} &\approx 160 \\
\mathrm{cond_2(A)} &\approx 5.000 & T_{G} &\approx 20.000, \quad T_{\mathrm{CG}} \approx 340
\end{aligned}$$

For the comparison of the various solution methods, we also have to take into account the work in each iteration step. For the number “OP” of “a. op.” (1 multiplication + 1 addition) per iteration step there holds:

$$\begin{aligned}
&\mathrm{OP_J} \approx \mathrm{OP_{H_1}} \approx \mathrm{OP_{H_\omega}} \approx 6\,\mathrm{n}, \\
&\mathrm{OP_G} \approx \mathrm{OP_{CG}} \approx 10\,\mathrm{n}.
\end{aligned}$$

As final result, we see that the computation of the approximate solution of the boundary value model problem (3.4.104) with a prescribed accuracy TOL by the Jacobi method, the Gauß-Seidel method and the gradient method requires $\mathcal{O}(n^2)$ a. op. In this case a direct method such as the Cholesky algorithm requires $\mathcal{O}(n^2) = \mathcal{O}(m^2 n)$ a. op. but significantly more storage space. The (optimal) SOR method and the CG method only require $\mathcal{O}(n^{3/2})$ a. op.

For the model problem with $n = 10^4$, we have the following total work “TW” required for the solution of the system (3.4.105) to discreetization accuracy $\varepsilon = 10^{-4}$:

$$\begin{aligned}
\mathrm{TW}_J(\mathrm{TOL}) &\approx 4 \cdot 3n^2 \approx 1,2 \cdot 10^9\,\text{a. op.}, \\
\mathrm{TW}_{GS}(\mathrm{TOL}) &\approx 4 \cdot 1,5n^2 \approx 6 \cdot 10^8\,\text{a. op.}, \\
\mathrm{TW}_{SOR}(\mathrm{TOL}) &\approx 4 \cdot 2n^{3/2} \approx 8 \cdot 10^6\,\text{a. op.}, \\
\mathrm{TW}_{CG}(\mathrm{TOL}) &\approx 4 \cdot 10n^{3/2} \approx 4 \cdot 10^7\,\text{a. op.}.
\end{aligned}$$

Remark 3.7: Using an appropriate preconditioning, e. g., the ILU preconditioning, in the CG method the work count can be reduced to $\mathcal{O}(n^{5/4})$. The same complexity can be achieved by Chebyshev acceleration of the (optimal) SOR method. Later, we will discuss a more sophisticated iterative method based on the “multi-level concept”, which has optimal solution complexity $\mathcal{O}(n)$. For such a multigrid ("MG") method, we can expect work counts like $\mathrm{TW}_{MG} \approx 4 \cdot 25n \approx 10^6\,$a. op..

Remark 3.8: For the 3-dimensional version of the above model problem, we have

$$\lambda_{max} \approx 12h^{-2}, \quad \lambda_{min} \approx 3\pi^2, \quad \kappa \approx \frac{8}{3\pi^2 h^2},$$

and consequently the same estimates for ρ_J, ρ_{GS} and ρ_{SOR} as well as for the iteration numbers T_J, T_{GS}, T_{SOR}, T_{CG}, as in the 2-dimensional case. In this case the total work per iteration step is $\mathrm{OP_J}$, $\mathrm{OP_{GS}}$, $\mathrm{OP_{SOR}} \approx 8\,\mathrm{N}$, $\mathrm{OP_{CG}} \approx 12\,\mathrm{N}$. Hence, the resulting total work amounts to

$$\begin{aligned}
\mathrm{TW}_J(\mathrm{TOL}) &\approx 4 \cdot 4n^2 \approx 1{,}6 \cdot 10^{13}\,\text{a. op.}\,,\\
\mathrm{TW}_{GS}(\mathrm{TOL}) &\approx 4 \cdot 2n^2 \approx 8 \cdot 10^{12}\,\text{a. op.}\,,\\
\mathrm{TW}_{SOR}(\mathrm{TOL}) &\approx 4 \cdot 3n^{3/2} \approx 1{,}2 \cdot 10^{10}\,\text{a. op.}\,,\\
\mathrm{TW}_{CG}(\mathrm{TOL}) &\approx 4 \cdot 12n^{3/2} \approx 4{,}8 \cdot 10^{10}\,\text{a. op.},
\end{aligned}$$

while that for the multigrid method increases only to $\mathrm{TW}_{MG} \approx 4 \cdot 50n \approx 2 \cdot 10^8\,$a. op..

Remark 3.9: For the interpretation of the above work counts, we have to consider the computing power of available computer cores, e.g., 200 MFlops (200 million "floating-point" oper./sec.) of a standard desktop computer. Here, the solution of the 3-dimensional model problem by the optimal SOR method takes about $1,5$ minutes while the multigrid method only needs less than 1 second.

3.5 Exercises

Exercise 3.1: Investigate the convergence of the fixed-point iteration $x^t = Bx^{t-1} + c$ with an arbitrary starting value $x^0 \in \mathbb{R}^3$ for the following matrices

$$i) \quad B = \begin{bmatrix} 0.6 & 0.3 & 0.1 \\ 0.2 & 0.5 & 0.7 \\ 0.1 & 0.1 & 0.1 \end{bmatrix}, \qquad ii) \quad B = \begin{bmatrix} 0 & 0.5 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

What are the limits of the iterates in case of convergence? (Hint: The eigenvalues of the matrices B are to be estimated. This can be done via appropriate matrix norms or also via the determinants.

Exercise 3.2: The linear system

$$\begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

is to be solved by the Jacobi and the Gauß-Seidel method. How many iterations are approximately (asymptotically) required for reducing the initial error $\|x^0 - x\|_2$ by the factor 10^{-6}? (Hint: Use the error estimate stated in the text.)

Exercise 3.3: Show that the two definitions of "irreducibility" of a matrix $A \in \mathbb{R}^{n\times n}$ given in the text are equivalent.

Hint: Use the fact that the definition of "reducibility" of the system $Ax = b$, i.e., the existence of simultaneous row and column permutations resulting in

$$PAP^T = \tilde{A} = \begin{bmatrix} \tilde{A}_{11} & 0 \\ 0 & \tilde{A}_{22} \end{bmatrix}, \quad \tilde{A}_{11} \in \mathbb{R}^{p\times p}, \quad \tilde{A}_{22} \in \mathbb{R}^{q\times q}, \quad n = p + q,$$

is equivalent to the existence of a non-trivial index partitioning $\{J, K\}$ of $N_n = \{1, \dots, n\}$, $J \cup K = N_n$, $J \cap K = \emptyset$, such that $a_{jk} = 0$ for $j \in J,\ k \in K$.

Exercise 3.4: Examine the convergence of the Jacobi and Gauss-Seidel methods for solving the linear system $A_i x = b$ $(i = 1, 2)$ for the following two matrices

$$A_1 = \begin{bmatrix} 2 & -1 & 2 \\ 1 & 2 & -2 \\ 2 & 2 & 2 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} 5 & 5 & 0 \\ -1 & 5 & 4 \\ 2 & 3 & 8 \end{bmatrix}.$$

(Hint: Use the convergence criteria stated in the text, or estimate the spectral radius)

Exercise 3.5: For the solution of the linear (2×2)-system

$$\begin{bmatrix} 1 & -a \\ -a & 1 \end{bmatrix} x = b, \quad x, b \in \mathbb{R}^2,$$

the following parameter-dependent fixed-point iteration is considered:

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

a) For which $a \in \mathbb{R}$ is this method with $\omega = 1$ convergent?

b) Determine for $a = 0.5$ the value

$$\omega \in \{0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4\},$$

for which the spectral radius of the iteration matrix B_ω becomes minimal and sketch the graph of the function $f(\omega) = \mathrm{spr}(B_\omega)$.

Exercise 3.6: Let $A \in \mathbb{R}^{n\times n}$ be a symmetric (and therefore diagonalizable) matrix with eigenvalues $\lambda_i \in \mathbb{R}$, $i = 1, \dots, n$. Show that for any polynomial $p \in P_k$ there holds

$$\operatorname{spr}(p(A)) = \max_{i=1,\dots,n} |p(\lambda_i)|.$$

(Hint: Use the fact that there exists an ONB of eigenvectors of A.)

Exercise 3.7: For the computation of the inverse A^{-1} of a regular matrix $A \in \mathbb{R}^{n\times n}$ the following two fixed-point iterations are considered:

$$\begin{aligned} &a) \quad X_t = X_{t-1}(I - AC) + C, \quad t = 1, 2, \dots, \quad C \in \mathbb{R}^{n\times n} \text{ a regular “preconditioner”},\\ &b) \quad X_t = X_{t-1}(2I - AX_{t-1}), \quad t = 1, 2, \dots . \end{aligned}$$

Give (sufficient) criteria for the convergence of these iterations. For this task (computation of a matrix inverse), how would the Newton iteration look like?

Exercise 3.8: Let B be an arbitrary $n \times n$-matrix, and let p by a polynomial. Show that

$$\sigma(p(B)) = p(\sigma(B)),$$

i. e., for any $\lambda \in \sigma(p(B))$ there exists a $\mu \in \sigma(B)$ such that $\lambda = p(\mu)$ and vice versa. (Hint: Recall the Schur or the Jordan normal form.)

Exercise 3.9: The method of Chebyshev acceleration can be applied to any convergent fixed-point iteration

$$x^t = Bx^{t-1} + c, \quad t = 1, 2, \dots ,$$

with symmetric iteration matrix B. Here, the symmetry of B guarantees the relation $\|p(B)\|_2 = \operatorname{spr}(p(B)) = \max_{\lambda\in\sigma(B)} |p(\lambda|$ for any polynomial $p \in P_k$, which is crucial for the analysis of the acceleration effect. In the text this has been carried out for the SSOR (**Symmetric** Successive Over-Relaxation) method. Repeat the steps of this analysis for the Jacobi method for solving the linear system $Ax = b$ with symmetric matrix $A \in \mathbb{R}^{n\times n}$.

Exercise 3.10: Consider the following symmetric “saddle point system”

$$\begin{bmatrix} A & B \\ B^T & O \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} b \\ c \end{bmatrix},$$

with a symmetric positive definite matrix $A \in \mathbb{R}^{n\times n}$ and a not necessarily quadratic matrix $B \in \mathbb{R}^{n\times m}$, $m \le n$. The coefficient matrix cannot be positive definite since some of its main diagonal elements are zero. Most of the iterative methods discussed in the text can directly be applied for this system.

i) Assume that the coefficient matrix is regular. Can the damped Richardson method,

$$\begin{bmatrix} x^t \\ y^t \end{bmatrix} = \left(\begin{bmatrix} I & O \\ O & I \end{bmatrix} - \theta \begin{bmatrix} A & B \\ B^T & O \end{bmatrix} \right) \begin{bmatrix} x^{t-1} \\ y^{t-1} \end{bmatrix} + \theta \begin{bmatrix} b \\ c \end{bmatrix},$$

be made convergent in this case for appropriately chosen damping parameter θ? (Hint: Investigate whether the coefficient matrix may have positive AND negative eigenvalues.)

ii) A classical approach to solving this saddle-point system is based on the equivalent "Schur-complement formulation":

$$B^TA^{-1}By = B^TA^{-1}b - c, \quad x = A^{-1}b - A^{-1}By,$$

in which the solution component y can be computed independently of x. The matrix $B^TA^{-1}B$ is called the "Schur complement" of A in the full block matrix. Show that the matrix $B^TA^{-1}B$ is symmetric and positive semi-definite and even positive definite if B has maximal rank. Hence the symmetrized Gauß-Seidel method with Chebyshev acceleration may be applied to this reduced system for y. Formulate this iteration!

Exercise 3.11: The general "descent method" for the iterative solution of a linear system $Ax = b$ with symmetric positive definite matrix $A \in \mathbb{R}^{N\times N}$ has the form

$$\begin{aligned} \text{starting value:} \quad & x^0 \in \mathbb{R}^n, \quad g^0 := Ax^0 - b\,, \\ \text{for } t \geq 0\text{:} \quad & \text{descent direction} \quad r^t\,, \\ & \alpha_t = -\frac{(g^t, r^t)_2}{(Ar^t, r^t)_2}\,, \\ & x^{t+1} = x^t + \alpha_t r^t\,, \quad g^{t+1} = g^t - \alpha_t Ar^t\,. \end{aligned}$$

The so-called "Coordinate Relaxation" uses descent directions r^t, which are obtained by cyclicing through the Cartesian unit vectors $\{e^1,\dots,e^n\}$. Verify that a full n-cycle of this method is equivalent to one step of the Gauß-Seidel iteration

$$\hat{x}^1 = D^{-1}b - D^{-1}(L\hat{x}^1 + Rx^0).$$

Exercise 3.12: The minimal squared-defect solution of an overdetermined linear system $Ax = b$ is characterized as solution of the normal equation

$$A^TAx = A^Tb.$$

The square matrix A^TA is symmetric and also positive definite, provided A has full rank. Formulate the CG method for solving the normal equation without explictly computing the matrix product A^TA. How many matrix-vector products with A are necessary per iteration (compared to the CG method applied to $Ax = b$)? Relate the convergence speed of this iteration to the singular values of the matrix A.

Exercise 3.13: For solving a linear system $Ax = b$ with symmetric positive definite coefficient matrix A one may use the Gauß-Seidel, the (optimal) SOR method, the gradient mathod, or the CG methods. Recall the estimates for the asymptotic convergence speed of these iterations expressed in terms of the spectral condition number $\kappa = \mathrm{cond}_2(A)$ and compare the corresponding performance results.

In order to derive convergence estimates for the Gauß-Seidel and (optimal) SOR method, assume that A is consistently ordered and that the spectral radius of the Jacobi iteration matrix is given by

$$\mathrm{spr(J)} = 1 - \frac{1}{\kappa}.$$

Discuss the pros and cons of the considered methods.

Exercise 3.14: Consider the symmetric "saddle point system" from Exercise 3.10

$$\begin{bmatrix} A & B \\ B^T & O \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} b \\ c \end{bmatrix},$$

with a symmetric positive definite matrix $A \in \mathbb{R}^{n\times n}$ and a not necessarily quadratic matrix $B \in \mathbb{R}^{n\times m}$, $m \le n$ with full rank. The coefficient matrix cannot be positive definite since some of its main diagonal elements are zero.

A classical approach of solving this saddle-point system is based on the equivalent "Schur-complement formulation":

$$B^T A^{-1} B y = B^T A^{-1} b - c, \quad x = A^{-1} b - A^{-1} B y,$$

in which the solution component y can be computed independently of x. The matrix $B^T A^{-1} B$ is called the "Schur complement" of A in the full block matrix.

In Exercise 3.10 it was shown that a symmetric variant of the Gauß-Seidel method with Chebyshev-acceleration can be applied to this system. However, this approach suffers from the severe drawback that $B^T A^{-1} B$ has to be explicitly known in order to construct the decomposition

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

Verify that, in contrast, the CG method applied to the Schur complement method does not suffer from this defect, i. e. that an explicit construction of A^{-1} can be avoided. Formulate the CG algorithm for above Schur complement and explain how to efficiently treat the explicit occurence of A^{-1} in the algorithm.

Exercise 3.15: For the gradient method and the CG method for a symmetric, positive definite matrix A there hold the error estimates

$$\|x^t_{\text{grad}} - x\|_A \le \Big(\frac{1 - 1/\kappa}{1 + 1/\kappa}\Big)^t \|x^0_{\text{grad}} - x\|_A,$$

$$\|x^t_{\text{cg}} - x\|_A \le 2\Big(\frac{1 - 1/\sqrt{\kappa}}{1 + 1/\sqrt{\kappa}}\Big)^t \|x^0_{\text{cg}} - x\|_A,$$

with the condition number $\kappa := \mathrm{cond}_2(A) = \lambda_{\max}/\lambda_{\min}$. Show that for reducing the initial error by a factor ε the following numbers of iteration are required:

$$t_{\text{grad}}(\varepsilon) \approx \tfrac{1}{2}\kappa \ln(1/\varepsilon), \qquad t_{\text{cg}}(\varepsilon) \approx \tfrac{1}{2}\sqrt{\kappa} \ln(2/\varepsilon).$$

Exercise 3.16: The SSOR preconditioning of the CG method for a symmetric, positive definite matrix A with the usual additive decomposition $A = L + D + L^T$ uses the parameter dependent matrix

$$C := \frac{1}{2-\omega}\Big(\frac{1}{\omega}D + L\Big)\Big(\frac{1}{\omega}D\Big)^{-1}\Big(\frac{1}{\omega}D + L^T\Big), \quad \omega \in (0,2).$$

Write this matrix in the form $C = KK^T$ with a regular, lower-triangular matrix K and explain why C^{-1} may be viewed as an approximation to A^{-1}.

Exercise 3.17: The model matrix $A \in \mathbb{R}^{n\times n}$, $n = m^2$, originating from the 5-point discretization of the Poisson problem on the unit square,

$$A = \left.\begin{bmatrix} B & -I & & \\ -I & B & -I & \\ & -I & B & \ddots \\ & & \ddots & \ddots \end{bmatrix}\right\} n \qquad B = \left.\begin{bmatrix} 4 & -1 & & \\ -1 & 4 & -1 & \\ & -1 & 4 & \ddots \\ & & \ddots & \ddots \end{bmatrix}\right\} m,$$

possesses an important property (of “nonnegative type” or a regular ”Z-matrix”):

$$a_{ii} > 0, \quad a_{ij} \le 0, \;\; i \ne j.$$

Show that the inverse $A^{-1} = (a_{ij}^{(-1)})_{i,j=1}^n$ has nonnegative elements $a_{ij}^{(-1)} \ge 0$, i. e., A is a so-called “M-matrix” (“(inverse) monotone” matrix). This implies that the solution x of a linear system $Ax = b$ with nonnegative right-hand side b, $b_i \ge 0$, is also nonnegative $x_i \ge 0$. (Hint: consider the Jacobi matrix $J = -D^{-1}(L+R)$ and the representation of the inverse $(I-J)^{-1}$ as a Neumann series.)

Exercise 3.18: In the text, we formulated the sequence of iterates $\{x^t\}_{t\ge1}$ of the CG-method formally as the solution x^t of the optimization problem

$$Q(x^t) = \min_{y\in x^0+K_t(d^0;A)} Q(y) \quad \leftrightarrow \quad \|Ax^t - b\|_{A^{-1}} = \min_{y\in x^0+K_t(d^0;A)} \|Ax^t - b\|_{A^{-1}},$$

with the Krylow spaces $K_t(d^0;A) = \operatorname{span}\{d^0, Ad^0, \cdots, A^{t-1}d^0\}$. The so called “Generalized minimal residual method” (GMRES), instead, formally constructs a sequence of iterates $\{x^t_{\mathrm{gmres}}\}_{t\ge1}$ by

$$\|Ax^t_{\mathrm{gmres}} - b\|_2 = \min_{y\in x^0+K_t(d^0;A)} \|Ay - b\|_2.$$

i) Prove that the GMRES method allows for an error inequality similar to the one that was derived for the CG method:

$$\|Ax^t_{\mathrm{gmres}} - b\|_2 \le \min_{p\in P_t, p(0)=1} \|p(A)\|_2 \, \|Ax^0 - b\|_2,$$

where P_t denotes the space of polynomials up to order t.

ii) Prove that in case of A being a symmetric, positive definite matrix, this leads to the same asymptotic convergence rate as for the CG method.

iii) Show that the result obtained in (i) can also be applied to the case of A being similar to a diagonal matrix $D = \mathrm{diag}_i(\lambda_i) \in \mathbb{C}^{n\times n}$, i. e.,

$$A = TDT^{-1},$$

with a regular matrix T. In this case there holds

$$\|x^t_{\mathrm{gmres}} - x\|_2 \le \kappa_2(T) \min_{p\in P_t, p(0)=1} \max_i |p(\lambda_i)|\, \|x^0 - x\|_2.$$

What makes this result rather cumbersome in contrast to the case of a symmetric, positive matrix discussed in (ii)?

Remark: The advantage of the GMRES method lies in the fact that it is, in principle, applicable to any regular matrix A. However, good convergence estimates for the general case are hard to prove.

Exercise 3.19: Repeat the analysis of the convergence properties of the various solution methods for the 3-dimensional version of the model problem considered in the text. The underlying boundary value problem has the form

$$-\Big(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\Big)u(x,y,z) = f(x,y,z), \quad (x,y,z) \in \Omega = (0,1)^3 \in \mathbb{R}^3,$$
$$u(x,y,z) = 0, \quad (x,y,z) \in \partial\Omega,$$

and the corresponding difference approximation (so-called "7-point approximation") at interior mesh points $(x,y,z) \in \{P_{ijk},\, i,j,k = 1,\dots,m\}$, reads

$$-h^{-2}\big(U(x\pm h,y,z) + U(x,y\pm h,z) + U(x,y,z\pm h) - 6U(x,y,z)\big) = f(x,y,z).$$

Using again row-wise numbering of the mesh points the resulting linear system for the mesh values $U_{ijk} \approx u(P_{ijk})$ takes the form

$$A = \underbrace{\begin{bmatrix} B & -I_{m^2} & \\ -I_{m^2} & B & \ddots \\ & \ddots & \ddots \end{bmatrix}}_{n=m^3} \qquad B = \underbrace{\begin{bmatrix} C & -I_m & \\ -I_m & C & \ddots \\ & \ddots & \ddots \end{bmatrix}}_{m^2} \qquad C = \underbrace{\begin{bmatrix} 6 & -1 & \\ -1 & 6 & \ddots \\ & \ddots & \ddots \end{bmatrix}}_{m}$$

In this case the corresponding eigenvalues and eigenvectors are explicitly given by

$$\lambda_{ijk} = 6 - 2\,\big(\cos[ih\pi] + \cos[jh\pi] + \cos[kh\pi]\big), \quad i,j,k = 0,\dots,m,$$
$$w^{ijk} = \big(\sin[pih\pi]\,\sin[qjh\pi]\,\sin[rkh\pi]\big)_{p,q,r=1}^m.$$

For the exact solution $u(x,y,z) = \sin(\pi x)\sin(\pi y)\sin(\pi z)$ there holds the error estimate

$$\max_{\Omega} |U_{ijk}) - u(P_{ijk})| \leq \frac{\pi^4}{8}\, h^2 + O(h^4),$$

which dictates a mesh size $h = 10^{-2}$ in order to guarantee a desired relative discretization accuracy of $\mathrm{TOL} = 10^{-3}$.

a) Determine formulas for the condition number $\mathrm{cond}_2(A)$ and the spectral radius $\mathrm{spr}(J)$ in terms of the mesh size h.

b) Give the number of iterations of the Jacobi, Gauß-Seidel and optimal SOR method as well as the gradient and CG method approximately needed for reducing the initial error to size $\varepsilon = 10^{-4}$ (including a small safety factor).

c) Give a rough estimate (in terms of h) of the total number of a. op. per iteration step for the methods considered.