Chapter 3

Classical Iterative Schemes

3.1 General Theory

Remark 3.1 *Basic idea, transform to a fixed point equation.* The construction of a classical iterative scheme for solving (1.1) starts with the decomposition

$$A = M - N, \quad M, N \in \mathbb{R}^{n\times n}, \quad M \text{ is non-singular}, \tag{3.1}$$

of the system matrix A. Using this decomposition, (1.1) can be transformed into the fixed point equation

$$M\mathbf{x} = \mathbf{b} + N\mathbf{x} \quad \Longleftrightarrow \quad \mathbf{x} = M^{-1}\left(\mathbf{b} + N\mathbf{x}\right). \tag{3.2}$$

Given an initial iterate $\mathbf{x}^{(0)} \in \mathbb{R}^n$, one can try to solve (3.2) with the fixed point iteration

$$\mathbf{x}^{(k+1)} = M^{-1}\left(\mathbf{b} + N\mathbf{x}^{(k)}\right), \quad k = 0, 1, 2, \ldots. \tag{3.3}$$

Banach's[1] fixed point theorem gives information on the convergence of this iteration. □

Theorem 3.2 Banach's fixed point theorem. *Let $(\mathcal{X}, d)$ be a complete metric space and let $f : \mathcal{X} \to \mathcal{X}$ be a contraction (f is Lipschitz[2] continuous with the Lipschitz constant $L < 1$). Then, the equation $x = f(x)$ possesses a unique solution $\bar{x} \in \mathcal{X}$ (a fixed point). The iterative scheme*

$$x^{(k+1)} = f\left(x^{(k)}\right), \quad k = 0, 1, 2, \ldots$$

converges to $\bar{x}$ for any initial iterate $x^{(0)} \in \mathcal{X}$.

Proof: Basic course on calculus. ■

Theorem 3.3 Condition on the iteration matrix of (3.3) **for convergence.** *The iterative scheme* (3.3) *converges to the solution* $\mathbf{x}$ *of* (1.1) *for any initial iterate* $\mathbf{x}^{(0)}$ *if and only if the spectral radius of the iteration matrix $G = M^{-1}N$ is smaller than one: $\rho(G) < 1$.*

Proof: i) The iteration (3.3) is a fixed point iteration with

$$f : \mathbb{R}^n \to \mathbb{R}^n, \quad \mathbf{x} \mapsto M^{-1}N\mathbf{x} + M^{-1}\mathbf{b}.$$

[1] Stefan Banach (1892 - 1945)
[2] Rudolf Lipschitz (1832 - 1903)

The operator $G = M^{-1}N$ is linear and bounded since $\|G\|_*$ is finite, where $\|\cdot\|_*$ is any matrix norm. Hence, G is continuous and even Lipschitz continuous. Since f is continuously differentiable, the Lipschitz constant is given by

$$L_* = \sup_{\mathbf{x}\in\mathbb{R}^n} \|J(f(\mathbf{x}))\|_* = \sup_{\mathbf{x}\in\mathbb{R}^n} \|G\|_* = \|G\|_*,$$

where $J(f(\mathbf{x}))$ is the (constant) Jacobian of $f(\mathbf{x})$.

ii) Let $\rho(G) < 1$. Then, it is possible to find a matrix norm $\|\cdot\|_*$ such that, according to Lemma 2.8, $\|G\|_* \leq \rho(G) + \varepsilon < 1$ with $\varepsilon > 0$. Hence $L_* < 1$ and $f(\mathbf{x})$ is a contraction.

iii) Let $\rho(G) \geq 1$. An initial guess will be constructed for which the fixed point iteration does not converge. Without loss of generality, consider the case $\mathbf{b} = \mathbf{0}$ such that the solution of (1.1) is $\mathbf{x} = \mathbf{0}$.

Since $\rho(G) \geq 1$, there is an eigenvalue $\lambda \in \mathbb{C}$ of G with $|\lambda| \geq 1$. The eigenvalue can be written in the form $\lambda = |\lambda|(\cos(\varphi) + i\sin(\varphi))$ where φ is the argument of λ. Let $\mathbf{z} \in \mathbb{C}^n, \mathbf{z} \neq \mathbf{0}$, be a corresponding eigenvector: $G\mathbf{z} = \lambda\mathbf{z}$. From the conjugate of this equation $\overline{G\mathbf{z}} = \overline{\lambda\mathbf{z}}$ it follows that $G\overline{\mathbf{z}} = \overline{\lambda}\overline{\mathbf{z}}$ since G is a real matrix.

Choose the initial iterate $\mathbf{x}^{(0)} = \mathbf{z} + \overline{\mathbf{z}} \in \mathbb{R}^n$. One has to exclude that $\mathbf{x}^{(0)} = \mathbf{0}$. If $\mathbf{x}^{(0)} = \mathbf{0}$ then $\mathbf{z} = i\mathbf{v}$ with $\mathbf{v} \in \mathbb{R}^n$. One obtains from the eigenvalue equation that $iG\mathbf{v} = i\lambda\mathbf{v}$ which is equivalent to $G\mathbf{v} = \lambda\mathbf{v}$. On the left hand side of this equation there is a real vector. Since $\mathbf{v}$ is a real vector, it follows that λ must be real, too. But in this case, the corresponding eigenvector is also real and it cannot be of form $\mathbf{z} = i\mathbf{v}$. Hence, an eigenvector of form $\mathbf{z} = i\mathbf{v}$ cannot appear and $\mathbf{x}^{(0)} \neq \mathbf{0}$.

Now, it follows that

$$\underbrace{G\left(G\left(\cdots G\mathbf{x}^{(0)}\right)\right)}_{k \text{ times}} = G^k\mathbf{x}^{(0)} = G^k\mathbf{z} + G^k\overline{\mathbf{z}} = \lambda^k\mathbf{z} + \overline{\lambda}^k\overline{\mathbf{z}} = 2\mathrm{Re}\left(\lambda^k\mathbf{z}\right),\ k = 0, 1, \ldots.$$

The iteration converges if

$$\begin{aligned}\mathbf{0} = &\lim_{k\to\infty} 2\mathrm{Re}\left(\lambda^k\mathbf{z}\right) = \lim_{k\to\infty} 2|\lambda|^k \mathrm{Re}\left((\cos(k\varphi) + i\sin(k\varphi))\,\mathbf{z}\right)\\ = &\lim_{k\to\infty} 2|\lambda|^k \left(\cos(k\varphi)\,\mathrm{Re}(\mathbf{z}) - \sin(k\varphi)\,\mathrm{Im}(\mathbf{z})\right).\end{aligned}$$

The factor $|\lambda|^k$ does not converge to zero since $|\lambda| \geq 1$. Note that the second factor is a vector. It converges to zero if and only if each of its components converges to zero. There is at least one component z_l with $z_l \neq 0$ since $\mathbf{z}$ is an eigenvector. Let ζ be the argument of z_l. It is

$$\begin{aligned}\cos(k\varphi)\,\mathrm{Re}(z_l) - \sin(k\varphi)\,\mathrm{Im}(z_l) &= |z_l|\left(\cos(k\varphi)\cos(\zeta) - \sin(k\varphi)\sin(\zeta)\right)\\ &= |z_l|\cos(k\varphi + \zeta). \qquad (3.4)\end{aligned}$$

If $\lambda \in \mathbb{R}$, i.e. $\varphi = \pm\pi$, then the eigenvector $\mathbf{z}$ is real, too, such that $\zeta = \pm\pi$. In this case, (3.4) takes the values $|z_l|$ or $-|z_l|$ and it does not tend to zero as $k \to \infty$. If $\lambda \notin \mathbb{R}$, then the period φ in (3.4) is not an integer multiple of π such that the argument of the cosine cannot tend to $\pi/2$ plus an integer multiple of π. Hence, also in this case, the second factor in (3.4) does not tend to zero.

In summary, the iteration (3.3) does not converge for the initial iterate $\mathbf{x}^{(0)} = \mathbf{z} + \overline{\mathbf{z}}$. That means, if the iteration (3.3) converges for all initial iterates, then $\rho(G) \geq 1$ cannot hold.

Note: the last part of the proof simplifies much if one considers complex-valued systems of linear equations. Then, one can take the initial iterate $\mathbf{x}^{(0)} = \mathbf{z} \neq \mathbf{0}$, finds that $\mathbf{x}^{(k)} = \lambda^k\mathbf{z}$, and concludes that $\left\|\mathbf{x}^{(k)}\right\|_2 = |\lambda|^k \|\mathbf{z}\|_2 \not\to 0$, since $\|\mathbf{z}\|_2 \neq 0$ and $|\lambda|^k \to \infty$. ■

3.2 Examples for classical iterative schemes

Remark 3.4 *Decomposition of the system matrix.* One uses for the definition of classical iterative schemes a decomposition of the matrix A into the form

$$A = D + L + U,$$

where D is the diagonal of A, L is its strict lower part and U its strict upper part. □

Example 3.5 *Jacobi method.* The Jacobi method is derived by setting

$$M = D, N = -(L+U).$$

A straighforward calculation reveals that the fixed point equation (3.2) has the form

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

This gives the following iterative scheme, called Jacobi[3] method

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + D^{-1}\left(\mathbf{b} - A\mathbf{x}^{(k)}\right), \quad k = 0, 1, 2, \ldots.$$

The iteration matrix is $G_{\text{Jac}} = -D^{-1}(L+U)$. □

Example 3.6 *Damped Jacobi method.* Let $\omega \in \mathbb{R}$, $\omega > 0$. The matrices which define the fixed point equation for the damped Jacobi method are given by

$$M = \omega^{-1}D, N = \omega^{-1}D - A.$$

The damped Jacobi method has the form

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \omega D^{-1}\left(\mathbf{b} - A\mathbf{x}^{(k)}\right), \quad k = 0, 1, 2, \ldots$$

and the iteration matrix is $G_{\text{dJac}} = I - \omega D^{-1}A$. □

Example 3.7 *Gauss–Seidel method.* In the Gauss[4]–Seidel[5] method, the invertible matrix M is a triangular matrix

$$M = D + L, N = -U.$$

It follows that

$$\mathbf{x}^{(k+1)} = (D+L)^{-1}\left(\mathbf{b} - U\mathbf{x}^{(k)}\right), \quad k = 0, 1, 2, \ldots,$$

such that the iteration matrix has the form $G_{\text{GS}} = -(D+L)^{-1}U$. Multiplying the equation for the Gauss–Seidel method by $(D+L)$ and rearranging terms gives the more familiar form of this iteration

$$\begin{aligned}\mathbf{x}^{(k+1)} &= D^{-1}\left(\mathbf{b} - L\mathbf{x}^{(k+1)} - U\mathbf{x}^{(k)}\right)\\ &= \mathbf{x}^{(k)} + D^{-1}\left(\mathbf{b} - L\mathbf{x}^{(k+1)} - (D+U)\mathbf{x}^{(k)}\right), \quad k = 0, 1, 2, \ldots.\end{aligned}$$

Writing this iteration for the components of the vector shows that the right hand side can be evaluated even if the new iterate appears there, since only already computed components of the new iterate are needed for this evaluation

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

□

[3] Carl Gustav Jacob Jacobi (1804 - 1851)
[4] Johann Carl Friedrich Gauss (1777 - 1855)
[5] Philipp Ludwig von Seidel (1821 - 1896)

Example 3.8 *SOR method.* The matrices which define the (forward) successive over relaxation (SOR) method are given by

$$M = \omega^{-1} D + L, N = \omega^{-1} D - (D + U),$$

where $\omega \in \mathbb{R}, \omega > 0$. This method can be written in the form

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \omega D^{-1} \left(\mathbf{b} - L\mathbf{x}^{(k+1)} - (D + U)\, \mathbf{x}^{(k)} \right), \quad k = 0, 1, 2, \ldots.$$

For $\omega = 1$, the Gauss–Seidel method is recovered. One obtains for the iteration matrix

$$\begin{aligned} G_{\mathrm{SOR}}(\omega) &= \left(\omega^{-1} D + L\right)^{-1} \left(\omega^{-1} D - (D + U)\right) \\ &= \omega \left(D + \omega L\right)^{-1} \left(\omega^{-1} D - (D + U)\right) \\ &= \left(D + \omega L\right)^{-1} \left((1 - \omega) D - \omega U\right). \end{aligned}$$

□

Example 3.9 *SSOR method.* In the SOR method, one can change the roles of L and U to obtain the backward SOR method

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \omega D^{-1} \left(\mathbf{b} - U\mathbf{x}^{(k+1)} - (D + L)\, \mathbf{x}^{(k)} \right), \quad k = 0, 1, 2, \ldots.$$

This method updates the unknowns in reverse order. The forward and backward SOR behave in general differently. There are cases in which one of them works much more efficient than the other one. However, in general one does not know a priori which is the better variant. The SSOR (symmetric SOR) combines both methods. One step of SSOR consists of two substeps, one forward SOR and one backward SOR step:

$$\begin{aligned} \mathbf{x}^{(k+1/2)} &= \mathbf{x}^{(k)} + \omega D^{-1} \left(\mathbf{b} - L\mathbf{x}^{(k+1/2)} - (D + U)\, \mathbf{x}^{(k)} \right) \\ \mathbf{x}^{(k+1)} &= \mathbf{x}^{(k)} + \omega D^{-1} \left(\mathbf{b} - U\mathbf{x}^{(k+1)} - (D + L)\, \mathbf{x}^{(k+1/2)} \right), \quad k = 0, 1, 2, \ldots. \end{aligned}$$

□

3.3 Some Convergence Results

Theorem 3.10 Strongly diagonally dominant matrices. *Let $A \in \mathbb{R}^{n \times n}$ be strongly diagonally dominant. Then, the Jacobi method and the Gauss–Seidel method converge for every initial iterate $\mathbf{x}^{(0)} \in \mathbb{R}^n$.*

Proof: Following Theorem 3.3, one has to show that the spectral radius of the iteration matrices is smaller than 1.

Jacobi method. Let $\mathbf{z} \in \mathbb{R}^n, \mathbf{z} \neq \mathbf{0}$. Then, the triangle inequality gives

$$\begin{aligned} \left|\left(G_{\mathrm{Jac}}\mathbf{z}\right)_i\right| &= \left|\left(-D^{-1}(L + U)\, \mathbf{z}\right)_i\right| = \left| \frac{1}{a_{ii}} \sum_{j=1, j\neq i}^{n} a_{ij} z_j \right| \leq \frac{1}{|a_{ii}|} \sum_{j=1, j\neq i}^{n} |a_{ij}|\, |z_j| \\ &\leq \frac{1}{|a_{ii}|} \underbrace{\sum_{j=1, j\neq i}^{n} |a_{ij}|}_{< |a_{ii}|} \|\mathbf{z}\|_\infty < \|\mathbf{z}\|_\infty. \end{aligned}$$

It follows that

$$\rho(G_{\mathrm{Jac}}) \leq \|G_{\mathrm{Jac}}\|_\infty = \max_{\mathbf{z} \in \mathbb{R}^n, \mathbf{z} \neq \mathbf{0}} \frac{\|G_{\mathrm{Jac}}\mathbf{z}\|_\infty}{\|\mathbf{z}\|_\infty} < 1.$$

Gauss–Seidel method. A direct calculation shows (exercise)

$$G_{\mathrm{GS}} = -D^{-1}\left(LG_{\mathrm{GS}} + U\right).$$

This relation gives for the first component and $\mathbf{z} \in \mathbb{R}^n, \mathbf{z} \neq \mathbf{0}$

$$\left|(G_{\mathrm{GS}}\mathbf{z})_1\right| \leq \frac{1}{|a_{11}|}\sum_{j=2}^{n}|a_{1j}|\,|z_j| \leq \frac{1}{|a_{11}|}\underbrace{\sum_{j=2,}^{n}|a_{1j}|}_{<|a_{11}|}\|\mathbf{z}\|_\infty < \|\mathbf{z}\|_\infty,$$

where the term with the factor LG_{GS} vanishes since the first row of L consists only of zeros. Using now the induction $\left|(G_{\mathrm{GS}}\mathbf{z})_j\right| < \|\mathbf{z}\|_\infty$, $j < i$, yields

$$\begin{aligned}\left|(G_{\mathrm{GS}}\mathbf{z})_i\right| &\leq \frac{1}{|a_{ii}|}\left(\sum_{j=1}^{i-1}|a_{ij}|\left|(G_{\mathrm{GS}}\mathbf{z})_j\right| + \sum_{j=i+1}^{n}|a_{ij}|\,|z_j|\right)\\ &\leq \frac{1}{|a_{ii}|}\underbrace{\sum_{j=1,j\neq i}^{n}|a_{ij}|}_{<|a_{ii}|}\|\mathbf{z}\|_\infty < \|\mathbf{z}\|_\infty, \quad i = 2,\ldots,n.\end{aligned}$$

The remainder of the proof is like for the Jacobi method. ■

Lemma 3.11 Eigenvalues of the iteration matrix of the damped Jacobi method. *Let $\omega > 0$, then $\lambda \in \mathbb{C}$ is an eigenvalue of G_{Jac} if and only if $\mu = 1-\omega+\omega\lambda$ is an eigenvalue of G_{dJac}.*

Proof: It is with $A = D + L + U$

$$G_{\mathrm{dJac}} = I - \omega D^{-1}A = I - \omega D^{-1}D - \omega\underbrace{D^{-1}\left(L+U\right)}_{-G_{\mathrm{Jac}}} = (1-\omega)\,I + \omega G_{\mathrm{Jac}}.$$

The statement of the lemma follows now from well known properties of eigenvalues. ■

Example 3.12 *Convergence of the damped Jacobi method where the Jacobi method fails.* If ω is chosen appropriately, there is the possibility that the damped Jacobi method converges for every initial guess whereas the Jacobi method does not.

Assume that G_{Jac} has only real eigenvalues. Denote by $\lambda_{\min}$ the smallest one and by $\lambda_{\max}$ the largest one. If

$$\lambda_{\min} < -1 < \lambda_{\max} < 1,$$

then there are initial iterates for which the Jacobi method does not converge, Theorem 3.3. From Lemma 3.11 one has

$$\mu_{\min} = (1-\omega) + \omega\lambda_{\min}, \quad \mu_{\max} = (1-\omega) + \omega\lambda_{\max}.$$

It follows that

$$-1 < \mu_{\min} \text{ if } \omega < \frac{2}{1-\lambda_{\min}} < 1, \qquad \mu_{\max} < 1 \text{ if } \omega > 0.$$

The choice of $\omega \in (0, 2/(1-\lambda_{\min}))$ ensures the convergence of the damped Jacobi method for each initial iterate.

Consider the case $\lambda_{\max} > 1$. Then

$$\mu_{\max} = (1-\omega) + \omega\lambda_{\max} = 1 + \omega(\lambda_{\max} - 1) > 1.$$

In this case, there are initial iterates for which the damped Jacobi method does not converge as well. □

Lemma 3.13 ***Parameter in the case that the SOR method converges, Lemma of Kahan***[6]**.** *If the SOR method converges for every initial iterates* $\mathbf{x}^{(0)} \in \mathbb{R}^n$ *then* $\omega \in (0,2)$.

Proof: Let $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ be the eigenvalues of $G_{\mathrm{SOR}}(\omega)$. It is

$$\begin{aligned}
\prod_{i=1}^{n} \lambda_i &= \det\left(G_{\mathrm{SOR}}(\omega)\right) = \det\left((D+\omega L)^{-1}\left((1-\omega)D - \omega U\right)\right) \\
&= \det\Big(\underbrace{(D+\omega L)^{-1}}_{\text{lower triangular matrix}}\Big) \det\Big(\underbrace{(1-\omega)D - \omega U}_{\text{upper triangular matrix}}\Big) \\
&= \det\left(D^{-1}\right)(1-\omega)^n \det(D) = (1-\omega)^n .
\end{aligned}$$

Hence

$$\prod_{i=1}^{n} |\lambda_i| = |1-\omega|^n .$$

There is at least one eigenvalue λ_i with $|\lambda_i| \geq |1-\omega|$ and it follows that $\rho\left(G_{\mathrm{SOR}}(\omega)\right) \geq |1-\omega|$. The application of Theorem 3.3 shows now that SOR cannot converge for all initial iterates if $\omega \notin (0,2)$, because if $\omega \notin (0,2)$ then $\rho\left(G_{\mathrm{SOR}}(\omega)\right) \geq |1-\omega| \geq 1$. ■

Theorem 3.14 Convergence of SOR for s.p.d. matrices. *Let* $A \in \mathbb{R}^{n\times n}$ *be s.p.d. Then the SOR method converges for all initial iterates* $\mathbf{x}^{(0)} \in \mathbb{R}^n$ *if* $\omega \in (0,2)$.

Proof: Let $\lambda \in \mathbb{C}$ be an eigenvalue of $G_{\mathrm{SOR}}(\omega)$ and let $\mathbf{z} \in \mathbb{C}^n$ be the corresponding eigenvector, i.e.

$$(D+\omega L)^{-1}\left((1-\omega)D - \omega U\right)\mathbf{z} = \lambda \mathbf{z}.$$

Following Theorem 3.3, one has to find a condition such that $|\lambda| < 1$. The following identities can be easily verified

$$\begin{aligned}
D + \omega L &= \left(1 - \frac{\omega}{2}\right) D + \frac{\omega}{2} A + \frac{\omega}{2}\left(L - U\right), \\
(1-\omega)D - \omega U &= \left(1 - \frac{\omega}{2}\right) D - \frac{\omega}{2} A + \frac{\omega}{2}\left(L - U\right).
\end{aligned}$$

Inserting these identities into the eigenvalue equation and multiplying this equation from the left hand side with the adjoint vector $\mathbf{z}^*$, one obtains

$$\lambda = \frac{\left(1 - \frac{\omega}{2}\right)\mathbf{z}^* D\mathbf{z} - \frac{\omega}{2}\mathbf{z}^* A\mathbf{z} + \frac{\omega}{2}\mathbf{z}^*\left(L-U\right)\mathbf{z}}{\left(1 - \frac{\omega}{2}\right)\mathbf{z}^* D\mathbf{z} + \frac{\omega}{2}\mathbf{z}^* A\mathbf{z} + \frac{\omega}{2}\mathbf{z}^*\left(L-U\right)\mathbf{z}}.$$

Now, the terms in this expression will be considered individually. The matrix $L-U$ is skew-symmetric since A is symmetric. It follows for all $\mathbf{x} \in \mathbb{R}^n$ that

$$\underbrace{\mathbf{x}^T(L-U)\mathbf{x}}_{\in \mathbb{R}} = \left(\mathbf{x}^T(L-U)\mathbf{x}\right)^T = \mathbf{x}^T\left(L-U\right)^T\mathbf{x} = -\mathbf{x}^T(L-U)\mathbf{x} \quad \in \mathbb{R},$$

consequently $\mathbf{x}^T(L-U)\mathbf{x} = 0$ for all $\mathbf{x} \in \mathbb{R}^n$, and

$$\mathrm{Re}\left(\mathbf{z}^*\left(L-U\right)\mathbf{z}\right) = \mathrm{Re}\left(\mathbf{z}^*\right)(L-U)\mathrm{Re}\left(\mathbf{z}\right) + \mathrm{Im}\left(\mathbf{z}^*\right)(L-U)\mathrm{Im}\left(\mathbf{z}\right) = 0.$$

Hence $\mathbf{z}^*\left(L-U\right)\mathbf{z} = ia$ with $a \in \mathbb{R}$. Since A is positive definite, its diagonal D is positive definite, too. The products $\mathbf{z}^* D\mathbf{z}$ and $\mathbf{z}^* A\mathbf{z}$ are positive real numbers since for $\mathbf{z} = \mathbf{u} + i\mathbf{v}, \mathbf{z}^* = \mathbf{u}^T - i\mathbf{v}^T$, $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, $\mathbf{z} \neq \mathbf{0}$ because it is an eigenvector, one obtains with the symmetry of A

$$\begin{aligned}
\mathbf{z}^* A\mathbf{z} &= \mathbf{u}^T A\mathbf{u} - i\mathbf{v}^T A\mathbf{u} + i\mathbf{u}^T A\mathbf{v} - i^2\mathbf{v}^T A\mathbf{v} \\
&= \mathbf{u}^T A\mathbf{u} - i\mathbf{v}^T A\mathbf{u} + i\mathbf{v}^T A\mathbf{u} + \mathbf{v}^T A\mathbf{v} > 0.
\end{aligned}$$

[6] William M. Kahan, born 1933

It follows that λ has the form

$$\lambda = \frac{b+ia}{c+ia} \quad a,b,c \in \mathbb{R}$$

with

$$b = \left(1-\frac{\omega}{2}\right)\mathbf{z}^* D\mathbf{z} - \frac{\omega}{2}\mathbf{z}^* A\mathbf{z}, \quad c = \left(1-\frac{\omega}{2}\right)\mathbf{z}^* D\mathbf{z} + \frac{\omega}{2}\mathbf{z}^* A\mathbf{z}.$$

Consequently

$$|\lambda|^2 = \frac{b^2+a^2}{c^2+a^2} = \frac{\left[\left(1-\frac{\omega}{2}\right)\mathbf{z}^* D\mathbf{z} - \frac{\omega}{2}\mathbf{z}^* A\mathbf{z}\right]^2 + a^2}{\left[\left(1-\frac{\omega}{2}\right)\mathbf{z}^* D\mathbf{z} + \frac{\omega}{2}\mathbf{z}^* A\mathbf{z}\right]^2 + a^2}.$$

Thus $|\lambda| < 1$ only if the numerator is smaller than the denominator. This is equivalent to

$$\begin{aligned} -\underbrace{\omega}_{>0}\left(1-\frac{\omega}{2}\right)\underbrace{\mathbf{z}^* D\mathbf{z}}_{>0}\underbrace{\mathbf{z}^* A\mathbf{z}}_{>0} &< \omega\left(1-\frac{\omega}{2}\right)\mathbf{z}^* D\mathbf{z}\,\mathbf{z}^* A\mathbf{z} \quad \Longleftrightarrow \\ -\left(1-\frac{\omega}{2}\right) &< \left(1-\frac{\omega}{2}\right) \quad \Longleftrightarrow \\ \omega &< 2. \end{aligned}$$

Hence, the SOR method converges for all initial iterates if $\omega \in (0,2)$. ■

Remark 3.15 *Difficulty of choosing ω in practice.* For choosing ω such that the SOR method converges as fast as possible, one needs information about the eigenvalues of A. However, the computation of these information is very costly, see Numerical Mathematics I. □

Remark 3.16 *Number of iterations in practice.* If classical iterative schemes are used for the solution of linear systems of equations which arise in discretizing partial differential equations, one finds that the number of iterations to fulfill a certain stopping criterion rapidly increases. One can show that the number of iteration depends on the condition number of the matrices and it scales linearly with the condition number. As example, the standard finite element discretization of the Laplace equation on an equidistant grid of size h leads to matrices with a condition number of $\mathcal{O}\left(h^{-2}\right)$, see homework problems. It follows that the number of iterations for the solution of the linear system increases approximately by the factor 4 with each refinement $h \to h/2$. For this reason, the classical iterative schemes are not useful as solver for such systems. They are important as preconditioner or as smoother in multigrid methods, see Chapter 9. □