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

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

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

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

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 $\hat{x} \in \mathcal{X}$ (a fixed point). The iterative scheme*

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

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

[1] Stefan Banach (1892 – 1945)

[2] Rudolf Lipschitz (1832 – 1903)

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 $\underline{x}$ of (1.1) for any initial iterate $\underline{x}^{(0)}$ 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

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

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

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

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

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 $f(\underline{x})$ is a contraction. Now, the statement follows with Theorem 3.2.

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 $\underline{b} = \underline{0}$ such that the solution of (1.1) is $\underline{x} = \underline{0}$.

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

$$\lambda = |\lambda|(\cos(\varphi) + i\sin(\varphi)), \tag{3.4}$$

where φ is the argument of λ. Let $\underline{z} \in \mathbb{C}^n, \underline{z} \neq \underline{0}$, be a corresponding eigenvector:

$$G\underline{z} = \lambda\underline{z}. \tag{3.5}$$

From the conjugate of this equation $\overline{G\underline{z}} = \overline{\lambda\underline{z}}$, it follows that $G\overline{\underline{z}} = \overline{\lambda}\overline{\underline{z}}$ since G is a real matrix.

Choose the initial iterate $\underline{x}^{(0)} = \underline{z} + \overline{\underline{z}} \in \mathbb{R}^n$.

One has to exclude that $\underline{x}^{(0)} = \underline{0}$. If $\underline{x}^{(0)} = \underline{0}$, then $\underline{z} = i\underline{v}$ with $\underline{v} \in \mathbb{R}^n$. One obtains from the eigenvalue equation that $iG\underline{v} = i\lambda\underline{v}$ which is equivalent to (3.5). On the left-hand side of the latter equation, there is a real vector. Since $\underline{v}$ is a real vector, it follows that λ must be real, too. But in this case, the corresponding eigenvector $\underline{z}$ is also real and it cannot be of the form $\underline{z} = i\underline{v}$. Hence, an eigenvector of form $\underline{z} = i\underline{v}$ cannot exist and $\underline{x}^{(0)} \neq \underline{0}$.

Using the definition of the initial iterate, the eigenvalue problem (3.5), and some basic properties of eigenvalues and complex numbers, it follows that

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

The iteration converges to the solution $\underline{x} = \underline{0}$ if

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

where (3.4) and basic properties of the real part of complex numbers were used. The factor $|\lambda|^k$ is always larger or equal to 1, since $|\lambda| \geq 1$. Hence, the second factor has to converge

to zero if the iteration should converge to $\underline{x} = \underline{0}$. 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 $\underline{z}$ is an eigenvector. Let ζ be the argument of z_l. Using the definition of the real and imaginary part of z_l and the trigometric identity for $\cos(\alpha + \beta)$ yields

$$\begin{aligned} \cos(k\varphi)\,\mathrm{Re}(z_l) - \sin(k\varphi)\,\mathrm{Im}(z_l) &= |z_l|\,(\cos(k\varphi)\cos(\zeta) - \sin(k\varphi)\sin(\zeta)) \\ &= |z_l|\cos(k\varphi + \zeta)\,. \end{aligned} \tag{3.6}$$

The only possibility to obtain convergence to zero for the periodic cosine function and for fixed increment φ is the case $\zeta = \pm\pi/2$ and $\varphi \in \{0, \pi\}$, because then $k\varphi + \zeta$ is $\pi/2$ plus an integer multiple of π. In this case, it follows that $\lambda \in \mathbb{R}$ and $\mathrm{Re}(z_l) = 0$. However, if $\lambda \in \mathbb{R}$, then the eigenvector $\underline{z}$ is real, too, which is a contradiction to $\mathrm{Re}(z_l) = 0$.

In summary, the iteration (3.3) does not converge for the initial iterate $\underline{x}^{(0)} = \underline{z} + \overline{\underline{z}}$. 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 $\underline{x}^{(0)} = \underline{z} \neq \underline{0}$, finds that $\underline{x}^{(k)} = \lambda^k \underline{z}$, and concludes that $\left\|\underline{x}^{(k)}\right\|_2 = |\lambda|^k \|\underline{z}\|_2 \not\to 0$, since $\|\underline{z}\|_2 \neq 0$ and $|\lambda|^k \geq 1$.

However, the considered situation is not a special case of the complex-valued case since in the considered situation the initial iterate must be a real vector. ■

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 in 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.

It turns out that the presented classical iterative schemes require that the diagonal of A does not contain a zero entry. □

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

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

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

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

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

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

[3] Carl Gustav Jacob Jacobi (1804 – 1851)

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 that define the fixed point equation for the damped Jacobi method are given by

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

The damped Jacobi method has the form

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

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

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

It follows that

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

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

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

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

$$x_i^{(k+1)} = x_i^{(k)} + \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i}^{n} a_{ij} x_j^{(k)} \right), \quad k = 0, 1, 2, \ldots .$$

□

[4] Johann Carl Friedrich Gauss (1777 – 1855)

[5] Philipp Ludwig von Seidel (1821 – 1896)

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

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

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

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

or

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

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

$$\begin{aligned} G_{\mathrm{SOR}}(\omega) &= \left(\omega^{-1} D + L \right)^{-1} \left(\omega^{-1} D - (D + U) \right) \\ &= \omega \left(D + \omega L \right)^{-1} \left(\omega^{-1} D - (D + U) \right) \\ &= \left(D + \omega L \right)^{-1} \left((1 - \omega) D - \omega U \right). \end{aligned} \tag{3.7}$$

□

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

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

This method updates the unknowns in reverse order. The forward and backward SOR behave in general differently. There are cases where 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) method combines both methods. One step of SSOR consists of two substeps, one forward SOR and one backward SOR step:

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

$k = 0, 1, 2, \ldots$. □

3.3 Some Convergence Results

Theorem 3.10. Convergence for strongly diagonally dominant matrices. *Let $A \in \mathbb{R}^{n \times n}$ be strongly diagonally dominant. Then, the Ja-*

cobi method and the Gauss–Seidel method converge for every initial iterate $\underline{x}^{(0)} \in \mathbb{R}^n$.

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

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

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

which is equivalent to $\|G_{\mathrm{Jac}}\underline{z}\|_\infty < \|\underline{z}\|_\infty$. With Remark 2.7 and the definition of $\|\cdot\|_\infty$, Remark 2.3, it follows that

$$\rho(G_{\mathrm{Jac}}) \leq \|G_{\mathrm{Jac}}\|_\infty = \max_{\underline{z}\in\mathbb{R}^n,\underline{z}\neq\underline{0}}\frac{\|G_{\mathrm{Jac}}\underline{z}\|_\infty}{\|\underline{z}\|_\infty} < 1. \tag{3.8}$$

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

$$G_{\mathrm{GS}} = -D^{-1}(LG_{\mathrm{GS}} + U)\,. \tag{3.9}$$

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

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

where the term with the factor LG_{GS} vanishes since the first row of L consists only of zeros. Using now the induction $\left|(G_{\mathrm{GS}}\underline{z})_j\right| < \|\underline{z}\|_\infty$, $j < i$, and (3.9) yields

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

The remainder of the proof is like for the Jacobi method, compare (3.8). ∎

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 $A = D + L + U$

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

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, compare Theorem 3.3. From Lemma 3.11, one has

$$\begin{aligned} \mu_{\min} &= (1-\omega) + \omega\lambda_{\min} = 1 - \omega\,(1-\lambda_{\min})\,, \\ \mu_{\max} &= (1-\omega) + \omega\lambda_{\max} = 1 - \omega\,(1-\lambda_{\max})\,. \end{aligned} \tag{3.10}$$

It follows that

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

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

Consider the case $\lambda_{\max} > 1$. Then, one gets from (3.10) that $\mu_{\max} > 1$. 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 $\underline{x}^{(0)} \in \mathbb{R}^n$, 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 with (3.7) and properties of the determinant

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

Hence, it follows that

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

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

[6] William M. Kahan, born 1933

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 $\underline{x}^{(0)} \in \mathbb{R}^n$ if $\omega \in (0,2)$.*

Proof. Let $\lambda \in \mathbb{C}$ be an eigenvalue of $G_{\mathrm{SOR}}(\omega)$, see (3.7), and let $\underline{z} \in \mathbb{C}^n$ be a corresponding eigenvector, i.e.,

$$(D+\omega L)^{-1}\left((1-\omega)D-\omega U\right)\underline{z} = \lambda \underline{z}. \tag{3.11}$$

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

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

Inserting these identities in the eigenvalue equation (3.11) and multiplying this equation from the left-hand side with $(D+\omega L)$ and with the adjoint vector $\underline{z}^*$, one obtains

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

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 $\underline{x} \in \mathbb{R}^n$ that

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

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

$$\mathrm{Re}\left(\underline{z}^*(L-U)\underline{z}\right) = \underbrace{\mathrm{Re}\left(\underline{z}^*\right)}_{\in\mathbb{R}^n}(L-U)\underbrace{\mathrm{Re}\left(\underline{z}\right)}_{\in\mathbb{R}^n} + \underbrace{\mathrm{Im}\left(\underline{z}^*\right)}_{\in\mathbb{R}^n}(L-U)\underbrace{\mathrm{Im}\left(\underline{z}\right)}_{\in\mathbb{R}^n} = 0.$$

Hence, $\underline{z}^*(L-U)\underline{z} = ia$ with $a \in \mathbb{R}$.

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

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

It follows that λ has the form

$$\lambda = \frac{b+ia}{c+ia} \quad a,b,c \in \mathbb{R},\ b,c > 0.$$

with

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

Consequently, it is

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

Thus $|\lambda| < 1$ holds only if the numerator is smaller than the denominator. The only difference of the numerator and the denominator is the mixed term in the square. Thus, this condition is equivalent to

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

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 or the literature. □

Remark 3.16. Number of iterations in practice. If classical iterative schemes are used for the solution of linear systems of equations that arise in discretizing partial differential equations, one finds that the number of iterations to fulfill a certain stopping criterion rapidly increases if the mesh is refined. 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 difference 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 8. □