Chapter 7
Generation of Iterations

Abstract The algebraic operations described in Chapter 5 are tools for generating linear iterations. In this chapter we discuss how these tools can be used to build new iterative methods. The product of iterations is recalled in Section 7.1 and refers to later applications in Part III. Many traditional iterations are constructed by the *additive splitting* technique of Section 7.2. The *regular splitting* and *weakly regular splitting* defined in §7.2.2 yield sufficient convergence criteria. Another kind of splitting is the *P-regular splitting* defined in §7.2.4. A special kind of additive splitting is the *incomplete triangular decomposition (ILU)* discussed in Section 7.3. The transformations introduced in §5.6 will reappear in Section 7.4 under the name *preconditioning*.

7.1 Product Iterations

We recall that new iterations can be constructed by the product of simpler ones:

$$\Pi := \Phi \circ \Psi \qquad \text{for } \Phi, \Psi \in \mathcal{L}.$$

Of particular interest are *symmetric iterations*. If Φ is not symmetric, it can be symmetrised: $\Phi^{\mathrm{sym}} := \Phi^* \circ \Phi$ (also $\Phi \circ \Phi^*$ would be possible). The Krylov methods of Part II are best to combine with *positive definite iterations*, for which $A > 0$ implies $N[A] > 0$.

Symmetric products of three factors will also appear (see, e.g., Lemma 11.44). Corollary 5.30 states that $\Phi^* \circ \Psi \circ \Phi$ is symmetric if Ψ is so. The corresponding statement about positive definiteness follows. Note that Criterion 5.10 yields a criterion for $\Phi^* \circ \Phi$ to be positive definite.

Lemma 7.1. *If $\Phi^* \circ \Phi \in \mathcal{L}_{\mathrm{pos}}$ and $\Psi \in \mathcal{L}_{\mathrm{semi}}$, then the product satisfies*

$$\Phi^* \circ \Psi \circ \Phi \in \mathcal{L}_{\mathrm{pos}}.$$

W. Hackbusch, *Iterative Solution of Large Sparse Systems of Equations*,
Applied Mathematical Sciences 95, DOI 10.1007/978-3-319-28483-5_7

Proof. Let $A > 0$. One verifies that

$$N_{\Phi^* \circ \Psi \circ \Phi} = M_{\Phi^*} N_\Psi M_{\Phi^*}^{\mathsf{H}} + N_{\Phi^* \circ \Phi} .$$

Positive semidefiniteness of Ψ yields $N_\Psi \geq 0$ and $M_{\Phi^*} N_\Psi M_{\Phi^*}^{\mathsf{H}} \geq 0$, while $N_{\Phi^* \circ \Phi} > 0$ follows since $\Phi^* \circ \Phi$ is positive definite. □

In §12 we shall produce iterations from A-orthogonal projections.

Definition 7.2. $\Phi \in \mathcal{L}$ is called an *A-orthogonal projection* if $\mathfrak{D}(\Phi) \ni A > 0$ implies that the matrix $A^{1/2} N[A]\, A^{1/2}$ is an orthogonal projection.

An orthogonal projection has a spectrum contained in $\{0, 1\}$. For our purpose, the following generalisation is sufficient:

$$\Phi \in \mathcal{L}_{\mathrm{sym}} \text{ with } \sigma(N[A]\, A) \subset [0, 2). \tag{7.1}$$

Definition 7.3. The iteration $\Phi(\cdot, \cdot, A) \in \mathcal{L}$ is called *nonexpansive* (with respect to an associated norm $\|\cdot\|$) if

$$\|M_\Phi[A]\| \leq 1.$$

Exercise 7.4. $A > 0$ and (7.1) imply that Φ is nonexpansive with respect to $\|\cdot\|_A$.

Lemma 7.5. *Assume that $A > 0$. Let $\Phi_i \in \mathcal{L}$ satisfy (7.1) for $1 \leq i \leq k$. Then the product iteration*

$$\Pi(\cdot, \cdot, A) := \Phi_k(\cdot, \cdot, A) \circ \ldots \circ \Phi_2(\cdot, \cdot, A) \circ \Phi_1(\cdot, \cdot, A)$$

converges if and only if (5.10) holds (cf. Proposition 5.23).

Proof. (i) Let $x \in \bigcap_{i=1}^k \ker(N_{\Phi_i})$. For an indirect proof, assume $x \neq 0$ and set $y := A^{-1} x \neq 0$. Since $x \in \ker(N_{\Phi_1})$, $y = M_{\Phi_1} y$ holds for the iteration matrix $M_{\Phi_1} = I - N_{\Phi_1} A$. By $y = M_{\Phi_2} y$, etc., we obtain $y = (M_\Pi) y$ for the iteration matrix $M_\Pi = \prod_{i=1}^k M_{\Phi_i}$ of $\Pi(\cdot, \cdot, A)$. The eigenvalue 1 of M_Π proves divergence of Π. Hence, convergence implies (5.10).

(ii) Assume that (5.10) holds and define

$$\hat{M}_i := A^{1/2} M_{\Phi_i} A^{-1/2} = I - A^{1/2} N_{\Phi_i} A^{1/2} \quad \text{and} \quad \hat{M}_\Pi := \prod_{i=1}^k \hat{M}_i .$$

The product iteration Π converges monotonically with respect to the energy norm if $\|\hat{M}_\Pi\|_2 < 1$. By (7.1), $\sigma(\hat{M}_i) \subset (-1, 1]$ and $\|\hat{M}_i x\|_2 \leq \|x\|_2$ hold for all $x \in \mathbb{K}^I$. In addition, $\|\hat{M}_i x\|_2 = \|x\|_2$ is equivalent to $A^{-1/2} x \in \ker(N_{\Phi_i})$. As a consequence $\|\hat{M}_\Pi x\|_2 \leq \|x\|_2$ holds for all $x \in \mathbb{K}^I$ and $\|\hat{M}_\Pi x\|_2 = \|x\|_2$ implies $A^{-1/2} x \in \bigcap_{i=1}^k \ker(N_{\Phi_i}[A])$. The assumption (5.10) yields $x = 0$. Hence $\|M_\Pi\|_A = \|\hat{M}_\Pi\|_2 < 1$ follows. □

7.2 Additive Splitting Technique

7.2.1 Definition and Examples

Most of the classical iterations are constructed by an additive[1] splitting as explained below. Given the system of equations

$$Ax = b \qquad (A \in \mathbb{K}^{I\times I},\ b \in \mathbb{K}^I), \tag{7.2}$$

we split A into the difference

$$A = W - R \qquad (W \text{ regular}). \tag{7.3}$$

The system (7.2) is equivalent to

$$Wx = Rx + b.$$

This suggests the iterative method

$$Wx^{m+1} = Rx^m + b \tag{7.4}$$

which is well defined since W is required to be regular.

Lemma 7.6. *(a) Assume (7.3). Then the iterative method (7.4) is consistent. The matrices of the first normal form (2.8) are*

$$M = W^{-1}R, \qquad N = W^{-1}.$$

The notation 'W' for the matrix in (7.3) is chosen because the third normal form (2.12),

$$W(x^m - x^{m+1}) = Ax^m - b,$$

is valid with the same matrix W.

(b) Vice versa, any iteration $\Phi \in \mathcal{L}$ with regular N can be obtained from an additive splitting (7.3).

Proof. (a) A comparison of the representation

$$x^{m+1} = W^{-1}Rx^m + W^{-1}b$$

derived from (7.4) with (2.8) shows that $M = W^{-1}R$ and $W = N^{-1}$.

(b) Choose $W := N_\Phi[A]^{-1}$ and $R := W - A$ in (7.3). □

Because of Lemma 7.6b, the additive splitting technique does not produce a special class of iterations but *all* linear iterations. This is a similar situation as the combination of the Richardson iteration Φ_1^{Rich} with a right transformation $T_\ell = N$

[1] The term 'additive' distinguishes this technique from the multiplicative factorisation in §7.3.

(cf. Proposition 5.44). In the case of the additive splitting, $W = N^{-1}$ is the primary quantity, whereas in the latter case, N determines the transformation.

Remark 7.7. The fact that the additive splitting can generate any linear iteration leads to the question: what are the data on which the choice of W can be based? The following cases can be distinguished:

(i) The choice is only based on the data of the matrix A. This means that there is an explicitly available mapping $A \mapsto W[A]$ or $A \mapsto N[A]$. In this case, the iteration is *algebraic* (cf. Definition 2.2b).

(ii) The matrix A may be the result of a discretised partial differential equation. Correspondingly, additional data of the partial differential equation not contained in the matrix data (e.g., geometric data, coarser discretisations, etc.) can be used for constructing W.

(iii) An intermediate situation between (i) and (ii) is the following one. The element matrices $B = \{B^{(\nu)} : \nu \in J\}$ introduced in §E.3 contain more data than A. Therefore a mapping $B \mapsto W[B]$ may be well defined, but cannot be obtained from A (cf. Remark E.8b).

In §7.4.5 we shall give an example for case (ii). There the proposed matrix W cannot be derived from the matrix A.

A typical example of cases (ii) or (iii) are domain decomposition iterations involving submatrices discretising Neumann boundary problems in subdomains (cf. §12.3). These subproblems lead to matrices A_1 and A_2 such that $A = A_1 + A_2$. Obviously, A is a result of A_1 and A_2, but these matrices cannot be determined from A.

All splittings discussed in this section and in §7.3 correspond to the case (i) of Remark 7.7.

Example 7.8. (a) A natural choice of W is some part of the matrix A. The splitting $A = D - (A - D)$ with the diagonal $W = D$ of A yields the Jacobi iteration.

(b) Starting from the splitting $A = D - E - F$ in (1.16), we choose $W = D - E$ and $R = F$. The resulting iteration (7.4) is the Gauss–Seidel iteration. Alternatively, the choice $W = D - F$ and $R = E$ yields the backward Gauss–Seidel method (cf. Proposition 5.1).

(c) Using the blockwise version of $A = D_{\mathrm{block}} - E_{\mathrm{block}} - F_{\mathrm{block}}$ in (3.19a–d), the respective splitting yields the block-Jacobi and block-Gauss–Seidel iterations.

Note that in the previous examples the matrices D, $D - E$, $D_{\mathrm{block}} - E_{\mathrm{block}}$ contain increasing parts of the matrix A. In Theorem 7.13 and §7.2.3 we shall see that this fact may improve the convergence. On the other hand, W must still be (easily) invertible, since we have to solve the system (2.12′).

The additive splitting can be combined with the summation introduced in §5.3 and yields the *multi-splitting method* (cf. O'Leary–White [296]).

7.2.2 Regular Splittings

In this section we shall make use of M-matrices (cf. §C.3). Accordingly, we use the notation

$$A < B, \quad A \leq B, \quad x < y, \quad x \leq y, \quad \dots$$

for matrices and vectors in the sense of *componentwise inequalities*. In particular, $A > 0$ denotes a positive matrix, not a positive definite one.

The following definition of a 'regular splitting' is due to Varga [375]. It allows not only qualitative convergence statements but also a comparison of different iterative methods.

Definition 7.9 (regular splitting). The real matrix $W \in \mathbb{R}^{I \times I}$ describes a *regular splitting* of $A \in \mathbb{R}^{I \times I}$ if

$$W \text{ regular}, \quad W^{-1} \geq 0, \quad W \geq A \text{ (i.e., } R := W - A \geq 0). \tag{7.5}$$

Condition (7.5) may be compared with (3.35g) in the positive definite case. The iteration matrix of the iteration (7.4) is

$$M = W^{-1}R \qquad \text{with} \quad R := W - A$$

(cf. Lemma 7.6a). Condition (7.5) implies that

$$M \geq 0 \qquad \text{for regular splittings} \tag{7.6}$$

because of $R \geq 0$. Using (7.6), we can weaken Definition 7.9 (cf. Ortega [298]).

Definition 7.10 (weakly regular splitting). The splitting (7.3) is *weakly regular* if

$$W \text{ regular}, \quad W^{-1} \geq 0, \quad M = W^{-1}R \geq 0. \tag{7.7}$$

Theorem 7.11 (convergence). *Let A be inverse positive: $A^{-1} > 0$ (a sufficient condition is that A be an M-matrix). Assume that W describes a weakly regular splitting of A. Then the induced iteration (7.4) converges:*

$$\rho(M) = \rho(W^{-1}R) = \frac{\rho(A^{-1}R)}{1 + \rho(A^{-1}R)} < 1. \tag{7.8}$$

Proof. (i) Obviously, it is sufficient to show $\rho(W^{-1}R) = \rho(C)/(1 + \rho(C))$ for $C := A^{-1}R$. The weak regularity (7.7) implies that

$$\begin{aligned} 0 \leq M = W^{-1}R &= [A^{-1}W]^{-1}A^{-1}R \\ &= [A^{-1}(A + R)]^{-1}A^{-1}R = [I + C]^{-1}C. \end{aligned}$$

By Theorem C.34 and $M \geq 0$, there is an eigenvector $x \gneqq 0$ belonging to the eigenvalue $\lambda = \rho(M) \in \sigma(M)$. Rewriting $\lambda x = Mx = (I + C)^{-1}Cx$, we obtain

$$\lambda x + \lambda Cx = Cx \tag{7.9a}$$

The value $\lambda = 1$ is excluded, since (7.9a) would yield $x = 0$. Hence,

$$Cx = \tfrac{\lambda}{1-\lambda} x \tag{7.9b}$$

follows. In part (iii) we shall show that $C \geq 0$. Equation (7.9b), together with $x \gneqq 0$ and $Cx \geq 0$, ensures the inequality $\frac{\lambda}{1-\lambda} \geq 0$, i.e., $0 \leq \lambda = \rho(M) < 1$.

(ii) (7.9b) proves that λ is an eigenvalue of M if and only if $\mu = \frac{\lambda}{1-\lambda}$ is an eigenvalue of C. The inequality $0 \leq \lambda < 1$ shows that $\mu \geq 0$. Since $\mu = \frac{\lambda}{1-\lambda}$ increases monotonically in λ, $|\mu| = \mu$ is maximal for $\lambda = \rho(M) \in \sigma(M)$. By Theorem C.34, $\mu = \rho(C) \in \sigma(C)$ is the maximal eigenvalue of C; therefore we have $\rho(C) = \rho(M)/[1 - \rho(M)]$. Solving this equation for $\rho(M)$, we arrive at assertion (7.8): $\rho(M) = \rho(C)/[1 + \rho(C)]$.

(iii) From

$$0 \leq \left[\sum_{\nu=0}^{m-1} M^\nu\right] W^{-1}, \quad W^{-1} = (I - M)A^{-1}, \text{ and}$$

$$\sum_{\nu=0}^{m-1} M^\nu (I - M) = I - M^m,$$

we conclude that

$$0 \leq (I - M^m)A^{-1} \leq A^{-1} \qquad \text{and} \qquad 0 \leq M^m A^{-1} \leq A^{-1}.$$

Therefore, M^m is bounded. This fact proves that $\kappa = \rho(M) < 1$. Since $\lambda = 1$ is already excluded, $\rho(M) < 1$ holds and implies

$$C = A^{-1}R = [W(I - M)]^{-1}R = (I - M)^{-1}W^{-1}R = \left[\sum_{\nu=0}^{\infty} M^\nu\right] M \geq 0. \qquad \square$$

It might be expected that the iteration converges faster the closer W is to A, i.e., the smaller the remainder $R = W - A$ is. This property is stated more precisely in the following *comparison theorem.*

Theorem 7.12. *Let A be inverse positive: $A^{-1} \geq 0$. Let W_1 and W_2 define two regular splittings. If W_1 and W_2 are comparable in the sense of*

$$A \leq W_1 \leq W_2\,, \tag{7.10a}$$

then the convergence rates satisfy the corresponding inequalities

$$0 \leq \rho(M_1) \leq \rho(M_2) < 1, \quad \textit{where } M_i := W_i^{-1}R_i, \quad R_i := W_i - A. \tag{7.10b}$$

Proof. The matrices $B := A^{-1}R_1$ and $C := A^{-1}R_2$ satisfy $0 \le B \le C$ and therefore $0 \le \rho(B) \le \rho(C)$ (cf. (C.15)). From representation (7.8) we obtain

$$0 \le \rho(M_1) = \rho(B)/[1+\rho(B)] \le \rho(C)/[1+\rho(C)] = \rho(M_2) < 1. \qquad \square$$

The comparisons in (7.10a,b) can be strengthened into strict inequalities.

Theorem 7.13. *From* $A^{-1} > 0$ *and*

$$A \lneqq W_1 \lneqq W_2, \qquad W_1, W_2 : \text{ regular splittings,} \tag{7.11a}$$

the strict inequalities

$$0 < \rho(M_1) < \rho(M_2) < 1 \quad \text{with } M_i := W_i^{-1}R_i\,, \;\; R_i := W_i - A \tag{7.11b}$$

follows.

Proof. Define B and C as in the previous proof. Since $B = A^{-1}R_1$ may be reducible, Theorem C.25 is not directly applicable. $R_1 \ge 0$ holds since the splitting is regular. Define

$$I_+ := \{\beta \in I : R_{1,\alpha\beta} > 0 \text{ for some } \alpha \in I\} \quad \text{and} \quad I_0 := I \backslash I_+.$$

Any column $s = (R_{1,\alpha\beta})_{\alpha\in I}$ of R_1 corresponding to an index $\beta \in I_+$ satisfies $s \gneqq 0$ and therefore $A^{-1}s > 0$ by Exercise C.20b. Hence, B has the form

$$B = \begin{bmatrix} B_1 & 0 \\ B_2 & 0 \end{bmatrix} \qquad \text{with positive blocks} \quad B_1 > 0 \quad \text{and} \quad B_2 > 0$$

corresponding to the block structure $\{I_+, I_0\}$. In particular,

$$\rho(B) = \rho(B_1) > 0 \tag{7.11c}$$

holds (cf. (C.11a)). Because of $R_2 - R_1 = W_2 - W_1 \gneqq 0$, there is a pair (α, β) with $(R_2 - R_1)_{\alpha\beta} > 0$. Hence, the column of $C - B = A^{-1}(R_2 - R_1)$ for the index β is positive. Assume $\beta \in I_+$. In this case, $C_1 \gneqq B_1$ and $C_2 \gneqq B_2$ hold for the blocks in $C = \begin{bmatrix} C_1 & C_3 \\ C_2 & C_4 \end{bmatrix}$. Lemma C.30 and (C.11c) yield the inequality

$$\rho(C) \ge \rho(C_1) > \rho(B_1).$$

In the remaining case of $\beta \in I_0$, we conclude that

$$C_3 \gneqq B_3 = 0, \qquad C_4 \gneqq B_4 = 0,$$

and

$$\rho(C) > \rho(C_1) \ge \rho(B_1)$$

(cf. Lemma C.30). In any case, using (7.11c), we arrive at the strict inequality $\rho(C) > \rho(B) > 0$, which via (7.8) leads us to the assertion. $\square$

7.2.3 Applications

Theorem 7.14. *Let A be an M-matrix. Then the point- and blockwise Jacobi iterations converge. Moreover, the blockwise iteration is faster:*

$$\rho(M^{\mathrm{blockJac}}) \le \rho(M^{\mathrm{Jac}}) < 1. \tag{7.12a}$$

Let D be the pointwise diagonal D^{ptw} or the block diagonal D^{block} of A. Then

$$D \text{ describes a regular splitting.} \tag{7.12b}$$

Assuming explicitly (7.12b), we may replace the assumption 'A is an M-matrix' by the inverse positivity: $A^{-1} \ge 0$. The strict inequality

$$0 < \rho(M^{\mathrm{blockJac}}) < \rho(M^{\mathrm{Jac}}) < 1$$

holds instead of (7.12a) if $A^{-1} > 0$ and $D^{\mathrm{ptw}} \neq D^{\mathrm{block}} \neq A$.

Proof. For an M-matrix A, the diagonals $D = D^{\mathrm{ptw}}$ and $D = D^{\mathrm{block}}$ satisfy the inequality $D > A$ and the sign condition (C.18b). By Theorem C.53, D is again an M-matrix, so that $D^{-1} \ge 0$ and (7.12b) follow. Because of $D^{\mathrm{ptw}} \ge D^{\mathrm{block}}$, Theorem 7.12 proves inequality (7.12a). Concerning the strict inequality, compare with Theorem 7.13. □

Theorem 7.15. *Split $A = D - E - F$ according to (3.11a–d) or (3.19a–d). The statements of Theorem 7.14 carry over to analogous ones for the pointwise and blockwise Gauss–Seidel iteration, where the statements (7.12a,b) become*

$$\rho(M^{\mathrm{blockGS}}) \le \rho(M^{\mathrm{GS}}) < 1, \qquad D - E \text{ describes a regular splitting.}$$

We omit the proof, since it is completely analogous to the previous one. The comparison between the Jacobi and Gauss–Seidel iteration is more interesting. The quantitative relation $\rho(M^{\mathrm{GS}}) = \rho(M^{\mathrm{Jac}})^2$, which according to Conclusion 4.30 holds for consistent orderings, can no longer be shown for the general case. However, a corresponding qualitative statement derived from $D - E \le D$ is valid.

Theorem 7.16. *For an M-matrix A, the following inequalities hold:*

$$\rho(M^{\mathrm{GS}}) \le \rho(M^{\mathrm{Jac}}) < 1, \quad \rho(M^{\mathrm{blockGS}}) \le \rho(M^{\mathrm{blockJac}}) < 1.$$

This statement can be generalised to other than M-matrices.

Theorem 7.17 (Stein–Rosenberg [352]). *Exactly one of the following four alternatives holds for the pointwise Jacobi and Gauss–Seidel iterations if A fulfils the sign condition (C.18b), $a_{\alpha\beta} \le 0$ for $\alpha \neq \beta$:*

$$\begin{aligned} 0 &= \rho(M^{\mathrm{GS}}) = \rho(M^{\mathrm{Jac}}), \\ 0 &< \rho(M^{\mathrm{GS}}) < \rho(M^{\mathrm{Jac}}) < 1, \\ \rho(M^{\mathrm{GS}}) &= \rho(M^{\mathrm{Jac}}) = 1, \\ \rho(M^{\mathrm{GS}}) &> \rho(M^{\mathrm{Jac}}) > 1. \end{aligned}$$

In particular, both methods converge or diverge simultaneously. The statement of the theorem remains valid if M^{Jac} *and* M^{GS} *are replaced by* $L+U$ *and* $(I-L)^{-1}U$ *with* $L \geq 0$ *being an arbitrary, strictly lower triangular matrix and* $U \geq 0$ *a strictly upper one.*

The proof can be found in Varga [375, §3.3] or in the original paper [352]. For generalisations, see Buoni–Varga [88, 89].

In the case of *over*relaxation (i.e., for $\omega > 1$), the SOR iteration does not lead to a regular splitting. To ensure regularity of the splitting, we have to restrict the parameter ω to $0 < \omega < 1$ (underrelaxation).

Exercise 7.18. Prove that the SOR iteration arises from a splitting (7.3) with $W = \omega^{-1}D - E$. Let A be an M-matrix and D its diagonal. For $0 < \omega \leq 1$, the matrix W describes a regular splitting. What conclusion can be drawn from $\omega^{-1}D - E \geq D - E$?

In the case of a regular splitting, the property (7.7) (i.e., $M \geq 0$) allows an enclosure of the solution $x = A^{-1}b$, provided that we find suitable starting iterates.

Theorem 7.19. *Let* $M \geq 0$ *be the iteration matrix of a convergent iteration. Starting with initial iterates* x^0 *and* y^0 *satisfying*

$$x^0 \leq x^1, \qquad x^0 \leq y^0, \qquad y^1 \leq y^0,$$

we obtain iterates x^m *and* y^m *with the enclosure property*

$$x^0 \leq x^1 \leq \ldots \leq x^m \leq \ldots \leq x = A^{-1}b \leq \ldots \leq y^m \leq \ldots \leq y^1 \leq y^0.$$

Proof. It follows from the estimates $x^{m+1} - x^m = M^m(x^1 - x^0) \geq 0$, and $y^m - y^{m+1} = M^m(y^0 - y^1) \geq 0$, $y^m - x^m = M^m(y^0 - x^0) \geq 0$ (cf. (2.16b)).□

We recall the generalisation of the M-matrices by the H-matrices in Definition C.60 and the definition of diagonal dominance in §C.3.3.

Theorem 7.20. *Each of the following conditions (7.13a,b) is sufficient for the convergence of the pointwise Jacobi and Gauss–Seidel iterations:*

$$A \text{ is an H-matrix,} \tag{7.13a}$$

$$A \text{ is strictly, irreducibly, or essentially diagonally dominant.} \tag{7.13b}$$

Exercise 7.21. Prove that (7.13b) implies (7.13a) and $\|M^{\mathrm{Jac}}\|_\infty \leq 1$, $\|M^{\mathrm{GS}}\|_\infty \leq 1$.

Proof. (i) The case (7.13b) is reduced to (7.13a) because of Exercise 7.21.

(ii) Define $B := |D| - |A - D|$ as in Definition C.60 and denote the iteration matrix of the Jacobi iteration for B by $M_B^{\mathrm{Jac}} := |D|^{-1}|A - D|$. Theorem 7.16 yields $\rho(M_B^{\mathrm{Jac}}) < 1$. By $|M_B^{\mathrm{Jac}}| = M_B^{\mathrm{Jac}}$, the convergence $\rho(M^{\mathrm{Jac}}) < 1$ follows from the next lemma, which remains to be proved.

Lemma 7.22. $\rho(A) \le \rho(|A|)$ *for all* $A \in \mathbb{C}^{I\times I}$.

(iii) Split $A = D - E - F$ according to (3.11a–d) and define $L := D^{-1}E$, $U := D^{-1}F$. Since $B = |D| - |E| - |F| = |D|(I - |L| - |U|)$, the iteration matrices belonging to A and B are:

$$M^{\mathrm{GS}} = (I-L)^{-1}U = \sum_{\nu=0}^{\infty} L^\nu U, \quad M_B^{\mathrm{GS}} = (I-|L|)^{-1}|U| = \sum_{\nu=0}^{\infty} |L|^\nu |U|$$

(cf. Lemma A.13). Hence, $|M^{\mathrm{GS}}| = |\sum_{\nu=0}^{\infty} L^\nu U| \le \sum_{\nu=0}^{\infty} |L|^\nu |U| = M_B^{\mathrm{GS}}$. From Lemma 7.22 and Theorem 7.15, we conclude that $\rho(M^{\mathrm{GS}}) \le \rho(M_B^{\mathrm{GS}}) < 1$. □

Proof of Lemma 7.22. By $\|A^\nu\|_\infty = \| \, |A^\nu| \, \|_\infty \le \| \, |A|^\nu \, \|_\infty$, Theorem B.27 yields

$$\rho(A) = \lim_{\nu\to\infty} \|A^\nu\|_\infty^{1/\nu} \le \lim_{\nu\to\infty} \| \, |A|^\nu \, \|_\infty^{1/\nu} = \rho(|A|). \qquad \square$$

The diagonal dominance in (7.13b) is often used as a convergence criterion since the proof becomes very simple. Strict diagonal dominance is historically the first convergence criterion for the Jacobi iteration (see the paper of R. von Mises and H. Pollaczek-Geiringer [381, Satz 2] from 1929).

Proposition 7.23. *If the strict diagonal dominance (C.16) can be quantified by a number* $q > 1$ *such that*

$$|a_{\alpha\alpha}| \ge q \sum_{\beta\in I\setminus\{\alpha\}} |a_{\alpha\beta}| \qquad \textit{for all } \alpha \in I, \tag{7.14a}$$

then the Jacobi and Gauss–Seidel iterations converge monotonically with respect to the maximum norm with the contraction numbers

$$\|M^{\mathrm{Jac}}\|_\infty, \|M^{\mathrm{GS}}\|_\infty \le 1/q < 1. \tag{7.14b}$$

Proof. Using (7.14a), the estimate of $M^{\mathrm{Jac}} = D^{-1}(A-D)$ by $\|M^{\mathrm{Jac}}\|_\infty \le 1/q$ follows immediately from (B.8).

In the Gauss–Seidel case, we use the description of the iteration by (1.15). The components of the error $e^m = x^m - x$ satisfy

$$e_i^{m+1} = -\left(\sum_{j=1}^{i-1} a_{ij} e_j^{m+1} + \sum_{j=i+1}^{n} a_{ij} e_j^m \right) / a_{ii}.$$

Induction on i yields $\|e^{m+1}\|_\infty \le \|e^m\|_\infty / q$. Since $e^{m+1} = M^{\mathrm{GS}} e^m$, the inequality (7.14b) follows. □

Concerning the convergence of the SSOR iteration for H-matrices, we refer to Alefeld–Varga [3] and Neumaier–Varga [289].

7.2.4 P-Regular Splitting

The P-regular splitting defined below is of different nature. In particular, it is based on the order relation of positive definite matrices (cf. §C.1). The term 'P-regular' is introduced by Ortega [298], but the following convergence statement goes back to Weissinger [392] in 1953 (see also Weissinger [391]).

Lemma 7.24. *Let X be any general matrix, while Z is positive definite; i.e., $Z > 0$. Then $Z - X^{\mathsf{H}}ZX > 0$ implies that*

$$\rho(X) < 1 \text{ and } \|Z^{1/2}XZ^{-1/2}\|_2 < 1.$$

Proof. Set $Y := Z^{1/2}XZ^{-1/2}$. Multiplying $Z - X^{\mathsf{H}}ZX > 0$ by $Z^{-1/2}$ from both sides yields $I - Y^{\mathsf{H}}Y > 0$ or $Y^{\mathsf{H}}Y < I$ (cf. (C.3a′)). Hence $\|Y\|_2^2 = \rho(Y^{\mathsf{H}}Y) < \rho(I) = 1$ proves the last statement. Since X and Y are similar matrices, $\rho(X) = \rho(Y) \le \|Y\|_2$ proves $\rho(X) < 1$. □

Definition 7.25. The splitting $A = W - R$ is called *P-regular* if W is regular and the Hermitian part $\frac{1}{2}(C + C^{\mathsf{H}})$ of $C := W + R$ is positive definite.

The last condition can be written as $0 < \frac{1}{2}(C + C^{\mathsf{H}}) = \frac{1}{2}(W + W^{\mathsf{H}} + R + R^{\mathsf{H}}) = W + W^{\mathsf{H}} - \frac{1}{2}(A + A^{\mathsf{H}})$, i.e.,

$$W + W^{\mathsf{H}} > \frac{1}{2}(A + A^{\mathsf{H}}) =: \hat{A}. \tag{7.15}$$

Theorem 7.26 (Weissinger [392]). *Assume $A + A^{\mathsf{H}} > 0$ and consider a P-regular splitting $A = W - R$. The corresponding iteration (7.4) converges monotonically with respect to the norm $\|\cdot\|_{\hat{A}}$ with $\hat{A}$ defined in (7.15):*

$$\rho(M) \le \|M\|_{\hat{A}} < 1 \qquad \text{for } M = I - W^{-1}A. \tag{7.16}$$

Proof. The splitting $A = W - R$ yields the iteration matrix $M = W^{-1}R$. Note that

$$\begin{aligned} A - M^{\mathsf{H}}AM &= A - (I - W^{-1}A)^{\mathsf{H}}A(I - W^{-1}A) \\ &= (W^{-1}A)^{\mathsf{H}}A + AW^{-1}A - (W^{-1}A)^{\mathsf{H}}A(W^{-1}A) \\ &= (W^{-1}A)^{\mathsf{H}}(W + W^{\mathsf{H}} - A)(W^{-1}A) =: B. \end{aligned}$$

Forming the expression $\frac{1}{2}(B + B^{\mathsf{H}})$ and using $\hat{A}$ in (7.15), we arrive at

$$\hat{A} - M^{\mathsf{H}}\hat{A}M = (W^{-1}A)^{\mathsf{H}}(W + W^{\mathsf{H}} - \hat{A})(W^{-1}A) > 0$$

because of (7.15). Lemma 7.24 with $Z := \hat{A}$ yields (7.16). □

7.3 Incomplete Triangular Decompositions

One learns from Theorem 7.13 that the convergence speed of Jacobi and Gauss–Seidel iterations could be improved if even larger parts of the matrix A were contained in W. The practical obstacle is that we must be able to solve the system $W\delta = d$ efficiently. In particular, this requirement seems to exclude splittings with W containing larger portions of A than the lower and upper triangular parts. However, if we are able to decompose W into triangular factors[2]

$$W = LU \qquad (L \text{ lower triangular, } U \text{ upper triangular matrix}),$$

the solution of $LU\delta = d$ can easily be performed using the forward and backward substitution (cf. Quarteroni–Sacco–Saleri [314, §3.2]).

Therefore, we are looking for a suitable matrix $W = LU$. In general, $W = A$ is not a good candidate since its LU decomposition leads to a fill-in, i.e., to larger nonzero parts of the matrix. In the case of sparse factors L, U with $A \neq W = LU$, this factorisation is called an *incomplete LU decomposition* of A and abbreviated as ILU.

Besides the use of ILU as a linear iteration (possibly accelerated by techniques of Part II), ILU is also of interest as smoothing iteration of the multigrid method (cf. §11.9.2).

7.3.1 Introduction and ILU Iteration

In the following, the index set I is ordered. Here the standard choice in the model case is the lexicographical ordering. By Conclusion 1.11, the LU decomposition $A = LU$ has proved to be inappropriate for sparse matrices, since the factors L and U contain many more nonzero entries than the original matrix A. Computing the LU decomposition is completely identical to Gauss elimination: U is the upper triangular matrix remaining after eliminating the entries below the diagonal, whereas L contains the elimination factors $L_{ji} = a_{ji}^{(i)} / a_{ii}^{(i)}$ $(j \geq i)$ (cf. Quarteroni–Sacco–Saleri [314, §3.3]). Instead of computing L and U by Gauss elimination, we may determine the $n^2 + n$ unknown entries L_{ji}, U_{ij} $(j \geq i)$ directly from the n normalisation conditions

$$L_{ii} = 1 \qquad (1 \leq i \leq n) \tag{7.17a}$$

and the n^2 equations involved in $A = LU$:

$$\sum_{j=1}^{n} L_{ij} U_{jk} = A_{ik} \qquad (1 \leq i, k \leq n). \tag{7.17b}$$

[2] In this section, L and U are general (nonstrict) triangular matrices and do not coincide with the matrices L, U defined in (3.15d).

The incomplete LU decomposition is based on the idea of not eliminating all matrix entries of A to avoid the fill-in of the matrix during the elimination process. Since, after an *incomplete elimination*, entries remain in the lower triangular part, an exact solution of the system is not possible. Instead, the previous equality $A = LU$ holds up to remainder R:

$$A = LU - R. \tag{7.18}$$

For the exact description of the ILU process, we choose a subset $E \subset I \times I$ of the product of the ordered index set $I = \{1, 2, \ldots, n\}$. The elimination is restricted to the pairs $(i, j) \in E$. Concerning E, we always require

$$(i, i) \in E \qquad \text{for all } i \in I. \tag{7.19a}$$

In general, one should choose E large enough, so that the graph $G(A)$ of A is contained in E (cf. Definition C.12):

$$G(A) \subset E. \tag{7.19b}$$

E is called the *(elimination) pattern* of the ILU decomposition. Examples of E will be given in §7.3.2. Through the definition of the triangular matrices, we have

$$L_{ij} = U_{ji} = 0 \qquad \text{for } 1 \le i < j \le n. \tag{7.20a}$$

To construct *sparse* matrices L and U, nonzero entries are allowed only at positions of the pattern E; otherwise, we require

$$L_{ij} = U_{ij} = 0 \qquad \text{for } (i, j) \notin E. \tag{7.20b}$$

Exercise 7.27. Prove that there are $\#E$ matrix entries of L and U which are not directly determined by (7.17a), (7.20a), and (7.20b).

In analogy to (7.17b), we pose $\#E$ equations for the same number of unknowns:

$$\sum_{j=1}^{n} L_{ij} U_{jk} = A_{ik} \qquad \text{for all } (i, k) \in E. \tag{7.20c}$$

The remainder $R = LU - A$ is obtained from (7.20d,e):

$$R_{ik} = 0 \qquad \text{for all } (i, k) \in E, \tag{7.20d}$$

$$R_{ik} = \sum_{j=1}^{n} L_{ij} U_{jk} - A_{ik} \qquad \text{for all } (i, k) \notin E. \tag{7.20e}$$

Under assumption (7.19b), the term A_{ik} in 7.20e may be omitted because of $A_{ik} = 0$.

The ILU factors satisfying (7.17a) and (7.20a–c) can, e.g., be constructed by the following algorithm:

$L := 0;\ U := 0;$
for $i := 1$ **to** n **do**
begin $L_{ii} := 1;$
  for $k := 1$ **to** $i-1$ **do if** $(i,k) \in E$ **then** $L_{ik} := \frac{A_{ik} - \sum' L_{ij} U_{jk}}{U_{kk}};$ (7.21a)
  for $k := 1$ **to** i **do if** $(k,i) \in E$ **then** $U_{ki} := A_{ki} - \sum'' L_{kj} U_{ji}$ (7.21b)
end;

The sums $\sum'$ and $\sum''$ are taken over all j with $j \neq k$. Since all indices referring to vanishing terms can be omitted, we may write:

$$\Sigma' = \sum_{j \in I \text{ with } j<k,\ (i,j)\in E,\ (j,k)\in E}, \quad \Sigma'' = \sum_{j \in I \text{ with } j<k,\ (k,j)\in E,\ (j,i)\in E}.$$

The definition of L_{ik} in (7.21a) is obtained from (7.20c). To prove (7.21b), interchange i and k in (7.20c). One verifies that only those components of L and U are involved in the right-hand sides of (7.21a,b) that are already computed. Remark 7.28 will enable a simplification of the algorithm.

Remark 7.28. The definitions $D := \mathrm{diag}\{U\},\ U' := U - D,\ L' := (L - I)D$ lead to a strictly lower triangular matrix L' and a strictly upper triangular matrix U'. Equation (7.18) rewritten with the new quantities becomes

$$A = (D + L')D^{-1}(D + U') - R. \tag{7.22}$$

The quantities D, L', and U' are the result of the following algorithm:

$D := 0;\ L' := 0;\ U' := 0;$
for $i := 1$ **to** n **do**
begin
  for $k := 1$ **to** $i-1$ **do if** $(i,k) \in E$ **then** $L'_{ik} := A_{ik} - \sum' L'_{ij} D_{jj}^{-1} U'_{jk};$ (7.23a)
  for $k := 1$ **to** $i-1$ **do if** $(k,i) \in E$ **then** $U'_{ki} := A_{ki} - \sum'' L'_{kj} D_{jj}^{-1} U'_{ji};$ (7.23b)
  $D_{ii} := A_{ii} - \sum'' L'_{ij} D_{jj}^{-1} U'_{ji}$ (7.23c)
end;

Hence, ILU iteration based on L', D, U' is algebraic.

Remark 7.29. (a) If A is Hermitian, (7.23a–c) immediately implies the symmetries $L' = U'^{\mathsf{H}}$ and $D = D^{\mathsf{H}}$.

(b) The incomplete Cholesky decomposition $A = L''L''^{\mathsf{H}} - R$ for positive definite matrices A follows from (7.22) with $L'' := (D + L')D^{-1/2}$.

Tacitly, we assume that the quantities U_{kk} (pivot entries) in (7.21a) and D_{jj} in (7.23a) do not vanish and that, in the case of Remark 7.29b, even $D_{jj} > 0$ holds. Concerning these assumptions, we refer to the analysis in §7.3.5.

Exercise 7.30. Complete LU decompositions are characterised by $R = 0$ in (7.18). Prove: (a) $R = 0$ holds for cases (i) $E = I \times I$ or (ii) $E = \{(i,j) : |i-j| \le w\}$ for band matrices of band width $w \ge 0$.
(b) $D = \mathrm{diag}\{A\}$ and $L' = U' = 0$ hold for the diagonal elimination pattern $E = \{\,(i,i) : i \in I\}$, which is the minimal pattern satisfying (7.19a).

The additive splitting $A = W - R$ of A given by (7.18) or (7.22) defines the corresponding ILU iteration:

$$W(x^m - x^{m+1}) = Ax^m - b \quad \text{with} \tag{7.24a}$$

$$W = LU \quad \text{or} \quad W = (D + L')D^{-1}(D + U'), \text{ respectively.} \tag{7.24b}$$

The matrices of the first and second normal forms are

$$M = NR \quad \text{with} \quad N = U^{-1}L^{-1} \quad \text{or} \quad N = (D + U')^{-1} D\, (D + L')^{-1}.$$

Remark 7.31. In addition to the factors L, U (or D, L', U', respectively), we can either store A and use (7.24a) or store R and apply the representation (7.25):

$$Wx^{m+1} = b + Rx^m. \tag{7.25}$$

Concerning the computational work, we recall §2.3.1: the decomposition (7.23a–c) defines the initialisation cost denoted by $\mathrm{Init}(\Phi^{\mathrm{ILU}}, A)$, while $\mathrm{Work}(\Phi^{\mathrm{ILU}}, A)$ is the cost required by (7.25).

7.3.2 Incomplete Decomposition with Respect to a Star Pattern

For the description of the pattern E, we should not use the ordered indices $1, \ldots, n$. In the case of the model problem, the pairs (i,j) for $1 \le i,j \le N-1$ are taken as indices of I. The edges of the graph $G(A)$ are described by the pairs $((i,j),(i \pm 1, j))$ (horizontal neighbours) and $((i,j),(i, j \pm 1))$ (vertical neighbours). For the case of a regular grid, the *star notation* was already used in §1.3.2 as short-hand notation of the matrices. In the following, we use the so-called star patterns. The entries '$*$' in the examples

$$\begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \end{bmatrix}, \quad \begin{bmatrix} * & * & \\ * & * & * \\ & * & * \end{bmatrix}, \quad \begin{bmatrix} & \cdots & * \\ & \cdots & \\ * & \cdots & \end{bmatrix}$$

refer to elements in the set E. If, for instance, '$*$' is the right neighbour of the midpoint, this means that for all $\alpha \in I$ having a right neighbour $\beta \in I$, the pair (α, β) belongs to E. Unmarked positions or the sign '$\cdot$' signify that the corresponding pairs (α, β) do not belong to E.

Remark 7.32. The 1×1 star $[*]$ characterises the minimal set $E = \{(i,i) : i \in I\}$ of Exercise 7.30b. The corresponding ILU iteration coincides with the Jacobi iteration.

7.3.3 Application to General Five-Point Formulae

Algorithm (7.23a,b) should be regarded more as a definition than a method for practically computing the matrices D, L', U'. For the example of a general five-point formula A, we demonstrate how to derive a cheaper computation. For the sake of convenience, we assume that the coefficients are constant:

$$A = \begin{bmatrix} & -e & \\ -a & d & -b \\ & -c & \end{bmatrix} \qquad \text{(cf. (1.13a))}. \tag{7.26}$$

To ensure that A be an M-matrix, we requite that

$$a, b, c, e \geq 0, \qquad d \geq a + b + c + e\,.$$

The smallest pattern satisfying (7.19b) is

$$E = G(A), \quad \text{i.e., } E = \begin{bmatrix} & * & \\ * & * & * \\ & * & \end{bmatrix} \qquad (\textit{five-point pattern}). \tag{7.27a}$$

Using lexicographical ordering, the strictly triangular matrix L' has the pattern

$$\begin{bmatrix} & \cdot & \\ * & \cdot & \cdot \\ & * & \end{bmatrix},$$

since the $*$-marked positions are the only matrix entries corresponding to the pattern E and located below the diagonal. Correspondingly, U' has the pattern

$$\begin{bmatrix} & * & \\ \cdot & \cdot & * \\ & \cdot & \end{bmatrix}.$$

In (7.23a,b), we replace the indices $i, j, k \in \{1, \ldots, n\}$ by $\alpha, \beta, \gamma \in I = \Omega_h$ and, subsequently, we identify $\alpha = (x, y) = (k_\alpha h, l_\alpha h) \in \Omega_h$ with the pair (k_α, ℓ_α), where now $1 < k_\alpha, \ell_\alpha < N - 1$ holds (cf. (1.3)). First, one has to discuss the sum Σ' in (7.23a). $L'_{\alpha\gamma} \neq 0$ can only be true for $\gamma = (k_\gamma, \ell_\gamma) = (k_\alpha - 1, \ell_\alpha)$ or $\gamma = (k_\alpha, \ell_\alpha - 1)$, whereas $U'_{\gamma\beta} \neq 0$ leads to $\beta = (k_\gamma + 1, \ell_\gamma)$ or $\beta = (k_\gamma, \ell_\gamma + 1)$. Hence,

$$L'_{\alpha\gamma} D^{-1}_{\gamma\gamma} U'_{\gamma\beta} \neq 0$$

requires $\beta = \alpha$ or $\beta = (k_\alpha + 1, \ell_\alpha - 1)$. Both possibilities contradict the inequality $\alpha \neq \beta$—in (7.23a) written as $k \leq i - 1$—and $(\alpha, \beta) \in E$. Therefore, Σ' is an empty sum and (7.23a) reduces to $L'_{\alpha\beta} = A_{\alpha\beta}$ for $\alpha > \beta$ and $(\alpha, \beta) \in E$. Hence, L' is the constant two-point star

$$L' = \begin{bmatrix} & 0 & \\ -a & 0 & 0 \\ & -c & \end{bmatrix}. \tag{7.27b}$$

Similarly, we obtain

$$U' = \begin{bmatrix} & -e & \\ 0 & 0 & -b \\ & 0 & \end{bmatrix}. \tag{7.27c}$$

Only for $\alpha = \beta$, is the sum Σ'' in (7.23c) not empty and does contain the two indices $\gamma = (i_\alpha - 1, j_\alpha)$ and $\gamma = (i_\alpha, j_\alpha - 1)$. We abbreviate the diagonal entry $D_{\alpha\alpha}$ by $d_\alpha = d_{i_\alpha, j_\alpha}$. Because of $A_{\alpha\alpha} = d$ and the already known values in (7.27b,c), definition (7.23c) can be rewritten as

$$d_{i,j} = d - \frac{ab}{d_{i-1,j}} - \frac{ce}{d_{i,j-1}} \qquad (1 \le i, j \le N-1), \tag{7.27d}$$

where the terms with $j - 1 = 0$ or $i - 1 = 0$ have to be ignored. In particular, we obtain $d_{11} = d$ for the first grid point. For the five-point formula (7.26), the double loop in (7.23a–c) is reduced to a simple loop over all $(i, j) \in I = \Omega_h$.

It is also possible to determine the remainder matrix R. Equations (7.20d,e) become

$$\begin{array}{ll} R_{\alpha\beta} = 0 & \text{for } (\alpha, \beta) \in E, \\ R_{\alpha\beta} = (L'D^{-1}U')_{\alpha\beta} & \text{for } (\alpha, \beta) \notin E. \end{array} \tag{7.27e}$$

One verifies that R has two (variable) coefficients per row:

$$R = \begin{bmatrix} r_{ij} & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & s_{ij} \end{bmatrix} \qquad \text{with } r_{ij} = \frac{ae}{d_{i-1,j}}, \quad s_{ij} = \frac{cb}{d_{i,j-1}}, \tag{7.27f}$$

where $r_{ij} = 0$ holds for $i = 1$ and $s_{ij} = 0$ for $j = 1$.

Remark 7.33. The ILU decomposition of a five-point formula with constant or variable coefficients requires $6n$ operations for computing the d_{ij} values in (7.27d). The solution of

$$W\delta = (D + L')D^{-1}(D + U')\delta = d$$

takes $10n$ operations; hence because of the additional $10n$ operations for computing $d = Ax^m - b$, one ILU iteration step (7.24a) requires, in total, $21n$ operations. Note that the d_{ij} values in (7.27d) have to be determined only once. An alternative is determining R by additional $4n$ operations. Afterwards, the iteration (7.25) requires only $14n$ operations. Together with $C_A = 5$, the following cost factors result:

$$C^{\mathrm{ILU}} = 4.2 \quad \text{or} \quad C^{\mathrm{ILU}} = 2.8 \quad \text{respectively for } E \text{ in (7.27a).}$$

7.3.4 Modified ILU Decompositions

So far we ignored matrix entries a_{ij} for $(i,j) \notin E$ completely. One may pose the question of whether or not this is a good strategy. The following approach will indirectly use all a_{ij}.

We recall the Gauss–Seidel iteration, where the matrix $W = D - E$ is changed into $W = \frac{1}{\omega} D - E$ for the SOR method. Hence, overrelaxation, which in general leads to improved convergence, corresponds to diminishing the diagonal in W. Following Wittum [403], we introduce a modification which also leads to a diminishing or enlargement of the diagonal depending on the choice of ω.

Let $\mathbf{1}$ be the vector $(1)_{\alpha \in I}$ consisting of the entries $\mathbf{1}_\alpha = 1$. Gustafsson [172] proposes replacing the equation $R_{ii} = 0$ (i.e., (7.20d) for $i = k$) by

$$A\mathbf{1} = W\mathbf{1}, \qquad \text{i.e.,} \quad R\mathbf{1} = 0. \tag{7.28}$$

One may view $\mathbf{1}$ as a *test vector*. By condition (7.28), W is gauged in such a way that A and W coincide with respect to their application to $\mathbf{1}$. We generalise the condition $R\mathbf{1} = 0$ by

$$R_{ii} = \omega \sum_{j \neq i} R_{ij} \qquad (\omega \in \mathbb{R}) \tag{7.29}$$

and denote the corresponding decomposition as ILU_ω decomposition (its existence is not yet claimed). The corresponding ILU_ω iteration is denoted by $\Phi_\omega^{\mathrm{ILU}}$.

Remark 7.34. (a) For $\omega = 0$, Eq. (7.29) coincides with (7.20d) for $i = k$: $R_{ii} = 0$. Hence, the unmodified ILU decomposition is the ILU_0 decomposition.

(b) For $\omega = -1$, the conditions (7.28) and (7.29) are identical, i.e., the ILU_{-1} decomposition describes the modification by Gustafsson [172].

In the case of the five-point formula (7.26) and the five-point pattern (7.27a), L' and U' are still obtainable from (7.27b,c), whereas recursion (7.27d) for the entries d_{ij} of D becomes

$$d_{ij} := d + \frac{(\omega e - b)a}{d_{i-1,j}} + \frac{(\omega b - e)c}{d_{i,j-1}} \tag{7.30}$$

(terms with $i - 1 = 0$ and $j - 1 = 0$ are again to be ignored).

7.3.5 Existence and Stability of the ILU Decomposition

In this section, the inequalities $A \leq B$ have to be understood in the sense of elementwise inequalities $A_{\alpha\beta} \leq B_{\alpha\beta}$ $(\alpha, \beta \in I)$ as in §C.3.

It is well known that the (complete) LU decomposition exists if and only if all principal submatrices $(a_{ij})_{1 \leq i,j \leq k}$ are regular for $1 \leq k \leq n$. However, even if the decomposition $A = LU$ exists, it can be useless since the solution process of

the equations $Ly = b$ and $Ux = y$ may be unstable. Choose, e.g., $A = LU$ with $U = L^\top$ and $L = \mathrm{tridiag}\{\alpha, 1, 0\}$ for $\alpha < 1$, and investigate the error propagation (cf. Elman [121]). The criterion involving the principal submatrices is satisfied for positive definite matrices. However, there are positive definite matrices for which the ILU decomposition fails because of $U_{kk} = 0$ in (7.21a). The first part of the following criterion is stated by Meijerink–van der Vorst [280], while the second part is due to Manteuffel [273].

Theorem 7.35. *Let $E \subset I \times I$ satisfy (7.19a). (a) M-matrices A permits an ILU decomposition $A = W - R$ with W in (7.24b), which, in addition, represents a splitting (7.4) in the sense of Definition 7.9.*

(b) If an H-matrix A has a positive diagonal D, the ILU decomposition

$$A = (D + L')D^{-1}(D + U')$$

exists. $\hat{A} := D - |A - D|$ (cf. Definition C.60) has also an ILU decomposition $(\hat{D} + \hat{L}')\hat{D}^{-1}(\hat{D} + \hat{U}')$. Then the following inequalities hold:

$$0 \le \hat{D} \le D, \quad \hat{L}'\hat{D}^{-1} \le -\left|L'D^{-1}\right| \le 0, \quad \hat{D}^{-1}\hat{U}' \le -\left|D^{-1}U'\right| \le 0.$$

Meijerink–van der Vorst [280] prove part (a) by interpreting the ILU decomposition as a sequence of Gauss elimination steps which conserve the M-matrix property (cf. Lemma C.59). We give another proof directly referring to the defining equations (7.20c) and requiring weaker assumptions.

X_E denotes the restriction of a matrix X to the index subset E:

$$(X_E)_{\alpha\beta} := \begin{cases} X_{\alpha\beta} & \text{if } (\alpha,\beta) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

The matrices denoted in the following by the letters D, L, and U with different indices should always be of diagonal structure or strictly lower or upper triangular structure, respectively. Note that the triple (D, L, U) is uniquely defined by the sum $X = D + L + U$. To express the single components of this triple, we write $X = \mathrm{diag}\{X\} + L(X) + U(X)$.

In the following, it is not necessarily assumed that $A_{\alpha\beta} \le 0$ holds for $\alpha \ne \beta$, as it is necessary for M-matrices. We define

$$(A_-)_{\alpha\beta} := \begin{cases} A_{\alpha\beta} & \text{if } \alpha = \beta \text{ or } A_{\alpha\beta} \le 0, \\ 0 & \text{otherwise.} \end{cases}$$

The matrix A is assumed to fulfil the following conditions:

$$A_{\alpha\beta} \le \left(L(A_-)_E \cdot \mathrm{diag}\{A\}^{-1} \cdot U(A_-)_E\right)_{\alpha\beta} \quad \text{for all } \begin{cases} \alpha \ne \beta, \\ (\alpha,\beta) \in E, \end{cases} \tag{7.31a}$$

A has a *complete* LU decomposition

$$A = (\underline{D} + \underline{L})\underline{D}^{-1}(\underline{D} + \underline{U}) = \underline{D} + \underline{L} + \underline{U} + \underline{L}\,\underline{D}^{-1}\underline{U} \tag{7.31b}$$

with $\underline{D} \ge 0$, $\underline{L} \le 0$, $\underline{U} \le 0$.

Remark 7.36. All M-matrices A satisfy the assumptions (7.31a,b). (7.31b) implies the inverse positivity of A, i.e., $A^{-1} \geq 0$. Condition (7.31a) is always satisfied if A fulfils the sign condition $A_{\alpha\beta} \leq 0$ $(\alpha \neq \beta)$ for all $(\alpha, \beta) \in E$.

Proof. Since the Gauss elimination yields the complete LU decomposition, the inequalities in (7.31b) follow from Lemma C.59. Vice versa, the inequalities in (7.31b) imply $(D+L)^{-1} \geq 0$, $D^{-1} \geq 0$, $(D+U)^{-1} \geq 0$, from which $A^{-1} \geq 0$ can be concluded. If A is an M-matrix and therefore $A_{\alpha\beta} \leq 0$ for $\alpha \neq \beta$, $(A_-)_{\alpha\beta} \leq 0 \leq (L(A_-)_E \operatorname{diag}\{A\}^{-1} U(A_-)_E)_{\alpha\beta}$ follows. □

Theorem 7.37. *Assume that $E \subset I \times I$ satisfies (7.19a) and that the matrix A fulfils (7.31a,b). Then A permits an ILU decomposition $A = W - R$ with W in (7.24b). $A = W - R$ is a regular splitting if $A_{\alpha\beta} \leq 0$ for $(\alpha, \beta) \notin E$ (the minimal condition (7.19b) is sufficient). The enclosure (7.32) holds with D, L, U from (7.31b):*

$$(\underline{D} + \underline{L} + \underline{U})_E \;\leq\; D + L' + U' \;\leq\; (A_-)_E. \tag{7.32}$$

Proof. The conditions (7.20d) can be written as $R_E = 0$. Inserting the remainder $R = D + L' + U' + L' D^{-1} U' - A$, we obtain $(D + L' + U' + L' D^{-1} U' - A)_E = 0$, i.e.,

$$(D + L' + U')_E = (A - L' D^{-1} U')_E. \tag{7.33}$$

Using the mapping

$$X \mapsto \Phi(X) := \big(A - L(X) \cdot \operatorname{diag}\{X\}^{-1} \cdot U(X)\big)_E \,, \tag{7.34a}$$

we may write the defining equation (7.33) as a *fixed-point equation*

$$D + L' + U' = \Phi(D + L' + U'). \tag{7.33'}$$

Assume the monotonicity properties

$$\left.\begin{array}{l} C_1 \leq C_2,\ \operatorname{diag}\{C_1\} \geq 0, \\ L(C_2) + U(C_2) \leq 0 \end{array}\right\} \quad \Longrightarrow \quad \Phi(C_1) \leq \Phi(C_2). \tag{7.34b}$$

Equation (7.31b) states that $A = \underline{D} + \underline{L} + \underline{U} + \underline{L}\,\underline{D}^{-1}\underline{U}$. We set

$$A_0 := \underline{D} + \underline{L} + \underline{U} \quad \text{and} \quad A^0 := (A_-)_E. \tag{7.34c}$$

$\underline{L}\,\underline{D}^{-1}\underline{U} \geq 0$ yields

$$A_0 = (A_0)_- \leq ((A_0)_-)_E \leq ((A_0 + \underline{L}\,\underline{D}^{-1}\underline{U})_-)_E = (A_-)_E = A^0,$$

i.e.,

$$A_0 \leq A^0. \tag{7.34d}$$

Next, we show that

$$A_0 \le \Phi(A_0) \quad \text{and} \quad \Phi(A^0) \le A^0. \tag{7.34e}$$

$\Phi(A_0) = (A - \underline{L}\,\underline{D}^{-1}\,\underline{U})_E = (\underline{D} + \underline{L} + \underline{U})_E = (A_0)_E \ge A_0$ holds because of $\underline{L}\,, \underline{U} \le 0$. The second inequality in (7.34e) is identical to (7.31a). Φ defines the following fixed-point iterations:

$$A_{m+1} := \Phi(A_m), \quad A^{m+1} := \Phi(A^m). \tag{7.34f}$$

The monotonicity (7.34b) and the inequalities (7.34d,e) lead to

$$A_0 \le A_1 \le \ldots \le A_m \le \ldots \le A^m \le \ldots \le A^1 \le A^0 \tag{7.34g}$$

(cf. Theorem 7.19). Hence, both sequences must converge to a unique limit $C = D + L' + U'$ satisfying the fixed-point equation (7.33′). (7.32) follows from (7.34c) and $A_0 \le D + L' + U' \le A^0$. $W^{-1} = (D + U')^{-1} D (D + L')^{-1} \ge 0$ is a consequence of the inequalities $D \ge 0$ and $L', U' \le 0$. Remainder R vanishes on E: $R_E = 0$; otherwise, $R_{\alpha\beta} = (L' D^{-1} U' - A)_{\alpha\beta}$ holds. The inequality $A_{\alpha\beta} \le 0$ for indices $(\alpha, \beta) \notin E$ implies $R_{\alpha\beta} \ge (L' D^{-1} U')_{\alpha\beta} \ge 0$. Hence, the splitting $A = W - R$ is regular. □

The *stability*[3] of the ILU decomposition is expressed in (7.32) by the estimate of the diagonal D from below by $\underline{D}$.

To generalise Theorem 7.37 to the ILU$_\omega$ decomposition with $\omega \ne 0$, we may write the equations $R_{ij} = 0$ for $i \ne j$, $(i, j) \in E$, and (7.29) as

$$R_E - \omega \operatorname{diag}\{R_{E'} \mathbf{1}\} = 0 \qquad \text{with } R = D + L + U + L D^{-1} U - A,$$

Here, $E' := (I \times I) \backslash E$ is the complement. For a vector $v = (v_1, \ldots, v_n)^{\mathsf{T}}$, $\operatorname{diag}\{v\}$ denotes the diagonal matrix $\operatorname{diag}\{v_1, \ldots, v_n\}$. Carrying over the proof technique, we are led to the fixed-point equation $C = \Phi_\omega(C)$ with

$$\Phi_\omega(C) := \Phi(C) - \omega \operatorname{diag}\{(A - L(C) \cdot \operatorname{diag}\{C\}^{-1} \cdot U(C))_{E'} \mathbf{1}\} \tag{7.35}$$

and Φ defined in (7.34a). In general, however, Φ_ω does not have the desired properties. The monotonicity corresponding to (7.34b) may be violated for $\omega > 0$, whereas for $\omega < 0$, it may happen that no A_0 exists with $\Phi_\omega(A_0) \ge A_0$ (and hence, no solution exists).

For a precise discussion, we study the five-point formula (7.26) with the five-point pattern (7.27a). Since L', U' are already uniquely determined (cf. (7.27b,c)), the fixed-point equation simplifies to a scalar equation for D:

$$\begin{aligned} D = \Phi_\omega(D) &:= \operatorname{diag}\{A - L' D^{-1} U'\} - \omega \operatorname{diag}\{(A - L' D^{-1} U')_{E'} \mathbf{1}\} \\ &= \operatorname{diag}\{d + (\omega a - c) e / D_{i-1,j} + (\omega c - a) b / D_{i,j-1}\} \end{aligned} \tag{7.36a}$$

[3] Concerning the problem that the solution of the systems $(D + L)x = b$ or $(D + U)x = b$ may lead to instabilities, we refer to Elman [121], where ILU decompositions for nonsymmetric matrices are discussed.

(cf. (7.30)). For analysing this equation, we investigate the one-dimensional fixed-point equation

$$d = \varphi_\omega(d) := d + [(\omega e - b)a + (vb - e)c]\,/d. \tag{7.36b}$$

A discussion of the function φ_ω, which is left to the reader, shows the following.

(i) The fixed-point equation (7.36b) is solvable if and only if

$$4\gamma < d^2 \qquad \text{for } \gamma := ce + ab - \omega(ae + cb). \tag{7.36c}$$

(ii) If (7.36c) is satisfied, the solutions of (7.36b) are

$$\delta_\pm = \frac{1}{2}\left(d \pm \sqrt{d^2 - 4\gamma}\right). \tag{7.36d}$$

(iii) δ_+ is the stable fixed point because (7.36e) leads to (7.36f):

$$\varphi_\omega(\delta) < \delta \text{ for } \delta > \delta_+, \qquad \varphi_\omega(\delta) > \delta \text{ for } \delta_- < \delta < \delta_+, \tag{7.36e}$$

$$\lim \delta_m = \delta_+ \qquad \text{for } \delta_0 > \delta_-,\ \ \delta_{m+1} := \varphi_\omega(\delta_m). \tag{7.36f}$$

(iv) On the other hand, starting values $\delta_0 < \delta_-$ generate sequences $\{\delta_m\}$ which contain at least one element $\delta_m \le 0$.

Exercise 7.38. Let A in (7.26) be diagonally dominant and symmetric:

$$a = b \ge 0,\ c = e \ge 0,\ s := a + c > 0,\ d = 2\sigma + \varepsilon \quad \text{with } \varepsilon \ge 0. \tag{7.37a}$$

Prove that for $\omega = -1$, the value δ_+ is obtained from (7.36d) with $\gamma = \sigma^2$. For small ε, this value has the expansion

$$\delta_+ = \sigma + \sqrt{\varepsilon\sigma} + \mathcal{O}(\varepsilon). \tag{7.37b}$$

Assuming (7.37a), we obtain for $\omega = 0$ that

$$\delta_+ = a + c + \sqrt{2}ac + \mathcal{O}(\sqrt{\varepsilon}).$$

Theorem 7.39. *Let $\omega \in [-1, \omega^*]$, where $\omega^* := \min\{\frac{c}{a}, \frac{a}{c}\}$. Assume that the matrix A in (7.26) satisfies (7.37a). Then the ILU$_\omega$ decomposition exists, and the entries d_{ij} of the diagonal D are enclosed by*

$$\delta_+ = \frac{d + \sqrt{d^2 - 4\,(c^2 + a^2 - 2\omega ac)}}{2} < d_{ij} \le d \quad \textit{for } (i,j) \in I. \tag{7.38}$$

The fixed-point iteration (7.36a) with the starting iterate $D^0 := \operatorname{diag}\{d\mathbf{1}\}$ converges from above to D.

Proof. (7.36c) is satisfied for $\omega > -1$, while Φ_ω is monotone for $\omega \le \omega^*$. One verifies that $D_0 \le \Phi_\omega(D_0)$ and $\Phi_\omega(D^0) \le D^0$ hold for $D_0 := \operatorname{diag}\{\delta_+\mathbf{1}\}$ and $D^0 := \operatorname{diag}\{d\mathbf{1}\}$. Hence, we can draw the same conclusions as in the proof of Theorem 7.37. □

7.3.6 Properties of the ILU Decomposition

An immediate consequence of Theorem 7.11 is the following convergence statement.

Theorem 7.40. *If A is an M-matrix or, if according to Theorem 7.37, $A = W - R$ describes a regular splitting, the ILU iteration (7.24a,b) converges with the convergence rate $\rho(A^{-1}R)/(1+\rho(A^{-1}R))$.*

In the standard case, one may assume $\|R\| = \mathcal{O}(\|A\|)$, so that $\rho(A^{-1}R) \le \|A^{-1}\|\,\|R\| \le C\|A^{-1}\|\,\|A\| = C\,\mathrm{cond}(A) \gg 1$ leads to the convergence rate $(1+1/\rho(A^{-1}R))^{-1} \approx 1 - \mathcal{O}(1/\mathrm{cond}(A))$. Hence, the ILU decomposition has the same order as the Jacobi or Gauss–Seidel iteration. A better result can be derived for the modified ILU_{-1} decomposition (cf. (7.28) or (7.29) with $\omega = -1$). We prepare its analysis with the following lemma (cf. Wittum [402]).

Lemma 7.41. *Assume (7.39a), where A, D_A, and D are positive definite:*

$$A = D_A - L - L^{\mathsf{H}}, \quad W = (D + L')D^{-1}(D + L'^{\mathsf{H}}). \tag{7.39a}$$

The spectrum $\sigma(W^{-1}A)$ is contained in $[0, \Gamma]$ if

$$(2 - \tfrac{1}{\Gamma})D - D_A + L + L^{\mathsf{H}} + L' + L'^{\mathsf{H}} \quad \textit{is positive semidefinite.} \tag{7.39b}$$

Proof. We write $D + L'$ as $\frac{1}{\Gamma}D + C$ with $C := (1 - \frac{1}{\Gamma})D + L'$. From

$$\begin{aligned}\Gamma W - A &= (\tfrac{1}{\Gamma}D + C)(\tfrac{1}{\Gamma}D)^{-1}(\tfrac{1}{\Gamma}D + C)^{\mathsf{H}} - A \;\ge\; \tfrac{1}{\Gamma}D + C + C^{\mathsf{H}} - A \\ &= (2 - \tfrac{1}{\Gamma})D - D_A + L + L^{\mathsf{H}} + L' + L'^{\mathsf{H}} \;\ge\; 0\end{aligned}$$

with '$\ge$' in the sense of semidefiniteness, it follows that $\sigma(W^{-1}A) \in [0, \Gamma]$. □

Theorem 7.42. *Let $-1 \le \omega \le \omega^*$ (cf. Theorem 7.39). The five-point formula (7.26) and the five-point pattern (7.27a) are assumed to satisfy (7.37a). Then the inequality*

$$\gamma W \le A \le \Gamma W \qquad \textit{with} \begin{cases} \gamma = 1/[1 + (1+\omega)\frac{2ac}{\delta_+ \lambda_{\min}}], \\ \Gamma = \delta_+ / [2\delta_+ - d]\,, \end{cases} \tag{7.40}$$

holds with '$\le$' in the sense of semidefiniteness, where δ_+ is defined in (7.38) and $\lambda_{\min} = \varepsilon + 4(a+c)\sin^2\frac{\pi h}{2}$ is the smallest eigenvalue of A. In particular, (7.41) holds:

$$\gamma = 1, \qquad \Gamma = \frac{1}{2}\sqrt{\frac{\sigma}{\varepsilon}} - \frac{1}{4} + \mathcal{O}\left(\sqrt{\frac{\varepsilon}{\sigma}}\right) \qquad \textit{for } \omega = -1. \tag{7.41}$$

Proof. (i) (7.39b) becomes $(2 - \frac{1}{\Gamma})D - D_A \ge 0$, since by (7.27b,c), $L = -L'$ holds in Lemma 7.41. Thanks to $D_A = dI$ and $\delta_+ I \le D$ (cf. (7.38)), Γ with $(2 - \frac{1}{\Gamma})\delta_+ = d$ is sufficient for (7.39b). Solving for Γ, we obtain $\Gamma = \delta_+/(2\delta_+ - d)$.

(ii) The entries r_{ij}, s_{ij} of R (cf. (7.27f)) are bounded from above by $2ac/\delta_+$. According to (7.29), the diagonal entries of R are equal to $\omega(r_{ij}+s_{ij})$. The eigenvalues of R lie in the Gershgorin circles around $\omega(r_{ij}+s_{ij})$ with the radius $r_{ij}+s_{ij}$ (cf. Hackbusch [193, Criterion 4.3.4], Varga [376]) and, hence, they are bounded by $(1+\omega)(r_{ij}+s_{ij}) \le 2\,(1+\omega)\,ac/\delta_+$, implying $R \le [\,2\,(1+\omega)\,ac/\delta_+\,]\,I$. From $\lambda_{\min} I \le A$, we deduce $R \le \rho A$ with $\rho := 2(1+\omega)ac/(\delta_+\lambda_{\min})$. $A = W - R \ge W - \rho A$ yields $W \le (1+\rho)A$. Hence, $\gamma = 1/(1+\rho)$ leads to the representation of γ.

(iii) For $\omega = -1$, insert the representation (7.37b) into (7.40). □

Conclusion 7.43. *(a) Replacing the Poisson equation $-\Delta u = f$ with the Helmholtz equation $-\Delta u + \varepsilon u = f$ with $\varepsilon > 0$, we obtain the coefficients $a = b = c = e = -h^{-2}$, $d = 4h^{-2} + \varepsilon$ in (7.37a). Equation (7.41) yields the bound and condition number $\Gamma = \Gamma/\gamma = h^{-1}/\sqrt{2\varepsilon} + \mathcal{O}(1)$ indicating the order improvement.*

(b) Let $\omega = -1$. The (modified) ILU$_{-1}$ iteration damped by $\vartheta_{\rm opt} = 2/(\gamma+\Gamma) = 2\sqrt{2\varepsilon}h + \mathcal{O}(h^2)$ has the convergence speed

$$\rho(M^{\rm ILU}_{\vartheta_{\rm opt}}) \le (\Gamma-1)/(\Gamma+1) \approx 1 - 2/\Gamma \approx 1 - 2\sqrt{2\varepsilon}\,h.$$

Hence, similar to the SSOR method with an optimal relaxation parameter $\omega_{\rm SSOR}$, it is of first order as long as $\varepsilon > 0$.

Proof. Use Theorem 6.7. □

The applicability of the ILU$_{-1}$ decomposition is not at all restricted to strict diagonal dominance in Theorem 7.42 and Remark 7.43b, as shown by the following remark.

Remark 7.44 (enlargement of the diagonal). Let $A = A_\varepsilon$ be a matrix satisfying (7.37a) with $\varepsilon > -4(a+c)\sin^2\frac{\pi h}{2}$ (i.e., $\lambda_{\min}(A) > 0$) instead of $\varepsilon > 0$. Then the ILU$_{-1}$ decomposition $A_\eta = W_\eta - R_\eta$ has to be applied to the matrix $A_\eta := A + (\eta-\varepsilon)I$ with $\eta > 0$ in order to re-establish diagonal dominance $d > 2\sigma$. W_η can be viewed as the ILU decomposition of $A = A_\varepsilon$ with remainder $R = W_\eta - A = R_\eta - (\eta-\varepsilon)I$. Conclusion 7.43 yields the spectral condition number $\kappa(W_\eta^{-1}A_\eta)$. Let $\lambda = \lambda_{\min}(A)$ and $\Lambda = \lambda_{\max}(A)$ be the extreme eigenvalues of A. Because of

$$\kappa(A_\eta^{-1}A) = \kappa(A_\eta^{-1}A_\varepsilon) = \frac{\Lambda\,(\lambda+\eta-\varepsilon)}{\lambda\,(\Lambda+\eta-\varepsilon)} \approx 1 + \frac{\eta-\varepsilon}{\lambda_{\min}(A)}\,,$$

Lemma 7.55 shows that

$$\kappa(W_\eta^{-1}A_\varepsilon) \lessapprox h^{-1}\left(1 + \frac{\eta-\varepsilon}{\lambda_{\min}(A)}\right)/\sqrt{2\eta}\,. \tag{7.42}$$

Exercise 7.45. Prove that the right-hand side of (7.42) becomes minimal for $\eta = 4(a+c)\sin^2\frac{\pi h}{2}$.

Exercise 7.46. Prove that the ILU decomposition coincides with the exact LU decomposition if A has the tridiagonal pattern $\begin{bmatrix} \cdot & & \\ * & * & * \\ & \cdot & \end{bmatrix}$ or $\begin{bmatrix} & * & \\ \cdot & * & \cdot \\ & * & \end{bmatrix}$. Then the ILU iteration solves $Ax = b$ directly.

7.3.7 ILU Decompositions Corresponding to Other Patterns

Strengthening (7.19b) by $E \supsetneqq G(A)$ is the minimal requirement to construct new methods. When choosing a pattern E larger than $G(A)$, we should add those positions where $R = 0$ is violated: According to (7.27f), these are the positions $\begin{bmatrix} * & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & * \end{bmatrix}$. Adding $\begin{bmatrix} * & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & * \end{bmatrix}$ to the five-point pattern, we obtain

$$E = \begin{bmatrix} * & * & \\ * & * & * \\ & * & * \end{bmatrix} \qquad (\text{'seven-point pattern'}). \tag{7.43}$$

Now the lower triangular matrix L' and upper triangular matrix U' have the form

$$L' = -\begin{bmatrix} 0 & 0 & \\ a_{ij} & 0 & 0 \\ & c & f_{ij} \end{bmatrix}, \qquad U' = -\begin{bmatrix} g_{ij} & e & \\ 0 & 0 & b_{ij} \\ & 0 & 0 \end{bmatrix},$$

whose coefficients result from the recursions

$$\begin{aligned} d_{ij} &= d - ec/d_{i,j-1} + a_{ij}(\omega g_{i-1,j} - b_{i-1,j})/d_{i-1,j} \\ &\quad + f_{ij}(\omega b_{i+1,j-1} - g_{i+1,j-1})/d_{i+1,j-1}, \\ a_{ij} &= a + g_{i,j-1}c/d_{i,j-1}, \qquad bij = b + ef_{ij}/d_{i+1,j-1}, \\ f_{ij} &= b_{i,j-1}c/d_{i,j-1}, \qquad g_{ij} = a_{ij}e/d_{i-1,j} \end{aligned} \tag{7.44}$$

for $1 \le i, j \le N-1$, where all terms with indices $i - 1 = 0$, $j - 1 = 0$, or $i + 1 = N$ have to be ignored. This seven-point ILU decomposition has properties similar to those of the five-point version in Theorem 7.42 (cf. Gustafsson [172], Axelsson–Barker [13, §7]).

Exercise 7.47. Prove: (a) For $-1 \le \omega \le 0$, the fixed-point iteration (7.35) converges for the starting iterate $C = A$ to values satisfying the inequalities $a_{ij} \le \alpha := a/\Delta$, $b_{ij} \le \beta := b/\Delta$, $f_{ij} \le \beta c/\gamma$, $g_{ij} \le \alpha e/\delta$, $d_{ij} \ge \delta$ with $\Delta := 1 - ec/\delta^2$, where d is the maximal solution of the fixed-point equation

$$\delta = \varphi(\delta) := d - \left[ec + \frac{ab}{\Delta^2}(1 + \frac{ec}{\delta^2}) - \frac{\omega}{\delta}(\alpha^2 e + \beta^2 c)\right]/\delta.$$

(b) For the next considerations, assume the symmetry $a = b$, $c = e$ as well as the diagonal dominance $d = 2(a + c) + \varepsilon$ with $\varepsilon \ge 0$. Furthermore, choose $\omega = -1$

(i.e., the modified ILU). Prove that the equation $\delta = \varphi(\delta)$ can be brought into the form $2a + \varepsilon = a(\xi + \xi^{-1})$ with $\xi := a\delta/(\delta - c)^2$. Hence, the solution is

$$\delta = c + a/(2\xi) + \sqrt{ac/\xi + a^2/(4\xi^2)}$$

with $\xi = 1 + \varepsilon/(2a) + \sqrt{\varepsilon/a + \varepsilon^2/(4a^2)}$.

(c) For $\varepsilon \geq 0$, a solution $\delta = \delta_0 + C\sqrt{\varepsilon} + \mathcal{O}(\varepsilon)$ exists.

(d) δ solves the equation $(\delta - \gamma - e - \beta)^2 = \varepsilon\delta$.

(e) The weak diagonal dominance, which is sufficient for (7.39b), leads to the condition $2\varphi + 2\,|a - \alpha| \leq (2 - \frac{1}{\Gamma})\delta - d$. Show that $\Gamma = \delta/\big(2\sqrt{\varepsilon\delta} - \varepsilon\big)$.

(f) As in (7.41), the estimate $\gamma W \leq A \leq \Gamma W$ holds with $\gamma = 1$.

Concerning ILU decompositions with a general k-point pattern, note that the amount of computational work increases more than linearly with the number k of pattern entries.

7.3.8 Approximative ILU Decompositions

The ILU decompositions, as defined in (7.27d) or (7.30), are strictly sequential algorithms. The same statement holds for solving the systems $(D + L)x = b$ and $(D + U)x = b$ arising during the solution of $W\delta = d$. This is a disadvantage for a parallel treatment. The parallel treatment of the systems is discussed by van der Vorst [371] (cf. also Ortega [298, §3.4]). Here we discuss the computation of the ILU decomposition. Note that the fixed-point iteration (7.34f) in the proof of Theorem 7.37 is suited to numerical computations. The upper starting iterate $A^0 = (A_-)_E$ (in general, $A^0 = A$) is available (in contrast to A_0), so that the iterates $A^{m+1} = \Phi(A^m)$ are computable.

Remark 7.48. The evaluation of the function Φ in (7.34a) can be performed in parallel for all coefficients $\Phi(X)_{\alpha\beta}$, $(\alpha, \beta) \in E$.

The equations (7.33′): $X = \Phi(X)$ or, more precisely, the recursions (7.30) and (7.44) represent simple systems of equations for the unknowns d_{ij} (and possibly a_{ij}, b_{ij}, f_{ij}, g_{ij}), which can be solved by backward substitutions. Independently of the starting iterate, the values for (i, j) with $\max\{i, j\} \leq m$ are exact after m iteration steps. If A and therefore also the starting iterate A^0 (cf. (7.34c)) have constant coefficients, the m-th iterate A^m has identical constant coefficients for all positions[4] (i, j) with $\min\{i, j\} \geq m$. Since the coefficients of A^m coincide for $\min\{i, j\} \geq m$, one need not calculate all of them. This consideration leads us to the truncated ILU version introduced by Wittum [400] for constant coefficients:

[4] At positions with $\min\{i, j\} < m$ other values are possible, since in (7.30) or (7.44) some terms may be absent because of $i - 1 = 0$ or $j - 1 = 0$.

$$\text{Compute } d_{ij} \text{ (and possibly } a_{ij},\, b_{ij},\, f_{ij},\, g_{ij}) \text{ from (7.30) or, respectively, (7.44) for all } i,j \text{ with } \max\{i,j\} = k \text{ for } k = 1,2,\ldots,m \text{ and continue these values constantly by means of } d_{ij} := d_{\min\{i,m\},\min\{j,m\}} \text{ for } \max\{i,j\} > m \tag{7.45}$$

(analogously for a_{ij}, b_{ij}, f_{ij}, g_{ij}). The amount of computational work is $\mathcal{O}(m^2)$ independent of dimension n of the matrix. The same statement holds for the storage requirement. The truncated ILU decomposition is a good substitute for the standard ILU decomposition and has favourable stability properties (cf. Wittum–Liebau [406]).

7.3.9 Blockwise ILU Decomposition

Choosing the row or column variables as blocks, A has a block structure with tridiagonal matrix blocks in diagonal position as shown in (3.17). In the decomposition ansatz (7.22):

$$A \;=\; (D + L')D^{-1}(D + U') - R \;=\; D + L' + U' + L'D^{-1}U' - R\,,$$

we may also require that D be a block-diagonal matrix with blocks of tridiagonal structure and that L' and U' be strictly (lower/upper) block-triangular matrices. The algorithm is similar to (7.23a,b) (cf. (11.95a–c)). With the increased amount of computational work, one gains, in general, more robust convergence properties. Block-ILU decompositions were introduced in the early 1980s (cf. §7.3.11).

7.3.10 Numerical Examples

Table 7.1 shows the errors $\|x^m - x\|_2$ after $m = 20$ iterations and the convergence factors for different ILU variants. *ILU_5* refers to the five-point ILU defined by (7.27a), while *ILU_7* refers to (7.43). The step size of the Poisson model problem is $h = 1/32$. For $\omega = 0$ and $\omega = 1$, the ILU iteration is applied to the original matrix, whereas for the modified method with $\omega = -1$ an enlargement of the diagonal by $A_\eta := A + 5I$ is chosen according to Remark 7.44. ϑ is the damping factor in (5.8).

version	ω	ϑ	$\|x^{20} - x\|_2$	$\frac{\|x^{20}-x\|_2}{\|x^{19}-x\|_2}$
ILU_5	0	1.66	1.617_{10}-1	0.9455
ILU_5	-1	0.25	1.628_{10}-3	0.7666
ILU_5	1	1.9	2.349_{10}-1	0.9617
ILU_7	0	1	8.904_{10}-2	0.9185
ILU_7	0	1.66	2.690_{10}-2	0.8646
ILU_7	-1	0.4	4.722_{10}-5	0.6254

Table 7.1 Results of the ILU iteration for the Poisson model case.

Exercise 7.49. Count the arithmetic operations (separately for the decompositions and the solution phase) and compare *ILU_5* and *ILU_7* with regard to the effective amount of work.

7.3.11 Remarks

ILU decompositions are first mentioned in 1960 by Varga [374, §6] and Buleev [84]. The first precise analysis is due to Meijerink–van der Vorst [280]. Here, we also mention Jennings–Malik [228]. ILU methods have proved to be very robust. This means that good convergence properties are not restricted to the Poisson model problem, but hold for a large class of problems. Since the existence of an ILU decomposition is not always ensured, there are many stabilising variants. Concerning literature about the ILU method, we refer to Axelsson–Barker [13], Axelsson [12, §7], and Beauwens [37].

Because of the improved condition number Γ/γ in (7.41), the modified version ($\omega = -1$) of Gustafsson [172] is the preferred basis for applications of the conjugate gradient technique (cf. §10) to ILU iterations. Because of the consistency condition $R\mathbf{1}=0$, this version is also called an ILU iteration of first order. A special decomposition for the Poisson model problem of second order is described by Stone [356]; however, because of other disadvantages, first-order variants are preferred.

The first publication of a blockwise ILU method in 1981 is due to Kettler [235], who refers to a 'publication in preparation' by Meijerink which appeared in [279] two years later. Additional early papers are those by Axelsson–Brinkkemper–Il'in [14] (1984 with a preprint in 1983) and Concus–Golub–Meurant [99] (1985 with a preprint in 1982).

In the literature, the distinction between SSOR and ILU methods is not very sharp. The SSOR method for $A = D + L' + U'$ corresponds to an ILU decomposition $W = (D + L')D^{-1}(D + U')$ with remainder $R = W - A = L'D^{-1}U'$. This R does not satisfy condition (7.20d); however, this condition is already weakened by (7.29) and addition of a diagonal part (cf. Remark 7.44). Vice versa, generalised SSOR methods have been introduced in which $D = \mathrm{diag}\{A\}$ is replaced by another diagonal (cf. Axelsson–Barker [13]). The ILU iteration based on a five-point pattern also falls into this category.

In the literature, one finds a lot of abbreviations for different ILU variants. 'IC' refers to the 'incomplete Cholesky' variant of the ILU decomposition. Additional numbers like '(5)' or '(7)' denote the respective five- or seven-point pattern. In other papers, '(0)' indicates the pattern $E = G(A)$, whereas '(1)' means the pattern which is enlarged by one level, etc. The supplement 'Tr' characterises the truncated version (7.45). The letter 'M' stands for the modified method with $\omega = -1$, whereas 'B' may indicate a block variant. If the block corresponds to a grid line (row or column), sometimes the symbol 'L' is used.

In particular concerning the ILU(p) variant, we refer to Saad [328, §§10.3]. The thresholding technique ILUT can also be found in [328, §§10.4]. See also Björck [48, §§4.4.3f]. Another kind of factorisation is proposed by Benzi–Tůma [42].

While ILU methods are less attractive as linear iterations, their combination with multigrid methods is successful (see §11.6.2 and Hackbusch–Wittum [208]).

7.4 Preconditioning

The term 'preconditioning' is rather ambiguous. In §7.4.1 we describe the preconditioning in the narrower sense. When it is used in the wider sense it is losing its original meaning and, in the extreme case, may mean any transformation in the sense of §5.6 (cf. §7.4.3).

7.4.1 Idea of Preconditioning

We recall the *spectral condition number* $\kappa(A) := \rho(A)\rho(A^{-1})$ of a regular[5] matrix defined in (B.13). In the case of $A > 0$, the spectral condition number $\kappa(A)$ simplifies to the ratio $\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$ of the extreme eigenvalues. Alternatively, for a given matrix norm we can define the condition $\mathrm{cond}(A) := \|A\|\,\|A^{-1}\|$. If A is normal, the Euclidean condition $\mathrm{cond}_2(A)$ with respect to the spectral norm coincides with $\kappa(A)$. This holds in particular under the assumption $A > 0$. Furthermore, we consider the simplest linear iteration: the Richardson iteration defined in §3.2.1. The convergence analysis in §3.5.1 shows that, for the optimal parameter Θ_{opt}, the convergence rate and contraction number coincide with

$$\rho(M^{\mathrm{Rich}}_{\Theta_{\mathrm{opt}}}) = \frac{\kappa(A)-1}{\kappa(A)+1} = \frac{1-\frac{1}{\kappa(A)}}{1+\frac{1}{\kappa(A)}}$$

(cf. (3.26c)). The essential observation is that $\rho\,(M^{\mathrm{Rich}}_{\Theta_{\mathrm{opt}}})$ depends only on the spectral number $\kappa(A)$ (cf. (B.13)).

If $\kappa(A)$ is very close to 1, we have very fast convergence. If $\kappa(A)$ is of moderate size, a moderate convergence speed results. If, however, $\kappa(A)$ is large, the asymptotic approximation $\rho(M^{\mathrm{Rich}}_{\Theta_{\mathrm{opt}}}) = 1 - 2/\kappa(A) + \mathcal{O}(\kappa(A)^{-2})$ shows that the convergence is rather slow.

Hence, one can try to choose a left transformation with $T_\ell = N = W^{-1}$ so that

$$\hat{A} := T_\ell\, A = W^{-1}A \quad \text{has a positive spectrum,} \tag{7.46a}$$

$$\kappa\left(W^{-1}A\right) \qquad \text{is as small as possible.} \tag{7.46b}$$

Note that under condition (7.46a) Theorem 6.7 implies that the optimally damped iteration

$$\Phi_W(x,b) := x - \vartheta_{\mathrm{opt}} W^{-1}(Ax-b) \tag{7.46c}$$

has the convergence rate

$$\rho(M_{\vartheta_{\mathrm{opt}}}) = \frac{\kappa(W^{-1}A)-1}{\kappa(W^{-1}A)+1}\,.$$

Often, the matrix W is called the *preconditioning matrix*, *preconditioning*, or *preconditioner*. Sometimes these names also refer to the matrix $N = W^{-1}$

[5] We may set $\kappa(A) = \infty$ for singular A. For certain purposes it makes sense to extend the spectral condition to singular matrices $A \neq 0$ by $\kappa_0(A) := \max_{\lambda\in\sigma(A)} |\lambda| \,/\, \min_{\lambda\in\sigma(A)\setminus\{0\}} |\lambda|$.

of the second normal form. The mapping

$$\Phi_{\Theta}^{\mathrm{Rich}} \mapsto \Phi_W = \Phi_{\Theta}^{\mathrm{Rich}} \circ W^{-1}$$

is also called 'preconditioning'. Note that this term expresses the intention to improve the condition, but it is not a concrete description of the mapping $A \mapsto W[A]$. Besides the size of the condition (and therefore the convergence speed) one must have in mind the related cost (cf. §2.3.2).

The condition numbers will also appear in Part II in connection with the semi-iterative method applied to the basic iteration Φ_W. Instead of a real spectrum contained in $[\lambda_{\min}(A), \lambda_{\max}(A)]$, we may replace the interval by an ellipse (cf. §8.3.6).

The construction of the iteration (7.46c) is not restricted to the left transformation $\Phi_W = \Phi^{\mathrm{Rich}} \circ W^{-1}$. The right transformation $T_r = W^{-1}$ applied to the Richardson method leads to the same iteration (5.41): $\Phi_W(x, b) = x - W^{-1}(Ax - b)$. The two-sided transformation $\Phi_W = W^{-1/2} \circ \Phi^{\mathrm{Rich}} \circ W^{-1/2}$ by (5.46) also leads to the same 'preconditioned' iteration.

7.4.2 Examples

As examples of preconditioning the positive definite matrix $A = D - E - F$ (cf. (1.16)) we recall the matrices W of the already described symmetric iterations:

$$\begin{aligned} W &= D = \mathrm{diag}\{A\} && \text{(Jacobi)}, \\ W &= (D - E)D^{-1}(D - F) && \text{(SSOR)}. \end{aligned}$$

Here, the methods can be understood pointwise or blockwise.

Since the choice of 'W = diagonal matrix' is especially simple and also computable in parallel, one might ask whether the Jacobi method with $D := \mathrm{diag}\{A\}$ represents the optimal diagonal preconditioning. The answer in given by Theorems 7.50 and 7.51: $D := \mathrm{diag}\{A\}$ is optimal in the 2-cyclic case, whereas D is close to the optimum in the general case (see also Higham [221, Theorem 7.5]).

Theorem 7.50 (Forsythe–Strauss [139]). *Assume that A is positive definite with $D := \mathrm{diag}\{A\}$ and $A - D$ is weakly 2-cyclic. Then $D := \mathrm{diag}\{A\}$ is the best diagonal preconditioner; i.e., $\kappa(D^{-1}A) \le \kappa(\Delta^{-1}A)$ for all diagonal matrices Δ.*

Theorem 7.51 (van der Sluis [368]). *Let the matrix A be positive definite with $D := \mathrm{diag}\{A\}$ and assume that each row of A contains at most C_A nonzero entries (cf. (2.28)). Then $\kappa(D^{-1}A) \le C_A \kappa(\Delta^{-1}A)$ holds for all diagonal matrices Δ.*

Bank–Scott [32] describe a related result about the condition of finite element matrices in the presence of local refinements.

Let Γ be the constant in (8.39c). The SSOR preconditioning improves the condition number from $\kappa(A)$ to $\frac{1}{2}\big(1 + \sqrt{\Gamma\kappa(A)}\big)$. The transition from $\kappa(A)$ to $\mathcal{O}\big(\sqrt{\kappa(A)}\big)$ corresponds to the improvement of the order (cf. Conclusion 6.29).

7.4.3 Preconditioning in the Wider Sense

Let $A = Q\,\mathrm{diag}\{\lambda_i : 1 \le i \le n\}\,Q^{\mathrm{H}}$ (Q is unitary) be any normal matrix with $0 < \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$. Obviously $\kappa(A) = \mathrm{cond}_2(A) = \lambda_n/\lambda_1$ is the condition. Now we replace λ_1 with $-\lambda_1$. $\hat{A} = Q\,\mathrm{diag}\{-\lambda_1, \lambda_2, \ldots, \lambda_n\}\,Q^{\mathrm{H}}$ is an indefinite matrix also satisfying $\kappa(\hat{A}) = \mathrm{cond}_2(\hat{A}) = \kappa(A)$. Although the condition is unchanged, the Richardson iteration has a problem because of Exercise 3.25. Obviously, it is not the condition which must be improved, but the indefinite matrix must be turned into a positive definite one. Again a transformation by $W^{-1} = \hat{A}$ helps: the resulting squared Richardson iteration $\hat{A}^2$ is positive definite. However, the condition $\kappa(A)$ is replaced with the larger condition $\kappa(\hat{A}^2) = \kappa(A)^2$. Calling $W^{-1} = \hat{A}$ a preconditioner, the original meaning of improving the condition is perverted. Nevertheless, we can try to precondition $\hat{A}^2$ in the narrower sense. One learns from this example that beside the condition other structural properties are important which may be improved by a transformation for which the name 'preconditioning' is not quite adequate.

Another systematic approach to indefinite Hermitian matrices A (cf. Remark 8.31) is the left transformation by a polynomial in A. Such 'preconditioners' are described, e.g., by Ashby–Manteuffel–Saylor [7]. Here the polynomial $T_\ell = p(A)$ should be close to the minimiser of $\min\{\rho(p(A)A) : \mathrm{degree}(p) = d\}$ for a fixed degree $d \ge 1$.

In the case of non-Hermitian matrices A, even the convergence of the Richardson iteration cannot be described by $\kappa(A)$ or $\mathrm{cond}_2(A)$. Hence the term 'preconditioning' loses its meaning. On the other hand, a large condition number is not necessarily a disadvantage (see the multigrid iteration in §11.4). For the extreme example of a diagonal matrix A, the system is exactly solvable independently of the condition.

7.4.4 Rules for Condition Numbers and Spectral Equivalence

The Euclidean condition $\mathrm{cond}_2(\cdot)$ and the spectral condition number $\kappa(\cdot)$ satisfy the following equations and inequalities (cf. (B.12), (B.13)).

Exercise 7.52. Let the matrices A, B, C be regular. Prove the following:

$$\kappa(A) = \kappa(A^{-1}), \quad \mathrm{cond}_2(A) = \mathrm{cond}_2(A^{-1}), \tag{7.47a}$$

$$\kappa(A) = \kappa(\lambda A), \quad \mathrm{cond}_2(A) = \mathrm{cond}_2(\lambda A) \text{ for } \lambda \in \mathbb{C}\backslash\{0\}, \tag{7.47b}$$

$$\kappa(A) = \mathrm{cond}_2(A) \qquad \text{for normal matrices } A, \tag{7.47c}$$

$$\mathrm{cond}_2(AB) \le \mathrm{cond}_2(A)\,\mathrm{cond}_2(B), \tag{7.47d}$$

$$\mathrm{cond}_2(C^{-1}A) \le \mathrm{cond}_2(C^{-1}B)\,\mathrm{cond}_2(B^{-1}A), \tag{7.47e}$$

$$\kappa(B^{-1}A) = \mathrm{cond}_2(B^{-1/2}AB^{-1/2}) \qquad \text{for } A, B > 0, \tag{7.47f}$$

$$\kappa(AB) = \kappa(BA). \tag{7.47g}$$

Following considerations are restricted to positive definite matrices. The next lemma shows that the spectral number can be formulated by matrix inequalities.

Lemma 7.53. *Let A and B be positive definite. Then $\kappa(B^{-1}A)$ can be represented as*

$$\kappa(B^{-1}A) = \bar{\alpha}/\underline{\alpha} \tag{7.48a}$$

where $\bar{\alpha}$ and $\underline{\alpha}$ are the best bounds in the inequality

$$\underline{\alpha}B \le A \le \bar{\alpha}B \qquad \text{with } \underline{\alpha} > 0. \tag{7.48b}$$

Vice versa, (7.48b) implies

$$\kappa(B^{-1}A) \le \bar{\alpha}/\underline{\alpha}. \tag{7.48c}$$

Proof. The best bounds in (7.48b) are the extreme eigenvalues of $B^{-1}A$. Hence, (7.48a) follows from (B.14). □

Exercise 7.54. Prove that (7.48b) is equivalent to either of the following inequalities:

$$\tfrac{1}{\underline{\alpha}}A \le B \le \tfrac{1}{\bar{\alpha}}A \qquad \text{with } \underline{\alpha} > 0, \tag{7.48d}$$

$$\underline{\alpha}A^{-1} \le B^{-1} \le \bar{\alpha}A^{-1} \qquad \text{with } \underline{\alpha} > 0, \tag{7.48e}$$

$$\underline{\alpha}\,\langle Bx, x\rangle \le \langle Ax, x\rangle \le \bar{\alpha}\,\langle Bx, x\rangle \qquad \text{for all } x \in \mathbb{K}^I. \tag{7.48f}$$

The inequalities (7.47e,f) yield the next lemma.

Lemma 7.55. *Let A, B, C be positive definite. Then*

$$\kappa(C^{-1}A) \le \kappa(C^{-1}B)\,\kappa(B^{-1}A). \tag{7.49}$$

Interpreting (7.47e) and (7.49) in the sense of preconditioning yields the following statement. If B is a good preconditioner for A and C is a good preconditioner for B, then C also represents a good preconditioning for A.

The following definition of spectral equivalence does not make sense for a single matrix. Instead we need two infinite families

$$\mathcal{A} = (A_\nu)_{\nu\in F}, \quad \mathcal{B} = (B_\nu)_{\nu\in F} \qquad (\#F = \infty)$$

of matrices (cf. §1.4). Usually, $\nu \in F = \mathbb{N}$ is related to a discretisation grid size h_ν with the property $h_\nu \to 0$. In this case, we prefer the notation $\mathcal{A} = (A_h)_{h\in H}$. Then the size of the matrices is increasing with $\nu \to \infty$. Another case may be a matrix depending on a parameter ν varying in an interval F.

Definition 7.56 (spectral equivalence). Let $\mathcal{A} = (A_\nu)_{\nu\in F}$ and $\mathcal{B} = (B_\nu)_{\nu\in F}$ be two families of positive semidefinite matrices. Then $\mathcal{A}$ and $\mathcal{B}$ are called *spectrally equivalent* if there is a constant $c > 0$ so that

$$\tfrac{1}{c}A_\nu \le B_\nu \le cA_\nu \qquad \text{for all } \nu \in F. \tag{7.50}$$

The explicit notation of the equivalence relation is

$$\mathcal{A} \sim \mathcal{B}.$$

Often, the less precise notation $A_\nu \sim B_\nu$ is used.

The characteristic properties of an equivalence relation are obviously satisfied: the symmetry $\mathcal{A} \sim \mathcal{B} \Leftrightarrow \mathcal{B} \sim \mathcal{A}$ and the transitivity $\mathcal{A} \sim \mathcal{B} \sim \mathcal{C} \Rightarrow \mathcal{A} \sim \mathcal{C}$. Gunn [171] used similar arguments in 1964 without mentioning the term *spectral equivalence*. This term is introduced by D'Yakonov [110] in 1966.

A more general definition of an equivalence relation can be based on $\mathrm{cond}(\cdot)$.[6]

Remark 7.57. (a) Assume that $A_\nu \geq 0$ but not $A_\nu > 0$. Then $A_\nu \sim B_\nu$ implies that B_ν is also semidefinite and that both matrices have coinciding kernels.

(b) If $A_\nu > 0$ and $B_\nu > 0$, then (7.50) is equivalent to $\sup_{\nu\in F} \kappa(A_\nu^{-1} B_\nu) < \infty$.

Proof. Rewriting (7.50) using (7.48f), part (a) is obvious. For part (b), use Lemma 7.53. □

Proposition 7.58. *Let the matrices A_ν, B_ν, C_ν, D_ν be positive semidefinite. The spectral equivalence relation satisfies the following rules:*

$$A_\nu \sim B_\nu \text{ and } \lambda \geq 0 \quad \Rightarrow \quad \lambda A_\nu \sim \lambda B_\nu\,, \tag{7.51a}$$

$$A_\nu \sim B_\nu \text{ and } C_\nu \sim D_\nu \quad \Rightarrow \quad A_\nu + C_\nu \sim B_\nu + D_\nu, \tag{7.51b}$$

$$A_\nu \sim B_\nu \quad \Rightarrow \quad A_\nu^{-1} \sim B_\nu^{-1} \qquad \text{if } A_\nu > 0, \tag{7.51c}$$

$$A_\nu \sim B_\nu \text{ and } C_\nu \in \mathbb{K}^{J\times I} \quad \Rightarrow \quad C_\nu A_\nu C_\nu^{\mathsf{H}} \sim C_\nu B_\nu C_\nu^{\mathsf{H}}, \tag{7.51d}$$

$$A_\nu \sim B_\nu \quad \Rightarrow \quad A_\nu^{-1/2} B_\nu A_\nu^{-1/2} \sim I \qquad \text{if } A_\nu > 0. \tag{7.51e}$$

In the cases (7.51a,b,d), the constant c in (7.50) is identical on both sides. The matrix C_ν in (7.51d) may be any rectangular matrix.

Proof. The implications (7.51a,b,d) are an immediate consequence of (7.48f). For (7.51c), use (7.48d,e). Statement (7.51d) implies (7.51e). □

We recall that the iteration $\Phi(x,b) = x - \vartheta W^{-1}(Ax - b)$ with optimal damping has the convergence rate $\rho(M_{\vartheta_{\mathrm{opt}}}) = \frac{\kappa-1}{\kappa+1}$ with $\kappa = \kappa(W^{-1}A)$.

Conclusion 7.59. *(a) Assume $A, W, W' > 0$ and $W \leq c'W'$, $W' \leq cW$. Then the linear iterations $\Phi(x,b) = x - \vartheta W^{-1}(Ax-b)$ and $\Phi'(x,b) = x - \vartheta W'^{-1}(Ax-b)$ with optimal damping have comparable convergence rates determined by*

$$\kappa = \kappa(W^{-1}A), \quad \kappa' = \kappa(W'^{-1}A) \quad \textit{with} \quad \frac{1}{c'}\kappa' \leq \kappa \leq c\kappa'.$$

[6] Consider families of regular matrices. Let $\mathrm{cond}(A) = \|A\|\|A^{-1}\|$ be defined with respect to some submultiplicative matrix norm. Analogously to Remark 7.57b, we define

$$\mathcal{A} \sim_{\mathrm{cond}} \mathcal{B} \quad :\Leftrightarrow \quad \sup_{\nu\in F} \mathrm{cond}(A_\nu^{-1} B_\nu) < \infty.$$

Also in this case, the properties (7.47a,e) prove that $\sim_{\mathrm{cond}}$ is an equivalence relation.

(b) If the family $\mathcal{A} = (A_h)_{h\in H}$ *is indexed by the step size and* $\kappa(W_h^{-1}A_h) = \mathcal{O}(h^{-\tau})$ *holds with* $\tau > 0$*, the convergence of* $\Phi_{\vartheta_{\mathrm{opt}}}$ *is of the order* τ*. All linear iterations* $\Phi'_h(x,b) = x - \vartheta W_h'^{-1}(A_h x - b)$ *with* $W'_h \sim W_h$ *have the same convergence order* τ*. Hence the convergence order is a property of the equivalence class.*

The optimal convergence order is $\tau = 0$ characterised by $\rho(M_h) \le c < 1$. In the case of linear iterations $\Phi(x,b) = x - \vartheta W^{-1}(Ax - b)$ with $A, W > 0$, the latter inequality can be ensured by the next statement.

Proposition 7.60. *The family of linear iterations* $\Phi_h(x,b) = x - \vartheta_{\mathsf{opt}} W_h^{-1}(A_h x - b)$ *with* $A_h, W_h > 0$ *satisfies*

$$\rho(M_h) \le c < 1 \qquad \textit{for all } h \in H$$

if and only if

$$A_h \sim W_h\,.$$

Proof. $A_h \sim W_h$ implies that $\kappa_h = \kappa(W_h^{-1}A_h) = \mathcal{O}(1)$. Hence $\rho(M_{\vartheta_{\mathrm{opt}}}) = \frac{\kappa_h - 1}{\kappa_h + 1} \le c < 1$ holds with $c := \sup\{2/(1+\kappa_h) : h \in H\}$. □

This result shows a way how to obtain optimal convergence, provided that $W_h^{-1}(A_h x - b)$ is easy to evaluate. As in Remark 7.7 we have to ask on what data the choice of W_h could be based. Using only the data of A_h, the traditional techniques do not lead to $A_h \sim W_h$ in general. In §13.4 we shall propose a new technique which is able to satisfy $A_h \sim W_h$.

7.4.5 Equivalent Bilinear Forms

We recall the Definition E.2 of coercive forms.

Definition 7.61. Two symmetric and coercive sesquilinear forms $a, b : V \times V \to \mathbb{C}$ are called *equivalent* (notation: $a \sim b$) if there is some $c > 0$ with

$$\frac{1}{c}\, a(u,u) \le b(u,u) \le c\, a(u,u) \qquad \text{for all } u \in V. \tag{7.52}$$

For simplicity, we use the term *bilinear* which suits for $\mathbb{K} = \mathbb{R}$. For $\mathbb{K} = \mathbb{C}$, the form must be sesquilinear (cf. Definition E.1).

The Galerkin matrix A_h corresponding to the bilinear form $a(\cdot,\cdot)$ satisfies

$$\langle A_h x, y\rangle = a(P_h x, P_h y) \qquad \text{for all } x, y \in \mathbb{K}^I,$$

where $\langle\cdot,\cdot\rangle$ is the Euclidean scalar product (cf. Exercise E.5). The mapping $P_h : \mathbb{K}^I \to V_n \subset V$ is defined in (E.6).

Applying (7.52) to $u = P_h x$, we obtain

$$\frac{1}{c}\,\langle A_h x, x\rangle \le \langle B_h x, x\rangle \le c\,\langle A_h x, x\rangle \qquad \text{for all } x \in \mathbb{K}^I, \tag{7.53a}$$

where B_h is the Galerkin matrix corresponding to the bilinear form $b(\cdot,\cdot)$. The symmetry of a and b implies that A_h and B_h are also Hermitian (cf. Exercise E.6a). Therefore the property (7.53a) is equivalent to the inequalities

$$\frac{1}{c} A_h \le B_h \le c\, A_h \tag{7.53b}$$

in the sense of §C.1.1.

Note that the constants in (7.52) and (7.53b) coincide. Therefore they hold for all discretisation parameters $h \in H$ which form the families $\mathcal{A} = (A_h)_{h\in H}$ and $\mathcal{B} = (B_h)_{h\in H}$. Using the notion of equivalence, we obtain the following statement.

Proposition 7.62. *Equivalent forms $a \sim b$ produce equivalent Galerkin matrix families $\mathcal{A} \sim \mathcal{B}$.*

A potential practical strategy is the following. Let a and $\mathcal{A}$ correspond to the problem to be solved. If there is a simpler but equivalent form b, it may be that the corresponding matrices B_h are easier to handle. Either $W_h^{-1}\delta = d$ can be solved for $W_h = B_h$ or for another choice $W_h \sim B_h$. By Proposition 7.62, $A_h \sim W_h$ holds as required in Proposition 7.60.

Conclusion 7.63. *Let the symmetric and coercive form $a(\cdot,\cdot)$ correspond to a boundary value problem for $u \in V := H_0^1(\Omega)$. Use the same finite element discretisation for $a(\cdot,\cdot)$ and the standard Poisson problem. Then both discretisation matrices are spectrally equivalent.*

7.5 Time-Stepping Methods

The term of a time-stepping method is used, in particular, in the engineering community. The function $x(t)$, $0 \le t < \infty$, is introduced as a solution of the system of ordinary differential equations

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = b - Ax \qquad \text{with the initial value } x(0) = x^0. \tag{7.54}$$

If A is positive definite (or if $\Re\mathfrak{e}(\lambda) > 0$ holds for all eigenvalues $\lambda \in \sigma(A)$), then $x(t)$ converges for $t \to \infty$ to the solution $x^* := A^{-1}b$, which is now interpreted as the stationary solution of (7.54). The time-stepping method tries to discretise the differential equation by a grid

$$0 = t_0 < t_1 < \ldots$$

and to approximate $x(t)$ for a large $t = t_m$. One explicit Euler step with the time step $\triangle t := t_{m+1} - t_m$ reads as

$$x(t_{m+1}) \approx x^{m+1} = x^m - \triangle t\,(Ax^m - b) \tag{7.55}$$

(cf. Quarteroni–Sacco–Saleri [314, §11.2]). For a fixed (or variable) step size $\triangle t$, recursion (7.55) describes the stationary (or instationary) Richardson method.

Often Runge–Kutta-like methods are proposed. For example, the Heun method becomes

$$x' := x^m - \alpha\triangle t\,(Ax^m - b), \quad x(t_{m+1}) \approx x^{m+1} = x^m - \beta\triangle t\,(Ax' - b) \tag{7.56}$$

with $\alpha = \frac{1}{2}$ and $\beta = 1$ (cf. Heun [220]; in the true Runge–Kutta case, there are four coefficients; cf. Runge [327] and Kutta [250]).

While the original discretisation methods try to achieve small discretisation errors $\|x^m - x(t_m)\|$ for all grid points t_m, the coefficients $\alpha,\ \beta$ are now chosen such that the convergence $x^m \to x^*$ is improved. The produced methods (as, e.g., (7.56)) are the semi-iterative variants of the Richardson iteration which will be described in §8.3.7.

In the language of ordinary differential equations, one explains the unfavourably slow convergence of the Richardson variants by the *stiffness* of the system. When preconditioning is introduced to speed up the convergence:

$$x^{m+1} = x^m - \triangle t\, W^{-1}(Ax^m - b),$$

this is called a *quasi-time stepping* method, which however does no longer approximate the equation (7.54) but only the same stationary solution x^*.

In essence, the interpretation by a time-stepping method is misleading (e.g., since the high consistency order of a Runge–Kutta method is given up for purposes which are not connected with this method). In particular, this concept is of no help for analysing the iteration or for constructing efficient iterations.

7.6 Nested Iteration

Three families of linear iterations, the multigrid iteration, the domain decomposition methods, and the hierarchical LU iteration will be described in Part III. The multigrid method is usually combined with the *nested iteration*. As shown in §11.5, the nested iteration technique can be combined with any linear or nonlinear iteration. It does not change the iteration, but yields advantageous starting values.