Chapter 8
Preconditioning

8.1 The General Approach

Remark 8.1. Motivation and idea. It was seen in Chapter 7 that the number of iterations might depend on the condition number of the matrix. In order to reduce the number of iterations, one wants to replace the original linear system of equations (1.1) by an equivalent system whose system matrix has a smaller condition number. This strategy is called preconditioning.

The main idea of preconditioning consists in applying the iterative method to the equivalent system

$$M^{-1}A\underline{x} = M^{-1}\underline{b} \quad \text{(preconditioning from left)}$$

or

$$AM^{-1}\underline{y} = \underline{b}, \ \underline{x} = M^{-1}\underline{y} \quad \text{(preconditioning from right).}$$

The non-singular matrix M is called preconditioner. This matrix should satisfy two requirements:

- The convergence of the iterative method for the system with the matrix $M^{-1}A$ or AM^{-1}, respectively, should be faster than for the original system with the matrix A. That means, M^{-1} should be a good approximation to A^{-1}.
- Linear systems with the matrix M should be solvable with low costs.

In general, one has to find a compromise between these two requirements.

Usually, left and right preconditioning lead to different methods which might behave sometimes quite differently. □

Remark 8.2. Some preconditioners. An easy way to construct preconditioners consists in starting with the decomposition $A = D + L + U$, see Section 3.2, and using parts of this decomposition which are easily invertible:

- $M = D$, diagonal preconditioner, Jacobi preconditioner,
- $M = D + L$, forward Gauss–Seidel preconditioner,

- $M = D + U$, backward Gauss–Seidel preconditioner,
- $M = (D + L)\, D^{-1} (D + U)$, symmetric Gauss–Seidel preconditioner.

Damped versions of the classical iterative schemes can be also used. A more advanced preconditioner will be presented in Section 8.3.

Note that M or M^{-1} do not need to be known explicitly. They can also stand for some numerical (iterative) method for solving linear systems of equations. Then, M^{-1} means that this method should be applied to a vector. □

Remark 8.3. Change in algorithms for general matrices if the preconditioner is applied. In algorithms for general matrices A, preconditioning from left consists in replacing A by $M^{-1}A$ and $\underline{r}^{(k)}$ by $M^{-1}\underline{r}^{(k)}$ in the algorithms. Then, e.g., GMRES computes the iterate

$$\underline{x}^{(k)} \in \underline{x}^{(0)} + K_k \left(M^{-1}\underline{r}^{(0)}, M^{-1}A \right)$$

such that $\left\| M^{-1}\underline{r}^{(k)} \right\|_2$ becomes minimal. □

8.2 Symmetric Matrices

Remark 8.4. A difficulty and its solution. A problem occurs if the matrix A is symmetric and the iterative method wants to exploit this property, e.g., using short recurrences, since in general neither $M^{-1}A$ nor AM^{-1} are symmetric. This problem can be solved by constructing the orthonormal basis of the Krylov subspace with respect to an appropriate inner product.

Let H be a Hilbert[1] space with the inner product $(\cdot,\cdot)_H$ and $\mathcal{L} : H \to H$ be a linear map. This map is called self-adjoint with respect to $(\cdot,\cdot)_H$ if

$$(\mathcal{L}v, w)_H = (v, \mathcal{L}w)_H \quad \forall\, v, w \in H.$$

In the case $H = \mathbb{R}^n$ equipped with the standard Cartesian basis and the Euclidean inner product $(\cdot,\cdot)$, a linear map, which is represented by a matrix A, is self-adjoint if

$$(A\underline{x}, \underline{y}) = (\underline{x}, A\underline{y}) \quad \forall\, \underline{x}, \underline{y} \in \mathbb{R}^n.$$

This condition is equivalent to A being symmetric.

If the preconditioner M is symmetric and positive definite, then

$$(\underline{x}, \underline{y})_M = (\underline{x}, M\underline{y}), \quad \forall\, \underline{x}, \underline{y} \in \mathbb{R}^n$$

[1] David Hilbert (1862 – 1943)

defines an inner product in $\mathbb{R}^n$. The induced norm is given by $\|\underline{x}\|_M = (\underline{x}, \underline{x})_M^{1/2}$.

Consider for the remainder of this section preconditioning from left. The matrix $M^{-1}A$ is self-adjoint with respect to this inner product since

$$(M^{-1}A\underline{x}, \underline{y})_M = (M^{-1}A\underline{x}, M\underline{y}) = (A\underline{x}, \underline{y}) = (\underline{x}, A\underline{y}) = (\underline{x}, M^{-1}A\underline{y})_M$$

for all $\underline{x}, \underline{y} \in \mathbb{R}^n$.

Now, one can generate an orthonormal basis with respect to the inner product $(\cdot,\cdot)_M$ of $K_k\left(M^{-1}\underline{r}^{(0)}, M^{-1}A\right)$ by an appropriate modification of the Lanczos algorithm. □

Algorithm 8.5. Preconditioned Lanczos algorithm for symmetric matrices. Given a symmetric matrix $A \in \mathbb{R}^{n\times n}$, a symmetric positive definite matrix $M \in \mathbb{R}^{n\times n}$, and $\underline{r}^{(0)} \in \mathbb{R}^n$.

```
$\underline{z} = M^{-1}\underline{r}^{(0)}$
$\underline{q}_1 = \dfrac{\underline{z}}{(\underline{r}^{(0)}, \underline{z})^{1/2}}$
$\beta_0 = 0$
$\underline{q}_0 = \underline{0}$
for $j = 1 : k$
    $\underline{s} = A\underline{q}_j$
    $\underline{z} = M^{-1}\underline{s}$
    $\alpha_j = \left(\underline{s}, \underline{q}_j\right)$
    $\underline{z} = \underline{z} - \alpha_j\underline{q}_j - \beta_{j-1}\underline{q}_{j-1}$
    $\beta_j = (\underline{s}, \underline{z})^{1/2}$
    $\underline{q}_{j+1} = \underline{z}/\beta_j$
endfor
```

□

Remark 8.6. On the preconditioned Lanczos algorithm for symmetric matrices.

- The vector $\underline{z}$ is computed by solving $M\underline{z} = \underline{s}$.
- The matrix form of the preconditioned Lanczos algorithm is

$$M^{-1}AQ_k = Q_{k+1}H_k \text{ with } H_k = \begin{pmatrix} \alpha_1 & \beta_1 & 0 & \cdots & 0 & 0 \\ \beta_1 & \alpha_2 & \beta_2 & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ & & & & & \\ 0 & 0 & 0 & \cdots & \alpha_{k-1} & \beta_{k-1} \\ 0 & 0 & 0 & \cdots & \beta_{k-1} & \alpha_k \\ 0 & 0 & 0 & \cdots & 0 & \beta_k \end{pmatrix} \in \mathbb{R}^{(k+1)\times k}. \tag{8.1}$$

The columns of Q_{k+1} are orthogonal with respect to $(\cdot,\cdot)_M$

$$Q_{k+1}^T M Q_{k+1} = I \in \mathbb{R}^{(k+1)\times(k+1)}. \tag{8.2}$$

□

Remark 8.7. On the orthogonality condition for the preconditioned conjugate gradient method. The preconditioned conjugate gradient method (PCG) is one of the most important algorithms for solving linear systems of equations with symmetric and positive definite matrix. Besides the preconditioned Lanczos algorithm, one needs to implement the orthogonality condition of the residual with respect to $(\cdot,\cdot)_M$ with a short recurrence. Concretely, one has to construct

$$\underline{x}^{(k)} \in \underline{x}^{(0)} + K_k\left(M^{-1}\underline{r}^{(0)}, M^{-1}A\right) \tag{8.3}$$

such that $M^{-1}\underline{r}^{(k)} = M^{-1}\left(\underline{b} - A\underline{x}^{(k)}\right)$ is orthogonal to $K_k\left(M^{-1}\underline{r}^{(0)}, M^{-1}A\right)$ with respect to $(\cdot,\cdot)_M$

$$M^{-1}\underline{r}^{(k)} \perp_M K_k\left(M^{-1}\underline{r}^{(0)}, M^{-1}A\right) \iff M^{-1}\underline{r}^{(k)} \perp_M Q_k. \tag{8.4}$$

Using the definition of $\underline{q}_1$, see Algorithm 8.5, lines 1 and 2, it is by (8.2)

$$\begin{aligned}\left(\underline{q}_k, \underline{r}^{(0)}\right) &= \left(\underline{q}_k, MM^{-1}\underline{r}^{(0)}\right) = \left\|M^{-1}\underline{r}^{(0)}\right\|_M \left(\underline{q}_k, M\underline{q}_1\right) \\ &= \left\|M^{-1}\underline{r}^{(0)}\right\|_M \delta_{1k},\end{aligned} \tag{8.5}$$

where δ_{ij} is the Kronecker symbol. Since by construction $\underline{x}^{(k)} = \underline{x}^{(0)} + Q_k\underline{y}_k$ for some $\underline{y}_k \in \mathbb{R}^k$, one obtains, for the desired orthogonality condition (8.4), with $\beta = \left\|M^{-1}\underline{r}^{(0)}\right\|_M$, (8.5), (8.1), and (8.2) the condition

$$\begin{aligned}\underline{0} &= \left(Q_k, M^{-1}\underline{r}^{(k)}\right)_M = \left(Q_k, MM^{-1}\underline{r}^{(k)}\right) \\ &= \left(Q_k, \underline{r}^{(k)}\right) = \left(Q_k, \underline{r}^{(0)} - AQ_k\underline{y}_k\right) \\ &= \beta\boldsymbol{e}_1 - Q_k^T AQ_k\underline{y}_k = \beta\boldsymbol{e}_1 - Q_k^T MQ_{k+1}H_k\underline{y}_k \\ &= \beta\boldsymbol{e}_1 - Q_k^T M\left[Q_k\underline{q}_{k+1}\right]H_k\underline{y}_k = \beta\boldsymbol{e}_1 - \left[I\underline{0}\right]H_k\underline{y}_k \\ &= \beta\boldsymbol{e}_1 - \tilde{H}_k\underline{y}_k,\end{aligned}$$

where $\tilde{H}_k \in \mathbb{R}^{k\times k}$ is the matrix consisting of the first k rows of H_k. Hence, $\underline{y}_k$ can be computed from $\tilde{H}_k$, which is known from the preconditioned Lanczos algorithm, from what follows, with analogous calculations as in Section 6.2, that $\underline{x}^{(k)}$ can be computed with a short recurrence. Finally, one obtains PCG. □

Algorithm 8.8. Preconditioned conjugate gradient (PCG). Given a symmetric positive definite matrix $A \in \mathbb{R}^{n \times n}$, a right-hand side $\underline{b} \in \mathbb{R}^n$, an initial iterate $\underline{x}^{(0)} \in \mathbb{R}^n$, a tolerance $\varepsilon > 0$, and a symmetric positive definite preconditioner $M \in \mathbb{R}^{n \times n}$.

```
 1.  $\underline{r}^{(0)} = \underline{b} - A\underline{x}^{(0)}$
 2.  solve $M\underline{z}_0 = \underline{r}^{(0)}$
 3.  $\underline{p}_1 = \underline{z}_0$
 4.  $k = 0$
 5.  while $(\underline{z}_k, \underline{r}^{(k)})^{1/2} > \varepsilon$
 6.     $k = k + 1$
 7.     $\underline{s} = A\underline{p}_k$
 8.     $\nu_k = \dfrac{(\underline{z}_{k-1}, \underline{r}^{(k-1)})}{(\underline{p}_k, \underline{s})}$
 9.     $\underline{x}^{(k)} = \underline{x}^{(k-1)} + \nu_k \underline{p}_k$
10.     $\underline{r}^{(k)} = \underline{r}^{(k-1)} - \nu_k \underline{s}$
11.     solve $M\underline{z}_k = \underline{r}^{(k)}$
12.     $\mu_{k+1} = \dfrac{(\underline{z}_k, \underline{r}^{(k)})}{(\underline{z}_{k-1}, \underline{r}^{(k-1)})}$
13.     $\underline{p}_{k+1} = \underline{z}_k + \mu_{k+1} \underline{p}_k$
14.  endwhile
```

□

Remark 8.9. On PCG. There exists also other ways to implement PCG. Compared with CG, one has to solve the linear system with the matrix M (to apply the preconditioner), line 11, and one has to store one additional vector (five vectors altogether). □

Remark 8.10. Preconditioners for PCG. If PCG should be applied for solving $A\underline{x} = \underline{b}$ with A being symmetric and positive definite, also M has to be symmetric and positive definite. Among the preconditioners given in Remark 8.2, the Jacobi and the symmetric Gauss–Seidel preconditioner possess this property.

For the Jacobi preconditioner, it follows from $\underline{x}^T A \underline{x} > 0$ that $\underline{x}^T D \underline{x} > 0$ for all $\underline{x} \in \mathbb{R}^n \setminus \{0\}$, see Remark 2.11.

For the symmetric Gauss–Seidel preconditioner, one has

$$\underline{x}^T (D + L) D^{-1} (D + U) \underline{x} = \underline{x}^T (D + U)^T D^{-1} \underbrace{(D + U) \underline{x}}_{\underline{y} \neq \underline{0}} = \underline{y}^T D^{-1} \underline{y} > 0,$$

since with $d_{ii} > 0$, it follows firstly that the matrix $D + U$ is non-singular such that $(D + U) \underline{x} \neq \underline{0}$ if $\underline{x} \neq \underline{0}$. Secondly, it follows that $d_{ii}^{-1} > 0$

In addition, multigrid methods and incomplete Cholesky factorizations, see Remark 8.23, can be also used as preconditioners. □

Theorem 8.11. Estimate of the rate of convergence for s.p.d. matrices, minimization of the error. *Let $A, M \in \mathbb{R}^{n\times n}$ be symmetric and positive definite. Then, the k-th iterate of the PCG method satisfies*

$$\frac{\left\| \underline{x} - \underline{x}^{(k)} \right\|_A}{\left\| \underline{x} - \underline{x}^{(0)} \right\|_A} \le 2 \left(\frac{\sqrt{\kappa_2\left(M^{-1/2}AM^{-1/2}\right)} - 1}{\sqrt{\kappa_2\left(M^{-1/2}AM^{-1/2}\right)} + 1} \right)^k .$$

The iterate of PCG method minimizes the error in the norm $\|\cdot\|_A$ within all vectors of form (8.3).

Proof. The proof follows the lines of the proof of Theorems 7.6 and 6.12. ■

Remark 8.12. On the spectral condition number for the preconditioned system. The matrix $M^{-1/2}$ is the inverse of $M^{1/2}$, compare Remark 2.12 for the definition of the square root of a symmetric positive definite matrix. The matrices $M^{-1/2}AM^{-1/2}$ and $M^{-1}A$ are similar, i.e., there is a non-singular matrix S such that $M^{-1}A = SM^{-1/2}AM^{-1/2}S^{-1}$. Obviously, it is $S = M^{-1/2}$. Since $M^{-1/2}AM^{-1/2}$ is symmetric and positive definite, cf. Remark 2.12, and similar matrices have the same eigenvalues, it follows that

$$\kappa_2\left(M^{-1/2}AM^{-1/2}\right) = \frac{\lambda_{\max}(M^{-1/2}AM^{-1/2})}{\lambda_{\min}(M^{-1/2}AM^{-1/2})} = \frac{\lambda_{\max}(M^{-1}A)}{\lambda_{\min}(M^{-1}A)}.$$

This formula means, if M is a good preconditioner, i.e., the ratio of the largest and smallest eigenvalue of $M^{-1}A$ is small, then the worst case upper bound for the number of iterations is reduced by using PCG instead of CG. In practice, also the number of iterations with PCG becomes usually smaller compared with the number for CG. □

8.3 Incomplete LU Factorization

Remark 8.13. Idea. One drawback of the application of direct solvers for linear systems of equations with a sparse matrix is the additional fill-in that occurs if a decomposition of the matrix, like the LU decomposition, is computed. A main part of popular direct solvers for sparse linear systems, like `UMFPACK`, see Davis (2004), which is the package behind the backslash command in MATLAB, is a reordering of the unknowns such that the fill-in is reduced. In the context of preconditioning, methods that are based on the LU decomposition can be constructed that respect the sparsity pattern or zero pattern of the matrix. □

Algorithm 8.14. Incomplete LU factorization (ILU). Given a matrix $A \in \mathbb{R}^{n\times n}$ and a zero pattern $P \subset \{(i,j) \ : \ i \neq j, 1 \le i,j \le n\}$.

```
for $k = 1 : n-1$     % loop over the rows
  for $i = k+1 : n$     % loop over rows below diagonal
    if $(i,k) \notin P$
      $a_{ik} = a_{ik}/a_{kk}$
      for $j = k+1 : n$     % columns right of diagonal
        if $(i,j) \notin P$
          $a_{ij} = a_{ij} - a_{ik}a_{kj}$
        endif
      endfor
    endif
  endfor
endfor
```

□

Remark 8.15. To Algorithm 8.14.

- For $P = \emptyset$, Algorithm 8.14 is just a standard LU factorization of the matrix A without pivot strategy. Then, Algorithm 8.14 replaces A by the factors L and U

$$\begin{aligned} u_{ij}&, \text{ for } 1 \le i \le j \le n, \text{ upper triangular matrix,} \\ l_{ij}&, \text{ for } 1 \le j < i \le n, \text{ lower triangular matrix,} \end{aligned}$$

  where the diagonal entries of L are all 1 and they are not stored. It holds that $A = LU$.
- If $P \neq \emptyset$, then it is

$$A = LU - N \text{ with } 0 \neq N \in \mathbb{R}^{n\times n}. \tag{8.6}$$

  In this case, one needs extra memory to store the factors L and U.
- Usually, one calls the algorithm ILU if the zero pattern P is chosen to be the zero pattern of A. Sometimes, this algorithm is called also ILU(0).
- ILU is a popular preconditioner with $M = LU$, where the factors are defined in (8.6). For using ILU as preconditioner, it is essential that the diagonal entries do not belong to P since a linear system of equations with matrix U has to be solved.

Now, some properties of ILU will be studied for an important class of matrices. □

Definition 8.16. Non-negative matrix. A matrix $B \in \mathbb{R}^{n\times n}$ is called non-negative, if

$$b_{ij} \ge 0, \quad \forall\, i,j = 1,\ldots,n.$$

The notation is $B \geq 0$ (and $B > 0$ if $b_{ij} > 0$ for all $i, j = 1, \ldots, n$). The same notations will be used to indicate vectors with non-negative or positive entries. □

Definition 8.17. M-matrix. A matrix $A \in \mathbb{R}^{n\times n}$ is called M-matrix if it satisfies the following conditions:

1. $a_{ij} \leq 0$ for $i, j = 1, \ldots, n$, $i \neq j$,
2. A is non-singular and A^{-1} is non-negative.

□

Remark 8.18. M-matrices and strongly diagonally dominant matrices. M-matrices arise in some discretizations of partial differential equations. It is an important class of matrices, which is closely connected to a certain class of diagonally dominant matrices.

A proper Minkowski matrix is a matrix with the following properties

i) The main diagonal entries are non-negative

$$a_{ii} \geq 0 \quad i = 1, \ldots, n.$$

ii) The off-diagonal entries are non-positive

$$a_{ij} \leq 0, \quad i, j = 1, \ldots, n, \ i \neq j.$$

iii) All row sums of A are positive.

It follows that a proper Minkowski matrix is a strongly diagonally dominant matrix. Diagonal dominant matrices are matrices with favorable mathematical properties, e.g., compare Theorem 3.11.

It can be shown that each M-matrix $A \in \mathbb{R}^{n\times n}$ can be obtained from a proper Minkowski matrix $\tilde{A}$ by scaling each column of $\tilde{A}$ with an appropriate positive number. In this way, M-matrices are in a somewhat hidden sense strongly diagonally dominant and they possess also favorable mathematical properties. □

Lemma 8.19. Alternative characterization of an M-matrix, majorizing element. *Let $A \in \mathbb{R}^{n\times n}$ with $a_{ij} \leq 0$ for $i \neq j$, then A is an M-matrix if and only if there is a vector $\underline{v} \in \mathbb{R}^n$, with $\underline{v} > 0$, such that $A\underline{v} > 0$. A vector $\underline{v}$ with this property is called majorizing element.*

Proof. Let A be an M-matrix. Then, there is a diagonal matrix D with $d_{ii} > 0$ such that $A = \tilde{A}D$, where $\tilde{A}$ is a proper Minkowski matrix, see Remark 8.18. Choosing $v_i = d_{ii}^{-1} > 0$, $i = 1, \ldots, n$, yields

$$A\underline{v} = \tilde{A}D\underline{v} = \tilde{A}\underline{1},$$

where $\underline{1}$ is the vector where all entries are 1. Since each row sum of a proper Minkowski matrix is positive, it follows that $A\underline{v} > 0$.

The proof of the other direction follows (Bohl, 1981, pp. 34). Let there be a vector $\underline{v} > 0$ such that $A\underline{v} > 0$, i.e., such that $\sum_{j=1}^{n} a_{ij}v_j > 0$ for $i = 1, \ldots, n$. Since all terms with off-diagonal entries are non-positive, it follows that $a_{ii} > 0$, for $i = 1, \ldots, n$. Consequently, the matrix $D = \mathrm{diag}(a_{ii})$ is non-singular. Define the matrix

$$B = D^{-1}(D - A) = I - D^{-1}A. \tag{8.7}$$

It is $B \ge 0$ since $D^{-1} \ge 0$ and $(D - A) \ge 0$. From (8.7), it follows that

$$A = D(I - B). \tag{8.8}$$

Applying the assumptions and multiplying with D^{-1} gives

$$A\underline{v} = D(I - B)\underline{v} > 0 \quad \Longleftrightarrow \quad (I - B)\underline{v} > 0 \quad \Longleftrightarrow \quad \underline{v} > B\underline{v}. \tag{8.9}$$

Define with the vector $\underline{v}$ a norm in $\mathbb{R}^n$, namely a weighted maximum norm,

$$\|\underline{w}\|_{\underline{v}} = \max_{i=1,\ldots,n} \left\{ |w_i| \, v_i^{-1} \right\}$$

and a corresponding induced matrix norm

$$\|P\|_{\underline{v}} = \max_{\underline{w}\in\mathbb{R}^n, \|\underline{w}\|_{\underline{v}}=1} \|P\underline{w}\|_{\underline{v}}.$$

For the matrix norm, one obtains

$$\begin{aligned} \|P\|_{\underline{v}} &= \max_{\underline{w}\in\mathbb{R}^n, \|\underline{w}\|_{\underline{v}}=1} \max_{i=1,\ldots,n} \left(\left| \sum_{j=1}^n p_{ij} w_j \right| v_i^{-1} \right) \\ &= \max_{i=1,\ldots,n} \left(\max_{\underline{w}\in\mathbb{R}^n, \|\underline{w}\|_{\underline{v}}=1} \left(\left| \sum_{j=1}^n p_{ij} w_j \right| \right) v_i^{-1} \right). \end{aligned}$$

The sum for each row is maximized when each entry of $\underline{w}$ has the maximal absolute value, so that $\|\underline{w}\|_{\underline{v}} = 1$ holds, and the appropriate sign, i.e., if $w_j = \mathrm{sgn}(p_{ij})v_j$, $j = 1, \ldots, n$. If $P \ge 0$, then the maximum is taken for $\underline{w} = \underline{v}$, i.e.,

$$\|P\|_{\underline{v}} = \max_{i=1,\ldots,n} \left(\left(\sum_{j=1}^n p_{ij} v_j \right) v_i^{-1} \right) = \|P\underline{v}\|_{\underline{v}}. \tag{8.10}$$

Applying this result to the matrix B gives with (8.9)

$$\|B\|_{\underline{v}} = \|B\underline{v}\|_{\underline{v}} < \|\underline{v}\|_{\underline{v}} = 1.$$

Since with (8.10) on has $\|I\|_{\underline{v}} = \|\underline{v}\|_{\underline{v}} = 1$, it follows now that the matrix $(I - B)$ is non-singular. Its inverse is given by $(I - B)^{-1} = \sum_{k=0}^{\infty} B^k$, because

$$(I - B)(I - B)^{-1} = \sum_{k=0}^{\infty} B^k - \sum_{k=1}^{\infty} B^k = I + \left(\sum_{k=1}^{\infty} B^k - \sum_{k=1}^{\infty} B^k \right) = I,$$

since the geometric series is absolutly convergent for $\|B\|_{\underline{v}} < 1$. From (8.8), one obtains that A is also non-singular with

$$A^{-1} = (I - B)^{-1} D^{-1} = \left(\sum_{k=0}^{\infty} B^k \right) D^{-1}.$$

Since all terms in the sum are non-negative and D^{-1} is non-negative, too, it follows that $A^{-1} \geq 0$. Hence, A satisfies all criteria of Definition 8.17, such that A is an M-matrix. ■

Lemma 8.20. M-matrices and Gaussian algorithm. *Let $A \in \mathbb{R}^{n\times n}$ be an M-matrix and let $A^{(1)} \in \mathbb{R}^{n\times n}$ be the matrix that is obtained as result of the first step of the Gaussian algorithm, which transforms A into an upper triangular matrix. Then, $A^{(1)}$ is also an M-matrix.*

Proof. The first step of the Gaussian algorithm can be written in the following form

$$A^{(1)} = L^{(1)}A = \begin{pmatrix} 1 & & & & \\ -a_{21}/a_{11} & 1 & & & \\ -a_{31}/a_{11} & 0 & 1 & & \\ \vdots & \vdots & & \ddots & \\ -a_{n1}/a_{11} & 0 & \cdots & \cdots & 1 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}.$$

For the off-diagonal elements of $A^{(1)}$, it is

$$\begin{aligned} a_{1j}^{(1)} &= a_{1j} \leq 0, \quad j > 1, \\ a_{i1}^{(1)} &= 0, \quad i > 1 \text{ (by construction)}, \\ a_{ij}^{(1)} &= \underbrace{a_{ij}}_{\leq 0} - \underbrace{\frac{a_{i1}}{a_{11}}}_{\leq 0} \underbrace{a_{1j}}_{\leq 0} \leq 0, \quad i,j > 1,\ i \neq j, \end{aligned}$$

since in the last line the first term is not positive and the second term is non-negative.

Because A is an M-matrix, there is by Lemma 8.19 a vector $\underline{v}$ with $\underline{v} > 0$, such that $A\underline{v} > 0$. Since $L^{(1)} \geq 0$, and there is a non-zero entry in each row of $L^{(1)}$, it follows that $A^{(1)}\underline{v} = L^{(1)}A\underline{v} > 0$. Since the off-diagonals of $A^{(1)}$ are not positive, it follows by Lemma 8.19 that $A^{(1)}$ is an M-matrix. ■

Lemma 8.21. Comparison criterion. *Let $A \in \mathbb{R}^{n\times n}$ be an M-matrix and let $B \in \mathbb{R}^{n\times n}$ be a matrix with*

$$\begin{aligned} a_{ij} &\leq b_{ij} \ \forall\ i,j = 1,\ldots,n, \\ b_{ij} &\leq 0 \quad \forall\ i,j = 1,\ldots,n, i \neq j. \end{aligned}$$

Then, B is also an M-matrix.

Proof. Since A is an M-matrix, there exists a vector $\underline{v} > 0$ such that $A\underline{v} > 0$. By assumption, it holds that $B = A + N$ with $N \geq 0$. It follows that

$$B\underline{v} = \underbrace{A\underline{v}}_{>0} + \underbrace{N\underline{v}}_{\geq 0} > 0,$$

hence $\underline{v}$ is a majorizing element for B. Because the off-diagonals of B are non-positive by assumption, one concludes from Lemma 8.19 that B is an M-matrix. ■

Theorem 8.22. Properties of an ILU decomposition of an M-matrix. *Let $A \in \mathbb{R}^{n\times n}$ be an M-matrix and let P be a given zero pattern that does not contain the diagonal entries. Then, there is a lower triangular matrix L*

with ones on the main diagonal and an upper triangular matrix U such that $A = LU - N$ with

$$l_{ij} = 0 \textit{ for } (i,j) \in P, \quad u_{ij} = 0 \textit{ for } (i,j) \in P.$$

The matrices L^{-1} and N are non-negative.

Proof. The principal ILU decomposition is computed analogously to the Gaussian elimination

```
1. $A^{(0)} = A$
2. for $k = 1 : n-1$     % loop over the rows
3.     $\tilde{A}^{(k)} = A^{(k-1)} + N^{(k)}$
4.     $A^{(k)} = L^{(k)} \tilde{A}^{(k)}$
5. endfor
```

In the k-th step of line 3, zero entries are generated in $\tilde{A}^{(k)}$ in the zero pattern of the k-th row and the k-th column, i.e., for $(k,j) \in P$ and $(i,k) \in P$. In line 4, the elimination step is applied to $\tilde{A}^{(k)}$. The elimination matrix has the form

$$L^{(k)} = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & -\tilde{a}^{(k)}_{k+1,k}/a^{(k)}_{k,k} & 1 & & \\ & & \vdots & & \ddots & \\ & & -\tilde{a}^{(k)}_{n,k}/a^{(k)}_{k,k} & & \dots & 1 \end{pmatrix}. \tag{8.11}$$

The matrix $A^{(0)} = A$ is an M-matrix. Hence, the off-diagonal entries of $A^{(0)}$ are non-positive from what follows, see line 3, that $N^{(1)}$ is non-negative, since the respective entry of $N^{(1)}$ is either zero or the negative of the same entry as that of $A^{(0)}$. The matrices $A^{(0)}$ and $\tilde{A}^{(1)}$ satisfy the assumptions of the comparison lemma, Lemma 8.21. It follows that $\tilde{A}^{(1)}$ is an M-matrix. All off-diagonal entries of $\tilde{A}^{(1)}$ are non-positive and $\tilde{a}_{11} > 0$. Thus, $L^{(1)}$ is non-negative, compare (8.11). Analogously as in the proof of Lemma 8.20, one can show now that $A^{(1)}$ is an M-matrix.

It can be shown now by induction that $A^{(k)}$ and $\tilde{A}^{(k)}$ are M-matrices for $k = 1, \dots, n-1$, and $N^{(k)}$ and $L^{(k)}$ are non-negative for $k = 1, \dots, n-1$.

By the form of the elimination matrix (8.11), the first k rows of $A^{(k)}$ and $\tilde{A}^{(k)}$ are the same, line 4. In particular, there are only non-zero entries in the first k rows of $A^{(k)}$ at entries which do not belong to the zero pattern. It follows that in the next step in line 3, the first k rows of $N^{(k+1)}$ are zero. From this form of the matrix $N^{(k+1)}$ and from the form of the elimination matrices, one obtains for $i < k+1$

$$L^{(i)} N^{(k+1)} = N^{(k+1)}. \tag{8.12}$$

With this relation, one gets

$$\begin{aligned} A^{(n-1)} &\overset{\text{line 4}}{=} L^{(n-1)} \tilde{A}^{(n-1)} \\ &\overset{\text{line 3}}{=} L^{(n-1)} \left(A^{(n-2)} + N^{(n-1)} \right) \\ &\overset{\text{line 4}}{=} L^{(n-1)} \left(L^{(n-2)} \tilde{A}^{(n-2)} + N^{(n-1)} \right) \end{aligned}$$

$$
\begin{aligned}
&\overset{\text{line 3}}{=} && L^{(n-1)}\left(L^{(n-2)}\left(A^{(n-3)}+N^{(n-2)}\right)+N^{(n-1)}\right) \\
&= && L^{(n-1)}L^{(n-2)}A^{(n-3)}+L^{(n-1)}L^{(n-2)}N^{(n-2)}+L^{(n-1)}N^{(n-1)} \\
&\overset{\text{line 4, 3}}{=} && \ldots \\
&= && \left(\prod_{j=1}^{n-1} L^{(n-j)}\right)A^{(0)}+\sum_{i=1}^{n-1}\left(\prod_{j=1}^{n-i} L^{(n-j)}\right)N^{(i)} \\
&\overset{\text{line 1,(8.12)}}{=} && \left(\prod_{j=1}^{n-1} L^{(n-j)}\right)A+\sum_{i=1}^{n-1}\left(\prod_{j=1}^{n-1} L^{(n-j)}\right)N^{(i)} \\
&= && \left(\prod_{j=1}^{n-1} L^{(n-j)}\right)\left(A+\sum_{i=1}^{n-1} N^{(i)}\right) \\
&=: && L^{-1}\left(A+N\right).
\end{aligned}
$$

Denoting $U = A^{(n-1)}$ yields $A = LU - N$. The matrix N is a sum of non-negative matrices such that it is non-negative. Similarly, the matrix L^{-1} is a product of non-negative matrices and hence it is non-negative, too. ■

Remark 8.23. ILU.

- For a given zero pattern P, the ILU decomposition is uniquely determined.
- Before or in the first iteration, one has to compute the incomplete decomposition.
- The application of ILU as preconditioner requires the solution of two sparse linear systems of equations with triangular matrices:
  1. solve the lower triangular system $L\underline{w} = \underline{r}$,
  2. solve the upper triangular system $U\underline{z} = \underline{w}$.
- The main costs of unpreconditioned iterative methods are the multiplications of the sparse matrix with a vector. If the zero pattern is appropriately given, then the costs for applying the ILU preconditioner are proportional to the costs of the matrix-vector multiplication. Very often, one takes the pattern of A, i.e., non-zero entries in L and U are allowed only for pairs of indices for which A has a non-zero entry.
- If A is symmetric and positive definite, then one obtains (if the non-zero pattern is chosen to be symmetric) an incomplete Cholesky decomposition. The precondition matrix $M = LL^T$ is also symmetric and positive definite and it can be applied in the PCG algorithm.

□