Chapter 5
Discrete Maximum Principles

A discretization of convection-diffusion-reaction equations should provide a numerical solution that is in some sense a good approximation of the analytic solution. In numerical analysis, the quality of approximation is usually measured in norms of function spaces. However, from the practical point of view, it is often of utmost importance that the numerical solution is physically consistent, i.e., that it possesses some basic physical properties that are valid in the same form for the solution of the continuous problem.

A very important property that should be reflected correctly by a numerical solution is the range of admissible physical values. From the mathematical point of view, this requirement can be formulated by the satisfaction of discrete maximum principles (DMPs). This chapter introduces DMPs for linear discretizations of the steady-state convection-diffusion-reaction problem (1.27).

5.1 Linear Discrete Problems

Since the convection-diffusion-reaction problem is a linear problem, it seems to be natural that a discretization of (1.27) leads also to a linear problem, i.e., to a linear system of equations. This section introduces two approaches for studying DMPs for linear systems.

In Section 5.1.1, an approach based on the concept of matrices of nonnegative type will be presented. This concept allows the formulation of sufficient conditions on the matrix of the linear system for the satisfaction of local and global DMPs.

The traditional approach for the study of global DMPs, which is based on the concept of monotone matrices, is the contents of Section 5.1.2. A special class of such matrices are M-matrices, which are a popular tool in the analysis of discretizations concerning the satisfaction of global DMPs. Section 5.1.3

contains a brief survey on this class of matrices, which, in particular, introduces all statements that were already used in the previous chapters.

The algebraic representation of a linear discretization of (1.27) is a linear system of equations. Consider a $m \times n$ matrix, $0 < m < n$, with entries a_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n$, given vectors $\underline{g} \in \mathbb{R}^m$ and $\underline{u}^{\mathrm{b}} \in \mathbb{R}^{n-m}$ with entries $g_1, \ldots, g_m$ and $u_{m+1}^{\mathrm{b}}, \ldots, u_n^{\mathrm{b}}$, respectively, then this system is of the form

$$\sum_{j=1}^{n} a_{ij}\, u_j = g_i\,, \quad i = 1, \ldots, m\,, \tag{5.1}$$

$$u_i = u_i^{\mathrm{b}}\,, \quad i = m+1, \ldots, n\,. \tag{5.2}$$

In matrix-vector notation, this system can be written in the form

$$A\underline{u} = \begin{pmatrix} A^{\mathrm{i}} & A^{\mathrm{b}} \\ 0 & I \end{pmatrix} \begin{pmatrix} \underline{u}^{\mathrm{i}} \\ \underline{u}^{\mathrm{b}} \end{pmatrix} = \begin{pmatrix} \underline{g} \\ \underline{u}^{\mathrm{b}} \end{pmatrix}, \tag{5.3}$$

with $A \in \mathbb{R}^{n\times n}$, $A^{\mathrm{i}} \in \mathbb{R}^{m\times m}$, $A^{\mathrm{b}} \in \mathbb{R}^{m\times(n-m)}$, I being the identity matrix of dimension $(n-m) \times (n-m)$, and $\underline{u}^{\mathrm{i}} \in \mathbb{R}^m$.

Lemma 5.1 (Non-singularity of A and A^{i}). *The matrix A is non-singular if and only if the matrix A^{i} is non-singular.*

Proof. If the matrix A is non-singular, then its inverse is given by

$$A^{-1} = \begin{pmatrix} \left(A^{\mathrm{i}}\right)^{-1} & -\left(A^{\mathrm{i}}\right)^{-1} A^{\mathrm{b}} \\ 0 & I \end{pmatrix}. \tag{5.4}$$

The statement of the lemma follows now directly from this representation. ■

Some notations will be introduced now. Let $\underline{v}, \underline{w} \in \mathbb{R}^n$. Then, one writes $\underline{v} \le \underline{w}$ (or $\underline{v} < \underline{w}$) if and only if $v_i \le w_i$ (or $v_i < w_i$) for all $i = 1, \ldots, n$. Analogously, the notation $A \ge 0$ (or $A > 0$) means for a matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ that $a_{ij} \ge 0$ (or $a_{ij} > 0$) for all $i, j = 1, \ldots, n$. A vector consisting only of zeros is denoted by $\underline{0}$ and a vector consisting only of ones by $\underline{1}$. If it becomes necessary for clarifying the presentation, the dimension of these vectors will be indicated by subscripts. The i-th Cartesian unit vector is denoted by $\underline{e}_i$. Let $\alpha \in \mathbb{R}$, then its positive part α^+ and negative part α^- are defined as follows

$$\alpha^+ := \max\{\alpha, 0\} \ge 0 \quad \text{and} \quad \alpha^- := \min\{\alpha, 0\} \le 0.$$

5.1.1 Local and Global DMPs Based on the Concept of Minkowski Matrices or Matrices of Non-Negative Type

This section presents general conditions for the satisfaction of local and global DMPs. The underlying theory is based on the concept of matrices of non-negative type. It turns out that the corresponding analysis utilizes quite elementary tools.

Definition 5.2 (Minkowski matrix, Ostrowski (1937), Matrix of non-negative type, Ciarlet & Raviart (1973)). A matrix $A = (a_{ij})_{j=1,\ldots,n}^{i=1,\ldots,m}$ is said to be a Minkowski matrix or a matrix of non-negative type if it satisfies the conditions

$$a_{ij} \le 0 \quad \forall\, i \ne j,\, i = 1,\ldots,m,\, j = 1,\ldots,n\,, \tag{5.5}$$

$$\sum_{j=1}^{n} a_{ij} \ge 0 \quad \forall\, i = 1,\ldots,m. \tag{5.6}$$

The notion of a matrix of non-negative type must not be confused with the notion of a non-negative matrix as it is studied, e.g., in (Varga, 2000, Chapter 2).

Theorem 5.3 (Local DMPs). *Consider a matrix* $A = (a_{ij})_{j=1,\ldots,n}^{i=1,\ldots,m}$ *with* $0 < m < n$*. Then, the local DMPs*

$$\sum_{j=1}^{n} a_{ij}\, u_j \le 0 \quad \Longrightarrow \quad u_i \le \max_{j\ne i,\, a_{ij}\ne 0} u_j^+\,, \tag{5.7}$$

$$\sum_{j=1}^{n} a_{ij}\, u_j \ge 0 \quad \Longrightarrow \quad u_i \ge \min_{j\ne i,\, a_{ij}\ne 0} u_j^- \tag{5.8}$$

hold for any $i \in \{1,\ldots,m\}$ *and any* $u_1,\ldots,u_n \in \mathbb{R}$ *if and only if* A *is a matrix of non-negative type with*

$$a_{ii} > 0 \quad \forall\, i = 1,\ldots,m\,. \tag{5.9}$$

Proof. Since the statements (5.7) and (5.8) are equivalent, because they imply each other by replacing u_j with $-u_j$, it suffices to consider (5.7).

i). (5.7) $\Longrightarrow$ (5.5), (5.6), (5.9). Assume that at least one of the conditions (5.5), (5.6), and (5.9) is not valid. Then, a counterexample to the validity of (5.7) will be constructed.

If (5.9) does not hold, i.e., if $a_{ii} \le 0$ for some $i \in \{1,\ldots,m\}$, set $u_i = 1$ and $u_j = 0$ for $j \ne i$. It follows that

$$\sum_{j=1}^{n} a_{ij}\, u_j = a_{ii}\, u_i \le 0 \quad \text{and} \quad u_i > 0 = \max_{j\ne i,\, a_{ij}\ne 0} u_j^+ \,,$$

so that (5.7) is not satisfied. Therefore, (5.9) will be assumed to be valid for the rest of the proof.

If (5.5) does not hold, i.e., if $a_{ik} > 0$ for some $i \in \{1,\dots,m\}$ and $k \in \{1,\dots,n\}$, $k \ne i$, then set

$$u_i = 1\,, \quad u_k = -\frac{a_{ii}}{a_{ik}}\,, \quad u_j = 0 \quad \forall\, j \in \{1,\dots,n\} \setminus \{i,k\}.$$

Then, it is $u_k < 0$ and hence $\max\{u_j^+\,;\, j \ne i,\, a_{ij} \ne 0\} = 0 < u_i$ whereas $\sum_{j=1}^n a_{ij}\, u_j = a_{ii}\, u_i + a_{ik}\, u_k = 0$, so that (5.7) does not hold.

If (5.6) is not valid, i.e., if $\sum_{j=1}^n a_{ij} < 0$ for some $i \in \{1,\dots,m\}$, then set

$$u_i = 1 - \frac{1}{a_{ii}} \sum_{j=1}^{n} a_{ij} > 1\,, \qquad u_j = 1 \quad \forall\, j \in \{1,\dots,n\} \setminus \{i\}.$$

Then, $\max\{u_j^+\,;\, j \ne i,\, a_{ij} \ne 0\} = 1 < u_i$, whereas $\sum_{j=1}^n a_{ij}\, u_j = \sum_{j=1}^n a_{ij} - \sum_{j=1}^n a_{ij} = 0$. One finds again that (5.7) does not hold.

In summary, this proves that the validity of (5.7) for any $i \in \{1,\dots,m\}$ and any $u_1,\dots,u_n \in \mathbb{R}$ implies (5.5), (5.6), and (5.9).

ii). (5.5), (5.6), (5.9) $\Longrightarrow$ (5.7). Assume that (5.5), (5.6), and (5.9) are satisfied. Consider any $i \in \{1,\dots,m\}$ and any $u_1,\dots,u_n \in \mathbb{R}$ such that $\sum_{j=1}^n a_{ij}\, u_j \le 0$. Setting

$$r := \max_{j\ne i,\, a_{ij}\ne 0} u_j^+ \,,$$

one has with the assumption of (5.7), (5.5), and (5.6),

$$\begin{aligned} a_{ii}\, u_i &\le \sum_{\substack{j=1\\ j\ne i}}^{n} (-a_{ij})\, u_j = \sum_{\substack{j=1\\ j\ne i}}^{n} (-a_{ij})\,(u_j - r) + \sum_{\substack{j=1\\ j\ne i}}^{n} (-a_{ij})\, r \\ &\le r \sum_{\substack{j=1\\ j\ne i}}^{n} (-a_{ij}) \le r\, a_{ii}\,, \end{aligned} \tag{5.10}$$

which implies, because of (5.9), that $u_i \le r$. ■

If a stronger assumption than (5.6) is satisfied, namely that all row sums vanish, a stronger form of the local DMPs can be proved.

Theorem 5.4 (Local DMPs). *Consider a matrix* $A = (a_{ij})_{j=1,\dots,n}^{i=1,\dots,m}$ *of non-negative type with* $0 < m < n$. *Then the local DMPs*

$$\sum_{j=1}^{n} a_{ij}\, u_j \le 0 \quad \Longrightarrow \quad u_i \le \max_{j\neq i,\, a_{ij}\neq 0} u_j\,, \tag{5.11}$$

$$\sum_{j=1}^{n} a_{ij}\, u_j \ge 0 \quad \Longrightarrow \quad u_i \ge \min_{j\neq i,\, a_{ij}\neq 0} u_j \tag{5.12}$$

hold for any $i \in \{1,\dots,m\}$ *and any* $u_1,\dots,u_n \in \mathbb{R}$ *if and only if the conditions* (5.9), (5.5), *and*

$$\sum_{j=1}^{n} a_{ij} = 0 \quad \forall\, i = 1,\dots,m \tag{5.13}$$

are satisfied.

Proof. *exercise problem* ■

Theorem 5.5 (Global DMPs). *Consider a matrix* $A = (a_{ij})_{j=1,\dots,n}^{i=1,\dots,m}$ *with* $0 < m < n$ *satisfying conditions* (5.5) *and* (5.6) *and let the matrix* $A^{\mathrm{i}} = (a_{ij})_{i,j=1}^{m}$ *be non-singular. Then, for any* $u_1,\dots,u_n \in \mathbb{R}$, *there hold the global DMPs*

$$\sum_{j=1}^{n} a_{ij}\, u_j \le 0,\ i = 1,\dots,m \quad \Longrightarrow \quad \max_{i=1,\dots,n} u_i \le \max_{i=m+1,\dots,n} u_i^+\,, \tag{5.14}$$

$$\sum_{j=1}^{n} a_{ij}\, u_j \ge 0,\ i = 1,\dots,m \quad \Longrightarrow \quad \min_{i=1,\dots,n} u_i \ge \min_{i=m+1,\dots,n} u_i^-\,. \tag{5.15}$$

If, in addition, condition (5.13) *is satisfied, then*

$$\sum_{j=1}^{n} a_{ij}\, u_j \le 0,\ i = 1,\dots,m \quad \Longrightarrow \quad \max_{i=1,\dots,n} u_i = \max_{i=m+1,\dots,n} u_i\,, \tag{5.16}$$

$$\sum_{j=1}^{n} a_{ij}\, u_j \ge 0,\ i = 1,\dots,m \quad \Longrightarrow \quad \min_{i=1,\dots,n} u_i = \min_{i=m+1,\dots,n} u_i\,. \tag{5.17}$$

Proof. For interested students only, not presented in the class.

Again, it suffices to prove (5.14) and (5.16), so that one can assume that

$$\sum_{j=1}^{n} a_{ij}\, u_j \le 0\,, \quad i = 1,\dots,m\,. \tag{5.18}$$

Consider (5.14) and let

$$s = \max_{i=1,\dots,n} u_i \quad \text{and} \quad J = \{i \in \{1,\dots,n\} \;:\; u_i = s\}\,. \tag{5.19}$$

It can be assumed that $s > 0$ since otherwise (5.14) trivially holds. Thus, let $s > 0$ and assume that $J \subset \{1, \ldots, m\}$. As first step, it is proved that

$$\exists k \in J \text{ such that } \mu_k := \sum_{j\in J} a_{kj} > 0\,. \tag{5.20}$$

The proof is performed by contradiction, showing that in case (5.20) is not valid, the matrix A^{i} is singular. Thus, assume that (5.20) does not hold. Combining (5.5) and (5.6) yields

$$\sum_{j\in J} a_{ij} = 0 \quad \forall\, i \in J\,,$$

such that the matrix $(a_{ij})_{i,j\in J}$ is singular because the sum of its columns is zero. Consequently, also its transposed $(a_{ji})_{i,j\in J}$ is singular. Hence, there exist $v_i, i \in J$, not all zero, such that

$$\sum_{i\in J} a_{ij} v_i = 0 \quad \forall\, j \in J\,. \tag{5.21}$$

Using that A is a matrix of non-negative type, one concludes that $a_{ij} = 0$ for all $i \in J$ and all $j \notin J$. With this property, (5.21), and the vector $\tilde{\boldsymbol{v}} = (\tilde{v}_i)_{i=1}^M$, where $\tilde{v}_i = v_i$ if $i \in J$, and $\tilde{v}_i = 0$ otherwise, one finds that

$$\sum_{i=1}^{M} a_{ij}\tilde{v}_i = \sum_{i\in J} a_{ij} v_i = 0 \quad \forall\, j \in \{1, \ldots, m\}.$$

This result implies that A^{i} is singular, which contradicts the hypothesis. Consequently, (5.20) holds.

Defining

$$r := \max_{i\notin J} u_i\,,$$

and utilizing (5.14), (5.5), and (5.6) leads to

$$\begin{aligned} s\mu_k &= \sum_{j\in J} a_{kj}u_j = f_k - \sum_{j\notin J} a_{kj}u_j \le -\sum_{j\notin J} a_{kj}u_j = \sum_{j\notin J}(-a_{kj})u_j \\ &\le r\sum_{j\notin J}(-a_{kj}) = r\left(\sum_{j=1}^{N}(-a_{kj}) + \sum_{j\in J} a_{kj}\right) \le r\mu_k\,, \end{aligned}$$

which implies that $s \le r$. Since this is a contradiction to the definition of s, it it can be inferred that $J \cap \{m+1, \ldots, n\} \ne \emptyset$, such that (5.14) follows.

The proof of (5.16) is performed analogously. If (5.13) is satisfied, one can also assume that $s > 0$, since (5.18) still holds if one adds a constant to all components of the vector $(u_1, \ldots, u_n)^T$. With the same argument, one can define r by the values of the function itself instead of their positive part. ■

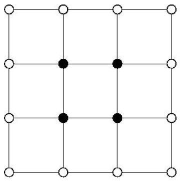

Fig. 5.1 Illustration to Remark 5.7.

Remark 5.6 (On Theorem 5.5).

- Note that if a matrix $(a_{ij})_{j=1,\dots,n}^{i=1,\dots,m}$ with $0 < m < n$ satisfies conditions (5.5) and (5.6) and, at the same time, $a_{kk} \leq 0$ for some $k \in \{1, \dots, m\}$, then the kth row of this matrix vanishes and hence the matrix $(a_{ij})_{i,j=1}^{m}$ is singular. Therefore, the assumptions of Theorem 5.5 imply the validity of (5.9).
- It will be stated below, Remark 5.21, that the matrix A^{i} from Theorem 5.5 is an M-matrix. Since M-matrices are a special class of monotone matrices, see Section 5.1.3, it turns out that the statement of Theorem 5.5, namely (5.14) and (5.15), is a special case of the statement given below in Theorem 5.13.

□

Remark 5.7 (Local DMPs do not imply global DMPs). Global DMPs do not follow from local DMPs, e.g., the validity of the right-hand sides of (5.11) does not imply the validity of the right-hand side of (5.16). This fact can be seen from the example depicted in Figure 5.1 where a mesh consisting of 4×4 vertices is shown. Assume that the values at interior vertices $\boldsymbol{x}_1, \dots, \boldsymbol{x}_4$ (denoted by black circles) are equal to 1 whereas the values at boundary vertices $\boldsymbol{x}_5, \dots, \boldsymbol{x}_{16}$ (denoted by white circles) are equal to 0. Typically, for each interior vertex $\boldsymbol{x}_i$, there is another interior vertex $\boldsymbol{x}_j$ such that $a_{ij} \neq 0$. Then, the right-hand sides of (5.7) and (5.11) hold for all interior vertices whereas the right-hand sides of (5.14) and (5.16) are not satisfied. □

5.1.2 Global DMPs Based on the Concept of Monotone Matrices

The concept of monotone matrices is a popular tool for investigating global DMPs. This section introduces the class of monotone matrices and states sufficient and necessary criteria for the satisfaction of global DMPs. An important subset of monotone matrices are M-matrices, which will be discussed in some detail in Section 5.1.3.

Definition 5.8 (Monotone matrix). A square matrix A is called monotone or inverse-monotone or inverse-positive if A is non-singular and $A^{-1} \geq 0$. □

The notion 'monotone matrix', which will be used in this course, comes from the following equivalent characterization.

Lemma 5.9 (Equivalent characterization of a monotone matrix). *A matrix $A \in \mathbb{R}^{n\times n}$ is monotone if and only if*

$$A\underline{v} \geq 0 \quad \Longrightarrow \quad \underline{v} \geq 0 \quad \forall\, \underline{v} \in \mathbb{R}^n. \tag{5.22}$$

Proof. *exercise problem.* ■

Lemma 5.10 (Discrete comparison principle). *Let $A \in \mathbb{R}^{n\times n}$ be a monotone matrix. If $A\underline{v} \leq A\underline{w}$ for $\underline{v}, \underline{w} \in \mathbb{R}^n$, then it follows that $\underline{v} \leq \underline{w}$.*

Proof. *exercise problem.* ■

Lemma 5.11 (Product of two monotone matrices). *The product of two monotone matrices is a monotone matrix.*

Proof. *exercise problem.* ■

Theorem 5.12 (Sufficient and necessary conditions for the satisfaction of a global DMP, Ciarlet (1970)). *Let $A \in \mathbb{R}^{n\times n}$ be given with the block structure* (5.3) *and let $\underline{u} \in \mathbb{R}^n$ with $\underline{u} = (\underline{u}^{\mathrm{i}}, \underline{u}^{\mathrm{b}})^T$, $\underline{u}^{\mathrm{i}} \in \mathbb{R}^m$, $\underline{u}^{\mathrm{b}} \in \mathbb{R}^{n-m}$. Then, the global DMP*

$$A^{\mathrm{i}}\underline{u}^{\mathrm{i}} + A^{\mathrm{b}}\underline{u}^{\mathrm{b}} \leq 0 \quad \Longrightarrow \quad \max_{i=1,\ldots,m} u_i \leq \max_{i=m+1,\ldots,n} u_i^+ \tag{5.23}$$

holds if and only if the following two conditions are satisfied:

1) A is monotone,

2) $-\left(A^{\mathrm{i}}\right)^{-1} A^{\mathrm{b}}\underline{1}_{n-m} \leq \underline{1}_m$, i.e., the row sums of $-\left(A^{\mathrm{i}}\right)^{-1} A^{\mathrm{b}}$, which is the right upper block of A^{-1}, see (5.4)*, are smaller than* 1.

Proof. The proof follows Ciarlet (1970).

i) Assume that A satisfies the global DMP, show conditions 1) and 2). These conditions will be derived by considering problem (5.3) with special right-hand sides.

Condition 1) will be shown. The first step consists in proving that A is non-singular. From Lemma 5.1 it is known that this statement is equivalent to the statement that A^{i} is non-singular. Let $\underline{u}^{\mathrm{i}} \in \mathbb{R}^m$ such that $A^{\mathrm{i}}\underline{u}^{\mathrm{i}} = \underline{0}$. Consider the vector $\underline{u} = \left(\underline{u}^{\mathrm{i}}, \underline{0}\right)^T \in \mathbb{R}^n$. Using the decomposition of A, one finds that $A\underline{u} = \underline{0} \in \mathbb{R}^n$. Since the global DMP is assumed to hold, it follows that $\underline{u}^{\mathrm{i}} \leq 0$. The same reasoning applied to $-\underline{u}^{\mathrm{i}}$ leads to the the conclusion that $-\underline{u}^{\mathrm{i}} \le 0$. In summary, it is $\underline{u}^{\mathrm{i}} = \underline{0}$, which proves the non-singularity of A^{i}.

In the next step, $A^{-1} \ge 0$ will be proved, which will be done for both upper blocks in (5.4) individually.

Consider problem (5.3) with the right-hand side $\underline{g}_j = (0,\ldots,0,-1,0,\ldots,0)^T$, $j = 1,\ldots,m$, where the non-zero entry is in the j-th component, and $\underline{u}^{\mathrm{b}} = \underline{0}$. Using the representation (5.4) of the inverse matrix, one finds that the unique solution of this problem is $-\left(a_{1j}^{\mathrm{inv}},\ldots,a_{mj}^{\mathrm{inv}},0,\ldots,0\right)^T$. By the global DMP, it follows that $-a_{ij}^{\mathrm{inv}} \le 0$, or $a_{ij}^{\mathrm{inv}} \ge 0$, for all $i,j = 1,\ldots,m$, which proves that $\left(A^{\mathrm{i}}\right)^{-1} \ge 0$.

Consider next (5.3) with $\underline{g} = \underline{0}$ and $\underline{u}^{\mathrm{b}} = (0,\ldots,0,-1,0,\ldots,0)^T$, $j = 1,\ldots,n-m$. Similarly as in the previous step, one calculates that the unique solution of this problem is $-\left(0,\ldots,0,a_{m+1,j}^{\mathrm{inv}},\ldots,a_{nj}^{\mathrm{inv}}\right)^T$. With the global DMP, it can be concluded that $-\left(A^{\mathrm{i}}\right)^{-1} A^{\mathrm{b}} \ge 0$. This step finishes the proof of condition 1).

The final part of this direction of the proof consists in showing condition 2). To this end, consider (5.3) with $\underline{g} = \underline{0}$ and $\underline{u}^{\mathrm{b}} = \underline{1}$. Applying the inverse matrix (5.4) yields for the unique solution that $\underline{u} = \left(-\left(A^{\mathrm{i}}\right)^{-1} A^{\mathrm{b}}\underline{1}_{n-m}, \underline{1}_{n-m}\right)^T$. From the satisfaction of the DMP, it follows that $\left(-\left(A^{\mathrm{i}}\right)^{-1} A^{\mathrm{b}}\underline{1}_{n-m}\right)_i \le 1$ for all $i = 1,\ldots,m$, which proves condition 2).

ii). Assume conditions 1) and 2) to hold, show the satisfaction of the global DMP. From condition 1), it follows that A is non-singular. Hence, the identity

$$\underline{u}^{\mathrm{i}} = \left(A^{\mathrm{i}}\right)^{-1}\left(A^{\mathrm{i}}\underline{u}^{\mathrm{i}} + A^{\mathrm{b}}\underline{u}^{\mathrm{b}}\right) - \left(A^{\mathrm{i}}\right)^{-1} A^{\mathrm{b}}\underline{u}^{\mathrm{b}}$$

is valid for any $\underline{u} = \left(\underline{u}^{\mathrm{i}}, \underline{u}^{\mathrm{b}}\right)^T \in \mathbb{R}^n$. Writing this identity component by component and using the form of the right upper block of the inverse of A given in (5.4) yields

$$u_i = \sum_{j=1}^{m} a_{ij}^{\mathrm{inv}}\left(A^{\mathrm{i}}\underline{u}^{\mathrm{i}} + A^{\mathrm{b}}\underline{u}^{\mathrm{b}}\right)_j + \sum_{j=1}^{n-m} a_{i,j+m}^{\mathrm{inv}}\left(\underline{u}^{\mathrm{b}}\right)_j, \quad i = 1,\ldots,m. \tag{5.24}$$

Consider a vector that satisfies the assumption of the DMP, i.e, that satisfies the left-hand side of (5.23). Then, the corresponding factor in the first sum is non-positive. By condition 1), all coefficients a_{ij}^{inv} are non-negative. Consequently, the first sum is non-positive.

If $\max_{j=1,\ldots,n-m}\left(\underline{u}^{\mathrm{b}}\right)_j \le 0$, then also the second sum in (5.24) is non-positive and $u_i \le 0$, consequently the right-hand side of (5.23) is satisfied.

In the case $\max_{j=1,\ldots,n-m}\left(\underline{u}^{\mathrm{b}}\right)_j > 0$, condition 2) has to be used, which states $\sum_{j=1}^{n-m} a_{i,j+m}^{\mathrm{inv}} \le 1$. Hence, one obtains from (5.24)

$$u_i \leq \sum_{j=1}^{n-m} a_{i,j+m}^{\text{inv}} \left(\underline{u}^{\text{b}}\right)_j \leq \max_{j=1,\ldots,n-m} \left(\underline{u}^{\text{b}}\right)_j \sum_{j=1}^{n-m} a_{i,j+m}^{\text{inv}} \leq \max_{j=1,\ldots,n-m} \left(\underline{u}^{\text{b}}\right)_j .$$

Also in this case, the right-hand side of (5.23) is satisfied. ■

Both conditions of Theorem 5.12 are based on the inverse of $\left(A^{\text{i}}\right)$, which is usually not available in practice. Consequently, these conditions cannot be checked. From the practical point of view, sufficient conditions are needed that can be used more easily to decide whether a discretization satisfies the DMP.

Theorem 5.13 (Sufficient condition for the satisfaction of the global DMP). *Let condition 1) of Theorem 5.12 be satisfied and let*

$$\sum_{j=1}^{n} a_{ij} \geq 0, \quad 1 \leq i \leq m, \tag{5.25}$$

then the matrix A satisfies the global DMP (5.23)*.*

Proof. Consider the vector $\underline{u} = \left(\underline{u}^{\text{i}}, \underline{u}^{\text{b}}\right) = \underline{1}$. From (5.25), one gets

$$A^{\text{i}}\underline{u}^{\text{i}} + A^{\text{b}}\underline{u}^{\text{b}} = \sum_{j=1}^{n} a_{ij} \geq 0. \tag{5.26}$$

By condition 1) from Theorem 5.12, it is known that A^{i} is invertible and the entries of its inverse are non-negative. Applying the inverse of this matrix from the left to (5.26) will not change the relation and one obtains

$$\underline{1}_m + \left(A^{\text{i}}\right)^{-1} A^{\text{b}}\underline{1}_{n-m} \geq 0,$$

which is exactly condition 2) of Theorem 5.12. Hence, both conditions of Theorem 5.12 are fulfilled, such that A satisfies the global DMP. ■

The conditions of Theorem 5.13 are not necessary for a matrix to fulfil the global DMPs.

Example 5.14 (An operator fulfilling the global DMPs that does not satisfy (5.25)*, Ciarlet (1970)).* Consider the matrix

$$A = \begin{pmatrix} -1 & 2 & 0 \\ 2 & -3 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

where $m = 2, n = 3$. Obviously, (5.25) is not satisfied. A direct calculation gives

$$A^{-1} = \begin{pmatrix} 3 & 2 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

such that A is monotone. In addition, it is $-\left(A^{\mathrm{i}}\right)^{-1} A^{\mathrm{b}} = \underline{0} < 1$, hence condition 2) from Theorem 5.12 is fulfilled. Theorem 5.12 gives now that A satisfies the global DMP. □

5.1.3 M-Matrices

The most important subset of monotone matrices is the class of M-matrices. This class is widely used for studying global DMPs. M-matrices are in a somewhat hidden sense diagonally dominant, compare a problem from the exercises. In this course, only M-matrices will be considered that are non-singular and these matrices will be called just M-matrices.

There are at least forty equivalent definitions of an M-matrix, see the survey Plemmons (1977). In connection with DMPs, most often the following definition is used.

Definition 5.15 (M-matrix). A matrix $A = (a_{ij})_{i,j=1}^{n}$ is an M-matrix if:

i) The off-diagonal entries are non-positive

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

ii) A is non-singular.

iii) It holds $A^{-1} \ge 0$.

□

With Definition 5.15 it becomes clear that the set of M-matrices is a subset of the set of monotone matrices since condition i) is an additional requirement to Definition 5.8. Example 5.14 shows that the set of M-matrices is a proper subset.

Corollary 5.16 (M-matrices and the global DMP). *A discretization leading to an M-matrix that has the additional property* (5.25) *gives a discrete solution that satisfies the global DMP.*

Proof. The statement follows from Theorem 5.13 and the fact that the set of M-matrices forms a subset of the set of monotone matrices. ■

Corollary 5.17 (Restriction to the inner nodes). *Let A be a non-singular matrix of form* (5.3) *that possesses that correct sign pattern for an M-matrix, i.e., Definition 5.15 i) is satisfied. Then, A is an M-matrix if and only if A^{i} is an M-matrix.*

Proof. By Lemma 5.1, it follows that A is non-singular if and only if A^{i} is non-singular. Noting that $A^{\mathrm{b}} \leq 0$ by assumption, the representation (5.4) of the inverse of A shows also that $A^{-1} \geq 0$ if and only if $\left(A^{\mathrm{i}}\right)^{-1} \geq 0$. Hence, both matrices A and A^{i} satisfy the requirements of Definition 5.15. ■

Given the sign property of the off-diagonal entries of an M-matrix, Corollary 5.17 states in particular that it is sufficient to prove that the restriction of the matrix to the inner nodes is an M-matrix in order to show that the complete matrix is an M-matrix.

The following theorem provides an explicit connection between general monotone matrices and M-matrices.

Theorem 5.18 (Connection of general monotone and M-matrices). *The matrix $A \in \mathbb{R}^{n\times n}$ is monotone if and only if there exist matrices $B_1 \geq 0$ and $B_2 \geq 0$, $B_1, B_2 \in \mathbb{R}^{n\times n}$, such that B_1AB_2 is an M-matrix.*

Proof. The proof follows Bramble & Hubbard (1964). Let A be a monotone matrix, then A is invertible. Now, one can choose $B_1 = I \geq 0$ and $B_2 = A^{-1} \geq 0$, such that $B_1AB_2 = I$, which is an M-matrix.

Let B_1AB_2 be an M-matrix, then this matrix is invertible with $(B_1AB_2)^{-1} \geq 0$ and in particular all factors are invertible. It follows that

$$A^{-1} = B_2\left(B_1AB_2\right)^{-1}B_1 \geq 0,$$

such that A is a monotone matrix, compare Definition 5.8. ■

Definition 5.19 (Proper Minkowski matrix, Ostrowski (1937)). A matrix of non-negative type defined in Definition 5.2 is is called a proper Minkowski matrix if all row sums are positive, i.e., the matrix is diagonally dominant. □

Theorem 5.20 (Connection of M-matrices and proper Minkowski matrices). *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.*

Proof. Exercise problem. ■

Here, just a summary of important algebraic properties of M-matrices is given.

Remark 5.21 (Properties of M-Matrices). The class of M-matrices is a subset of monotone matrices. A matrix is a monotone matrix if and only if it can be represented as a product of an M-matrix and two non-negative matrices, see Theorem 5.18.

Let $A \in \mathbb{R}^{n\times n}$ with $a_{ii} > 0$ and $a_{ij} \leq 0$, $i, j = 1, \ldots n$, $i \neq j$.

- The matrix A is an M-matrix if and only if all principal minors of A are positive. For example, let

$$A = \begin{pmatrix} 4 & -8 \\ -2 & 5 \end{pmatrix}, \tag{5.27}$$

  then the principal minors of first order are just the diagonal entries, which are both positive. The principal minor of second order is $\det(A) = 4$, which is also positive.
- The matrix A is an M-matrix if and only if A is non-singular and $A^{-1} \geq 0$ (Definition 5.15).
- If A is of block form (5.3), then A is an M-matrix if and only if A^{i} is an M-matrix (Corollary 5.17).
- If A is an M-matrix, then there is a proper Minkowski matrix $\tilde{A}$ and a diagonal matrix D with $d_{ii} > 0$, $i = 1, \ldots, n$, such that $A = \tilde{A}D$ (Theorem 5.20).
- If A is irreducible and weakly diagonally dominant, i.e., all row sums are non-negative and at least one row sum is positive, then A is an M-matrix.
- If A is a non-singular matrix of non-negative type, then A is an M-matrix. Not every M-matrix is a matrix of non-negative type, e.g., see the matrix given in (5.27).
- The matrix A is an M-matrix, if and only if there is a vector $\underline{v} \in \mathbb{R}^n$, $\underline{v} > 0$, with $A\underline{v} > 0$. The vector $\underline{v}$ is called majorizing element.

Let $A \in \mathbb{R}^{n\times n}$ be an M-matrix. Then, the following properties hold.

- It is $a_{ii}^{\mathrm{inv}} > 0$, $i = 1, \ldots, n$, exercise problem.
- Multiplying arbitrary rows or columns of A with positive constants gives an M-matrix, exercise problem.
- Let $B \in \mathbb{R}^{n\times n}$ with $a_{ij} \leq b_{ij}$ for $i, j = 1, \ldots, n$ and $b_{ij} \leq 0$ for $i, j = 1, \ldots, n, i \neq j$, then B is an M-matrix, comparison criterion.
- Let $D \geq 0$ be a diagonal matrix, then $A + D$ is an M-matrix.
- The sum of two M-matrices is generally not a monotone matrix, exercise problem.
- The product of two M-matrices is a monotone matrix, exercise problem.
- The row sum norm of A^{-1} can be estimated by the maximum norm of a majorizing element of A (Lemma 5.22).

□

The following estimate is sometimes used for obtaining a bound for the stability of a discretization, compare Lemma 3.9 for finite difference methods.

Lemma 5.22 (Bound of a norm of the inverse of an M-matrix by a norm of a majorizing element). *Let $A \in \mathbb{R}^{n\times n}$ be an M-matrix and let $\underline{v} \in \mathbb{R}^n$ be a majorizing element. Then, it holds that*

$$\left\|A^{-1}\right\|_\infty \leq \frac{\|\underline{v}\|_\infty}{\min_{j=1,\ldots,n} \left(A\underline{v}\right)_j}. \tag{5.28}$$

Proof. From $\underline{v} = A^{-1}A\underline{v}$, one gets, because $\underline{v} > 0$,

$$0 < v_i = a_{i1}^{\text{inv}}(A\underline{v})_1 + \cdots + a_{in}^{\text{inv}}(A\underline{v})_n.$$

Since all terms are non-negative, one obtains for all $i = 1, \ldots, n$,

$$v_i \ge \left(a_{i1}^{\text{inv}} + \ldots + a_{in}^{\text{inv}}\right) \min_{j=1,\ldots,n} (A\underline{v})_j\,,$$

such that

$$\begin{aligned} \|\underline{v}\|_\infty = \max_{i=1,\ldots,n} v_i &\ge \max_{i=1,\ldots,n} \left(a_{i1}^{\text{inv}} + \ldots + a_{in}^{\text{inv}}\right) \min_{j=1,\ldots,n} (A\underline{v})_j \\ &= \left\|A^{-1}\right\|_\infty \min_{j=1,\ldots,n} (A\underline{v})_j\,. \end{aligned}$$

This inequality is just the statement of the lemma because $A\underline{v} > 0$. ■

Remark 5.23 (Constructing a majorizing element). Let A be a an M-matrix that represents a discretization of a linear differential operator L. The following approach is often successful for the construction of a majorizing element.

- Find a function $v(\boldsymbol{x}) > 0$ such that $(Lv)(\boldsymbol{x}) > 0$ for $\boldsymbol{x} \in \Omega$. This function is a majorizing element of L.
- Interpolate $v(\boldsymbol{x})$ with a corresponding discrete function $v_h(\boldsymbol{x})$, which is represented by a vector $\underline{v}$. For finite difference methods, one takes usually the values of $v(\boldsymbol{x})$ in the nodes. In finite element methods, $\underline{v}$ depends on the chosen basis. Using for continuous Lagrangian finite elements a local basis, then the Lagrangian interpolation operator can be used, which also takes the values at the positions of the degrees of freedom.

If the first step of this approach is possible and if the discretization of L is consistent, then this approach generally works, at least if the mesh width is sufficiently small. □