Numerical Linear Algebra with Applications, Vol. 1(1), 1–7 (1993)

On the Convergence of Basic Iterative Methods for Convection-Diffusion Equations

Jürgen Bey

Lehrstuhl für Numerische Mathematik, RWTH Aachen, Templergraben 55, D-52056 Aachen, Germany

and

Arnold Reusken

Lehrstuhl für Numerische Mathematik, RWTH Aachen, Templergraben 55, D-52056 Aachen, Germany

In this paper we analyze convergence of basic iterative Jacobi and Gauss-Seidel type of methods for solving linear systems which result from finite element or finite volume discretization of convection-diffusion equations on unstructured meshes. In general the resulting stiffness matrices are neither M-matrices nor satisfy a diagonal dominance criterion. We introduce two new matrix classes and analyze the convergence of the Jacobi and Gauss-Seidel methods for matrices from these classes. A new convergence result for the Jacobi method is proved and negative results for the Gauss-Seidel method are obtained. For a few well-known discretization methods it is shown that the resulting stiffness matrices fall into the new matrix classes.

KEY WORDS convection-diffusion equation, basic iterative methods, convergence analysis.

1. Introduction

In this paper we study the convergence of basic iterative Jacobi and Gauss-Seidel type of methods for solving large sparse linear systems. This is a classical topic which is treated already in detail in [22,23]. More recent references concerning this subject are [1,13,18]. In these references one can find convergence analyses of Jacobi and Gauss-Seidel type of methods applied to matrices from certain standard classes. The main classes for which convergence results are known are: symmetric positive definite (spd) matrices, M-matrices, matrices with a diagonal dominance property, and positive definite matrices (i.e. matrices A with $A + A^T$ spd). A rather complete overview of the main known convergence results is given in [13].

1070–5325/93/010001–07$8.50

Received
Revised

In this paper we focus on linear systems resulting from discretization of a scalar convection-diffusion problem. If for the discretization one applies the usual *finite difference* techniques, then in many cases the resulting matrix is an M-matrix or satisfies a diagonal dominance criterion. In these cases known convergence analyses apply. However, if higher order finite difference methods are applied, then in general the resulting matrix is not an M-matrix and convergence results for Jacobi or Gauss-Seidel type of methods are known only in special cases (see [20], for example). If one uses *finite element* (FE) or *finite volume* (FV) techniques on *unstructured meshes*, then in general the resulting stiffness matrix does also not fall into one of the above-mentioned standard matrix classes. If the stiffness matrix resulting from FE or FV discretization on an unstructured mesh lies in one of the standard classes, then often rather special assumptions are used, for example the assumption that the underlying triangulation is of weakly acute type (which in practice is usually not the case).

If one compares the algebraic properties (with respect to e.g. diagonal dominance, symmetry, sign properties) of the stiffness matrices resulting from standard FE or FV discretization methods for convection-diffusion problems with the assumptions that are used in the known convergence analyses for basic iterative methods, it follows that often these do not match. Hence there are still many open problems related to the convergence of Jacobi and Gauss-Seidel type of iterative methods applied to these stiffness matrices. As a concrete example, consider the upwind triangle finite element method of Tabata (explained in Section 4.3.) applied to a model convection-diffusion problem. On an unstructured mesh (not necessarily of weakly acute type) the stiffness matrix resulting from this method is in general neither positive definite nor weakly diagonally dominant nor an M-matrix.

Motivated by algebraic properties of stiffness matrices resulting from standard FE or FV discretization methods we will introduce two nonstandard matrix classes. The first, seemingly natural, class consists of all matrices which can be represented as a sum of an spd matrix and an M-matrix:

$$SPD.M := \{\, A \in \mathbb{R}^{n\times n} \mid A = A_d + A_c \text{ with } A_d \text{ spd and } A_c \text{ an M-matrix} \,\}.$$

Under reasonable assumptions it follows that the stiffness matrix resulting from the Tabata FE method is an element of this matrix class. As far as we know, the convergence of Jacobi and Gauss-Seidel type of methods has not yet been analyzed for the matrix class $SPD.M$. Hence the question arises whether one can prove convergence of Jacobi and Gauss-Seidel type of methods for all matrices in the class $SPD.M$. In this paper this question is answered. As a second nonstandard matrix class we consider

$$SPD.M_0 := \{\, A \in \mathbb{R}^{n\times n} \mid A = A_d + A_c \text{ with } A_d \text{ spd and } A_c \in PD \cap Z \,\},$$

where $Z := \{\, A \in \mathbb{R}^{n\times n} \,|\, a_{ij} \leq 0 \text{ for all } i \neq j \,\}$ and where PD denotes the class of positive definite matrices. We will study convergence of Jacobi and Gauss-Seidel type of methods for matrices from this class.

In Section 2. we discuss relations between the different matrix classes that are considered in this paper (M-matrices, $SPD.M$, $SPD.M_0$, PD).

In Section 3. we consider the matrix classes $SPD.M$, $SPD.M_0$ and PD and derive convergence properties of Jacobi and Gauss-Seidel type of methods when applied to matrices from these classes. It will be shown that both for the Jacobi and Gauss-Seidel method there are matrices in $SPD.M$ for which the method (even with optimal damping) is not convergent. Hence the favourable properties w.r.t. convergence of the damped Jacobi and Gauss-Seidel methods which hold in the class of spd matrices and in the class of M-matrices are lost in the class $SPD.M$. For matrices from the class $SPD.M_0$ we will prove

a new contraction result for the damped Jacobi method and compare this result with a result from the literature concerning convergence of the damped Jacobi method for positive definite matrices. For positive definite matrices we introduce and analyze a hybrid method (which has features both from the Jacobi and Gauss-Seidel method) in which downwind numbering techniques on unstructured meshes (cf. [7,14]) play a role.

In Section 4. we consider a few known FE and FV discretization methods for convection-diffusion problems. We analyze the resulting stiffness matrix w.r.t. algebraic properties and show in which of the classes considered in Section 2. and 3. this matrix lies.

In our opinion there are still many open problems in this field of convergence of basic iterative methods applied to discretized convection-diffusion equations. We briefly comment on this in Section 5..

2. Classes of matrices

We introduce the following notation for a few well-known classes of matrices:

$$Z := \{\, A \in \mathbb{R}^{n\times n} \mid a_{ij} \le 0 \text{ for all } i \neq j \,\}, \quad \textit{(Z-matrix)} \tag{2.1}$$
$$SPD := \{\, A \in \mathbb{R}^{n\times n} \mid A = A^T > 0 \,\}, \quad \textit{(symmetric positive definite)} \tag{2.2}$$
$$PD := \{\, A \in \mathbb{R}^{n\times n} \mid A + A^T > 0 \,\}, \quad \textit{(positive definite)} \tag{2.3}$$
$$M := \{\, A \in \mathbb{R}^{n\times n} \mid A \in Z \text{ and } \mathrm{Re}(\lambda) > 0 \text{ for all } \lambda \in \sigma(A) \,\}. \quad \textit{(M-matrix)} \tag{2.4}$$

In (2.4) one of the many characterizations of M-matrices is used (cf. [5]). We will also use the matrix class

$$M_0 := PD \cap Z\,. \tag{2.5}$$

Note that

$$M_0 \subset M \tag{2.6}$$

holds. The discretization of convection-diffusion problems often results in matrices of the form

$$A = A_d + A_c\,, \tag{2.7}$$

where $A_d \in SPD$ represents the discrete diffusion term and A_c results from a 'stable' discretization of the convection term (cf. Section 4.). Based on this we introduce two more classes of matrices:

$$SPD.M := \{\, A \in \mathbb{R}^{n\times n} \mid A = A_d + A_c \text{ with } A_d \in SPD,\, A_c \in M \,\}, \tag{2.8}$$
$$SPD.M_0 := \{\, A \in \mathbb{R}^{n\times n} \mid A = A_d + A_c \text{ with } A_d \in SPD,\, A_c \in M_0 \,\}. \tag{2.9}$$

Remark 2.1. For the matrix classes $SPD.M$ and $SPD.M_0$ the *strict* inequalities in (2.3) and (2.4) are not essential. The matrix classes in (2.8), (2.9) can also be characterized by

$$SPD.M = \{\, A = A_d + A_c \in \mathbb{R}^{n\times n} \mid A_d \in SPD,\, A_c \in Z \text{ and } \mathrm{Re}(\lambda) \ge 0,\, \lambda \in \sigma(A_c)\},$$
$$SPD.M_0 = \{\, A = A_d + A_c \in \mathbb{R}^{n\times n} \mid A_d \in SPD,\, A_c \in Z \text{ and } A_c + A_c^T \ge 0\,\}.$$

Lemma 2.1. *The following relations hold:*

$$\begin{array}{ccccc} M_0 & \subset & M & \subset & Z \\ \cap & & \cap & & \\ SPD.M_0 & \subset & SPD.M & & \\ \cap & & & & \\ PD & & & & \end{array} \tag{2.10}$$

In this diagram all inclusions are strict.

Proof The results $M_0 \subset M \subset Z$ follow from (2.6) and from the definition of M in (2.4). $M_0 \neq M$ follows from the example

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

For $A \in M_0$ the splitting $A = \varepsilon I + (A - \varepsilon I) =: A_d + A_c$, with $\varepsilon > 0$ sufficiently small, shows that $M_0 \subset SPD.M_0$ holds. The same argument yields $M \subset SPD.M$. $M_0 \neq SPD.M_0$ and $M \neq SPD.M$ can be seen from the example

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

For $A = A_c + A_d \in SPD.M_0$ we have

$$A + A^T = 2\,A_d + A_c + A_c^T > 0\,,$$

since $A_d = A_d^T > 0$ and $A_c + A_c^T > 0$. This proves $SPD.M_0 \subset PD$. Finally, consider

$$A = \begin{pmatrix} 2 & 0 \\ 3 & 2 \end{pmatrix} \in PD$$

and assume a splitting

$$A = \begin{pmatrix} \beta_1 & \alpha \\ \alpha & \beta_2 \end{pmatrix} + A_c =: A_s + A_c\,,$$

with $A_s \in SPD$, $A_c \in M_0$. The off-diagonal entries of A_c are nonpositive and the diagonal entries are strictly positive, and thus $0 < \beta_1, \beta_2 < 2$, $\alpha \geq 3$ must hold. This implies that $\det A_s = \beta_1\,\beta_2 - \alpha^2$, which equals the product of the eigenvalues of A_s, is negative. Hence we have a contradiction and $PD \neq SPD.M_0$ holds. ∎

For the class of M-matrices the theory of regular splittings (cf. [13,22]) yields that both the Jacobi and the Gauss-Seidel method are convergent iterations:

$$\varrho(M_J) < 1\,, \qquad \varrho(M_{GS}) < 1\,, \tag{2.11}$$

where M_J and M_{GS} are the iteration matrices of the Jacobi and Gauss-Seidel method, respectively. In the case of finite difference discretizations of convection-diffusion problems often the resulting matrix is an M-matrix and the results in (2.11) apply. In case of finite element and finite volume discretizations on irregular grids, however, usually the sign of off-diagonal entries varies and the resulting matrix is not an M-matrix in general. In Section 4. it is shown that often these matrices are elements of the classes $SPD.M_0$, $SPD.M$, or PD. Here we briefly discuss a typical example:

Example 2.1. We consider an elliptic boundary value problem of the form

$$-\varepsilon\,\Delta u + \nabla\cdot(b\,u) = f \qquad \text{in } \Omega \subset \mathbb{R}^n, \tag{2.12}$$

$$u = 0 \qquad \text{on } \partial\Omega, \tag{2.13}$$

with a constant $\varepsilon > 0$ and a function b which is sufficiently smooth. We use a finite element or finite volume discretization with piecewise linear ansatz functions on a consistent triangulation of Ω. In Section 4. it is shown that under certain reasonable conditions on the function b the following holds:

(a) the upwind triangle finite element method of Tabata yields a matrix $A = A_d + A_c \in SPD.M$.
(b) the finite volume schemes of Bank et al. [2] and Bey [6] yield a matrix $A = A_d + A_c \in SPD.M$. If $\nabla\cdot b = 0$ then even $A \in SPD.M_0$ holds.
(c) the streamline diffusion method yields a matrix $A \in PD$.

We note that in all these cases, in general the resulting matrix is not an M-matrix. This M-matrix property can be proved for the cases (a), (b) if one assumes that the triangulation is of *weakly acute type* (cf. Section 4.). This assumption, however, is usually not fulfilled in practice.

3. Convergence of Jacobi and Gauss-Seidel type of methods

In the subsections below we consider the matrix classes $SPD.M_0$, $SPD.M$, and PD (cf. diagram (2.10)) and derive convergence properties of Jacobi and Gauss-Seidel type of methods when applied to matrices from these classes.

3.1. Convergence analysis in the matrix class $SPD.M_0$

We consider

$$A \;=\; A_d + A_c$$

with $A_d \in SPD$ and $A_c \in M_0 = PD \cap Z$. We use the notation

$$D_d \;:=\; \mathrm{diag}(A_d), \qquad D_c \;:=\; \mathrm{diag}(A_c), \qquad D_A \;:=\; \mathrm{diag}(A) \;=\; D_d + D_c\,.$$

In the first part of this section we analyze the convergence of the damped Jacobi method. We will prove a contraction result with respect to the Euclidean norm (Theorem 3.1.). In the analysis we use the numerical radius (cf. [15])

$$r(B) \;:=\; \sup\{\,|\,x^*Bx\,|\,;\; x\in\mathbb{C}^n,\;\|\,x\,\|_2 = 1\,\}$$

for a matrix $B \in \mathbb{R}^{n\times n}$.

We collect a few results concerning the numerical radius from the literature (cf. [10,15]):

$$r(B+C) \;\le\; r(B) + r(C) \qquad (B, C \in \mathbb{R}^{n\times n}) \tag{3.1}$$

$$r(\alpha B) \;=\; |\,\alpha\,|\,r(B) \qquad (\alpha \in \mathbb{C}) \tag{3.2}$$

$$r(B) \;=\; \varrho(B) \qquad \text{if } B \text{ is normal} \tag{3.3}$$

$$\varrho(B) \;\le\; r(B) \tag{3.4}$$

$$\begin{aligned} \tfrac{1}{2} \parallel B \parallel_2 &\leq r(B) \leq \parallel B \parallel_2 && (3.5)\\ r(B^k) &\leq r(B)^k \qquad (k \in \mathbb{N}) && (3.6)\\ r(B) &= \varrho(\tfrac{1}{2}(B + B^T)) \qquad \text{if } B \geq 0 \text{ elementwise} && (3.7) \end{aligned}$$

Furthermore we use the following convergence result for the damped Jacobi method applied to the matrix $A_d \in SPD$ (cf. [13]):

$$\varrho(I - \theta D_d^{-1} A_d) < 1 \qquad \text{for all } \theta \in \Big(0, \frac{2}{\varrho(D_d^{-1} A_d)}\Big). \tag{3.8}$$

Applying this result to the matrix $\frac{1}{2}(A_c + A_c^T) \in SPD$ yields

$$\varrho\Big(I - \tfrac{\theta}{2} D_c^{-1}(A_c + A_c^T)\Big) < 1 \qquad \text{for all } \theta \in \Big(0, \frac{4}{\varrho(D_c^{-1}(A_c + A_c^T))}\Big). \tag{3.9}$$

Lemma 3.1. *Let $\Lambda \geq \lambda > 0$ be such that*

$$\lambda I \leq D_d^{-1} D_c \leq \Lambda I. \tag{3.10}$$

We define

$$k_1 := \frac{1}{\lambda + 1}, \qquad k_2 := \frac{\Lambda}{\Lambda + 1},$$

and

$$\theta_d := \frac{2}{\varrho(D_d^{-1} A_d)}, \qquad \theta_c := \min\left\{1, \frac{4}{\varrho(D_c^{-1}(A_c + A_c^T))}\right\}. \tag{3.11}$$

Then

$$r(I - \theta D_A^{-1/2} A D_A^{-1/2}) < 1 \qquad \textit{for all} \quad \theta \in (0, \theta_A), \tag{3.12}$$

where

$$\theta_A = \frac{\theta_d \theta_c}{k_1 \theta_c + k_2 \theta_d}. \tag{3.13}$$

Proof Note that $D_d > 0$ and, due to $A_c + A_c^T \in SPD$, also $D_c > 0$. For $\alpha \in (0,1)$ and $\theta \in (0, \theta_A)$ we have

$$\begin{aligned} r(I - \theta D_A^{-1/2} A D_A^{-1/2}) &= r\Big(\alpha I + (1-\alpha) I - \theta D_A^{-1/2}(A_d + A_c) D_A^{-1/2}\Big)\\ &\leq \alpha\, r\Big(I - \frac{\theta}{\alpha} D_A^{-1/2} A_d D_A^{-1/2}\Big)\\ &\quad + (1-\alpha)\, r\Big(I - \frac{\theta}{1-\alpha} D_A^{-1/2} A_c D_A^{-1/2}\Big). \end{aligned} \tag{3.14}$$

We use the notation $\tilde{D} := k_1 (I + D_d^{-1} D_c)$, $\hat{D} := k_2 (I + D_c^{-1} D_d)$ and note that

$$I \leq \tilde{D}, \qquad I \leq \hat{D} \tag{3.15}$$

holds. We first consider the term $r(I - \frac{\theta}{\alpha} D_A^{-1/2} A_d D_A^{-1/2})$ in (3.14). For the symmetric positive definite matrix $D_A^{-1/2} A_d D_A^{-1/2}$ we obtain (cf. (3.15))

$$\begin{aligned} D_A^{-1/2} A_d D_A^{-1/2} &= (D_d + D_c)^{-1/2} A_d (D_d + D_c)^{-1/2}\\ &= k_1 \tilde{D}^{-1/2} D_d^{-1/2} A_d D_d^{-1/2} \tilde{D}^{-1/2} \leq k_1 D_d^{-1/2} A_d D_d^{-1/2}. \end{aligned}$$

Hence, $\sigma(D_A^{-1/2} A_d D_A^{-1/2})$, the spectrum of the matrix $D_A^{-1/2} A_d D_A^{-1/2}$, is contained in the interval $(\,0,\, 2\,k_1/\theta_d\,)$. It follows that with

$$\alpha \;:=\; \frac{k_1\,\theta_c}{k_1\,\theta_c + k_2\,\theta_d} \tag{3.16}$$

we have

$$\sigma\Big(\frac{\theta}{\alpha}\,D_A^{-1/2} A_d D_A^{-1/2}\Big) \;\subset\; \Big(0,\, \frac{\theta_A}{\alpha}\cdot\frac{2\,k_1}{\theta_d}\Big) \;=\; (\,0,2\,)$$

and thus

$$r\Big(I - \frac{\theta}{\alpha}\,D_A^{-1/2} A_d D_A^{-1/2}\Big) \;=\; \varrho\Big(I - \frac{\theta}{\alpha}\,D_A^{-1/2} A_d D_A^{-1/2}\Big) \;<\; 1\,. \tag{3.17}$$

We now consider the term $r\Big(I - \frac{\theta}{1-\alpha} D_A^{-1/2} A_c D_A^{-1/2}\Big)$ with α as in (3.16). First note that due to $A_c \in Z$ and $D_A > 0$ we have $D_A^{-1/2} A_c D_A^{-1/2} \in Z$. For the diagonal of $D_A^{-1/2} A_c D_A^{-1/2}$ we obtain

$$\mathrm{diag}\,(D_A^{-1/2} A_c D_A^{-1/2}) \;=\; D_A^{-1} D_c \;=\; (D_d + D_c)^{-1} D_c \;=\; k_2\,\hat{D}^{-1}$$

and thus (cf. (3.15))

$$\begin{aligned}
\mathrm{diag}\,\Big(I - \frac{\theta}{1-\alpha} D_A^{-1/2} A_c D_A^{-1/2}\Big) \;&=\; \mathrm{diag}\,\Big(I - \frac{\theta\,k_2}{1-\alpha}\hat{D}^{-1}\Big) \\
&\geq\; \Big(1 - \frac{\theta\,k_2}{1-\alpha}\Big)\,I \;\geq\; \Big(1 - \frac{\theta_A\,k_2}{1-\alpha}\Big)\,I \\
&=\; (1-\theta_c)\,I \;\geq\; 0\,.
\end{aligned}$$

We conclude that $I - \frac{\theta}{1-\alpha} D_A^{-1/2} A_c D_A^{-1/2} \geq 0$ elementwise and thus (cf. (3.7))

$$r\Big(I - \frac{\theta}{1-\alpha} D_A^{-1/2} A_c D_A^{-1/2}\Big) \;=\; \varrho\Big(I - \frac{1}{2}\,\frac{\theta}{1-\alpha} D_A^{-1/2}(A_c + A_c^T) D_A^{-1/2}\Big)\,. \tag{3.18}$$

The term on the right hand side in (3.18) can be treated along the same lines as the diffusion term above. For the symmetric positive definite matrix $D_A^{-1/2}(A_c + A_c^T) D_A^{-1/2}$ we obtain

$$\begin{aligned}
D_A^{-1/2}(A_c + A_c^T) D_A^{-1/2} \;&=\; k_2\,\hat{D}^{-1/2} D_c^{-1/2}(A_c + A_c^T) D_c^{-1/2}\hat{D}^{-1/2} \\
&\leq\; k_2\,D_c^{-1/2}(A_c + A_c^T) D_c^{-1/2}\,.
\end{aligned}$$

and hence

$$\begin{aligned}
\sigma\Big(\frac{\theta}{2\,(1-\alpha)}\,D_A^{-1/2}(A_c + A_c^T) D_A^{-1/2}\Big) \;&\subset\; \Big(0,\, \frac{\theta_A}{2\,(1-\alpha)}\cdot k_2\,\varrho(D_c^{-1}(A_c + A_c^T))\Big) \\
&\subset\; \Big(0,\, \frac{\theta_A}{1-\alpha}\cdot\frac{2\,k_2}{\theta_c}\Big) \;=\; (\,0,2\,)\,.
\end{aligned}$$

It follows that

$$r\Big(I - \frac{\theta}{1-\alpha}\,D_A^{-1/2} A_c D_A^{-1/2}\Big) \;<\; 1 \tag{3.19}$$

holds. Combination of the results (3.14), (3.17), (3.19) completes the proof. ■

Remark 3.1. We briefly comment on the maximal damping parameter $\theta_A = \frac{\theta_d\,\theta_c}{k_1\,\theta_c + k_2\,\theta_d}$ in Lemma 3.1.. The scalars k_1, k_2 can be interpreted as a measure for the convection-diffusion ratio. If, for example, $D_d = \alpha_d\,I$, $D_c = \alpha_c\,I$, then

$$\lambda = \Lambda = \frac{\alpha_c}{\alpha_d}\,, \qquad k_1 = \frac{\alpha_d}{\alpha_c + \alpha_d}\,, \qquad k_2 = 1 - k_1\,,$$

and

$$\theta_A = \Big(k_1\,\frac{1}{\theta_d} + (1-k_1)\,\frac{1}{\theta_c}\Big)^{-1}\,,$$

i.e. θ_A is a weighted harmonic average of θ_d and θ_c. In this case, if $\alpha_c/\alpha_d << 1$ then $k_1 \approx 1$ and $\theta_A \approx \theta_d$, and if $\alpha_c/\alpha_d >> 1$ then $k_2 \approx 1$ and $\theta_A \approx \theta_c$.

In the general case we have $k_1 + k_2 \in [\,1, 2\,)$ and thus

$$\frac{1}{2}\,\min\{\,\theta_d, \theta_c\,\} \;\le\; \theta_A \;\le\; \max\{\,\theta_d, \theta_c\,\}\,. \tag{3.20}$$

In our applications we often have $\theta_d \approx \theta_c \approx 1$, in which case the damping parameter θ_A is also of order 1 (cf. Example 3.4.).

Using Lemma 3.1. we obtain a convergence result for the damped Jacobi method:

Theorem 3.1. *For $\theta \in (\,0, \theta_A\,)$, with θ_A as in (3.13), define*

$$C_\theta := r(I - \theta\,D_A^{-1/2} A D_A^{-1/2}) < 1\,.$$

Then the estimate

$$\|\,(I - \theta\,D_A^{-1} A)^k\,\|_2 \;\le\; 2\,\sqrt{\kappa(D_A)}\;C_\theta^k\,, \qquad k \in \mathbb{N}\,, \tag{3.21}$$

holds, where $\kappa(D_A)$ denotes the condition number of D_A with respect to the Euclidean norm.

Proof

$$\begin{aligned}
\|\,(I - \theta\,D_A^{-1} A)^k\,\|_2 &= \|\,D_A^{-1/2}(I - \theta\,D_A^{-1/2} A D_A^{-1/2})^k D_A^{1/2}\,\|_2 \\
&\le \sqrt{\kappa(D_A)}\,\|\,(I - \theta\,D_A^{-1/2} A D_A^{-1/2})^k\,\|_2 \\
&\le 2\,\sqrt{\kappa(D_A)}\,r\Big((I - \theta\,D_A^{-1/2} A D_A^{-1/2})^k\Big) \qquad \text{(cf. (3.5))} \\
&\le 2\,\sqrt{\kappa(D_A)}\,\Big(r(I - \theta\,D_A^{-1/2} A D_A^{-1/2})\Big)^k\,. \qquad \text{(cf. (3.6))}
\end{aligned}$$

■

If in our applications we restrict ourselves to boundary value problems with smoothly varying coefficients, then the term $\sqrt{\kappa(D_A)}$ is usually harmless. Theorem 3.1. then yields a contraction in the Euclidean norm for the damped Jacobi method. Note that for many strongly nonsymmetric problems such a contraction result is very different from the asymptotic convergence result in (2.11).

Remark 3.2. In Theorem 3.1. convergence of the damped Jacobi method is proved for $A_c \in M_0$. In practice the condition $A_c + A_c^T > 0$ is often too restrictive, for example if the domain under consideration contains regions where the convection term vanishes. If we allow $\lambda = 0$ in (3.10) and if D_c^{-1} is replaced by the pseudo-inverse D_c^+ of D_c in (3.11), then Theorem 3.1. still holds for matrices $A_c \in Z$ satisfying $A_c + A_c^T \geq 0$. This can be shown by a simple perturbation argument. In Section 4.4. it is shown that for problems in which the convection field b is incompressible, i.e. $\nabla \cdot b = 0$, the finite volume schemes of Bank et al. [2] and Bey [6] yield matrices $A = A_d + A_c$ satisfying $A_d \in SPD$, $A_c \in Z$, and $A_c + A_c^T \geq 0$.

We now consider the damped Gauss-Seidel method for an arbitrary matrix $A = A_d + A_c$ in $SPD.M_0$. The example below shows that, even with optimal damping, the Gauss-Seidel method is not convergent for all $A \in SPD.M_0$. For the formulation of the Gauss-Seidel method we use a splitting

$$A \;=\; L_A + D_A + U_A\,,$$

with L_A strictly lower triangular, U_A strictly upper triangular, and $D_A = \mathrm{diag}(A)$.

Example 3.1. Define $\mathbb{1} := (1,1,1,1,1)^T$ and, with I the identity matrix in $\mathbb{R}^{5\times 5}$ and $\varepsilon > 0$:

$$A_d \;:=\; \mathbb{1}\mathbb{1}^T + \varepsilon\, I \;\in\; SPD\,.$$

For $\delta > 0$ we define

$$A_c \;:=\; \delta \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix} \;\in\; M_0$$

and $A = A_{\varepsilon,\delta}$ by $A := A_d + A_c \in SPD.M_0$. Then for all $\delta_0 > 0$ sufficiently small there exists $\varepsilon_0 = \varepsilon_0(\delta) > 0$ such that for all $\varepsilon \in (\,0, \varepsilon_0\,]$:

$$\varrho\Big(I - \theta\,(L_A + D_A)^{-1} A\Big) \;>\; 1 \qquad\qquad \text{for all } \theta \in \mathbb{R} \setminus \{0\}.$$

Proof We consider A_d with $\varepsilon = 0$, i.e. $A_d = \mathbb{1}\mathbb{1}^T$, and use a continuity argument to obtain a result for $\varepsilon > 0$. Let $B := \delta^{-1} A_c$ and $w := B^{-1}\mathbb{1} = (1,2,3,4,5)^T$. Note that

$$L_A + D_A \;=\; L_{A_d} + D_{A_d} + A_c \;=\; B^{-1} + \delta\, B$$

holds. We first consider the matrix $\delta\, A^{-1}(L_A + D_A)$:

$$\begin{aligned} \delta\, A^{-1}(L_A + D_A) &= \delta\,(\mathbb{1}\mathbb{1}^T + \delta\, B)^{-1}(B^{-1} + \delta\, B) \\ &= \Big(I + \frac{1}{\delta}\, w\, \mathbb{1}^T\Big)^{-1}(B^{-2} + \delta\, I) \\ &= \Big(I - \frac{1}{\delta + \mathbb{1}^T w}\, w\, \mathbb{1}^T\Big)\,(B^{-2} + \delta\, I) \\ &= \Big(I - \frac{1}{15}\, w\, \mathbb{1}^T\Big)\, B^{-2} + O(\delta) \qquad (\delta \to 0)\,. \end{aligned}$$

The matrix $E := (I - \frac{1}{15}\, w\, \mathbb{1}\mathbb{1}^T)\, B^{-2}$ has eigenvalues

$$\sigma(E) \;=\; \{\, 0,\; -0.0968 \pm 0.8122i,\; 0.2635 \pm 0.1738i \,\}\,.$$

Hence the matrix $A^{-1}(L_A+D_A)$ has both eigenvalues with positive real part and eigenvalues with negative real part if δ is sufficiently small. Since $\text{sign}(\text{Re}(\lambda)) = \text{sign}(\text{Re}(\lambda^{-1}))$ for $\lambda \in \mathbb{C}$, it follows that $(L_A + D_A)^{-1}A$ has both eigenvalues with positive real part and eigenvalues with negative real part if δ is sufficiently small. This holds for $A_d = \mathbb{1}\mathbb{1}^T + \varepsilon\, I$ with $\varepsilon = 0$. From a continuity argument it follows that for $\varepsilon(\delta) > 0$ sufficiently small the matrix $(L_A + D_A)^{-1}A$ has both eigenvalues with positive real part and eigenvalues with negative real part. ■

Remark 3.3. It turns out that if we consider dimension $n \leq 4$ then the matrix E used in the proof of Example 3.1. is positive semidefinite: $E + E^T \geq 0$. This explains why we consider dimension $n = 5$ in Example 3.1.. A straightforward calculation yields that for *every* 2×2 matrix in $SPD.M_0$ the Gauss-Seidel method without damping ($\theta = 1$) is convergent. Hence, for a negative result as in Example 3.1. we need dimension $n \geq 3$.

3.2. Convergence analysis in the matrix class $SPD.M$

The embedding $SPD.M_0 \subset SPD.M$ and the negative result in Example 3.1. for the class $SPD.M_0$ show that even with optimal damping the Gauss-Seidel method is not convergent for all $A \in SPD.M$. In the following example we show that a similar negative result holds for the Jacobi method. We conclude that the favourable properties w.r.t. convergence of the damped Jacobi and Gauss-Seidel methods which hold both in the class SPD and in the class M are lost in the class $SPD.M$.

Example 3.2. Consider $A = A_d + A_c$ with

$$A_d := \begin{pmatrix} 1 & -\frac{5}{2} \\ -\frac{5}{2} & 25 \end{pmatrix} \in SPD\,, \qquad A_c := \begin{pmatrix} 1 & -\frac{1}{2} \\ -50 & 50 \end{pmatrix} \in M \setminus M_0\,.$$

Then

$$\varrho(I - \theta\, D_A^{-1}A) \;>\; 1 \qquad \text{for all } \theta \in \mathbb{R} \setminus \{0\}.$$

Proof The matrix

$$D_A^{-1}A \;=\; \begin{pmatrix} 1 & -\frac{3}{2} \\ -\frac{7}{10} & 1 \end{pmatrix}$$

has determinant < 0. Hence this matrix has two real eigenvalues with opposite sign. ■

3.3. Convergence analysis in the class PD

The embedding $SPD.M_0 \subset PD$ and the negative result in Example 3.1. for the class $SPD.M_0$ imply that even with optimal damping the Gauss-Seidel method is not convergent for all $A \in PD$.

For the Jacobi method applied to matrices $A \in PD$ one can find convergence results in the literature, cf. [13,19] and the references therein. Here, we present one typical result ([13], Thm. 4.4.16):

Theorem 3.2. *Let $A \in PD$, $D_A := diag(A)$, and $0 < \lambda \leq \Lambda$, $\tau \geq 0$ be constants such that*

$$\lambda\, D_A \,\leq\, \frac{1}{2}\,(A + A^T) \,\leq\, \Lambda\, D_A\,, \qquad -\tau\, D_A \,\leq\, \frac{1}{2i}\,(A - A^T) \,\leq\, \tau\, D_A\,. \quad (3.22)$$

Then

$$\| \, I - \theta\, D_A^{-1} A \, \|_{D_A} \,:=\, \| \, I - \theta\, D_A^{-1/2} A D_A^{-1/2} \, \|_2 \,<\, 1 \qquad \textit{for all } \theta \in (\,0,\, \tilde{\theta}_A\,),$$

with $\tilde{\theta}_A$ defined by

$$\tilde{\theta}_A \;:=\; \frac{2\,\lambda}{\lambda\,\Lambda + \tau^2}\,. \quad (3.23)$$

Proof Given in [13]. ∎

We now compare the damping parameters $\tilde{\theta}_A$ (cf. Thm. 3.2.) and θ_A (cf. Thm. 3.1.) which are used for the Jacobi method in the matrix classes PD and $SPD.M_0 \subset PD$, respectively. First note that in practical applications it is easy to obtain a reasonable estimate of the parameter θ_A (cf. Example 3.4. below). For the parameter $\tilde{\theta}_A$ it is often much harder to obtain a reasonable estimate. This is caused by the factor λ occurring in the formula for $\tilde{\theta}_A$, which is an estimate for the *smallest* eigenvalue of the matrix $\frac{1}{2}\, D_A^{-1}(A+A^T)$. Also note that in Thm. 3.2. the damping is very strong if the nonsymmetric part $\frac{1}{2}\,(A-A^T)$ is "much larger" than the symmetric part $\frac{1}{2}\,(A + A^T)$: $\tilde{\theta}_A << 1$ if $\tau^2 >> \lambda$. The ratio between the size of the symmetric and nonsymmetric part does not play an important role in the damping parameter θ_A (cf. Remark 3.1.). To illustrate these phenomena, we consider two examples:

Example 3.3. Let $B \in \mathbb{R}^{n\times n}$ be skew-symmetric: $B^T = -B$. Consider the matrix

$$A \,=\, \varepsilon\, I + B \in\; PD$$

with $\varepsilon > 0$. The matrix $D_A^{-1/2} A D_A^{-1/2} \,=\, I + \frac{1}{\varepsilon}\, B$ is normal and a simple calculation yields

$$\| \, I - \theta\, D_A^{-1/2} A D_A^{-1/2} \, \|_2 \,=\, \varrho(I - \theta\, D_A^{-1/2} A D_A^{-1/2}) \,<\, 1$$

iff

$$0 < \theta < \frac{2\,\varepsilon^2}{\varepsilon^2 + \varrho^2}\,, \quad (3.24)$$

where $\varrho = \varrho(B)$ denotes the spectral radius of B. Since

$$\frac{1}{2}\, D_A^{-1/2}(A + A^T) D_A^{-1/2} \,=\, I, \qquad \frac{1}{2i}\, D_A^{-1/2}(A - A^T) D_A^{-1/2} \,=\, -\frac{i}{\varepsilon}\, B,$$

for the constants in Thm. 3.2. we can take $\lambda = \Lambda = 1$, $\tau = \varrho/\varepsilon$, and we then obtain a maximal damping parameter

$$\tilde{\theta}_A \,=\, \frac{2\,\varepsilon^2}{\varepsilon^2 + \varrho^2}\,.$$

Comparison with (3.24) shows that for this example the result of Thm. 3.2. is sharp. It is clear that if $\tau = \varrho/\varepsilon >> 1$ then $\tilde{\theta}_A << 1$.

We note that for $\varrho/\varepsilon > 2$ we have $A \notin SPD.M_0$. This can be shown as follows: Assume that

$$A \,=\, \varepsilon\, I + B \,=\, A_d + A_c\,, \qquad\qquad A_d \in SPD,\ A_c \in M_0\,.$$

From this we obtain

$$B \,=\, \tfrac{1}{2}\,(A_c - A_c^T)\,, \qquad\qquad \varepsilon\, I \,=\, A_d + \tfrac{1}{2}\,(A_c + A_c^T)\,,$$

and hence, it follows from $A_c + A_c^T \geq 0$ that $\varrho(D_c) \leq \varepsilon$ and $\frac{1}{2}\,\varrho(A_c + A_c^T) \leq \varepsilon$. For $R_c = D_c - A_c$, which is ≥ 0 elementwise, we have (cf. (3.1) to (3.7))

$$\begin{aligned} r(R_c) \;&=\; \tfrac{1}{2}\,\varrho(R_c + R_c^T) \,=\, \parallel D_c - \tfrac{1}{2}\,(A_c + A_c^T)\,\parallel_2 \\ &\leq\; \parallel D_c\,\parallel_2 + \tfrac{1}{2}\parallel A_c + A_c^T\,\parallel_2 \,=\, \varrho(D_c) + \tfrac{1}{2}\,\varrho(A_c + A_c^T)\,. \end{aligned}$$

Combination of these results yields

$$\varrho \,=\, \varrho(B) \,=\, r(B) \,=\, \tfrac{1}{2}\,r(R_c^T - R_c) \,\leq\, r(R_c) \,\leq\, \varrho(D_c) + \tfrac{1}{2}\,\varrho(A_c + A_c^T)_2 \,\leq\, 2\,\varepsilon\,.$$

We conclude that if the skew-symmetric part of A is large compared to the symmetric part ($\varrho/\varepsilon >> 1$) then we need strong damping and the matrix A does not lie in the class $SPD.M_0$.

Example 3.4. We consider the elliptic boundary value problem in Example 2.1. and assume that the flow field b is incompressible: $\nabla \cdot b = 0$. Then the finite volume discretization described in Section 4.4. yields a matrix $A = A_d + A_c \in SPD.M_0$ (cf. Lemma 4.5.(h)). We further assume that the corresponding triangulations $\mathcal{T}_h$ are *quasi-uniform* and *stable*, i.e., the elements in each triangulation $\mathcal{T}_h$ are of comparable size and the elements do not degenerate for $h \to 0$. Using well-known estimates from the theory of finite element methods (cf. [8,12]) it then follows that $\varrho(D_d^{-1}A_d)$ is bounded independently of ε, h and hence, the parameter θ_d in (3.11) is bounded away from zero independently of ε, h. Note that θ_d represents the maximal damping parameter of the Jacobi method applied to the matrix A_d. Since the matrix A_c is weakly diagonally dominant with respect to its rows and columns (cf. Lemma 4.5.(c),(f)), we obtain $\theta_c = 1$ in (3.11). Using (3.13), (3.20), and Theorem 3.1., we conclude that the damped Jacobi method applied to the matrix $A = A_d + A_c$ converges for $\theta \in (\,0, \theta_A\,)$, where

$$\theta_A \;\geq\; \frac{\theta_d}{2}$$

is bounded away from zero independently of ε, b and h. On the other hand, if we apply Theorem 3.2. in this situation then the result is less satisfactory. Assuming $|\,b\,| > 0$, for the parameters λ, Λ and τ in (3.22) one can show

$$\lambda \,\sim\, h^2\,, \qquad \Lambda \,\sim\, 1\,, \qquad \tau \,\sim\, \frac{1}{1 + \frac{\varepsilon}{|\,b\,|\,h}}\,.$$

Hence, for the maximal damping parameter $\tilde{\theta}_A$ in (3.23) we obtain

$$\tilde{\theta}_A \;\sim\; \frac{(\frac{\varepsilon}{|\,b\,|} + h)^2}{(\frac{\varepsilon}{|\,b\,|} + h)^2 + 1}\,.$$

It follows that $\tilde{\theta}_A$ approaches zero for $\frac{\varepsilon}{|\,b\,|} + h \to 0$. We conclude that for this example the damping resulting from Thm. 3.2. is much too strong if both $\frac{\varepsilon}{|\,b\,|}$ and h are small.

Due to the negative result in Example 3.1., convergence of the damped Gauss-Seidel method cannot be proved for arbitrary $A \in SPD.M_0$. On the other hand, for convection-dominated problems the Gauss-Seidel method with "downwind numbering" (cf. [7,14]) seems to be an efficient solver (smoother) in many practical applications. In order to fill partly this gap between theory and practice we consider a hybrid method which is motivated by the following result:

Theorem 3.3. *For $A \in PD$ consider a splitting*

$$A \;=\; A_d + A_c \qquad \textit{with} \qquad A_d \in SPD,\; A_c + A_c^T \geq 0\,. \tag{3.25}$$

Let $W \in SPD$ be such that

$$2\,W - A_d \;>\; 0\,. \tag{3.26}$$

Then the following holds:

$$\| \, I - (W + A_c)^{-1} A \, \|_W \;\leq\; \| \, I - W^{-1} A_d \, \|_W \;<\; 1\,. \tag{3.27}$$

Proof Note that

$$\begin{aligned} \| \, I - (W + A_c)^{-1} A \, \|_W &= \| \, I - (W + A_c)^{-1} (A_d - W + W + A_c) \, \|_W \\ &= \| \, (W + A_c)^{-1} (A_d - W) \, \|_W \\ &= \| \, (I + W^{-1} A_c)^{-1} (I - W^{-1} A_d) \, \|_W \\ &\leq \| \, (I + W^{-1} A_c)^{-1} \, \|_W \, \| \, I - W^{-1} A_d \, \|_W\,. \end{aligned} \tag{3.28}$$

Using the notation $\tilde{A}_c := W^{-1/2} A_c W^{-1/2}$, for the first term on the right hand side in (3.28) we have

$$\begin{aligned} \| \, (I + W^{-1} A_c)^{-1} \, \|_W &= \| \, (I + \tilde{A}_c)^{-1} \, \|_2 \\ &= \varrho \left((I + \tilde{A}_c)^{-1} (I + \tilde{A}_c^T)^{-1} \right)^{1/2} \\ &= \varrho \left((I + \tilde{A}_c + \tilde{A}_c^T + \tilde{A}_c^T \tilde{A}_c)^{-1} \right)^{1/2} \;\leq\; 1\,. \end{aligned} \tag{3.29}$$

The latter result follows from the fact that both $\tilde{A}_c + \tilde{A}_c^T$ and $\tilde{A}_c^T \tilde{A}_c$ are symmetric positive semi-definite. Using the result (3.29) in (3.28) yields

$$\begin{aligned} \| \, I - (W + A_c)^{-1} A \, \|_W &\leq \| \, I - W^{-1} A_d \, \|_W \\ &= \varrho(I - W^{-1/2} A_d W^{-1/2}) \;<\; 1\,. \end{aligned}$$

The latter inequality results from $0 < W^{-1/2} A_d W^{-1/2} < 2\,I$ (cf. (3.26)). ■

Remark 3.4. It is easy to show that the matrix class

$$V \;:=\; \{\; A \in \mathbb{R}^{n \times n} \mid A = A_d + A_c \text{ with } A_d \in SPD,\; A_c + A_c^T \geq 0 \;\}$$

equals the class of positive definite matrices: $V = PD$.

We now discuss how for a discretized convection-diffusion problem with matrix A in the class PD Thm. 3.3. can yield convergence of a feasible method.

As an example we consider the stabilized box method described in Section 4.4., applied to the problem (2.12), (2.13). We assume incompressibility, i.e. $\nabla \cdot b = 0$. Then the resulting discretization has a matrix $A = A_d + A_c$ with $A_d \in SPD$, $A_c + A_c^T \geq 0$ (cf. Lemma 4.5.). If the directed graph corresponding to the convection matrix A_c is acyclic, i.e. does not contain any cycles, then using numbering algorithms as in [7,14] one can reorder the unknowns such that the resulting permuted convection matrix is lower triangular. We take $W := \theta \,\mathrm{diag}(A_d)$ with $\theta > 0$ such that $2\,W - A_d > 0$ holds. For this example the assumptions of Theorem 3.3. are satisfied and we obtain a feasible method since systems with the matrix $W + A_c = \theta \,\mathrm{diag}(A_d) + A_c$ can be solved with acceptable computational costs. If the directed graph corresponding to the convection matrix A_c contains cycles, then after suitable reordering (cf. [6,14]) the matrix A_c is block-lower triangular. Depending on the size of the diagonal blocks this may still result in a feasible method.

4. Algebraic properties of stiffness matrices resulting from the discretization of convection-diffusion problems

Let $\Omega \subset \mathbb{R}^n$ be a polyhedral domain with boundary Γ. We consider elliptic boundary value problems of the form

$$-\varepsilon\,\Delta u + \nabla \cdot (b\,u) + c\,u = f \quad \text{in } \Omega, \tag{4.1}$$

$$u = 0 \quad \text{on } \Gamma. \tag{4.2}$$

For simplicity we assume that $\varepsilon > 0$ is constant and that the vector field b and the scalar functions c, f are sufficiently smooth, e.g. $b \in H^{1,\infty}(\Omega)^n$, $c \in L^\infty(\Omega)$ and $f \in L^2(\Omega)$. If $\nabla \cdot b = 0$ then problem (4.1), (4.2) is called *incompressible*. The weak formulation of (4.1), (4.2) reads: *Find* $u \in H_0^1(\Omega)$ *such that*

$$a(u,v) = \int_\Omega \varepsilon\,\nabla u \cdot \nabla v\; dx + \int_\Omega v \nabla \cdot (b\,u)\; dx + \int_\Omega c\,u\,v\; dx = \int_\Omega f\,v\; dx \tag{4.3}$$

for all $v \in H_0^1(\Omega)$. If $c + \frac{1}{2}\nabla \cdot b \geq 0$ then the bilinear form $a(\cdot,\cdot)$ is coercive in $H_0^1(\Omega)$, i.e.

$$a(u,u) \geq \alpha \parallel u \parallel^2, \qquad u \in H_0^1(\Omega), \tag{4.4}$$

with coercivity constant $\alpha > 0$ and energy norm $\parallel u \parallel^2 := \int_\Omega |\nabla u|^2\; dx$ for $u \in H_0^1(\Omega)$. In this case it follows from the Lax-Milgram Lemma that (4.3) has a unique solution u in $H_0^1(\Omega)$. Otherwise, if $a(\cdot,\cdot)$ fails to be coercive, the Fredholm alternative applies and hence (4.3) has a unique solution iff the homogeneous problem ($f = 0$) has only the trivial solution $u = 0$ (cf. [9]).

In the subsections below we consider a few discretization methods for the problem (4.1), (4.2), which are known from the literature. In the setting of this paper we are interested in the algebraic properties of the resulting stiffness matrices.

4.1. Finite element discretization

We first discretize (4.1), (4.2) by piecewise linear finite elements based on the standard Galerkin approach. Let $\mathcal{T}_h$ be a consistent triangulation of Ω, let $x_1, \ldots, x_N$ be the vertices

of $\mathcal{T}_h$ not lying on Γ, and let V_h be the corresponding space of continuous piecewise linear functions which are zero on Γ. Then the finite element discretization of the continuous problem (4.3) leads to the discrete problem: *Find* $u_h \in V_h$ *such that*

$$\begin{aligned} a(u_h, v_h) &:= \int_\Omega \varepsilon \nabla u_h \cdot \nabla v_h \, dx + \int_\Omega v_h \, b \cdot \nabla u_h \, dx + \int_\Omega (c + \nabla \cdot b) \, u_h \, v_h \, dx \\ &= \int_\Omega f \, v_h \, dx \end{aligned} \tag{4.5}$$

for all $v_h \in V_h$. If $c + \frac{1}{2} \nabla \cdot b \geq 0$ then $a(\cdot, \cdot)$ is coercive in V_h and (4.5) has a unique solution. Denoting by $\Phi_h := \{\varphi_1, \ldots, \varphi_N\}$ the standard nodal basis of V_h, (4.5) is equivalent to the linear system

$$Ax \; = \; b \tag{4.6}$$

with matrix coefficients

$$A_{ij} \; = \; a(\varphi_j, \varphi_i) \, , \qquad 1 \leq i, j \leq N \, . \tag{4.7}$$

The solution x of (4.6) is related to the solution u_h of (4.5) by $u_h = \sum_{j=1}^N x_j \, \varphi_j$. Corresponding to the left three integrals in (4.5), the *stiffness matrix* A can be split into a diffusion part A_d, a convection part A_c, and a reaction part A_r :

$$A \; = \; A_d + A_c + A_r \, . \tag{4.8}$$

Often for the reaction term a so called *lumping* procedure is used in which the matrix A_r is approximated by a certain diagonal matrix. To be more precise, let Ω_i be the union of all simplices in $\mathcal{T}_h$ sharing the vertex x_i, for $1 \leq i \leq N$. The volume of Ω_i is denoted by $| \, \Omega_i \, |$. The reaction term in (4.5) can then be approximated by

$$\int_\Omega (c + \nabla \cdot b) \, u_h \, v_h \, dx \; \approx \; \sum_{i=1}^N \frac{| \, \Omega_i \, |}{n+1} \, (c + \nabla \cdot b)(x_i) \, u_h(x_i) \, v_h(x_i) \, . \tag{4.9}$$

If $c + \nabla \cdot b \geq 0$ this lumping procedure results in a nonnegative diagonal matrix $\tilde{A}_r$ (cf. Lemma 4.2.(f)). Note that the approximation order of the finite element method is not affected by the lumping procedure.

Before we summarize properties of these matrices, we introduce a few definitions:

Definition 4.1. *$A \in \mathbb{R}^{N \times N}$ satisfies the strong (weak) sign condition if $A \in Z$ and $A_{ii} > 0$ ($A_{ii} \geq 0$) for $1 \leq i \leq N$. The matrix A is called weakly diagonally dominant w.r.t. its rows (columns) if*

$$\sum_{j \neq i} | \, A_{ij} \, | \leq | \, A_{ii} \, | \quad \textit{for} \quad 1 \leq i \leq N \qquad \Big(\sum_{i \neq j} | \, A_{ij} \, | \leq | \, A_{jj} \, | \quad \textit{for} \quad 1 \leq j \leq N \Big).$$

Definition 4.2. *Let $x_{N+1}, \ldots, x_{\hat{N}}$ be the vertices of $\mathcal{T}_h$ which lie on Γ. Let $\hat{V}_h \supset V_h$ be the space of continuous piecewise linear functions corresponding to $\mathcal{T}_h$, i.e. the space of all functions $v_h \in C(\overline{\Omega})$ such that the restriction of v_h to any element $T \in \mathcal{T}_h$ is a linear polynomial on T. We denote by $\hat{\Phi}_h := \{\varphi_1, \ldots, \varphi_{\hat{N}}\}$ the standard nodal basis of $\hat{V}_h$, defined by*

$$\varphi_i(x_j) \; = \; \delta_{ij} \, , \qquad 1 \leq i, j \leq \hat{N} \, .$$

Assume that the coefficients of the matrix $A \in \mathbb{R}^{N\times N}$ are defined by an expression of the form

$$A_{ij} \;=\; a_h(\varphi_j, \varphi_i)\,, \qquad 1 \le i,j \le N\,, \tag{4.10}$$

with some bilinear form $a_h(\cdot,\cdot)$ defined on $\hat{V}_h \times \hat{V}_h$. Then A is said to have the zero row (column) sum property if

$$\sum_{j=1}^{\hat{N}} a_h(\varphi_j, \varphi_i) = 0 \quad \textit{for} \quad 1 \le i \le N \qquad \Big(\sum_{i=1}^{\hat{N}} a_h(\varphi_j, \varphi_i) = 0 \quad \textit{for} \quad 1 \le j \le N \Big).$$

Example 4.1. Let A be the finite element matrix in (4.6), and let $\hat{A}$ be the matrix resulting from the finite element discretization of the corresponding Neumann problem. Clearly, A satisfies (4.10) with $a_h(\cdot,\cdot) = a(\cdot,\cdot)$ (cf. (4.7)). If A has the zero row (column) sum property, this means that the sum of the entries in each row (column) of $\hat{A}$ vanishes. This implies that

$$\sum_{j=1}^{N} A_{ij} = 0 \qquad \Big(\sum_{i=1}^{N} A_{ij} = 0 \Big) \tag{4.11}$$

for each row i (column j) corresponding to a vertex x_i (x_j) which is not connected by an edge to any boundary vertex. Note, however, that in general (4.11) does not hold for the remaining rows (columns) of A. This is due to the fact that "couplings" to boundary vertices are contained in $\hat{A}$ but not in A.

Definition 4.3. *A triangulation $\mathcal{T}_h$ in $\mathbb{R}^n$ is of weakly acute type if the maximum angle between two $(n-1)$-subsimplices of any simplex $T \in \mathcal{T}_h$ is bounded by $\pi/2$.*

In the literature the condition that a triangulation is of weakly acute type is often used as a *sufficient* condition for an M-matrix property of the corresponding stiffness matrix. A typical result is given in the following Lemma:

Lemma 4.1. *Consider the finite element stiffness matrix corresponding to the diffusion part of the bilinear form in (4.5), represented by the symmetric positive definite matrix A_d in (4.8). If the triangulation $\mathcal{T}_h$ is of weakly acute type then $A_d \in M_0$ holds.*

Proof Let $\nabla\varphi_i(T)$ denote the restriction of $\nabla\varphi_i$ to the element $T \in \mathcal{T}_h$. Note that $\nabla\varphi_i$ is constant on every element $T \in \mathcal{T}_h$. Using $|\,T\,|$ to denote the volume of T, we obtain for $1 \le i,j \le N$

$$\begin{aligned} A_{d,i,j} &= \varepsilon \int_\Omega \nabla\varphi_i \cdot \nabla\varphi_j \, dx \\ &= \varepsilon \sum_{T\in\mathcal{T}_h} \int_T \nabla\varphi_i \cdot \nabla\varphi_j \, dx \;=\; \varepsilon \sum_{T\in\mathcal{T}_h} |\,T\,|\, \nabla\varphi_i(T) \cdot \nabla\varphi_j(T). \end{aligned} \tag{4.12}$$

Now consider an arbitrary element $T \in \mathcal{T}_h$. For any vertex x_j of T let $S_j(T)$ be the $(n-1)$-subsimplex of T opposite to x_j. For any two such subsimplices $S_i(T)$, $S_j(T)$ let $\alpha_{ij}(T)$ be the angle between $S_i(T)$ and $S_j(T)$ inside T. If $\mathcal{T}_h$ is of weakly acute type, then we have $\alpha_{ij}(T) \le \pi/2$, hence $\cos\,\alpha_{ij}(T) \ge 0$ for each such angle $\alpha_{ij}(T)$ in $\mathcal{T}_h$.

It is not difficult to see that $\nabla\varphi_j(T)$ is orthogonal to $S_j(T)$. Moreover, $\nabla\varphi_j(T)$ points from there into the direction of the interior of T. It follows that

$$\nabla\varphi_i(T)\cdot\nabla\varphi_j(T) \;=\; -\cos\,\alpha_{ij} \;\leq\; 0 \tag{4.13}$$

for any two vertices $x_i \neq x_j$ of T. On the other hand, if either x_i or x_j is not a vertex of T, then we have $\int_T \nabla\varphi_i\cdot\nabla\varphi_j\;dx = 0$. In combination with (4.12) and (4.13) this proves $A_d \in Z$. It is clear that $A_d \in SPD$ holds, hence we conclude that $A_d \in M_0$. ■

Remark 4.1. Under the assumption that $\mathcal{T}_h$ is of weakly acute type, one can show that the matrix A_d is *essentially diagonally dominant* (cf. [13] for definition). Note that the latter property in combination with the strong sign condition implies the M-matrix property (cf. Theorem 6.4.10 in [13]).

The results in Lemma 4.1. or Remark 4.1. show that A_d is an M-matrix if $\mathcal{T}_h$ is of weakly acute type. This assumption, however, is unrealistic in particular if $\mathcal{T}_h$ is generated by an adaptive refinement process. If $\mathcal{T}_h$ is not of weakly acute type then in general $A_d \notin Z$ and hence $A_d \notin M$. Even worse, A_d in general fails to be inverse monotone, i.e., some entries of A^{-1} may be negative.

In the remainder of this paper for a few well-known discretization methods we derive properties of the corresponding stiffness matrix for the case that $\mathcal{T}_h$ is *not* necessarily of weakly acute type. We start with a few properties of the different components of the finite element matrix A from (4.6) and of the matrix A_r resulting from the lumping procedure (4.9):

Lemma 4.2. *For the matrices A_d, A_c, A_r in (4.8) and for the matrix $\tilde{A}_r$ we have:*

(a) $A_d \in SPD$,
(b) A_d *has both the zero row sum and zero column sum property,*
(c) A_c *has the zero row sum property,*
(d) A_r, $\tilde{A}_r$ *are symmetric.*

Using additional assumptions, we obtain the following results:

(e) if $\nabla\cdot b = 0$ *then* A_c *has the zero column sum property,*
(f) if $c+\nabla\cdot b \geq 0$ *then* A_r, $\tilde{A}_r$ *are positive semi-definite and* ≥ 0 *elementwise,*
(g) if $c+\nabla\cdot b \geq 0$ *then* $A_d + A_r \in SPD$, $A_d + \tilde{A}_r \in SPD$ *hold,*
(h) if $c+\frac{1}{2}\nabla\cdot b \geq 0$ *then* $A \in PD$.

Proof Clearly, A_d, A_r and $\tilde{A}_r$ are symmetric. The coercivity of the diffusion part of the bilinear form $a(\cdot,\cdot)$ implies that A_d is positive definite. The zero row and column sum properties of A_d follow from the fact that the diffusion term vanishes if either u or v is constant. A similar argument for the convection term shows that A_c has the zero row sum property. In case of $\nabla\cdot b = 0$ the zero column sum property of A_c easily follows by partial integration. The diagonal matrix $\tilde{A}_r$, defined by the lumping procedure (4.9), is positive semi-definite and ≥ 0 elementwise if $c+\nabla\cdot b \geq 0$. If the latter condition is satisfied then $\varphi_j \geq 0$ for $1 \leq j \leq N$ implies $A_r \geq 0$ elementwise. In this case we also have $\int_\Omega (c+\nabla\cdot b)\,v^2\;dx \geq 0$ for arbitrary $v \in H_0^1(\Omega)$. Hence A_r is positive semi-definite. The properties in (g) follow from (a), (d), and (f). If $c+\frac{1}{2}\nabla\cdot b \geq 0$ then the coercivity of $a(\cdot,\cdot)$ and the relation $x^T(A+A^T)\,x = 2\,x^T Ax$, $x \in \mathbb{R}^N$, imply $A \in PD$. ■

Unlike A_d and A_r the convection matrix A_c has a skew-symmetric character. This skew-symmetric character will dominate the total stiffness matrix A in regions where the *mesh Peclet number* $|\,b\,|\,h/\varepsilon$ is large. It is well known that the discretization (4.5) in general is highly unstable in case of dominating convection and may produce unphysical oscillations in the discrete solution, in particular if the solution of the continuous problem contains internal or boundary layers.

In the 1D case, if the triangulation $\mathcal{T}_h$ is equidistant, the discretization of the convection term in (4.5) corresponds to the use of central finite differences to approximate the first order derivative of u_h. In this case the above difficulties can be overcome by using backward finite differences instead. This *upwind differencing* procedure preserves the M-matrix property and hence also the inverse monotonicity property. On the other hand, the use of upwind finite differences results in a loss of accuracy since the method is only first order accurate.

In the following sections we consider some well known stabilization methods for the higher-dimensional case. We focus on the algebraic properties of the resulting stiffness matrices and do not investigate the approximation order of the schemes. A rigorous error analysis of these upwind methods can be found in [17].

4.2. Artificial diffusion

The artificial diffusion method is probably the simplest higher dimensional upwind scheme. It is equivalent to application of the standard finite element discretization (4.5) to a modified problem with a larger diffusion constant $\tilde{\varepsilon}$, which can defined, for example, by

$$\tilde{\varepsilon} \;:=\; \varepsilon + h \parallel b \parallel_\infty . \tag{4.14}$$

With this modified diffusion constant, the resulting discrete problem reads: *Find* $u_h \in V_h$ *such that*

$$\int_\Omega \tilde{\varepsilon}\, \nabla u_h \cdot \nabla v_h \; dx + \int_\Omega v_h\, b \cdot \nabla u_h \; dx + \int_\Omega (c + \nabla \cdot b)\, u_h\, v_h \; dx \;=\; \int_\Omega f\, v_h \; dx \tag{4.15}$$

for any $v_h \in V_h$. Note that (4.15) differs from (4.5) only by the diffusion constant. Let now A^{ad} be the corresponding stiffness matrix for the artificial diffusion method. We consider a splitting $A^{ad} = A_d + A_c^{ad} + A_r$, that is, the artificial diffusion term is added to the convection matrix while the diffusion and reaction matrices remain unchanged. We summarize the main properties of this splitting in the following Lemma:

Lemma 4.3. *All results of Lemma 4.2. hold with* A *and* A_c *replaced by* A^{ad} *and* A_c^{ad}*, respectively.*

Proof $A_c^{ad} - A_c$ is the discretization matrix corresponding to the artificial diffusion term. It follows that $A_c^{ad} - A_c$ has both the zero row sum and the zero column sum property. This implies that A_c^{ad} has the zero row sum or zero column sum property if and only if A_c has the same property. This proves (c) and (e). If $c + \frac{1}{2}\nabla \cdot b \geq 0$, then $A^{ad} - A = A_c^{ad} - A_c \in SPD$ in combination with Lemma 4.2.(h) proves $A^{ad} \in PD$. ∎

A main disadvantage of the artificial diffusion method is that the artificial diffusion acts in all directions, i.e., the method introduces artificial diffusion not only in streamline direction but also perpendicular to the streamlines, in crosswind direction. Below we will consider upwind schemes which in general produce less crosswind diffusion.

Remark 4.2. One of the widely used finite element discretization schemes for convection-diffusion problems is the streamline diffusion method (SDFEM). For a description and analysis of this method we refer to [17]. In the SDFEM artificial diffusion is added mainly in the direction of streamlines in such a way that the resulting stabilized discretization still satisfies a favourable consistency condition (cf. [17]). In general this method has much better approximation quality than the simple artificial diffusion method based on (4.14), (4.15). However, in the algebraic setting of Section 2 the SDFEM stiffness matrix A^{sd} does not seem to have more structure than the stiffness matrix A^{ad} of the simple artificial diffusion method. For the SDFEM as described in [17] a basic result on discrete coercivity of the stabilized bilinear form (Lemma 3.28 in [17]) immediately yields: If $c + \frac{1}{2}\nabla \cdot b \geq 0$ then $A^{sd} \in PD$. A result $A^{sd} \in S$ with S one of the other matrix classes in the diagram (2.10) does not seem to hold under reasonable assumptions.

4.3. The upwind triangle method of Tabata

The upwind triangle method of Tabata introduces, in a certain sense, upwind finite differences into the finite element method on unstructured grids [21]. Although originally formulated for the 2D case, the method can easily be generalized to higher dimensions. Hence, from now on we use the name *upwind simplex method*.

This method works as follows: The diffusion and reaction terms are discretized as in the finite element method, resulting in the matrices A_d, A_r, or A_d, $\tilde{A}_r$ if lumping is applied. The approximation of the convection term is based on a special lumping procedure. Each vertex x_i is associated with a simplex T_i, the *upwind simplex* or the *upwind triangle* in the 2D case, such that (i) x_i is a vertex of T_i and (ii), the vector $-b(x_i)$ points from x_i into T_i. If $b(x_i) = 0$, any simplex with vertex x_i can be chosen as upwind simplex T_i. Otherwise, if $-b(x_i)$ points into the direction of an edge, then the upwind simplex is chosen from the set of simplices sharing that edge.

Now suppose we have chosen exactly one upwind simplex T_i for every vertex x_i, $1 \leq i \leq N$. Then the convection term is approximated by (cf. (4.5))

$$\int_\Omega \varphi_i \, b \cdot \nabla u_h \, dx \;\approx\; \frac{|\,\Omega_i\,|}{n+1}\, b(x_i)\, \nabla u_h|_{T_i} \tag{4.16}$$

for $u_h \in V_h$ and $1 \leq i \leq N$. The resulting convection matrix is denoted by A_c^{ut}. Properties of A_c^{ut}, $A^{ut} = A_d + A_c^{ut} + A_r$, and $\tilde{A}^{ut} = A_d + A_c^{ut} + \tilde{A}_r$ are summarized in the following Lemma:

Lemma 4.4. *The matrices A_c^{ut}, A^{ut} and $\tilde{A}^{ut}$ have the following properties:*

(a) A_c^{ut} satisfies the weak sign condition,
(b) A_c^{ut} has the zero row sum property,
(c) A_c^{ut} is weakly diagonally dominant w.r.t. its rows,
(d) if $c + \nabla \cdot b \geq 0$ then $A^{ut}, \tilde{A}^{ut} \in SPD.M$.

Proof First note that on each simplex $T \in \mathcal{T}_h$ and for each vertex x_j of T the gradient $\nabla\varphi_j$ is orthogonal to the $(n-1)$-subsimplex opposite x_j. Further note that $\nabla\varphi_j$ points from there into the direction of x_j. This implies that for any vertex x_i with upwind simplex T_i we have $b(x_i) \cdot \nabla\varphi_i|_{T_i} \geq 0$, and for any other vertex x_j we have $b(x_i) \cdot \nabla\varphi_j|_{T_i} \leq 0$ if x_j is a vertex of T_i and $b(x_i) \cdot \nabla\varphi_j|_{T_i} = 0$ otherwise. Hence A_c^{ut} satisfies the weak sign condition.

Property (b) follows from the fact that the right hand side of (4.16) vanishes for constant functions u_h. Together (a) and (b) imply that A_c^{ut} is weakly diagonally dominant w.r.t. its rows. Using the Gerschgorin circle theorem we conclude that $\mathrm{Re}(\lambda) \geq 0$ holds for all $\lambda \in \sigma(A_c^{ut})$. Hence, if $c + \nabla \cdot b \geq 0$, it follows from Lemma 4.2.(g) and from the characterization of the matrix class $SPD.M$ in Remark 2.1. that $A^{ut}, \tilde{A}^{ut} \in SPD.M$ holds. ∎

Related to diagram (2.10) the result in Lemma 4.4.(d) is relevant. Note that in general A_c^{ut} does not have the zero column sum property, even not if b is constant. It is therefore not clear whether A_c^{ut} is positive semidefinite or if $A^{ut}, \tilde{A}^{ut} \in PD$ holds if $c + \nabla \cdot b \geq 0$.

Remark 4.3. If $\mathcal{T}_h$ is of weakly acute type, then using the results in Remark 4.1. one can prove the following:

- if $c + \nabla \cdot b = 0$ then $A^{ut} \in M$,
- if $c + \nabla \cdot b \geq 0$ then $\tilde{A}^{ut} \in M$.

4.4. The box method

A number of upwind schemes can be derived from a certain class of finite volume discretizations of (4.3). Finite volume methods applied to elliptic equations are often based on a *dual box mesh* constructed from a usual finite element triangulation $\mathcal{T}_h$. To be precise, let $\mathcal{T}_h$ be a consistent triangulation of Ω whose vertices $x_1, \ldots, x_{\hat{N}}$ are numbered such that $x_1, \ldots, x_N$ lie in the interior of Ω while $x_{N+1}, \ldots, x_{\hat{N}}$ belong to Γ (cf. Definition 4.2.). A *dual box mesh* for $\mathcal{T}_h$ is a partition $\mathcal{B}_h = \{B_1, \ldots, B_{\hat{N}}\}$ of Ω into $\hat{N}$ closed Lipschitz sets B_i such that $x_i \in B_i$ and $B_i \subset \Omega_i$ holds for $1 \leq i \leq \hat{N}$, cf. [6]. The sets B_i are called *boxes* and can be constructed in different ways. The two best known methods for the construction of dual box meshes are the *center-of-mass* method (see [6,11], for example) and the method of *perpendicular bisectors* (cf. [2]). In both cases the boxes B_i are polyhedra.

Now suppose that a consistent triangulation $\mathcal{T}_h$ and a corresponding dual box mesh $\mathcal{B}_h$ are given. Denoting again by V_h the space of continuous piecewise linear functions corresponding to $\mathcal{T}_h$ (!) which are zero on Γ, the simplest finite volume discretization of the continuous problem (4.3) reads: *Find* $u_h \in V_h$ *such that*

$$\int_{\partial B_i} \varepsilon \nabla u_h \cdot d\sigma + \int_{\partial B_i} u_h \, b \cdot d\sigma + \int_{B_i} c \, u_h \, dx = \int_{B_i} f \, dx \tag{4.17}$$

for $1 \leq i \leq N$. Note that the boxes $B_{N+1}, \ldots, B_{\hat{N}}$, corresponding to boundary vertices of $\mathcal{T}_h$, are not used in (4.17). Taking again the standard nodal basis $\Phi_h := \{\varphi_1, \ldots, \varphi_N\}$ of V_h, (4.17) results in a linear system with stiffness matrix A^{box} defined by

$$A_{ij}^{box} = -\int_{\partial B_i} \varepsilon \nabla \varphi_j \cdot d\sigma + \int_{\partial B_i} \varphi_j \, b \cdot d\sigma + \int_{B_i} c \, \varphi_j \, dx, \qquad 1 \leq i, j \leq N. \tag{4.18}$$

We use the splitting $A^{box} = A_d^{box} + A_c^{box} + A_r^{box}$ corresponding to the three integrals in (4.18). It is shown in [4,6,11] that the diffusion matrix A_d^{box} coincides with the diffusion matrix obtained by the finite element method: $A_d^{box} = A_d$.

In general, the reaction matrix A_r^{box} is non-symmetric and the sign of the non-zero entries in A_r^{box} is determined by the sign of c. If $c \geq 0$ then A_r^{box} contains only non-negative

entries. Note that in contrast to the finite element case the reaction term does not contain $\nabla \cdot b$. The reaction term is often discretized using the following lumping procedure:

$$\int_{B_i} c\, u_h \, dx \;\approx\; |\, B_i \,|\, c(x_i)\, u_h(x_i)\,, \qquad 1 \le i \le N\,. \tag{4.19}$$

If $c \ge 0$, this results in a non-negative diagonal matrix $\tilde{A}_r^{box}$. Hence, for the case with $b \equiv 0$ and $c \ge 0$, the matrix $\tilde{A}^{box} = A_d + \tilde{A}_r^{box}$ is symmetric positive definite, and if A_d is an M-matrix, then $\tilde{A}^{box} = A_d + \tilde{A}_r^{box}$ is an M-matrix, too.

The convection matrix A_c^{box} has the same skew-symmetric character as the convection matrix of the finite element method. Hence, in case of dominating convection the discretization (4.17) is in general unstable. The finite volume formulation, however, gives rise to a new class of upwind schemes. Such schemes have been presented, for example, by Bank et al. [2], and by Bey [6]. These methods only differ by the way in which the dual box mesh is constructed. In [2] the method of perpendicular bisectors is used, which in practice is restricted to the 2D case. In [6] the center-of-mass method is considered, which can be applied to triangulations of arbitrary dimension.

We now describe the basic idea of these upwind schemes without making any assumption on the construction of the dual boxmesh. Hence, the upwind scheme presented below has both the method in [2] and the one in [6] as special cases.

For $1 \le i \le \hat{N}$ let Λ_i be the set of indices $j \in \{1, \ldots, \hat{N}\}$ such that x_i, x_j are endpoints of a common edge in $\mathcal{T}_h$. For each $j \in \Lambda_i$ let $\Gamma_{ij} := \partial B_i \cap \partial B_j$ be the common boundary of B_i and B_j. With this notation the convection term in (4.17) can be represented as

$$\int_{\partial B_i} u_h\, b \cdot d\sigma \;=\; \sum_{j \in \Lambda_i} \int_{\Gamma_{ij}} u_h\, b \cdot d\sigma_i\,. \tag{4.20}$$

Here, the index i in $d\sigma_i$ indicates that the outer normal $\vec{n}$ used in the definition of the line integral $\int_{\Gamma_{ij}} u_h\, b \cdot d\sigma_i := \int_{\Gamma_{ij}} u_h\, b \cdot \vec{n}\, ds$ is the outer normal of the Box B_i. The total flux of the convection field b from box B_i into box B_j over the common boundary Γ_{ij} is given by the value

$$b_{ij} \;:=\; \int_{\Gamma_{ij}} b \cdot d\sigma_i\,. \tag{4.21}$$

Note that $b_{ij} = -b_{ji}$ for all $i \ne j$. Using the *upwind vertices*

$$x_{ij} \;:=\; \begin{cases} x_i & \text{if } b_{ij} \ge 0\,, \\ x_j & \text{if } b_{ij} < 0\,, \end{cases} \qquad i \ne j\,, \tag{4.22}$$

the integrals on the right hand side of (4.20) can be approximated by

$$\int_{\Gamma_{ij}} u_h\, b \cdot d\sigma_i \;\approx\; u_h(x_{ij}) \int_{\Gamma_{ij}} b \cdot d\sigma_i \;=\; u_h(x_{ij})\, b_{ij}\,, \qquad j \in \Lambda_i\,. \tag{4.23}$$

We denote the resulting convection matrix by A_c^{bup}. Some properties of the matrices A_c^{bup} and $\tilde{A}^{bup} = A_d + A_c^{bup} + \tilde{A}_r^{box}$ are summarized in the following Lemma:

Lemma 4.5. *The matrix A_c^{bup} has the following properties:*

(a) A_c^{bup} *satisfies the weak sign condition,*
(b) A_c^{bup} *has the zero column sum property,*
(c) A_c^{bup} *is weakly diagonally dominant w.r.t. its columns.*
(d) if $c \geq 0$ *then* $\tilde{A}^{bup} \in SPD.M$.

If $\nabla \cdot b = 0$ *then in addition we have*

(e) A_c^{bup} *has zero row sum property,*
(f) A_c^{bup} *is weakly diagonally dominant w.r.t. its rows,*
(g) A_c^{bup} *is positive semidefinite:* $A_c^{bup} + (A_c^{bup})^T \geq 0$,
(h) if $c \geq 0$ *then* $\tilde{A}^{bup} \in SPD.M_0$.

Proof The weak sign condition for A_c^{bup} follows from the construction, cf. (4.20) – (4.23). For $1 \leq j \leq N$ we have

$$\sum_i (A_c^{bup})_{ij} \;=\; (A_c^{bup})_{jj} + \sum_{i \in \Lambda_j} (A_c^{bup})_{ij} \;=\; \sum_{\substack{i \in \Lambda_j, \\ b_{ji} \geq 0}} b_{ji} + \sum_{\substack{i \in \Lambda_j, \\ b_{ij} < 0}} b_{ij} \;=\; \sum_{\substack{i \in \Lambda_j, \\ b_{ij} < 0}} (b_{ji} + b_{ij}) \;=\; 0\,.$$

Hence A_c^{bup} has the zero column sum property. Together (a) and (b) imply that that A_c^{bup} is weakly diagonally dominant w.r.t. its columns. Using the the same arguments as in the proof of Lemma 4.4. we obtain $\tilde{A}^{bup} \in SPD.M$. If in addition $\nabla \cdot b = 0$ then it follows from

$$\begin{aligned} \sum_j (A_c^{bup})_{ij} \;&=\; (A_c^{bup})_{ii} + \sum_{j \in \Lambda_i} (A_c^{bup})_{ij} \;=\; \sum_{\substack{j \in \Lambda_i, \\ b_{ij} \geq 0}} b_{ij} + \sum_{\substack{j \in \Lambda_i, \\ b_{ij} < 0}} b_{ij} \\ &=\; \sum_{i \in \Lambda_j} b_{ij} \;=\; \int_{\partial B_i} b \cdot d\sigma_i \;=\; 0\,, \end{aligned}$$

that A_c^{bup} has the zero row sum property, too. This together with (a) yields (f). Properties (c) and (f) imply that that the symmetric matrix $A_c^{bup} + (A_c^{bup})^T$ is weakly diagonally dominant. Using the Gerschgorin cycle theorem we conclude that the eigenvalues of $A_c^{bup} + (A_c^{bup})^T$ are nonnegative. Hence we have $A_c^{bup} + (A_c^{bup})^T \geq 0$. This in combination with the characterization of the matrix class $SPD.M_0$ given in Remark 2.1. proves (h). ■

Remark 4.4. If $\mathcal{T}_h$ is of weakly acute type, then using the results in Remark 4.1. one can prove that if $c \geq 0$ then $\tilde{A}^{bup} \in M$ holds, or even $\tilde{A}^{bup} \in M_0$ if $\nabla \cdot b = 0$ and $c \geq 0$.

5. Concluding remarks

In this paper we obtained a satisfactory contraction result for the Jacobi method applied to matrices from the class $SPD.M_0$ (Theorem. 3.1.). On the other hand, for the Gauss-Seidel method only negative results are presented (Example 3.1.).For the class of positive definite matrices a hybrid method, which converges without any damping, is introduced (Theorem 3.3.). Furthermore, a few well-known finite element and finite volume discretization

methods are analyzed with respect to algebraic properties of the resulting stiffness matrices.

In our opinion, there are still quite a few interesting open problems in this field. Here we mention two of these. The convergence analysis of the Gauss-Seidel method is an interesting topic for further research. In the literature we did not find convergence results for the Gauss-Seidel method applied to matrices in $PD \setminus M_0$. A few results for the SOR method are known ([16]). These results, however, are comparable to the result in Theorem 3.2. (but now for the SOR instead of the Jacobi method) and hence not very satisfactory when applied to discrete convection-diffusion problems (cf. Example 3.4.).

A second question which seems to be of interest is whether one can define a suitable subclass of PD (different from $SPD.M_0$) which contains the stiffness matrices resulting from popular discretization methods for convection-diffusion equations (e.g. SDFEM, Tabata-scheme, box-scheme) and also allows a satisfactory convergence analysis of Jacobi and Gauss-Seidel type of methods. Maybe, in the definition of such a subclass, a zero row sum or column sum property as explained in Section 4.1. will play a role.

REFERENCES

1. O. Axelsson. *Iterative Solution Methods*. Cambridge University Press, NY, 1994.
2. R. E. Bank, J. F. Bürgler, W. Fichtner, and R. K. Smith. Some upwinding techniques for finite element approximations of convection-diffusion equations. *Numer. Math.*, 58:185–202, 1990.
3. R. E. Bank and T. F. Chan. A composite step biconjugate gradient method. *Numer. Math.*, 66:295–319, 1994.
4. R. E. Bank and D. J. Rose. Some error estimates for the box method. *SIAM J. Numer. Anal.*, 24(4):777–787, 1987.
5. A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. Academic Press, NY, 1979.
6. J. Bey. *Finite-Volumen- und Mehrgitterverfahren für elliptische Randwertprobleme*. Advances in Numerical Mathematics. B. G. Teubner, Stuttgart, Leipzig, 1998.
7. J. Bey and G. Wittum. Downwind numbering: Robust multigrid for convection-diffusion problems. *Appl. Numer. Math.*, 23:177–192, 1997.
8. P. G. Ciarlet. Basic error estimates for elliptic problems. In P. G. Ciarlet and J. L. Lions, editors, *Handbook of Numerical Analysis, Volume II: Finite Element Methods (Part 1)*. North Holland, Amsterdam, 1991.
9. D. Gilbarg and N. S. Trudinger. *Elliptic Partial Differential Equations of Second Order*, volume 224 of *A Series of Comprehensive Studies in Mathematics*. Springer, Berlin, Heidelberg, 1977.
10. M. Goldberg, E. Tadmor, and G. Zwas. Numerical radius for positive matrices. *Linear Algebra Appl.*, 12:209–214, 1975.
11. W. Hackbusch. On first and second order box schemes. *Computing*, 41:277–296, 1989.
12. W. Hackbusch. *Elliptic Differential Equations: Theory and Numerical Treatment*, volume 18 of *Springer Series in Computational Mathematics*. Springer, Berlin, 1992.
13. W. Hackbusch. *Iterative solution of large sparse systems of equations*, volume 95 of *Applied Mathematical Sciences*. Springer, New York, 1994.
14. W. Hackbusch and T. Probst. Downwind Gauß-Seidel smoothing for convection dominated problems. *Numerical Linear Algebra with Applications*, 4(2):85–102, 1997.
15. R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, New York, 1991.
16. W. Niethammer. Relaxation bei nichtsymmetrischen Matrizen. *Math. Zeitschr.*, 85:319–327, 1964.
17. H.-G. Roos, M. Stynes, and L. Tobiska. *Numerical Methods for Singularly Perturbed Differential Equations. Convection-Diffusion and Flow Problems*, volume 24 of *Springer Series in Computational Mathematics*. Springer, Berlin, Heidelberg, 1996.

18. Y. Saad. *Iterative Methods for Sparse Linear Systems*. PWS Publishing Company, London, 1996.
19. A. A. Samarskij and E. S. Nikolaev. *Numerical Methods for Grid Equations. Vol. II: Iterative Methods*. Birkhäuser, Basel, 1989.
20. D. Sang, J. Zhang, and S. Zhang. Convergence proof of Jacobi iterative method for a discretized 2d convection-diffusion equation. Technical Report 276–98, Department of Computer Science, University of Kentucky, Lexington, 1998.
21. M. Tabata. A finite element approximation corresponding to the upwind differencing. *Memoirs of Numerical Mathematics*, 1:47–63, 1977.
22. R. S. Varga. *Matrix Iterative Analysis*. Prentice Hall, Englewood Cliffs, New Jersey, 1962.
23. D. M. Young. *Iterative Solutions of Large Linear Systems*. Academic Press, NY, 1971.