Numer. Math. (2007) 106:255–288
DOI 10.1007/s00211-006-0060-z

Numerische Mathematik

Monotonicity of control volume methods

J. M. Nordbotten · I. Aavatsmark · G. T. Eigestad

Received: 13 December 2005 / Revised: 11 July 2006 /
Published online: 11 January 2007

Abstract Robustness of numerical methods for multiphase flow problems in porous media is important for development of methods to be used in a wide range of applications. Here, we discuss monotonicity for a simplified problem of single-phase flow, but where the simulation grids and media are allowed to be general, posing challenges to control-volume methods. We discuss discrete formulations of the maximum principle and derive sufficient criteria for discrete monotonicity for arbitrary nine-point control-volume discretizations for conforming quadrilateral grids in 2D. These criteria are less restrictive than the M-matrix property. It is shown that it is impossible to construct nine-point methods which unconditionally satisfy the monotonicity criteria when the discretization satisfies local conservation and exact reproduction of linear potential fields. Numerical examples are presented which show the validity of the criteria for monotonicity. Further, the impact of nonmonotonicity is studied. Different behavior for different discretization methods is illuminated, and simple ideas are presented for improvement in terms of monotonicity.

Mathematics Subject Classification (1991) 65N12 · 65N06 · 76S05 · 35R05

J. M. Nordbotten · G. T. Eigestad
Department of Mathematics, University of Bergen, Bergen, Norway
e-mail: janmn@mi.uib.no

G. T. Eigestad
e-mail: geirte@mi.uib.no

I. Aavatsmark (✉)
Centre for Integrated Petroleum Research, University of Bergen, Bergen, Norway
e-mail: ivar.aavatsmark@cipr.uib.no

1 Introduction

Control-volume discretizations on two-dimensional quadrilateral grids are discussed for linear second-order elliptic equations on heterogeneous media. Convergence of such methods are investigated and proved in e.g. [4–6,15–18,24,25,33,34]. This paper is concerned with undesirable oscillations in the discrete solution, and how they can be avoided.

The monotonicity of discrete solutions has a history of research starting with the observation that most spatial finite difference discretizations yield M-matrices [9,11,21]. This guarantees the monotonicity of the solution. For discretizations not leading to M-matrices, far less is known. An early analysis in one dimension was conducted by Bramble and Hubbard [10], and Varga [32] analyzed a weaker form of the maximum principle for more general problems.

Our applications are solution of multiphase flow equations in subsurface flow [19,30]. These equations contain an elliptic operator similar to the one occurring in our simple model equation. The equations have properties which constrain the choice of grid and discretization technique used for the elliptic operator. For multiphase flow, some variables (saturations) behave like solutions of hyperbolic equations, while the pressure behaves like a solution of an elliptic equation. Phase transitions which are strongly pressure dependent may occur.

Due to the hyperbolicity and the strongly nonlinear saturation-dependent terms, the discretization scheme should be locally conservative. In a fully implicit formulation, an explicit expression for the flux is required. This motivates the application of control-volume formulations with flux expressions honoring the heterogeneities of the medium.

Unphysical oscillations in the discrete solution are undesirable in any application, but for multiphase flow, such oscillations may have serious consequences. If the computed pressure lies below the bubble-point curve of the mixture, while the actual pressure lies above it, artificial gas may be liberated, yielding a strongly diverging solution. Therefore, an oscillation-free solution of the elliptic model equation is desired.

Our treatment begins with the relationship between the maximum principle and monotonicity. Then monotone matrices are discussed, and local criteria for the monotonicity are derived. The discussion is general, with application to any conservative nine-point scheme for the second-order elliptic operator in two dimensions. This includes the discretizations discussed in [1–6,13–15,18,24,25,34].

In the case of homogeneous medium and uniform parallelogram grid, we use the derived local criteria to determine sufficient conditions for monotone control-volume schemes. Finally, the local criteria are tested numerically, both for homogeneous and heterogeneous media.

2 Maximum principle and monotonicity

We consider an operator L, defined by

$$Lu = -\operatorname{div}(\boldsymbol{K} \operatorname{grad} u), \tag{1}$$

where the tensor $\boldsymbol{K}$ is a symmetric positive definite matrix. For a source term q, let

$$Lu = q \tag{2}$$

in some (open) domain D. Suppose that in the domain D, the source term is nonnegative, $q \geq 0$, and $\boldsymbol{K}$ is sufficiently smooth. Then u has no minimum in D. More precisely, if there is a point $\boldsymbol{x}_0 \in D$ such that $u(\boldsymbol{x}_0) \leq u(\boldsymbol{x})$ for all other $\boldsymbol{x} \in D$, then u is constant in D. This version of the maximum principle is known as Hopf's lemma [22,31] and proved in [23] under the condition that $\boldsymbol{K}$ is continuously differentiable. We shall make use of a slightly weaker form of this principle: If $q \geq 0$ in D, there is no point $\boldsymbol{x}_0 \in D$ such that $u(\boldsymbol{x}_0) < u(\boldsymbol{x})$ for all other $\boldsymbol{x} \in D$. Obviously, since Hopf's lemma can be stated for any subdomain in D, this also means that if $q \geq 0$ in D, u can have no local minima in D. We formulate this property as

Property MC If $Lu = q \geq 0$ in D, there is no point $\boldsymbol{x}_0 \in D$ such that $u(\boldsymbol{x}_0) < u(\boldsymbol{x})$ for all other points $\boldsymbol{x}$ in a neighborhood of $\boldsymbol{x}_0$.

Suppose that on some domain $\Omega \subset D$ the potential u satisfies Eq. (2) with homogeneous Dirichlet boundary conditions

$$u = 0 \quad \text{on } \partial\Omega. \tag{3}$$

If the tensor $\boldsymbol{K}$ and the boundary $\partial\Omega$ are sufficiently smooth, the solution is given by

$$u(\boldsymbol{x}) = \int_{\Omega} G_{\Omega}(\boldsymbol{x}, \boldsymbol{\xi}) q(\boldsymbol{\xi})\, d\tau_{\boldsymbol{\xi}}, \tag{4}$$

where $G_{\Omega}(\boldsymbol{x}, \boldsymbol{\xi})$ is Green's function for the given boundary value problem, i.e., $G_{\Omega}(\boldsymbol{x}, \boldsymbol{\xi})$ is the solution of (2), (3) on Ω with $q(\boldsymbol{x}) = \delta(\boldsymbol{x} - \boldsymbol{\xi})$. Applying Green's formula with delta functionals as source terms, it follows that Green's function is symmetric, $G_{\Omega}(\boldsymbol{x}, \boldsymbol{\xi}) = G_{\Omega}(\boldsymbol{\xi}, \boldsymbol{x})$. Below, we assume that $\boldsymbol{K}$ and $\partial\Omega$ are sufficiently smooth to make Green's function $G_{\Omega}(\boldsymbol{x}, \boldsymbol{\xi})$ continuous at all points but $\boldsymbol{\xi}$. Then the following inequality holds:

$$G_{\Omega}(\boldsymbol{x}, \boldsymbol{\xi}) \geq 0 \quad \text{for } \boldsymbol{x}, \boldsymbol{\xi} \in \Omega. \tag{5}$$

Inequality (5) and its significance follow from Theorem 1 below. An immediate consequence of (4) and (5) is that

$$q \geq 0 \quad \Rightarrow \quad u \geq 0 \quad \text{in } \Omega. \tag{6}$$

Utilizing (4), it is straight foreward to prove

Theorem 1 *Property MC holds if and only if inequality* (5) *holds for all* $\Omega \subset D$, *where* $G_{\Omega}(\boldsymbol{x}, \boldsymbol{\xi})$ *is Green's function with homogeneous Dirichlet boundary conditions on* $\partial\Omega$.

When discretizing the operator (1), one would like the discrete operator to satisfy a discrete version of the maximum principle. However, it is not obvious how the discrete version should be formulated. In the formulation below, we will make use of the monotonicity property (6).

Suppose that the discretization of (2) with homogeneous Dirichlet boundary conditions on D leads to a system of equations

$$\boldsymbol{A}\boldsymbol{u} = \boldsymbol{q}, \tag{7}$$

where each component of $\boldsymbol{u}$ is the value of u at a grid point, and each component of $\boldsymbol{q}$ is the source term integrated over the associated grid cell. Let $\boldsymbol{O}$ be the zero matrix. If

$$\boldsymbol{A}^{-1} \geq \boldsymbol{O}, \tag{8}$$

i.e., if each element of $\boldsymbol{A}^{-1}$ is nonnegative, then the discrete system satisfies the same monotonicity property as the continuous system:

$$\boldsymbol{q} \geq \boldsymbol{0} \quad \Rightarrow \quad \boldsymbol{u} \geq \boldsymbol{0}. \tag{9}$$

A matrix $\boldsymbol{A}^{-1}$ with the sign property (8) is called *monotone* [12]. Note that the matrix $\boldsymbol{A}^{-1}$ is a discrete version of the integral operator which uses Green's function as kernel.

However, while the sign property (8) excludes negative solutions for nonnegative source terms and homogeneous Dirichlet boundary conditions, it does not exclude local minima. As Theorem 1 shows, a natural discrete maximum principle is achieved by requiring that the monotonicity property (8) holds when $\boldsymbol{A}$ is constructed for any subset of the grid points. Therefore, our discrete maximum principle will be:

Property MD For a given grid in D, let (7) be a discretization of (2) with homogeneous Dirichlet boundary conditions on any subgrid. Then the matrix of coefficients $\boldsymbol{A}$ of the subgrid boundary value problem must satisfy the monotonicity property (8).

Remark 1 The discrete maximum principle Property MD builds on Theorem 1 which was derived under sufficient smoothness conditions on $\boldsymbol{K}$ and $\partial\Omega$. However, Property MD should be valid for any $\boldsymbol{K}$ which is acceptable for the discretization (in practice, any piecewise constant $\boldsymbol{K}$).

Remark 2 Different formulations of the discrete maximum principle are possible since they can be derived from different formulations of the continuous maximum principle. One example is the study by Varga [32], where the method is said to satisfy a maximum principle for problems where the source term $q = 0$ and $\|u\|_\infty$ on D is less than or equal to $\|u\|_\infty$ on ∂D. Another discrete maximum principle follows from requiring $\boldsymbol{A}^{-1} \geq \boldsymbol{O}$ for the given grid in D. Both these discrete maximum principles are weaker than Property MD and allow for oscillating solutions.

The definition of the discrete maximum principle leads to our definition of a

Monotone Method A method which defines a discretization satisfying Property MD is said to be monotone.

3 Monotone matrices

An M-matrix is defined as a nonsingular matrix $\boldsymbol{A} = \{a_{i,j}\}$ whose off-diagonal elements are nonpositive ($a_{i,j} \leq 0$ for $i \neq j$) and whose inverse is monotone ($\boldsymbol{A}^{-1} \geq \boldsymbol{O}$) [11,21]. The matrices we consider do in general have some positive off-diagonal elements, and therefore we need to consider a wider class of matrices.

In this section, we generalize the theory of M-matrices to a class suitable for our purposes. More precisely, we show that a matrix has a monotone inverse, subject to the properties of a splitting of the matrix. This will be valuable, since we can analyze a particular splitting in the next section and see that such a splitting leads to local monotonicity criteria.

A splitting $\boldsymbol{A} = \boldsymbol{B} - \boldsymbol{C}$ is called *weakly regular* if $\boldsymbol{B}$ is nonsingular, $\boldsymbol{B}^{-1} \geq \boldsymbol{O}$, and $\boldsymbol{B}^{-1}\boldsymbol{C} \geq \boldsymbol{O}$ [21,29]. A matrix $\boldsymbol{A}$ with a weakly regular splitting $\boldsymbol{A} = \boldsymbol{B} - \boldsymbol{C}$ has a monotone inverse, $\boldsymbol{A}^{-1} \geq \boldsymbol{O}$, if and only if the spectral radius ρ of $\boldsymbol{B}^{-1}\boldsymbol{C}$ satisfies the inequality

$$\rho(\boldsymbol{B}^{-1}\boldsymbol{C}) < 1. \tag{10}$$

When $\boldsymbol{A}$ stems from a discretization which yields nonpositive off-diagonal elements, the matrix $\boldsymbol{A}$ is often irreducibly diagonally dominant. This property can be utilized to demonstrate the inequality (10). Matrices whose off-diagonal elements have different signs, cannot be expected to be diagonally dominant. However, for matrices with a weakly regular splitting, the inequality

$$\forall i: \quad \sum_j a_{i,j} \geq 0 \tag{11}$$

can be utilized to prove (10).

Theorem 2 *Suppose that the matrix* $\boldsymbol{A}$ *has a weakly regular splitting* $\boldsymbol{A} = \boldsymbol{B} - \boldsymbol{C}$, *and suppose that inequality* (11) *holds. Then* $\left\|\boldsymbol{B}^{-1}\boldsymbol{C}\right\|_\infty \leq 1$. *Assume, in addition, that either the inequality* (11) *is strict for all i or* $\boldsymbol{B}^{-1}\boldsymbol{C}$ *is irreducible and the inequality* (11) *is strict for at least one i. Then the inequality* (10) *holds.*

Proof Let $\boldsymbol{e} = [1, \ldots, 1]^{\mathrm{T}}$. Then inequality (11) can be written $\boldsymbol{A}\boldsymbol{e} \geq \boldsymbol{0}$. Thus,

$$\boldsymbol{C}\boldsymbol{e} = \boldsymbol{B}\boldsymbol{e} - \boldsymbol{A}\boldsymbol{e} \leq \boldsymbol{B}\boldsymbol{e}. \tag{12}$$

Hence, utilizing $\boldsymbol{B}^{-1} \geq \boldsymbol{O}$,

$$\boldsymbol{0} \leq \boldsymbol{B}^{-1}\boldsymbol{C}\boldsymbol{e} \leq \boldsymbol{B}^{-1}\boldsymbol{B}\boldsymbol{e} = \boldsymbol{e}. \tag{13}$$

Hence, it follows that $\left\|\boldsymbol{B}^{-1}\boldsymbol{C}\right\|_\infty \leq 1$.

If inequality (11) is strict for all i, then $\left\|\boldsymbol{B}^{-1}\boldsymbol{C}\right\|_\infty < 1$ and the theorem is proved. To prove the rest of the second part of the theorem, assume that $\sum_j a_{i,j} > 0$ for $i = k$, i.e., $[\boldsymbol{A}\boldsymbol{e}]_k > 0$. Then

$$[\boldsymbol{C}\boldsymbol{e}]_k = [\boldsymbol{B}\boldsymbol{e}]_k - [\boldsymbol{A}\boldsymbol{e}]_k < [\boldsymbol{B}\boldsymbol{e}]_k. \tag{14}$$

Since $\boldsymbol{B}^{-1}$ is nonsingular, the k-th column in $\boldsymbol{B}^{-1}$ must have at least one positive element. Let us assume that $[\boldsymbol{B}^{-1}]_{l,k} > 0$. Then

$$[\boldsymbol{B}^{-1}\boldsymbol{C}\boldsymbol{e}]_l < [\boldsymbol{B}^{-1}\boldsymbol{B}\boldsymbol{e}]_l = 1. \tag{15}$$

The case when $\left\|\boldsymbol{B}^{-1}\boldsymbol{C}\right\|_\infty < 1$ for all i has already been considered, so assume that $\max_i[\boldsymbol{B}^{-1}\boldsymbol{C}\boldsymbol{e}]_i = 1$. With the assumption that $\boldsymbol{B}^{-1}\boldsymbol{C}$ is irreducible, we can apply the Perron-Frobenius theorem [8,11] on the matrix $\boldsymbol{B}^{-1}\boldsymbol{C}$. It follows that the spectral radius $\rho(\boldsymbol{B}^{-1}\boldsymbol{C})$ satisfies (10). This proves the theorem. □

Remark 3 Choosing $\boldsymbol{B}$ as the diagonal part of $\boldsymbol{A}$, Theorem 2 gives sufficient conditions for $\boldsymbol{A}$ to be an M-matrix.

If the transpose of a matrix $\boldsymbol{A}$ has a weakly regular splitting $\boldsymbol{A}^{\mathrm{T}} = \boldsymbol{B}^{\mathrm{T}} - \boldsymbol{C}^{\mathrm{T}}$, then $\boldsymbol{B}^{-\mathrm{T}} \geq \boldsymbol{O}$ and $\boldsymbol{B}^{-\mathrm{T}}\boldsymbol{C}^{\mathrm{T}} \geq \boldsymbol{O}$. Hence, $\boldsymbol{B}^{-1} \geq \boldsymbol{O}$ and $\boldsymbol{C}\boldsymbol{B}^{-1} \geq \boldsymbol{O}$. With the inequality

$$\forall j: \quad \sum_i a_{i,j} \geq 0, \tag{16}$$

the theory of weakly regular splittings now yields the

Corollary 1 *Suppose that the matrix $\boldsymbol{A}$ has a splitting $\boldsymbol{A} = \boldsymbol{B} - \boldsymbol{C}$, such that $\boldsymbol{B}$ is nonsingular, $\boldsymbol{B}^{-1} \geq \boldsymbol{O}$, and $\boldsymbol{C}\boldsymbol{B}^{-1} \geq \boldsymbol{O}$, and suppose that inequality (16) holds. Assume further that either the inequality (16) is strict for all j or $\boldsymbol{C}\boldsymbol{B}^{-1}$ is irreducible and the inequality (16) is strict for at least one j. Then $\boldsymbol{A}$ is nonsingular with a monotone inverse, i.e., $\boldsymbol{A}^{-1} \geq \boldsymbol{O}$.*

4 Monotonicity criteria for general quadrilateral grids

In this section, we apply the theory of Sect. 3 to determine conditions under which a control-volume discretization of the operator (1) fulfills Property MD. The discretizations which we consider are nine-point stencils on a two-dimensional quadrilateral grid.

The global numbering of the cells is given by the indices (i, j), where i is the column number and j the row number in the grid. The local cell numbering in a stencil is shown in Fig. 1. The elements in the stencil of cell (i, j) are denoted by $m_k^{i,j}$, $k = 1, \ldots, 9$, where k is the local index of Fig. 1. The cell stencil approximates the integral of the elliptic expression (1) over the cell (i, j), i.e., the outflux

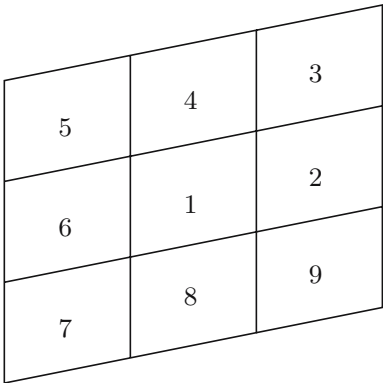

Fig. 1 Local cell numbering in a nine-point stencil

out of cell (i,j),

$$-\int_{\Omega(i,j)} \operatorname{div}(\boldsymbol{K} \operatorname{grad} u)\, d\Omega \approx \sum_{k=1}^{9} m_k^{i,j} u_k. \tag{17}$$

Applying homogeneous Dirichlet boundary conditions for the domain $\Omega = \bigcup_{i,j} \Omega(i,j)$, we obtain a linear system of equations of the form (7), where the matrix of coefficients $\boldsymbol{A}$ is irreducible. Each equation of the system (7) has the form

$$\sum_{k=1}^{9} m_k^{i,j} u_k = \int_{\Omega(i,j)} q\, d\Omega. \tag{18}$$

If the potential values u_k are constant, there should be no flow. Therefore, for each cell (i,j) not at the boundary, the coefficients $m_k^{i,j}$ must satisfy

$$\sum_{k=1}^{9} m_k^{i,j} = 0. \tag{19}$$

Hence, the elements of the matrix $\boldsymbol{A}$ satisfy (11) as an equality for all inner cells. If we assume that the inequality (11) is fulfilled for the boundary cells, we may use the theory of weakly regular splittings to establish conditions which ensure monotonicity of $\boldsymbol{A}^{-1}$.

In a control-volume method, the coefficients $m_k^{i,j}$ are constructed such that the flux across an interface is the same for both two cells sharing this interface, and the outflux out of a cell equals exactly the source term of this cell. Therefore, control-volume methods are locally conservative. For conservative methods, the sum of the source terms is zero whenever the outflux out of the domain is zero. In the following, we only consider control-volume methods where the flux across an interface is zero if all the cells sharing a corner with this interface have the same potential values. For example, in Fig. 1, the flux across the interface between cell 1 and cell 2 is zero if the cells 1, 2, 3, 4, 8, and 9 have the same potential values.

Now consider the system of Eq. (7). Let $\boldsymbol{u} = \boldsymbol{e}_j$, where $\boldsymbol{e}_j$ is the unit vector with the jth component equal to 1 and all other components equal to 0. Then $\boldsymbol{A}\boldsymbol{e}_j$ is the jth column of $\boldsymbol{A}$. If the index j corresponds to an inner cell in the grid, then the potential is zero in the boundary cells as well as on the boundary, due to the homogenous Dirichlet conditions. Therefore, the flux across the boundary is zero, and hence, the sum of the source terms must vanish. It follows that the sum of the elements of the jth column of $\boldsymbol{A}$ is zero. Hence, in a conservative method the elements of the matrix $\boldsymbol{A}$ satisfy (16) as an equality for all inner cells. Assuming that (16) is fulfilled for the boundary cells, we may use Corollary 1 to establish conditions which ensure monotonicity of $\boldsymbol{A}^{-1}$.

For any splitting $\boldsymbol{A} = \boldsymbol{B} - \boldsymbol{C}$, we may now establish monotonicity of $\boldsymbol{A}^{-1}$ in two ways: either by determining the conditions under which $\boldsymbol{B}^{-1} \geq \boldsymbol{O}$ and $\boldsymbol{B}^{-1}\boldsymbol{C} \geq \boldsymbol{O}$, or by determining the conditions under which $\boldsymbol{B}^{-1} \geq \boldsymbol{O}$ and $\boldsymbol{C}\boldsymbol{B}^{-1} \geq \boldsymbol{O}$. In the following, we will apply the latter approach. The question of irreducibility of $\boldsymbol{C}\boldsymbol{B}^{-1}$ is discussed in Remark 4.

Different splittings may lead to different sets of monotonicity conditions. However, since the set of conditions derived for each splitting is only sufficient, we may pick the weakest set of conditions.

Since the elements $m_k^{i,j}, k \geq 2$, in general have different signs, we cannot apply a splitting where $\boldsymbol{B}$ consists of the diagonal part of $\boldsymbol{A}$. This would in fact be equivalent to requiring that $\boldsymbol{A}$ is an M-matrix, which is too restrictive [1].

In [26,28] a splitting was chosen which yields remarkably good monotonicity conditions. This splitting will be discussed in the section below.

4.1 Local conditions for monotonicity of $\boldsymbol{A}^{-1}$

We choose the natural ordering of the unknowns, first counting the unknowns in column i of row 1, and then proceeding by counting the unknowns of each subsequent row j. In a grid with m columns and n rows, the matrix $\boldsymbol{A}$ then has a block-tridiagonal structure (see Fig. 2), each block being an $m \times m$ tridiagonal matrix. $\boldsymbol{A}$ has n diagonal blocks and $n - 1$ blocks in the upper and lower block diagonal.

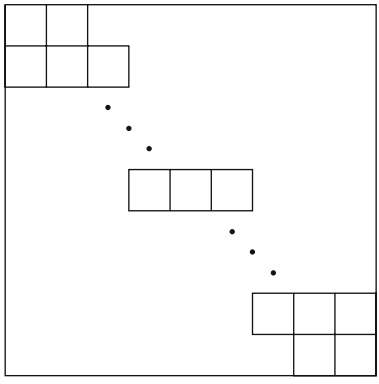

Fig. 2 Block-tridiagonal structure

Let $\boldsymbol{B}$ consist of the diagonal blocks of $\boldsymbol{A}$, and let $\boldsymbol{C} = \boldsymbol{B} - \boldsymbol{A}$. The diagonal blocks of $\boldsymbol{B}$ will be denoted $\boldsymbol{B}_j, j = 1, \ldots, n$. The blocks of the first lower block diagonal of $\boldsymbol{C}$ will be denoted $\boldsymbol{C}_j^L, j = 2, \ldots, n$, while the blocks of the first upper block diagonal of $\boldsymbol{C}$ will be denoted $\boldsymbol{C}_j^U, j = 1, \ldots, n-1$. The matrices $\boldsymbol{B}_j, \boldsymbol{C}_j^L$ and $\boldsymbol{C}_j^U$ are tridiagonal. These matrices are shown by displaying the nonzero elements around the ith diagonal element:

$$\boldsymbol{B}_j = \begin{bmatrix} \ddots & m_2^{i-1,j} & \\ m_6^{i,j} & m_1^{i,j} & m_2^{i,j} \\ & m_6^{i+1,j} & \ddots \end{bmatrix}, \tag{20}$$

$$\boldsymbol{C}_j^L = -\begin{bmatrix} \ddots & m_9^{i-1,j} & \\ m_7^{i,j} & m_8^{i,j} & m_9^{i,j} \\ & m_7^{i+1,j} & \ddots \end{bmatrix}, \tag{21}$$

$$\boldsymbol{C}_j^U = -\begin{bmatrix} \ddots & m_3^{i-1,j} & \\ m_5^{i,j} & m_4^{i,j} & m_3^{i,j} \\ & m_5^{i+1,j} & \ddots \end{bmatrix}.$$

We now derive conditions which ensure that $\boldsymbol{B}^{-1} \geq \boldsymbol{O}$ and $\boldsymbol{C}\boldsymbol{B}^{-1} \geq \boldsymbol{O}$.

Consider the matrix $\boldsymbol{B}_j$ given by (20). Using the diagonal to split this matrix, Remark 3 shows that $\boldsymbol{B}_j^{-1} \geq \boldsymbol{O}$ if

$$\begin{aligned} &\text{(A0):} && m_1^{i,j} > 0, \\ &\text{(A1a):} && m_2^{i,j} < 0, \\ &\text{(A1c):} && m_6^{i,j} < 0, \\ &\text{(A2):} && m_1^{i,j} + m_2^{i,j} + m_6^{i,j} > 0. \end{aligned}$$

In fact, these conditions ensure that $\boldsymbol{B}_j$, and thereby $\boldsymbol{B}$, are M-matrices. Also, these conditions imply that $\boldsymbol{B}_j$ is irreducible, and hence, $\boldsymbol{B}_j^{-1} > \boldsymbol{O}$ [20,21].

To derive the conditions which also guarantee that $\boldsymbol{C}\boldsymbol{B}^{-1} \geq \boldsymbol{O}$, we introduce the matrices

$$\boldsymbol{D}_j = \boldsymbol{B}_j^{-1}, \quad \boldsymbol{E}_{j+1}^L = \boldsymbol{C}_{j+1}^L \boldsymbol{D}_j, \quad \boldsymbol{E}_{j-1}^U = \boldsymbol{C}_{j-1}^U \boldsymbol{D}_j. \tag{22}$$

The matrices $\boldsymbol{E}_j^L$ and $\boldsymbol{E}_j^U$ are the nonzero blocks of $\boldsymbol{C}\boldsymbol{B}^{-1}$. Thus, we have to derive conditions which ensure that $\boldsymbol{E}_j^L \geq \boldsymbol{O}$ and $\boldsymbol{E}_j^U \geq \boldsymbol{O}$. Let $\boldsymbol{D}_j = \{d_{i,k}^j\}$,

$\boldsymbol{E}_j^L = \{e_{i,k}^{j,L}\}$, and $\boldsymbol{E}_j^U = \{e_{i,k}^{j,U}\}$. From the equation $\boldsymbol{B}_j\boldsymbol{D}_j = \boldsymbol{I}$, i.e., from

$$m_6^{i,j} d_{i-1,k}^j + m_1^{i,j} d_{i,k}^j + m_2^{i,j} d_{i+1,k}^j = \delta_{i,k}, \tag{23}$$

it follows that

$$d_{i,k}^j = \frac{\delta_{i,k}}{m_1^{i,j}} - \frac{m_6^{i,j}}{m_1^{i,j}} d_{i-1,k}^j - \frac{m_2^{i,j}}{m_1^{i,j}} d_{i+1,k}^j. \tag{24}$$

The definition of $\boldsymbol{E}_{j+1}^L$ and Eq. (24) yield

$$\begin{aligned} e_{i,k}^{j+1,L} &= -m_7^{i,j+1} d_{i-1,k}^j - m_8^{i,j+1} d_{i,k}^j - m_9^{i,j+1} d_{i+1,k}^j \\ &= -\frac{m_8^{i,j+1}}{m_1^{i,j}} \delta_{i,k} + \left(\frac{m_6^{i,j}}{m_1^{i,j}} m_8^{i,j+1} - m_7^{i,j+1} \right) d_{i-1,k}^j \\ &\quad + \left(\frac{m_2^{i,j}}{m_1^{i,j}} m_8^{i,j+1} - m_9^{i,j+1} \right) d_{i+1,k}^j. \end{aligned} \tag{25}$$

Likewise, the definition of $\boldsymbol{E}_{j-1}^U$ and Eq. (24) yield

$$\begin{aligned} e_{i,k}^{j-1,U} &= -m_5^{i,j-1} d_{i-1,k}^j - m_4^{i,j-1} d_{i,k}^j - m_3^{i,j-1} d_{i+1,k}^j \\ &= -\frac{m_4^{i,j-1}}{m_1^{i,j}} \delta_{i,k} + \left(\frac{m_6^{i,j}}{m_1^{i,j}} m_4^{i,j-1} - m_5^{i,j-1} \right) d_{i-1,k}^j \\ &\quad + \left(\frac{m_2^{i,j}}{m_1^{i,j}} m_4^{i,j-1} - m_3^{i,j-1} \right) d_{i+1,k}^j. \end{aligned} \tag{26}$$

Since $d_{i,k}^j \geq 0$, it follows that $e_{i,k}^{j,L} \geq 0$ and $e_{i,k}^{j,U} \geq 0$ if all the terms in the expressions (25) and (26) are nonnegative. Hence, $\boldsymbol{C}\boldsymbol{B}^{-1} \geq \boldsymbol{O}$ if

$$\begin{aligned} &\text{(A1b):} && m_4^{i,j} < 0, \\ &\text{(A1d):} && m_8^{i,j} < 0, \\ &\text{(A3a):} && m_2^{i,j} m_4^{i,j-1} - m_3^{i,j-1} m_1^{i,j} > 0, \\ &\text{(A3b):} && m_6^{i,j} m_4^{i,j-1} - m_5^{i,j-1} m_1^{i,j} > 0, \\ &\text{(A3c):} && m_2^{i,j} m_8^{i,j+1} - m_9^{i,j+1} m_1^{i,j} > 0, \\ &\text{(A3d):} && m_6^{i,j} m_8^{i,j+1} - m_7^{i,j+1} m_1^{i,j} > 0, \end{aligned}$$

where the inequalities in the conditions A3 are omitted whenever $j \pm 1$ lies outside the index domain. The above arguments imply

Lemma 1 *The inverse of the matrix* $\boldsymbol{A}$ *arising from a locally conservative nine-point discretization in 2D is monotone if conditions A0 through A3 defined above hold for all pairs* (i, j).

Remark 4 To achieve monotonicity for $\boldsymbol{A}^{-1}$, Corollary 1 requires that $\boldsymbol{C}\boldsymbol{B}^{-1}$ is irreducible. In general, $\boldsymbol{A}$ is irreducible, but this does not necessarily imply that $\boldsymbol{C}\boldsymbol{B}^{-1}$ is irreducible. However, the strict inequality in the conditions A1 and A3 yields irreducibility for the matrix $\boldsymbol{C}\boldsymbol{B}^{-1}$. Therefore, the demand for irreducibility does not lead to new inequalities.

Remark 5 Conditions A0–A3 guarantee the monotonicity of the matrix $\boldsymbol{A}^{-1}$ through *local* conditions. For control-volume discretizations with local flux approximations, the local discretization is independent of the domain. Thus, it follows that conditions A are sufficient for Property MD to hold.

4.2 Reordering the system matrix

Conditions A2 and A3 above are not symmetric, in the sense that they place more restrictions on $m_2^{i,j}$ and $m_6^{i,j}$ than on $m_4^{i,j}$ and $m_8^{i,j}$. This is not a property of the discretization method. Rather, it appears as a consequence of the choice of cell numbering or, equivalently, the choice of splitting $\boldsymbol{A} = \boldsymbol{B} - \boldsymbol{C}$.

In this section, we state conditions similar to A0 through A3 for a case where the cells have been reordered. For the case considered, the unknowns in each column i are counted first, and then the columns are counted. The matrix $\boldsymbol{A}$ then has the same block-tridiagonal structure as before, but now the blocks are $n \times n$ matrices, and there are m diagonal blocks.

Similar to the previous section, we denote the nonzero block matrices of $\boldsymbol{A}$ by $\boldsymbol{B}_i, -\boldsymbol{C}_i^L$ and $-\boldsymbol{C}_i^U$. These matrices are shown by displaying the nonzero elements around the jth diagonal element:

$$\begin{aligned}
\boldsymbol{B}_i &= \begin{bmatrix} \ddots & m_4^{i,j-1} & \\ m_8^{i,j} & m_1^{i,j} & m_4^{i,j} \\ & m_8^{i,j+1} & \ddots \end{bmatrix}, \\
\boldsymbol{C}_i^L &= -\begin{bmatrix} \ddots & m_5^{i,j-1} & \\ m_7^{i,j} & m_6^{i,j} & m_5^{i,j} \\ & m_7^{i,j+1} & \ddots \end{bmatrix}, \\
\boldsymbol{C}_i^U &= -\begin{bmatrix} \ddots & m_3^{i,j-1} & \\ m_9^{i,j} & m_2^{i,j} & m_3^{i,j} \\ & m_9^{i,j+1} & \ddots \end{bmatrix}.
\end{aligned} \tag{27}$$

We also introduce the matrices $\boldsymbol{D}_i = \boldsymbol{B}_i^{-1}$, $\boldsymbol{E}_{i+1}^L = \boldsymbol{C}_{i+1}^L \boldsymbol{D}_i$ and $\boldsymbol{E}_{i-1}^U = \boldsymbol{C}_{i-1}^U \boldsymbol{D}_i$. Instead of the Eqs. (24), (25) and (26), we now get

$$d_{j,k}^i = \frac{\delta_{j,k}}{m_1^{i,j}} - \frac{m_8^{i,j}}{m_1^{i,j}} d_{j-1,k}^i - \frac{m_4^{i,j}}{m_1^{i,j}} d_{j+1,k}^i, \tag{28}$$

$$\begin{aligned} e_{j,k}^{i+1,L} &= -m_7^{i+1,j} d_{j-1,k}^i - m_6^{i+1,j} d_{j,k}^i - m_5^{i+1,j} d_{j+1,k}^i \\ &= -\frac{m_6^{i+1,j}}{m_1^{i,j}} \delta_{j,k} + \left(\frac{m_8^{i,j}}{m_1^{i,j}} m_6^{i+1,j} - m_7^{i+1,j} \right) d_{j-1,k}^i \\ &\quad + \left(\frac{m_4^{i,j}}{m_1^{i,j}} m_6^{i+1,j} - m_5^{i+1,j} \right) d_{j+1,k}^i, \end{aligned} \tag{29}$$

$$\begin{aligned} e_{j,k}^{i-1,U} &= -m_9^{i-1,j} d_{j-1,k}^i - m_2^{i-1,j} d_{j,k}^i - m_3^{i-1,j} d_{j+1,k}^i \\ &= -\frac{m_2^{i-1,j}}{m_1^{i,j}} \delta_{j,k} + \left(\frac{m_8^{i,j}}{m_1^{i,j}} m_2^{i-1,j} - m_9^{i-1,j} \right) d_{j-1,k}^i \\ &\quad + \left(\frac{m_4^{i,j}}{m_1^{i,j}} m_2^{i-1,j} - m_3^{i-1,j} \right) d_{j+1,k}^i. \end{aligned} \tag{30}$$

Hence, the local criteria similar to A0 through A3 for the case when the unknowns are ordered along the columns, are

$$\begin{aligned} &\text{(B0):} && m_1^{i,j} > 0, \\ &\text{(B1a):} && m_2^{i,j} < 0, \\ &\text{(B1b):} && m_4^{i,j} < 0, \\ &\text{(B1c):} && m_6^{i,j} < 0, \\ &\text{(B1d):} && m_8^{i,j} < 0, \\ &\text{(B2):} && m_1^{i,j} + m_4^{i,j} + m_8^{i,j} > 0, \\ &\text{(B3a):} && m_4^{i,j} m_2^{i-1,j} - m_3^{i-1,j} m_1^{i,j} > 0, \\ &\text{(B3b):} && m_4^{i,j} m_6^{i+1,j} - m_5^{i+1,j} m_1^{i,j} > 0, \\ &\text{(B3c):} && m_8^{i,j} m_2^{i-1,j} - m_9^{i-1,j} m_1^{i,j} > 0, \\ &\text{(B3d):} && m_8^{i,j} m_6^{i+1,j} - m_7^{i+1,j} m_1^{i,j} > 0, \end{aligned}$$

where, as above, the inequalities in the conditions B3 are omitted whenever $i \pm 1$ lies outside the index domain.

Since each of the conditions A and B is sufficient for the monotonicity of $\boldsymbol{A}^{-1}$, only the weakest set of conditions has to be satisfied to achieve monotonicity. Hence, we have proved the following

Theorem 3 *The inverse of the matrix* $\boldsymbol{A}$ *arising from a locally conservative nine-point discretization in 2D is monotone if conditions A or B defined above hold for all pairs* (i, j).

Note that the conditions A and B are defined locally at each cell. This implies that these conditions may be used in a grid generation to, if possible, achieve a monotone scheme for a given discretization.

Remark 6 As mentioned at the beginning of Sect. 4, we may for any splitting $\boldsymbol{A} = \boldsymbol{B} - \boldsymbol{C}$, either determine conditions which guarantee that $\boldsymbol{C}\boldsymbol{B}^{-1} \geq \boldsymbol{O}$ or determine conditions which guarantee that $\boldsymbol{B}^{-1}\boldsymbol{C} \geq \boldsymbol{O}$. Applying the same procedure as above, we find that the rowwise ordering of Sect. 4.1 with the demand $\boldsymbol{B}^{-1}\boldsymbol{C} \geq \boldsymbol{O}$ results in the conditions B, whereas the columnwise ordering of Sect. 4.2 with the demand $\boldsymbol{B}^{-1}\boldsymbol{C} \geq \boldsymbol{O}$ results in the conditions A. Therefore, these derivations do not yield any new, and hence, no weaker sufficient monotonicity conditions.

Remark 7 Other orderings of the cells may also be considered, e.g. the chessboard ordering. However, our numerical tests indicate that the conditions derived from the chosen orderings are good.

4.3 Nonlocal criteria

The conditions A3 and B3 are used to ensure that $\boldsymbol{C}\boldsymbol{B}^{-1} \geq \boldsymbol{O}$. To check if these criteria are sharp, one could test numerically whether all the elements of the nonzero block matrices of $\boldsymbol{C}\boldsymbol{B}^{-1}$ are nonnegative. Below, we show that this sign test can be reduced to only a limited number of these elements. We show this for the matrices of Section 4.1 only. The reduction of the number of sign tests is then from $\mathcal{O}(nm^2)$ to $\mathcal{O}(nm)$.

When $\boldsymbol{B}_j$ is an irreducible M-matrix, its inverse satisfies $\boldsymbol{D}_j > \boldsymbol{O}$. The tridiagonal structure of $\boldsymbol{B}_j$ then implies that the elements of $\boldsymbol{D}_j$ satisfy

$$d^j_{i,k} = \begin{cases} \mu^j_{k,1} d^j_{i,1} & \forall i \geq k, \\ \mu^j_{k,m} d^j_{i,m} & \forall i \leq k, \end{cases} \tag{31}$$

where $\mu^j_{k,1}$ and $\mu^j_{k,m}$ are independent of i. Further, $\mu^j_{k,1} > 0$ and $\mu^j_{k,m} > 0$. The property (31) is easily seen from the $\boldsymbol{L}\boldsymbol{U}$ decomposition of $\boldsymbol{B}_j$ [7].

When $\boldsymbol{B}_j$ is a rowwise diagonally dominant, tridiagonal M-matrix, the elements of the columns of the inverse $\boldsymbol{D}_j$ are nonincreasing as one moves away from the diagonal. This is easily seen by examining the equation $\boldsymbol{B}_j\boldsymbol{D}_j = \boldsymbol{I}$. The diagonal dominance of $\boldsymbol{B}_j$ implies that the columns of $\boldsymbol{D}_j$ cannot have any local inner extrema except at the diagonal and must have a minimum at the boundary. Hence,

$$\left.\begin{aligned} d^j_{i,1} &\leq d^j_{i-1,1}, \\ d^j_{i-1,m} &\leq d^j_{i,m}, \end{aligned}\right\} \quad i = 2, \ldots, m. \tag{32}$$

We will apply the properties (31) and (32) to reduce the necessary number of sign tests in $\boldsymbol{CB}^{-1}$. Let S be one of the superscripts L and U in Sect. 4.1. From the first line in the Eqs. (25) and (26) it follows that

$$\begin{aligned} e_{i,k}^{j,S} &= \mu_{k,1}^{j} e_{i,1}^{j,S} \quad \text{for } i > k, \\ e_{i,k}^{j,S} &= \mu_{k,m}^{j} e_{i,m}^{j,S} \quad \text{for } i < k. \end{aligned} \tag{33}$$

Hence, to check if $\boldsymbol{E}_j^S \geq \boldsymbol{O}$, we do not have to check the sign of all the elements $e_{i,k}^{j,S}$. It is sufficient to check if $e_{i,1}^{j,S} \geq 0$, $e_{i,m}^{j,S} \geq 0$ and $e_{i,i}^{j,S} \geq 0$ for all i. However, also the sign test of the diagonal elements with indices $i = 2, \ldots, m-1$ are superfluous. We show this for the elements $e_{i,i}^{j+1,L}$. From Eqs. (25) and (31) it follows that

$$\begin{aligned} \frac{e_{i,i}^{j+1,L}}{d_{i,i}^{j}} &= -\left(m_7^{i,j+1} \frac{d_{i-1,i}^{j}}{d_{i,i}^{j}} + m_8^{i,j+1} + m_9^{i,j+1} \frac{d_{i+1,i}^{j}}{d_{i,i}^{j}} \right) \\ &= -\left(m_7^{i,j+1} \frac{d_{i-1,m}^{j}}{d_{i,m}^{j}} + m_8^{i,j+1} + m_9^{i,j+1} \frac{d_{i+1,1}^{j}}{d_{i,1}^{j}} \right). \end{aligned} \tag{34}$$

Introducing the quantities $\nu_{i,1}^{j}$ and $\nu_{i,m}^{j}$, $i = 2, \ldots, m-1$, defined by the pair of equations

$$\begin{aligned} \nu_{i,1}^{j} d_{i-1,1}^{j} + \nu_{i,m}^{j} d_{i-1,m}^{j} &= d_{i-1,m}^{j} / d_{i,m}^{j}, \\ \nu_{i,1}^{j} d_{i+1,1}^{j} + \nu_{i,m}^{j} d_{i+1,m}^{j} &= d_{i+1,1}^{j} / d_{i,1}^{j}, \end{aligned} \tag{35}$$

the expression (34) may be rewritten as

$$\frac{e_{i,i}^{j+1,L}}{d_{i,i}^{j}} = \nu_{i,1}^{j} e_{i,1}^{j+1,L} + \nu_{i,m}^{j} e_{i,m}^{j+1,L} - \left(1 - \nu_{i,1}^{j} d_{i,1}^{j} - \nu_{i,m}^{j} d_{i,m}^{j}\right) m_8^{i,j+1}. \tag{36}$$

Clearly, the first two terms in the right-hand side of (36) are nonnegative if $e_{i,1}^{j+1,L} \geq 0$ and $e_{i,m}^{j+1,L} \geq 0$ and both $\nu_{i,1}^{j} \geq 0$ and $\nu_{i,m}^{j} \geq 0$. The last term of (36) is nonnegative if $m_8^{i,j+1} \leq 0$ and

$$1 - \nu_{i,1}^{j} d_{i,1}^{j} - \nu_{i,m}^{j} d_{i,m}^{j} \geq 0. \tag{37}$$

To check these sign properties, we investigate the solution of (35),

$$v_{i,1}^j = d_{i-1,m}^j \frac{d_{i+1,m}^j/d_{i,m}^j - d_{i+1,1}^j/d_{i,1}^j}{d_{i-1,1}^j d_{i+1,m}^j - d_{i+1,1}^j d_{i-1,m}^j},$$
$$v_{i,m}^j = d_{i+1,1}^j \frac{d_{i-1,1}^j/d_{i,1}^j - d_{i-1,m}^j/d_{i,m}^j}{d_{i-1,1}^j d_{i+1,m}^j - d_{i+1,1}^j d_{i-1,m}^j}. \tag{38}$$

If follows that due to the property (32), both $v_{i,1}^j \geq 0$ and $v_{i,m}^j \geq 0$ as required above.

It remains to show that inequality (37) holds. From (23) it follows that for $i \neq 1$ and $i \neq m$,

$$\begin{aligned}
1 &= m_6^{i,j} d_{i-1,i}^j + m_1^{i,j} d_{i,i}^j + m_2^{i,j} d_{i+1,i}^j \\
&\quad - d_{i,i}^j v_{i,1}^j \left(m_6^{i,j} d_{i-1,1}^j + m_1^{i,j} d_{i,1}^j + m_2^{i,j} d_{i+1,1}^j\right) \\
&\quad - d_{i,i}^j v_{i,m}^j \left(m_6^{i,j} d_{i-1,m}^j + m_1^{i,j} d_{i,m}^j + m_2^{i,j} d_{i+1,m}^j\right) \\
&= m_1^{i,j} d_{i,i}^j \left(1 - v_{i,1}^j d_{i,1}^j - v_{i,m}^j d_{i,m}^j\right) \\
&\quad + m_6^{i,j} \left[d_{i-1,i}^j - d_{i,i}^j \left(v_{i,1}^j d_{i-1,1}^j + v_{i,m}^j d_{i-1,m}^j\right)\right] \\
&\quad + m_2^{i,j} \left[d_{i+1,i}^j - d_{i,i}^j \left(v_{i,1}^j d_{i+1,1}^j + v_{i,m}^j d_{i+1,m}^j\right)\right].
\end{aligned} \tag{39}$$

The last two terms vanish by (35), and since $m_1^{i,j} \geq 0$, inequality (37) is valid.

Using the property (32), we have now shown that the inequalities $e_{i,1}^{j,L} \geq 0$, $e_{i,m}^{j,L} \geq 0$ and $m_8^{i,j} \leq 0$ imply that $e_{i,i}^{j,L} \geq 0$. An analogous derivation holds for the diagonal elements of $\boldsymbol{E}_j^U$, but here with the condition $m_4^{i,j} \leq 0$. Thus, applying the properties of $\boldsymbol{B}_j$ (irreducible, rowwise diagonally dominant, tridiagonal M-matrix), the following theorem appears.

Theorem 4 *The inverse of the matrix* $\mathbf{A}$ *arising from a locally conservative nine-point discretization in 2D is monotone if conditions A0, A1 and A2 hold, and the first and last column of the matrices* $\boldsymbol{E}_j^L$ *and* $\boldsymbol{E}_j^U$ *have only nonnegative elements.*

This shows that we can verify monotonicity in $\mathcal{O}(mn)$ operations, equivalent to the number of grid cells, which is optimal complexity.

Remark 8 Theorem 4 is established by using the condition $\boldsymbol{C}\boldsymbol{B}^{-1} \geq \boldsymbol{O}$ in the derivation of Sect. 4.1. The rowwise diagonal dominance of $\boldsymbol{B}_j$ yields the columnwise property (32), and this is used to reduce the sign test to the first and last columns. If instead the condition $\boldsymbol{B}^{-1}\boldsymbol{C} \geq \boldsymbol{O}$ had been used, the theorem would have had to be formulated with a reduction to the first and last row. This would require that $\boldsymbol{B}_j$ is columnwise diagonally dominant.

Remark 9 As demonstrated in Sect. 4.2, complementary versions of Theorem 4 can trivially be obtained using different cell orderings.

Remark 10 Note that Theorem 4 only gives monotonicity criteria for a given grid. Remark 5 does not apply to Theorem 4 due to the nonlocal terms in the matrices $\boldsymbol{E}_j^L$ and $\boldsymbol{E}_j^U$. Thus, Theorem 4 is not sufficient for Property MD.

5 Case with homogeneous medium and uniform grid

In this section, we consider the case where the medium is homogeneous and the grid is uniform. In this special case, the conditions simplify considerably.

If the medium is homogeneous and the grid is uniform, then $m_k^{i,j} = m_k$, independent of i and j. Also, $m_k = m_{k+4}$ for $k = 2, \ldots, 5$. In this case, the conditions A and B can be combined and stated as

$$\begin{aligned} &\text{(C0):} && m_1 > 0, \\ &\text{(C1):} && \max\{m_2, m_4\} < 0, \\ &\text{(C2):} && m_1 + 2\max\{m_2, m_4\} > 0, \\ &\text{(C3):} && m_2 m_4 - \max\{m_3, m_5\} \cdot m_1 > 0. \end{aligned}$$

Remark 11 The conditions C are identical to the monotonicity conditions previously developed in [27], except for the inequality C1 which in [27] only is required for m_2. However, in the derivation of the conditions of [27], also $m_4 < 0$ was demanded, but this was unfortunately omitted in the final summary of the conditions.

5.1 A class of control-volume methods

For the case of a uniform parallelogram grid in a homogeneous medium, we will develop a class of control-volume methods constrained by three desirable properties: it is locally conservative, it has a local flux representation, and it is exact for linear potential fields.

We study methods with a local flux expression. Consider the fluxes across the interfaces in Fig. 1 which separate cell 1 from cell 2 and cell 1 from cell 4, respectively. These fluxes can be written as the weighted sum of potentials:

$$f_1 = \sum_{k=1,2,3,4,8,9} t_{1,k} u_k, \quad f_2 = \sum_{k=1}^{6} t_{2,k} u_k. \tag{40}$$

When we look at the special case of uniform parallelogram grids on homogeneous media, the local grid symmetry implies that

$$\begin{aligned} &t_{1,1} = -t_{1,2}, \quad t_{1,3} = -t_{1,8}, \quad t_{1,4} = -t_{1,9}, \\ &t_{2,1} = -t_{2,4}, \quad t_{2,2} = -t_{2,5}, \quad t_{2,3} = -t_{2,6}. \end{aligned} \tag{41}$$

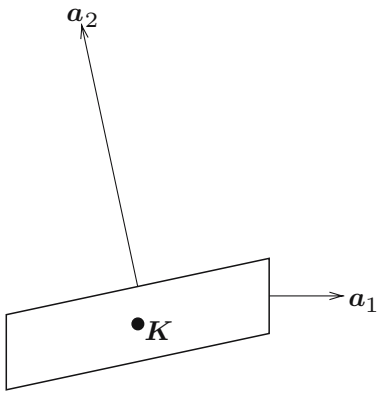

Fig. 3 The vectors $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$

Equations (40) can now be written as

$$f_1 = t_{1,1}(u_1 - u_2) + t_{1,3}(u_3 - u_8) + t_{1,4}(u_4 - u_9), \tag{42}$$
$$f_2 = t_{2,1}(u_1 - u_4) + t_{2,2}(u_2 - u_5) + t_{2,3}(u_3 - u_6). \tag{43}$$

For any grid cell, let $\boldsymbol{a}_i$, $i = 1, 2$, be the normal vector of edge i, having length equal to the length of the edge, see Fig. 3. Further, let V be the area of the parallelogram cell, and define the quantities a, b and c by [1,27]

$$\begin{bmatrix} a & c \\ c & b \end{bmatrix} = \frac{1}{V} \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 \end{bmatrix}^{\mathrm{T}} \boldsymbol{K} \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 \end{bmatrix}. \tag{44}$$

These quantities cannot attain arbitrary values since the positive definiteness of $\boldsymbol{K}$ implies that $a > 0$, $b > 0$, and

$$|c| < \sqrt{ab}. \tag{45}$$

In the case of a linear potential field, the fluxes can be expressed by [1]

$$\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \begin{bmatrix} a & c \\ c & b \end{bmatrix} \begin{bmatrix} u_1 - u_2 \\ u_1 - u_4 \end{bmatrix} = \begin{bmatrix} a(u_1 - u_2) + c(u_1 - u_4) \\ c(u_1 - u_2) + b(u_1 - u_4) \end{bmatrix}. \tag{46}$$

To determine the transmissibility coefficients of the Eqs. (42) and (43), we may apply linear potential fields in two independent directions. Let us first determine the coefficients of (42). In this case, we choose the potential field such that either $f_1 = 0$ or $\operatorname{grad} u \parallel \boldsymbol{a}_1$. When $f_1 = 0$, Eq. (46) yields

$$u_1 - u_2 = -\frac{c}{a}(u_1 - u_4). \tag{47}$$

From Fig. 1 it follows that for linear potential fields,

$$u_3 - u_8 = -2(u_1 - u_4) - (u_1 - u_2) = -\left(2 - \frac{c}{a}\right)(u_1 - u_4), \tag{48}$$

$$u_4 - u_9 = -2(u_1 - u_4) + (u_1 - u_2) = -\left(2 + \frac{c}{a}\right)(u_1 - u_4). \tag{49}$$

Applying the expressions (47), (48) and (49) in Eq. (42), it follows that

$$f_1 = -\left[\frac{c}{a}t_{1,1} + \left(2 - \frac{c}{a}\right)t_{1,3} + \left(2 + \frac{c}{a}\right)t_{1,4}\right](u_1 - u_4) = 0. \tag{50}$$

In the case when $\operatorname{grad} u \parallel \boldsymbol{a}_1$,

$$u_1 - u_2 = -(u_3 - u_8) = u_4 - u_9, \tag{51}$$

and thus, from (42),

$$f_1 = [t_{1,1} - t_{1,3} + t_{1,4}](u_1 - u_2) = a(u_1 - u_2). \tag{52}$$

The Eqs. (50) and (52) yield the pair of equations

$$\begin{aligned} ct_{1,1} + (2a - c)t_{1,3} + (2a + c)t_{1,4} &= 0, \\ t_{1,1} - t_{1,3} + t_{1,4} &= a. \end{aligned} \tag{53}$$

Similarly, we may determine the transmissibility coefficients of (43) by choosing the potential field such that either $f_2 = 0$ or $\operatorname{grad} u \parallel \boldsymbol{a}_2$. These cases yield the following pair of equations,

$$\begin{aligned} ct_{2,1} + (2b + c)t_{2,2} + (2b - c)t_{2,3} &= 0, \\ t_{2,1} + t_{2,2} - t_{2,3} &= b. \end{aligned} \tag{54}$$

The pair of Eqs. (53) determines the coefficients $t_{1,1}$, $t_{1,3}$ and $t_{1,4}$ up to one undetermined parameter α. Likewise, the pair of Eqs. (54) determines the coefficients $t_{2,1}$, $t_{2,2}$ and $t_{2,3}$ up to one undetermined parameter β. The solutions read

$$\begin{aligned} t_{1,1} &= a - \alpha, & t_{2,1} &= b - \beta, \\ t_{1,3} &= -\frac{c}{4} - \frac{\alpha}{2}, & t_{2,2} &= -\frac{c}{4} + \frac{\beta}{2}, \\ t_{1,4} &= -\frac{c}{4} + \frac{\alpha}{2}, & t_{2,3} &= -\frac{c}{4} - \frac{\beta}{2}. \end{aligned} \tag{55}$$

These transmissibilities can be combined into a nine-point stencil given by

$$\sum_{k=1}^{9} m_k u_k = f_1 + f_2 - f_3 - f_4, \tag{56}$$

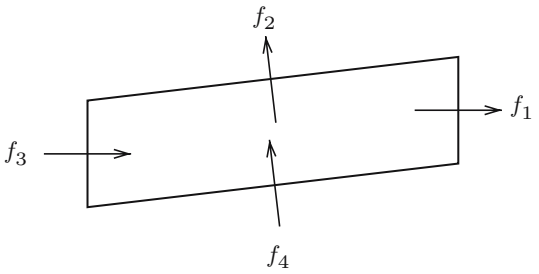

Fig. 4 Fluxes through the edges of a cell

where the fluxes f_1, f_2, f_3, and f_4 are shown in Fig. 4. Introducing the parameter $\gamma = \alpha + \beta$, the elements of the nine-point stencil read

$$\begin{aligned} m_2 = m_6 &= -t_{1,1} + t_{2,2} - t_{2,3} = -a + \gamma, \\ m_3 = m_7 &= t_{1,3} + t_{2,3} = -\frac{c}{2} - \frac{\gamma}{2}, \\ m_4 = m_8 &= -t_{1,3} + t_{1,4} - t_{2,1} = -b + \gamma, \\ m_5 = m_9 &= -t_{1,4} - t_{2,2} = \frac{c}{2} - \frac{\gamma}{2}, \\ m_1 = -\sum_{k=2}^{9} m_k &= 2(t_{1,1} + t_{2,1}) = 2(a + b - \gamma). \end{aligned} \tag{57}$$

Observe that these expressions imply that there is only one degree of freedom in choosing a control-volume discretization which reproduces linear potential fields for the elliptic problem (2). However, we see that although the system matrix only depends on the parameter γ, the local flux approximations themselves are functions of the parameters α and β.

5.2 Monotone regions

In this section, we investigate the implications of the conditions C on the class of nine-point schemes defined by (57). We will show that there is a range of parallelogram grids for which it is impossible to satisfy the conditions C over the full range of these control-volume methods. This is not due to the strictness of the criteria, but it appears to be a fundamental property of the discrete representation.

The parameter γ, appearing in the coefficients m_k in (57), defines all possible conservative nine-point discretizations on uniform parallelogram grids in homogeneous media, where the discretization method has an explicit flux representation which is exact for linear potential fields. Several discretization methods have this property, including the multi-point flux approximation (MPFA) class of methods [1,2,14]. In this section, we will discuss this parameter further, and obtain an explicit value for γ for some common discretization methods.

Following Remark 3, the coefficients (57) yield an M-matrix if and only if

$$|c| \leq \gamma \leq \min\{a, b\}. \tag{58}$$

These inequalities are derived from the sign property $m_k \leq 0$ for $k = 2, \ldots, 9$. The inequalities (58) imply that

$$|c| \leq \min\{a, b\}. \tag{59}$$

Hence, the following lemma applies.

Lemma 2 *For control-volume methods with local flux approximations which yield exact solutions of linear potential fields, it is impossible to define a nine-point scheme resulting in an M-matrix, on grids violating inequality* (59).

The conditions C yield a wider class of methods than those resulting in an M-matrix. Inserting the coefficients (57) into conditions C0 through C3, one gets a class of monotone control-volume methods for uniform grids on homogeneous media. The conditions C1 and C3 read

$$\gamma < \min\{a, b\}, \tag{60}$$

$$(\gamma - a)(\gamma - b) - (\gamma - |c|)\big(\gamma - (a + b)\big) > 0. \tag{61}$$

Condition C2 is trivial, while condition C0 follows from inequality (60). Hence, for any given combination of a, b and c, the conditions C restrict the parameter γ to

$$a + b - \frac{ab}{|c|} < \gamma < \min\{a, b\}. \tag{62}$$

The inequalities (62) imply inequality (59). This makes it tempting to extend Lemma 2 from M-matrix nine-point schemes to monotone nine-point schemes. However, since the conditions C are only sufficient, such an extension cannot be claimed. It is, however, our experience that inequality (59) is a practical bound for monotone nine-point schemes.

The parameter domain defined by (59) is smaller than the domain defined by (45), see Fig. 5. A method possesses good monotonicity properties if it is monotone in large areas of the domain defined by (59).

The expressions (57) show that the choice $\gamma = |c|$ yields a seven-point method, in which either $m_3 = m_7 = 0$ or $m_5 = m_9 = 0$. This choice satisfies the inequalities (62), provided (59) is satisfied with strict inequality. Hence, the method defined by $\gamma = |c|$ has optimal monotonicity properties with respect to the conditions C in the sense that no other nine-point method satisfies the monotonicity conditions over a larger range of parameters. The inequalities (58) show that this choice of the parameter γ also leads to an M-matrix discretization, provided (59) is satisfied.

In the MPFA methods, we denote the choice of continuity point of the potential by η. The parameter η is defined as the fraction of the cell half-edge

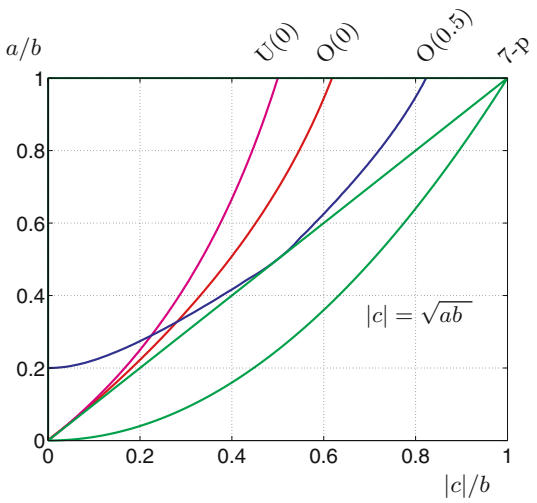

Fig. 5 Monotonicity regions defined by (62) for some O(η)- and U(η)-methods as well as the seven-point stencil for the case $a \leq b$. The monotone regions are above the curves in question. The elliptic bound is $|c| = \sqrt{ab}$

Table 1 The parameter γ calculated for some well-known discretization schemes

Method	Expression for γ	Optimal
O(η)	$\gamma = (ab\eta + c^2)/(d(1+\eta))$	For e.g. $\eta = \lvert c\rvert / \max\{a, b\}$
O(0)	$\gamma = c^2/d$	No
O(0.5)	$\gamma = (ab + 2c^2)/3d$	No
U(η)	$\gamma = (a+b)\eta/2$	For e.g. $\eta = 2\lvert c\rvert /(a+b)$
U(0)	$\gamma = 0$	No
7-point	$\gamma = \lvert c\rvert$	Yes

Here, $d = 2ab/(a+b)$. Optimal is here in the sense of the monotonicity conditions C being satisfied for all grids with $|c| < \min\{a, b\}$

from the boundary of the interaction region. The MPFA O-method with a specific choice of η is denoted by O(η). The O(0)-method is discussed in [1,2], while the O(η)-method with special emphasis on the O(0.5)-method has been investigated in [14]. The expression for γ of the O(η)-method is derived in Appendix 7.

Similarly, we denote the MPFA U-method with a specific choice of η by U(η). For a general η, the transmissibility coefficients (55) with $\alpha = \eta a/2$ and $\beta = \eta b/2$ are easily derived. The common choice here is the U(0)-method, and in fact, this is the only choice previously studied [2].

For purpose of illustration, we give the γ expressions for some common MPFA discretization methods in Table 1. For each method, the conditions C are satisfied when a, b and c satisfy inequality (62). This defines the monotonicity region in terms of these parameters. For the O(0)-method, the monotonicity region is

$$|c| \left(2 - \frac{c^2}{ab}\right) < \frac{2ab}{a+b}. \tag{63}$$

For the O(0.5)-method, the monotonicity region is

$$|c|\left(5-\frac{2c^2}{ab}\right)<\frac{6ab}{a+b} \quad \text{and} \quad c^2<\frac{1}{2}\frac{ab}{a+b}(5a-b). \tag{64}$$

Finally, for the U(0)-method, the monotonicity region is

$$|c|<\frac{ab}{a+b}. \tag{65}$$

The monotonicity regions of these methods as well as the seven-point stencil which has the monotonicity bound (59) are shown in Fig. 5. Each method is monotone in the region above the curve in question. Note that these monotonicity regions are the regions where the sufficient conditions C are satisfied.

We see from Table 1 that neither the O(0)-method nor the O(0.5)-method is optimal, where optimal refers to the conditions C being satisfied in the full range where inequality (59) is satisfied. Indeed, for any choice of η equal to a constant, the resulting O(η)-method will never be optimal. However, by choosing a case-dependent continuity point, optimal methods can be obtained, and one example of such a choice is given in the table.

Similarly, the U(η)-method will never be optimal for any single choice of η, but optimal methods can be obtained by choosing a case-dependent continuity point. An example of such a choice is given in the table, yielding $\gamma = |c|$. This choice does not only satisfy the conditions C, but it also leads to an M-matrix discretization, provided inequality (59) is satisfied.

6 Numerical examples

6.1 Cases with homogeneous medium and uniform grid

In this section, we demonstrate by numerical examples the validity of the curves of Fig. 5. We have calculated the same curves numerically, by testing Property MD on a 19×19 grid. All tests are performed on homogeneous media with uniform parallelogram grids. Figure 6 shows the numerically computed monotonicity regions for different methods. These are to be compared with analytical regions of Fig. 5. As seen in the figures, for small values of $|c|/b$, the sufficient conditions C are approximately necessary to guarantee monotonicity. For larger $|c|/b$, the conditions C are still good, but no longer sharp, and monotonicity is ensured in a somewhat larger domain.

Figures 7 and 8 show cases outside the monotonicity regions for a 19×3 grid. The boundary value problem (2), (3) is solved with a delta functional as a source term in block $(10, 3)$. As can be seen from the figures, the solution is negative in some areas of the domain, i.e., monotonicity is lost.

Within the monotonicity regions of Fig. 6, the inverse of the coefficient matrix, $\boldsymbol{A}^{-1}$, has only nonnegative elements for all subgrids of a 19×19 grid. Figure 9

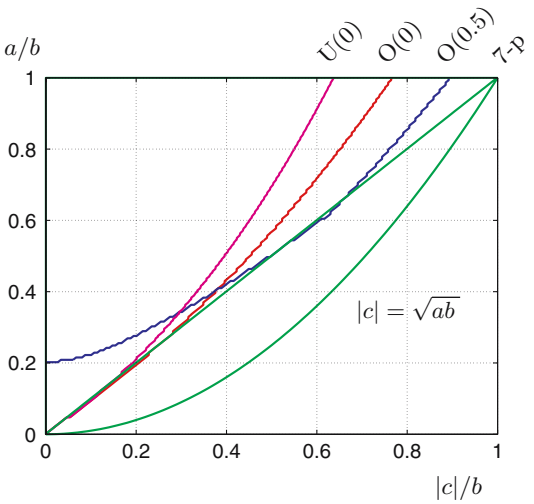

Fig. 6 Numerically computed monotonicity regions for some O(η)- and U(η)-methods as well as the seven-point stencil for homogeneous media and uniform parallelogram grids. The elliptic bound is $|c| = \sqrt{ab}$

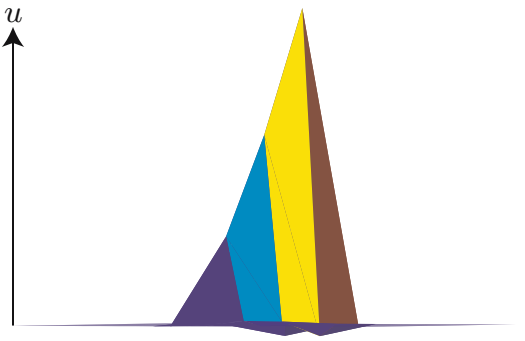

Fig. 7 Solution u with a delta functional as source term at the boundary. O(0.5)-method. $a/b = 0.05$, $c/b = 0$

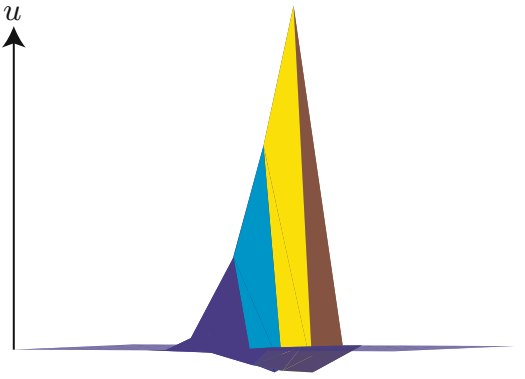

Fig. 8 Solution u with a delta functional as source term at the boundary. O(0)-method. $a/b = 0.05$, $c/b = 0.15$

shows the monotonicity regions of the inverse of the coefficient matrix, $\boldsymbol{A}^{-1}$, for a 19×19 grid without considering the subgrid cases. Obviously, the regions of Fig. 9 are larger than the regions of Fig. 6. To avoid oscillations, the boundary

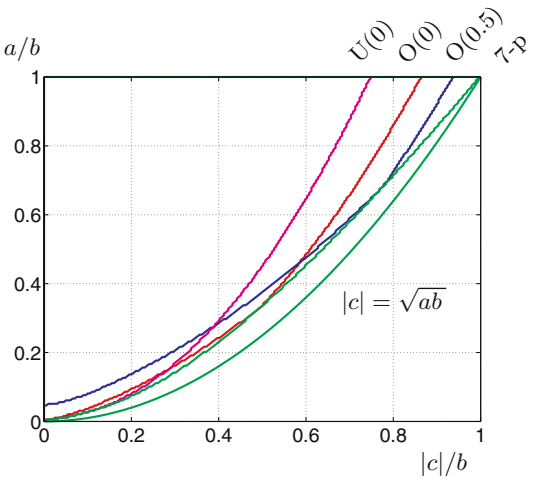

Fig. 9 Numerically computed monotonicity regions of $\boldsymbol{A}^{-1}$ for some O(η)- and U(η)-methods for homogeneous media and a uniform parallelogram 19×19 grid. The elliptic bound is $|c| = \sqrt{ab}$

value problem (2), (3) for all subgrids must yield matrices of coefficients $\boldsymbol{A}$ with $\boldsymbol{A}^{-1} \geq \boldsymbol{O}$.

To show the impact of the loss of monotonicity for a sequence of O(η)-methods, we consider a case with $c = 0$. As seen from Table 1, for $c = 0$ the parameter γ of the O(η)-methods is $\gamma = \frac{1}{2}(a+b)\eta/(1+\eta)$. Inequality (61) is now always fulfilled, and from inequality (60) it follows that monotonicity is ensured if

$$\eta < \frac{2\min\{a,b\}}{|b-a|}. \tag{66}$$

As in the examples of Figs. 7 and 8, we choose $a/b = 0.05$. The monotonicity limit (66) then yields $\eta < 0.105$. We apply a 19×1 grid, and for $j = 10$ we compute

$$\epsilon = \frac{\min_i [\boldsymbol{A}^{-1}]_{i,j}}{\max_i [\boldsymbol{A}^{-1}]_{i,j}}. \tag{67}$$

The quantity ϵ is nonnegative for monotone methods and negative for non-monotone cases. It is computed for a sequence of the parameter $\eta \in [0, 1)$, i.e., for a sequence of O(η)-methods. The function $\epsilon(\eta)$ is shown in Fig. 10. As seen from the figure, $\epsilon = 0$ in the monotonicity region $\eta < 0.105$. As η increases beyond this limit, the significance of the loss of monotonicity gets more and more severe, with increasing negative values of ϵ. For $\boldsymbol{K}$-orthogonal grids, i.e., for $c = 0$, unconditional monotonicity is lost for all O(η)-methods with $\eta \neq 0$.

The test runs above are for small values of a/b and $|c|/b$, i.e., parameters in the bottom left part of the diagram of Fig. 5. For large values of a/b and $|c|/b$, i.e., in the top right part of the diagram of Fig. 5, the impact of a violation of monotonicity is less severe. Typically, the ratio (67) for the column index j which gives the largest negative ratio is two orders of magnitude smaller in the top right part of the diagram than in the bottom left part.

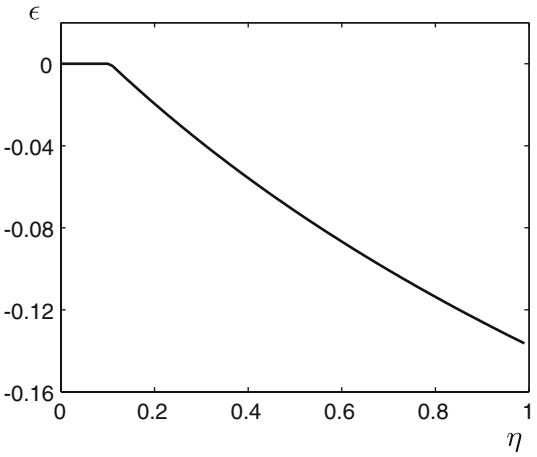

Fig. 10 Function (67) for a sequence of O(η)-methods on a 19×1 grid. $a/b = 0.05, c = 0$

6.2 Heterogeneous cases on uniform grids

In this section we investigate the monotonicity regions for heterogeneous cases. The local criteria of Theorem 3 still hold, and we now apply them to layered media and uniform parallelogram grids using the O(0)-method. The conductivity of the medium is isotropic, and every second layer of the medium has the same conductivity given by $k_1\boldsymbol{I}$ and $k_2\boldsymbol{I}$, respectively. Two cases are investigated, one case with $k_1/k_2 = 2$ and one case with $k_1/k_2 = 100$. In the first case, clearly different curves occur, depending on whether the layers are parallel to the i-lines or the j-lines of the grid.

The quantities a, b and c, defined in (44), now vary from layer to layer, but the ratios a/b and c/b are constant throughout the grid. Figure 11 shows the

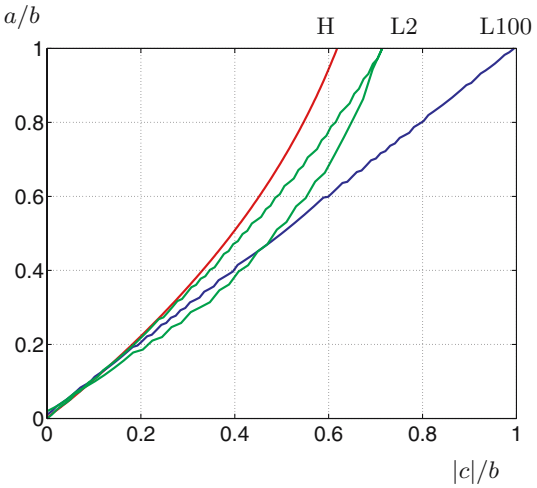

Fig. 11 Monotonicity regions for a layered medium with the O(0)-method, defined by the criteria of Theorem 3. *H* = homogeneous; *L2* = layered with conductivity ratio 2, layers aligned with the *j*-lines and *i*-lines, respectively; *L100* = layered with conductivity ratio 100

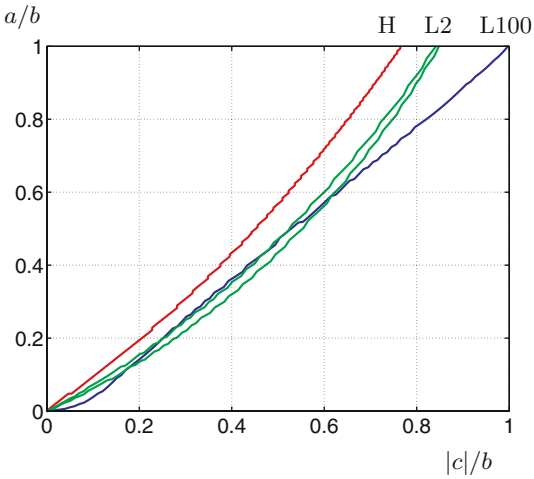

Fig. 12 Numerically computed monotonicity regions for a layered medium with the O(0)-method. *H* = homogeneous; *L2* = layered with conductivity ratio 2, layers aligned with the *j*-lines and *i*-lines, respectively; *L100* = layered with conductivity ratio 100

sufficient monotonicity regions determined by the local criteria of Theorem 3. As before, the monotone regions are above the curves in question. The regions of Fig. 11 may be compared to the monotonicity regions of Fig. 12. Here, the curves are determined by verifying Property MD on a 19×19 grid.

As seen from Fig. 11, the local criteria reveal a significant difference in the monotonicity regions for the O(0)-method when the difference in conductivities in the layers is increased. A similar, but weaker difference in the monotonicity regions is seen in Fig. 12.

The increasing difference in conductivity in the layers makes the solution of the potential differ from the homogeneous case. Due to the preferred flow pattern which high-conductive layers create, the flow within a high-conductive layer will be less affected by the low-conductive cells. The cell stencil will for cells in these layers be dominated by the elements of the three cells which have the largest conductivity. Cells in low-conductive layers will still be affected by the high-conductive cells, and the monotonicity will be lost for combinations of grid skewness and grid aspect ratio. The curves for the two cases of layered media show the combined monotonicity behavior of the high- and low-conductive layers.

The monotonicity regions are somewhat larger for layers aligned with the *i*-lines compared to layers aligned with the *j*-lines. This is explained by the fact that $a/b < 1$ such that the aspect ratios imply thinner cell cross sections in the *i*-direction. The application of the criteria and the numerical tests then suggest that the discretization is somewhat more robust when the layers are aligned with the longest edges of the grid cells.

6.3 Heterogeneous cases on nonuniform grids

One strength of the monotonicity criteria of Theorem 3 lies in their determination of problem areas in the grid/medium, for which violation of monotonicity

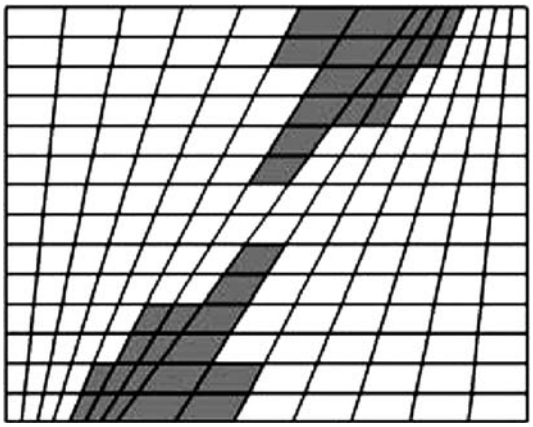

Fig. 13 Localization of regions of nonmonotonicity; O(0)-method

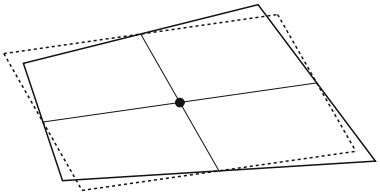

Fig. 14 Quadrilateral with associated parallelogram. Center evaluation

may be detected. This is illustrated in Fig. 13, where the criteria of Theorem 3 have been applied to a case where the grid is forced to adapt to an inner edge, and where the medium is homogeneous. The grid is specified as a 14×14 grid on a domain $[0, 1] \times [0, 0.8]$ for which the angle between the inner edge and the x-axis is $\arctan(4\sqrt{3}/5)$.

Cells for which the a-priori monotonicity criteria are not satisfied are identified as the darker cells. Note the way the grid is generated here: due to the structure of the grid, there will exist regions where neighboring cells have a jump in grid aspect ratio. The combination of skewness and different aspect ratios then make the criteria fail for grid cells close to the inner edge, and monotonicity is possibly violated.

We further use the ideas that lead to optimal continuity points for the generalized O(η)-method in Table 1. A general nine-point stencil is obtained for all grid cells when we a-priori use alternative continuity points based on the optimized points found for homogeneous cases on uniform grids. The continuity points may be chosen as averages of cellwise calculated optimized continuity points η. One possible way is to define an associated parallelogram grid cell for each physical grid cell, which corresponds to a transformation to computational space [5]. Such a cell is depicted in Fig. 14. These parallelograms and associated conductivities provide local parameters a, b and c for each grid cell, as defined by Eq. (44), and will in general be different from cell to cell. The application of the local criteria for this discretization gave no violation of the monotonicity criteria. Note, however, that there in general is no guarantee that the

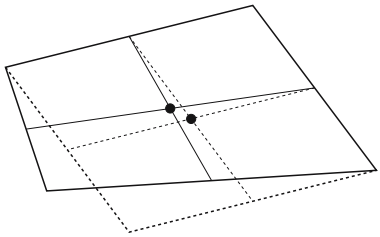

Fig. 15 Quadrilateral with associated parallelogram. Corner evaluation

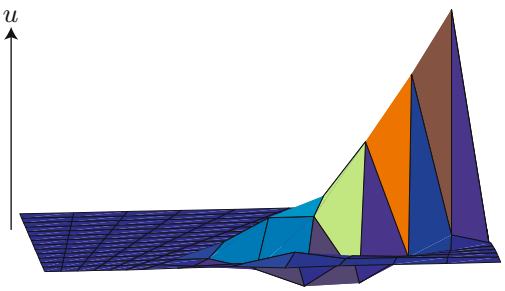

Fig. 16 Local oscillation due to violation of monotonicity criteria. Source located in cell where criteria are violated and medium heterogeneity occurs

monotonicity will be improved by this simple choice of continuity point. When applied to more general cases of skewness and heterogeneity, a more detailed analysis must be done to choose improved points in terms of monotonicity.

An alternative construction of an associated parallelogram is shown in Fig. 15. Here, the parallelogram is determined by the edges of the corner at the center of the interaction region, yielding one set of parameters a, b and c for each subcell [4,13].

The above case may be generalized to include discontinuous conductivity. A case where corner discontinuities arise is created by letting the 7×7 subdomain in the top right corner of the medium depicted in Fig. 13 have conductivity $\boldsymbol{K} = \text{diag}(10^{-4}, 10^{-2})$, whereas the remaining cells have conductivity $\boldsymbol{K} = \boldsymbol{I}$. The cells in the subdomain with anisotropic conductivity will no longer have the a-priori monotonicity criteria satisfied. When sources are located in cells in this region, we may observe large oscillations in the potential, applying the O(0)-method. This is depicted in Fig. 16, where a source is located in cell $(12, 8)$, and the numerical potential is clearly nonphysical.

In Fig. 17, a source has been inserted in cell $(8, 2)$ which is away from the subdomain where the monotonicity criteria are violated. No oscillations in the potential solution are observed here. These two different source locations illustrate the effect which the nonlinearity of the potential can have on oscillations. In regions where the potential is almost linear, oscillations may not arise due to the fact that the methods are exact for linear potential fields.

It should be noted that the observed oscillations are largest in cells for which the monotonicity criteria are violated, and where in addition the local medium heterogeneities occur.

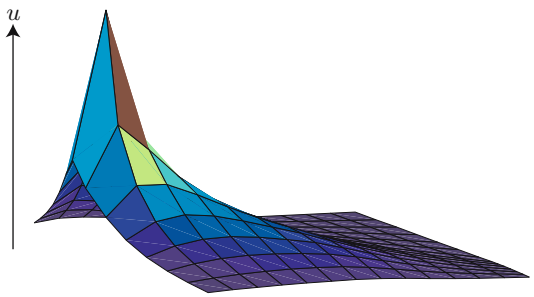

Fig. 17 Source away from cells where criteria are violated. Oscillations are not observed

7 Discussion

This paper discusses a-priori criteria for monotonicity of discretization methods where the discretizations do not yield M-matrices, but general nine-point cell stencils. The criteria may be used locally in the grid, and allow for general geometry, heterogeneity and anisotropy.

The a-priori monotonicity criteria are not satisfied unconditionally for discretization methods which reproduce linear potential fields exactly and have local conservation.

The local monotonicity criteria have good potential as an aid for grid generation.

The fact that the seven-point stencil yields the optimal monotonicity region in the case of homogeneous medium and uniform grid, indicates that fewer grid cells in the cell stencil may give better monotonicity properties. For general cases, modified transmissibility calculations can possibly be performed, based on interaction of fewer grid cells. This is a topic for future research.

Appendix: Flux expressions for the general O-method

In this appendix, we derive the flux expressions of the general O(η)-method in the case of homogeneous medium and uniform parallelogram grid. The derivation is done for a general location of the continuity points of the potential. As before, the parameter η, defined as the fraction of the cell half-edge from the boundary of the interaction region, is used to describe the location of the continuity points. The case $\eta = 0$ has been derived before [1], and the case below with a general $\eta \in [0, 1)$ follows that derivation closely.

As in [1], we construct around each corner of the grid an interaction region shown by the dashed line in Fig. 18. The fluxes through the half edges in the interaction region are determined by applying linear potentials in each subcell. The flux is required to be continuous, and the potential is required to be continuous at the continuity points $\bar{\boldsymbol{x}}_k$, $k = 1, 2, 3, 4$. The flux through a half edge with normal vector $\boldsymbol{n}_i$ is [1]

$$f_i = -\frac{1}{T_j} \sum_{k=k_1,k_2} \boldsymbol{n}_i^{\mathrm{T}} \boldsymbol{K}_j \boldsymbol{v}_{j,k} (\bar{u}_k - u_j). \tag{68}$$

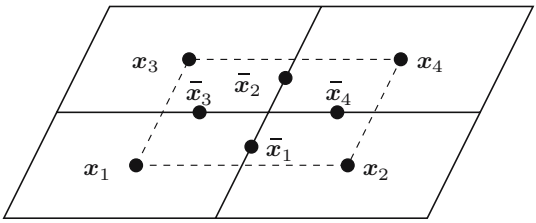

Fig. 18 Local numbering in the interaction region

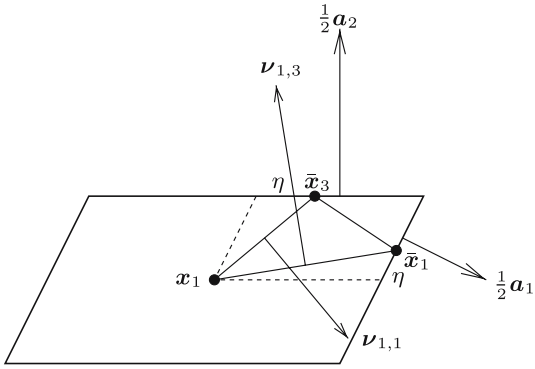

Fig. 19 Vectors of cell 1

Here, u_j is the potential at the cell center $\boldsymbol{x}_j$, and $\bar{u}_k$ is the potential at continuity point $\bar{\boldsymbol{x}}_k$. Further, $\boldsymbol{v}_{j,k}$ are the inner normal vectors of the triangle spanned by the cell center and the continuity points, see Fig. 19, and T_j is twice the area of this triangle. For cell 1, shown in Fig. 19, where $k_1 = 1$ and $k_2 = 3$, these vectors are

$$\boldsymbol{v}_{1,1} = \tfrac{1}{2}(\boldsymbol{a}_1 - \eta\boldsymbol{a}_2), \qquad \boldsymbol{v}_{1,3} = \tfrac{1}{2}(\boldsymbol{a}_2 - \eta\boldsymbol{a}_1), \tag{69}$$

where, as before, $\boldsymbol{a}_i$ are the normal vectors of the parallelogram cell, having length equal to the length of the edges, see Fig. 3. The area T_j is for uniform grids independent of the cell index j. Introducing the "cross-product" matrix

$$\boldsymbol{R} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \tag{70}$$

this area is

$$T = \boldsymbol{v}_{1,1}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{v}_{1,3} = \tfrac{1}{4}\boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{R}\boldsymbol{a}_2(1-\eta^2) = \tfrac{1}{4}V(1-\eta^2), \tag{71}$$

where V is the area of the parallelogram cell. Hence, the flux through the edge at $\bar{\boldsymbol{x}}_1$ in Fig. 19 is

$$\begin{aligned} f_1 &= -\frac{4}{V(1-\eta^2)}\left[\frac{\boldsymbol{a}_1^{\mathrm{T}}}{2}\boldsymbol{K}\frac{\boldsymbol{a}_1 - \eta\boldsymbol{a}_2}{2}(\bar{u}_1 - u_1)\right. \\ &\qquad \left. + \frac{\boldsymbol{a}_1^{\mathrm{T}}}{2}\boldsymbol{K}\frac{\boldsymbol{a}_2 - \eta\boldsymbol{a}_1}{2}(\bar{u}_3 - u_1)\right] \\ &= -\frac{1}{1-\eta^2}[(a-\eta c)(\bar{u}_1 - u_1) + (c-\eta a)(\bar{u}_3 - u_1)], \end{aligned} \tag{72}$$

where the quantities a, b and c are defined in (44). By the same procedure, we may construct all the fluxes through the half edges of the interaction region of Fig. 18. Equating the fluxes of each edge, the following system of equations appears,

$$\begin{aligned}
(1-\eta^2)f_1 &= -(a-\eta c)(\bar{u}_1-u_1)-(c-\eta a)(\bar{u}_3-u_1)\\
&= (a+\eta c)(\bar{u}_1-u_2)-(c+\eta a)(\bar{u}_4-u_2),\\
(1-\eta^2)f_2 &= (a-\eta c)(\bar{u}_2-u_4)+(c-\eta a)(\bar{u}_4-u_4)\\
&= -(a+\eta c)(\bar{u}_2-u_3)+(c+\eta a)(\bar{u}_3-u_3),\\
(1-\eta^2)f_3 &= -(c+\eta b)(\bar{u}_2-u_3)+(b+\eta c)(\bar{u}_3-u_3)\\
&= -(c-\eta b)(\bar{u}_1-u_1)-(b-\eta c)(\bar{u}_3-u_1),\\
(1-\eta^2)f_4 &= (c+\eta b)(\bar{u}_1-u_2)-(b+\eta c)(\bar{u}_4-u_2)\\
&= (c-\eta b)(\bar{u}_2-u_4)+(b-\eta c)(\bar{u}_4-u_4).
\end{aligned} \tag{73}$$

The equations to the right in this system may be written $\boldsymbol{A}\boldsymbol{v}=\boldsymbol{B}\boldsymbol{u}$, where $\boldsymbol{v}=[\bar{u}_1,\bar{u}_2,\bar{u}_3,\bar{u}_4]^{\mathrm{T}}$ and $\boldsymbol{u}=[u_1,u_2,u_3,u_4]^{\mathrm{T}}$. Introducing the 2×2 identity matrix $\boldsymbol{I}$ and the 2×2 matrices

$$\boldsymbol{E}=\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix},\quad \boldsymbol{F}=\begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix}, \tag{74}$$

the matrix $\boldsymbol{A}$ reads

$$\boldsymbol{A}=\begin{bmatrix}2a\boldsymbol{I} & c\boldsymbol{F}-\eta a\boldsymbol{E}\\ c\boldsymbol{F}-\eta b\boldsymbol{E} & 2b\boldsymbol{I}\end{bmatrix}. \tag{75}$$

Further,

$$\boldsymbol{B}=(1-\eta)\begin{bmatrix}a+c & a-c & 0 & 0\\ 0 & 0 & a-c & a+c\\ b+c & 0 & b-c & 0\\ 0 & b-c & 0 & b+c\end{bmatrix}. \tag{76}$$

The inverse of the matrix $\boldsymbol{A}$ is

$$\begin{aligned}
\boldsymbol{A}^{-1}=&\begin{bmatrix}\dfrac{1}{2a}\boldsymbol{I}+\dfrac{c^2}{4a(ab-c^2)}\boldsymbol{F} & -\dfrac{c}{4(ab-c^2)}\boldsymbol{F}\\ -\dfrac{c}{4(ab-c^2)}\boldsymbol{F} & \dfrac{1}{2b}\boldsymbol{I}+\dfrac{c^2}{4b(ab-c^2)}\boldsymbol{F}\end{bmatrix}\\
&+\frac{\eta}{1-\eta^2}\begin{bmatrix}\dfrac{\eta}{4a}\boldsymbol{E} & \dfrac{1}{4b}\boldsymbol{E}\\ \dfrac{1}{4a}\boldsymbol{E} & \dfrac{\eta}{4b}\boldsymbol{E}\end{bmatrix}.
\end{aligned} \tag{77}$$

Thus, $\boldsymbol{v} = \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{u}$, where

$$\begin{aligned}\boldsymbol{A}^{-1}\boldsymbol{B} = &\frac{1}{4}\begin{bmatrix} 2-\eta & 2-\eta & \eta & \eta \\ \eta & \eta & 2-\eta & 2-\eta \\ 2-\eta & \eta & 2-\eta & \eta \\ \eta & 2-\eta & \eta & 2-\eta \end{bmatrix} \\ &+ \frac{c}{4a(1+\eta)}\begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ \eta & -\eta & -\eta & \eta \\ \eta & -\eta & -\eta & \eta \end{bmatrix} + \frac{c}{4b(1+\eta)}\begin{bmatrix} \eta & -\eta & -\eta & \eta \\ \eta & -\eta & -\eta & \eta \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \end{bmatrix}. \end{aligned} \tag{78}$$

The equations to the left in the system of Eqs. (73) may be written $\boldsymbol{f} = \boldsymbol{C}\boldsymbol{v} + \boldsymbol{D}\boldsymbol{u}$, where $\boldsymbol{f} = [f_1, f_2, f_3, f_4]^{\mathrm{T}}$. Here,

$$\boldsymbol{C} = \frac{1}{1-\eta^2}\begin{bmatrix} -(a-\eta c) & 0 & -(c-\eta a) & 0 \\ 0 & a-\eta c & 0 & c-\eta a \\ 0 & -(c+\eta b) & b+\eta c & 0 \\ c+\eta b & 0 & 0 & -(b+\eta c) \end{bmatrix} \tag{79}$$

and

$$\boldsymbol{D} = \frac{1}{1+\eta}\begin{bmatrix} a+c & 0 & 0 & 0 \\ 0 & 0 & 0 & -(a+c) \\ 0 & 0 & -(b-c) & 0 \\ 0 & b-c & 0 & 0 \end{bmatrix}. \tag{80}$$

Hence, the fluxes through the half edges of the interaction region have the form $\boldsymbol{f} = \boldsymbol{T}\boldsymbol{u}$, where

$$\begin{aligned}\boldsymbol{T} = \{\tau_{i,j}\} &= \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B} + \boldsymbol{D} \\ &= \frac{1}{2(1+\eta)}\begin{bmatrix} a & -a & 0 & 0 \\ 0 & 0 & a & -a \\ b & 0 & -b & 0 \\ 0 & b & 0 & -b \end{bmatrix} \\ &\quad + \frac{1}{4(1+\eta)}\begin{bmatrix} c-c^2/b & c+c^2/b & -c+c^2/b & -c-c^2/b \\ c+c^2/b & c-c^2/b & -c-c^2/b & -c+c^2/b \\ c-c^2/a & -c+c^2/a & c+c^2/a & -c-c^2/a \\ c+c^2/a & -c-c^2/a & c-c^2/a & -c+c^2/a \end{bmatrix} \\ &\quad + \frac{\eta}{4(1+\eta)}\begin{bmatrix} a+c & -(a-c) & a-c & -(a+c) \\ a+c & -(a-c) & a-c & -(a+c) \\ b+c & b-c & -(b-c) & -(b+c) \\ b+c & b-c & -(b-c) & -(b+c) \end{bmatrix}. \end{aligned} \tag{81}$$

Having found the fluxes of the half edges, we may express the fluxes through the entire edges by adding the half-edge fluxes. Numbering the cells as in Fig. 1,

the fluxes through the right and the top edges of the central cell in Fig. 1 are

$$\begin{aligned} f_1 &= (\tau_{1,1}+\tau_{2,3})u_1 + (\tau_{1,2}+\tau_{2,4})u_2 + \tau_{1,4}u_3 + \tau_{1,3}u_4 + \tau_{2,1}u_8 + \tau_{2,2}u_9 \\ &= \frac{1}{4(1+\eta)}\Bigg[\left(a(4+2\eta) - \frac{2c^2}{b}\right)(u_1-u_2) - \left(c(1+\eta) + \frac{c^2}{b} + \eta a\right)(u_3-u_8) \\ &\qquad - \left(c(1+\eta) - \frac{c^2}{b} - \eta a\right)(u_4-u_9)\Bigg] \end{aligned} \tag{82}$$

and

$$\begin{aligned} f_2 &= (\tau_{3,1}+\tau_{4,2})u_1 + (\tau_{3,3}+\tau_{4,4})u_4 + \tau_{4,3}u_5 + \tau_{4,1}u_6 + \tau_{3,2}u_2 + \tau_{3,4}u_3 \\ &= \frac{1}{4(1+\eta)}\Bigg[\left(b(4+2\eta) - \frac{2c^2}{a}\right)(u_1-u_4) - \left(c(1+\eta) - \frac{c^2}{a} - \eta b\right)(u_2-u_5) \\ &\qquad - \left(c(1+\eta) + \frac{c^2}{a} + \eta b\right)(u_3-u_6)\Bigg], \end{aligned} \tag{83}$$

respectively. The expressions (82) and (83) may be compared with the expressions (55). Thus, we find for the general O(η)-method,

$$\alpha = \frac{ab\eta + c^2}{2b(1+\eta)}, \quad \beta = \frac{ab\eta + c^2}{2a(1+\eta)}. \tag{84}$$

Adding these expressions, we get

$$\gamma = \alpha + \beta = \frac{ab\eta + c^2}{d(1+\eta)}, \tag{85}$$

where, as before, $d = 2ab/(a+b)$. The expression (85) is used in Table 1.

References

1. Aavatsmark, I.: An introduction to multipoint flux approximations for quadrilateral grids. Comput. Geosci. **6**, 405–432 (2002)
2. Aavatsmark, I., Barkve, T., Bøe, Ø., Mannseth, T.: Discretization on non-orthogonal, quadrilateral grids for inhomogeneous, anisotropic media. J. Comput. Phys. **127**, 2–14 (1996)
3. Aavatsmark, I., Barkve, T., Bøe, Ø., Mannseth, T.: Discretization on unstructured grids for inhomogeneous, anisotropic media. Part I: Derivation of the methods. Part II: Discussion and numerical results. SIAM J. Sci. Comput. **19**, 1700–1736 (1998)
4. Aavatsmark, I., Eigestad, G.T., Klausen, R.A., Wheeler, M.F., Yotov, I.: Convergence of a symmetric MPFA method on quadrilateral grids. Comput. Geosci. (submitted) (2005)
5. Aavatsmark, I., Eigestad, G.T., Klausen, R.A.: Numerical convergence of the MPFA O-method for general quadrilateral grids in two and three dimensions. In: Arnold, D.N., Bochev, P.B., Lehoucq, R.B., Nicolaides, R.A., Shashkov, M. (eds.) Compatible Spatial Discretizations, IMA Vol. Ser., pp. 1–21. Springer, Heidelberg (2006)
6. Arbogast, T., Dawson, C.N., Keenan, P.T., Wheeler, M.F., Yotov, I.: Enhanced cell-centered finite differences for elliptic equations on general geometry. SIAM J. Sci. Comput. **19**, 404–425 (1998)

7. Björk, Å., Dahlquist, G.: Numerical Methods. Prentice-Hall, Englewood Cliffs (1974)
8. Blum, E.K.: Numerical Analysis and Computation. Theory and Practice. Addison-Wesley, Reading (1972)
9. Bramble, J.H., Hubbard, B.E.: On the formulation of finite difference analogues of the Dirichlet problem for Poisson's equation. Numer. Math. **4**, 313–327 (1962)
10. Bramble, J.H., Hubbard, B.E.: On a finite difference analogue of an elliptic boundary problem which is neither diagonally dominant nor of non-negative type. J. Math. Phys. **43**, 117–132 (1964)
11. Bunse, W., Bunse-Gerstner, A.: Numerische Lineare Algebra. Teubner, Stuttgart (1985)
12. Collatz, L.: Funktionalanalysis und numerische Mathematik. Springer, Heidelberg (1964) (English translation, Academic Press, 1966)
13. Edwards, M.G.: Unstructured, control-volume distributed, full-tensor finite-volume schemes with flow based grids. Comput. Geosci. **6**, 433–452 (2002)
14. Edwards, M.G., Rogers, C.F.: Finite volume discretization with imposed flux continuity for the general tensor pressure equation. Comput. Geosci. **2**, 259–290 (1998)
15. Eigestad, G.T., Klausen, R.A.: On the convergence of the multi-point flux approximation O-method: Numerical experiments for discontinuous permeability. Numer. Methods Partial Diff. Equ. **21**, 1079–1098 (2005)
16. Eymard, R., Gallouët, T., Herbin, R.: Convergence of finite volume schemes for semilinear convection diffusion equations. Numer. Math. **82**, 91–116 (1999)
17. Eymard, R., Gallouët, T., Herbin, R.: Finite volume approximation of elliptic problems and convergence of an approximate gradient. Appl. Numer. Math. **37**, 31–53 (2001)
18. Faille, I.: A control volume method to solve an elliptic equation on a two-dimensional irregular mesh. Comput. Methods Appl. Mech. Eng. **100**, 275–290 (1992)
19. Freeze, R.A., Cherry, J.A.: Groundwater. Prentice-Hall, Englewood Cliffs (1979)
20. Graham, A.: Nonnegative matrices and applicable topics in linear algebra. Ellis Horwood, Chichester (1987)
21. Hackbusch, W.: Iterative Lösung großer schwachbesetzter Gleichungssysteme. Teubner, Stuttgart (1991) (English translation, Springer, 1994)
22. Hellwig, G.: Partielle Differentialgleichungen. Teubner, Stuttgart, (1960) (English translation, Blaisdell, 1964)
23. Hopf, E.: Elementare Bemerkungen über die Lösungen partieller Differentialgleichungen zweiter Ordnung vom elliptischen Typus. Sitzungsber. Preuß. Akad. Wiss. **19**, 147–152 (1927)
24. Klausen, R.A., Winther, R.: Convergence of multipoint flux approximations on quadrilateral grids. Numer. Methods Partial Diff. Eq. **22**, 1438–1454 (2006)
25. Klausen, R.A., Winther, R.: Robust convergence of multi point flux approximation on rough grids. Numer. Math. **104**, 317–337 (2006)
26. Nordbotten, J.M.: Sequestration of carbon in saline aquifers; Mathematical and numerical analysis. Thesis, University of Bergen (2004)
27. Nordbotten, J.M., Aavatsmark, I.: Monotonicity conditions for control volume methods on uniform parallelogram grids in homogeneous media. Comput. Geosci. **9**, 61–72 (2005)
28. Nordbotten, J.M., Eigestad, G.T., Monotonicity conditions for control volume methods on general quadrilateral grids; Application to MPFA. In: Proceedings of the 16th Nordic Seminar on Computational Mechanics, Trondheim (2003)
29. Ortega, J.M., Rheinboldt, W.C.: Iterative solution of nonlinear equations in several variables. Academic, New York (1970)
30. Peaceman, D.W.: Fundamentals of numerical reservoir simulation. Elsevier, Amsterdam (1977)
31. Protter, M.H., Weinberger, H.F.: Maximum principles in differential equations. Springer, Heidelberg (1984)
32. Varga, R.S.: On a discrete maximum principle. SIAM J. Numer. Anal. **3**, 355–359 (1966)
33. Weiser, A., Wheeler, M.F.: On convergence of block-centered finite differences for elliptic problems. SIAM J. Numer. Anal. **25**, 351–375 (1988)
34. Wheeler M.F., Yotov I.: A cell-centered finite difference method on quadrilaterals. In: Arnold, D.N., Bochev, P.B., Lehoucq, R.B., Nicolaides, R.A., Shashkov, M. (eds.) Compatible spatial discretizations, IMA Vol. Ser., pp. 189–207. Springer, Heidelberg (2006)