A small stencil and extremum-preserving scheme for anisotropic diffusion problems on arbitrary 2D and 3D meshes

Zhiming Gao, Jiming Wu

Institute of Applied Physics and Computational Mathematics, P.O. Box 8009, Beijing 100088, PR China

Abstract

In this paper a nonlinear extremum-preserving scheme for the heterogeneous and anisotropic diffusion problems is proposed on general 2D and 3D meshes through a certain linearity-preserving approach. The so-called harmonic averaging points located at the interface of heterogeneity are employed to define the auxiliary unknowns. This new scheme is locally conservative, has only cell-centered unknowns and possesses a small stencil, which is five-point on the structured quadrilateral meshes and seven-point on the structured hexahedral meshes. The stability result in $H_1$ norm is obtained under quite general assumptions. Numerical results show that our scheme is robust and extremum-preserving, and the optimal convergence rates are verified on general distorted meshes in case that the diffusion tensor is taken to be anisotropic, at times discontinuous.

1. Introduction

Anisotropic diffusion problem arises in a wide range of scientific fields such as hydrogeology, oil reservoir simulations, plasma physics, semiconductor modeling, and so on. In many cases, diffusion equation is coupled with some other physical processes such as the Lagrange approach in radiation hydrodynamics. In this case, the solution should respect the physical bounds and the computational mesh may be nonconforming and highly distorted, which makes the design of numerical schemes even difficult. Accurate modeling of diffusion processes in these applications requires reliable discretization methods.

In this article, we are interested in constructing a cell-centered finite volume scheme which satisfies the following:

- it is locally conservative;
- it satisfies discrete extremum principle;
- it has a local and small stencil;
- it is simple and easy for coding especially in three dimensions;
- it must be reliable on unstructured anisotropic meshes that may be highly distorted;
- it allows heterogeneous full diffusion tensors;
- it has a second-order accuracy for smooth solutions.

© 2013 Published by Elsevier Inc.
The second property is the most difficult one for a discrete scheme to satisfy, which includes the maximum principle and minimum principle. A closed related concept is the so-called monotone scheme that preserves the nonnegativity of the solution, a special case of the minimum principle. Extremum principle or monotonicity is very important, for example, for diffusion terms in modeling two-phase flows in porous media [22] and for coupling transport equation with a chemical model.

It is well known that classical linear schemes do not always satisfy extremum principle or monotonicity for distorted meshes or with high anisotropy ratio [15]. To our knowledge, there are no linear schemes unconditionally satisfying all of the above requirements. There are several linear schemes [1,7,13,25,28] satisfying some requirements above, but not all of them. The schemes which satisfy the discrete extremum principle impose severe restrictions on the meshes or the diffusion coefficients. To enlarge the class of admissible problems and meshes, some schemes such as the multi-point flux approximation methods (MPFA) [2] use built-in flexibility to increase their monotonicity regions. For classical finite elements, it is explained in [9], that for the Laplacian, the resulting global matrix is an M-matrix if some geometrical constraints are satisfied. Monotone schemes based on slope limiters are proposed in [25]. Two approaches based on repair technique and constrained optimization have been introduced to enforce discrete extremum principle for linear finite element solutions on triangular meshes [20]. A linear scheme [24] satisfying a maximum principle for anisotropic diffusion operators on distorted grids is developed, but this method is generally only first order accurate for smooth solutions. The sufficient condition to ensure the monotonicity of the mimetic finite difference method is analyzed in [17].

Recently, a few nonlinear schemes [6,8,23,19] have been proposed. A nonlinear finite volume scheme proposed in [23] satisfying either the minimum or the maximum principle but not the both simultaneously. The monotonicity of this method for steady state diffusion problems was proven in [18], and 3D extension for this method has been proposed and analyzed in [16]. Further development of the method was made in [26,27,33]. A common property of these methods is that in addition to primary unknowns defined at cell centers, solution values at mesh vertices or edge midpoints are involved. These auxiliary unknowns are usually interpolated from primary cell-centered unknowns. The interpolation technique proves to be a very important issue in these methods and it becomes a very difficult one when the diffusion tensor is discontinuous or polyhedral meshes are involved. The authors in [33] used vertex unknowns as auxiliary unknowns and suggested a piecewise linear approximation to the solution around points where the coefficient is discontinuous. However, as shown in [18,33], the choice of the interpolation method affects the accuracy of the nonlinear scheme even in the case of a constant diffusion coefficient. In [26], a nonlinear extremum-preserving finite volume scheme is proposed, and the edge midpoints are used to define auxiliary unknowns. The interpolation procedure requires that each edge midpoint is located within a triangle formed by three cell centers and located in the smooth area of the solution. When the edge midpoint is on the discontinuity, some special technique is required. Therefore, as pointed out in Remark 1 of [26], this interpolation procedure can be used to deal with only a part of discontinuous diffusion problems. In [27], the cell edge unknowns were used once again as auxiliary ones in the construction of a monotone finite volume scheme. This time, following the idea of MPFA [1], the cell edge unknowns are eliminated by solving a local linear system. However, the solvability of such local linear systems cannot be always guaranteed theoretically [28]. The most important thing is that we do not know whether this interpolation method is positivity-preserving, which is one of the fundamental requirements in the construction of the extremum-preserving or monotone schemes. In most cases, the construction of a second-order positivity-preserving interpolation algorithm is a more challenging task than the design of interpolation-based monotone schemes itself. This is the reason for the authors in [19] to suggest a certain interpolation-free monotone scheme, however, this alternative approach introduces a constraint on the choice of cell centers.

In this article, we further develop and analyze the nonlinear finite volume scheme proposed in [26,27,33]. One key characteristic of our new scheme is that a new and simple interpolation technique is employed to improve the robustness of the scheme for strong heterogeneous and anisotropic diffusion problems and make our scheme to have a small stencil. We introduce, for each cell facet, the so-called harmonic averaging point suggested in [4,13]. Unknowns located at these particular cell facet points are utilized as auxiliary unknowns, and can be interpolated from the two cell-centered unknowns which share this cell facet. The use of harmonic averaging points not only simplifies the interpolation procedure, but also assures it to be a positivity-preserving one. Usually each cell facet has a harmonic averaging point, which allows our one-side flux expression to process a small stencil involving only the present cell and the cells having a common facet with it. This nature makes it easy to implement our scheme on arbitrary and unstructured polygonal meshes or to extend the scheme to three-dimensional polyhedral meshes. The implementation for both 2D and 3D case is almost the same, most of the program modules are in common use, and only some particular modules such as the computation of cell matrix, cell volume and facet area should be written separately. Moreover, the present approach allows us to obtain the stability result in $H_1$ norm for the new scheme.

Our new scheme satisfies all the seven properties mentioned above, and it should be noted that: (1) our scheme has a small stencil for the interior cell-centered unknowns, i.e., a four-point stencil on the triangle meshes, a five-point stencil on the structured quadrilateral meshes, and a seven-point stencil on structured hexahedral meshes; (2) it is linearity-preserving, which means that the scheme provides the exact solution whenever, on each mesh cell, the solution is linear and the diffusion coefficient is constant. This property can be found in many articles for instance in [1,4,13,14,28–31].

The outline of the paper is organized as follows. In Section 2, we state the diffusion problem and give some notations. In Section 3, A small stencil and extremum-preserving finite volume scheme is constructed in five steps. The stability analysis in $H_1$ norm is given in Section 4. Then in Section 5, we present some 2D and 3D numerical experiments to illustrate the
features of the scheme. Finally, some conclusions are given in Section 6 and a new and simple derivation for the harmonic averaging point is given in Appendix A.

2. Steady diffusion problem and notations

We consider a diffusion problem on an open bounded subset \( \Omega \subset \mathbb{R}^d \), \( d = 2 \) or 3,

\[
-\text{div}(\Lambda \nabla u) = f \quad \text{in} \ \Omega, \\
u = g_D \quad \text{on} \ \Gamma_D, \\
-\Lambda \nabla u \cdot n = g_N \quad \text{on} \ \Gamma_N, 
\]

where \( \Lambda(x) \) is a \( d \times d \) diffusion tensor, \( f \in L^2(\Omega) \) is a volumetric source term, \( \partial \Omega = \Gamma_D \cup \Gamma_N \) is the boundary of \( \Omega \), \( n \) denotes the outward unit vector normal to the boundary \( \partial \Omega \) and \( g_D, g_N \) are given scalar functions which are almost everywhere defined on \( \Gamma_D, \Gamma_N \), respectively.

Throughout this paper, a finite volume discretization of \( \Omega \), denoted by \( \mathcal{D} \), is defined as the quaternion \( \mathcal{D} = (\mathcal{M}, \mathcal{E}, \mathcal{O}, \mathcal{P}) \), where

- \( \mathcal{M} = \{K\} \) is a finite family of disjoint open polygonal or polyhedral cells in \( \Omega \) such that \( \Omega = \cup_{K \in \mathcal{M}} K \). For \( K \in \mathcal{M} \), let \( \partial K \) and \( |K| \) denote the cell boundary and measure, respectively.
- \( \mathcal{E} = \{\sigma\} \) is a finite family of disjoint facets in \( \Omega \) such that for \( \sigma \in \mathcal{E} \), \( \sigma \) has a strictly positive \( (d - 1) \)-dimensional measure denoted as \( |\sigma| \). Specifically, these facets are line segments in 2D case and planar polygons in 3D case. Let \( \mathcal{E}^\text{ext} = \mathcal{E} \cap \partial \Omega \). For \( K \in \mathcal{M} \), there exists a subset \( \mathcal{E}_K \) of \( \mathcal{E} \) such that \( \partial K = \cup_{\sigma \in \mathcal{E}_K} \sigma \). \( \mathbf{n}_\sigma \) denotes the unit vector normal to \( \sigma \in \mathcal{E}_K \) outward to \( K \). For \( \sigma \in \mathcal{E}_K \), notation \( \sigma \) may denote either a generic facet on \( \partial K \) or the local number of this same facet in cell \( K \), depending on the context. This slight abuse of notation should not be a source of confusion.
- \( \mathcal{O} = \{x_K, K \in \mathcal{M}\} \) is a set of points, known as cell centers, where \( x_K \in K \).
- \( \mathcal{P} = \cup_{K \in \mathcal{M}} \mathcal{P}_K \), where \( \mathcal{P}_K = \{x_K, \sigma \in \mathcal{E}_K\} \) consists of the interpolation points and \( x_K, \sigma \) is associated with cell \( K \) and facet \( \sigma \).

Approximation of the solution \( u \) at the cell center (or the so-called collocation point) \( x_K \) is known as the primary variable and denoted as \( u_K \). By contrast, \( u_{x_K, \sigma} \), known as the auxiliary variable, is the approximation of \( u \) at the interpolation point \( x_K, \sigma \). As for the flux, we denote by \( F_{x_K, \sigma} \) the approximation of \( -\int_{\sigma} (\Lambda_K \nabla u) \cdot \mathbf{n}_\sigma \, d\mathcal{L} \), where we assume that \( \Lambda \) is constant on each cell \( K \in \mathcal{M} \) with \( \Lambda_K \) denoting the restriction of \( \Lambda \) on \( K \). Throughout, the hollow letters \( \mathcal{A}, \mathcal{F}, \mathcal{I}, \mathcal{M}, \mathcal{N}, \ldots \) will be used to denote matrices with column numbers greater than one, while the bold ones \( \mathbf{F}, \mathbf{U}, \mathbf{I}, \mathbf{n}, \mathbf{x} \), etc., will be employed to denote column vectors.

3. A small stencil and extremum-preserving scheme

In this section, we will construct a nonlinear extremum-preserving finite volume scheme, which consists of five steps (see Section 3.1–3.5).

3.1. Definition of the primary and auxiliary variables

The primary variables are defined at the cell centers. In this article, the barycenter and the geometric center of \( K \) are the usual candidates for the cell center \( x_K \) but not the necessary ones.

The auxiliary variables are defined at the interpolation points, which are usually located on the cell facets. For any facet \( \sigma \in \mathcal{E} \), we associate it with an interpolation point \( y_\sigma \). Specifically, for a boundary facet \( \sigma \in \mathcal{E}^\text{ext} \), let \( y_\sigma \) be the center (midpoint in 2D) of \( \sigma \); for an interior facet \( \sigma \in \mathcal{E}_K \cap \mathcal{E}_L \), define

\[
y_\sigma = \frac{d_{x_K, \sigma} \hat{\lambda}_K^{(0)} \mathbf{x}_K + d_{x_L, \sigma} \hat{\lambda}_L^{(0)} \mathbf{x}_L + d_{x_K, \sigma} d_{x_L, \sigma} \left( \Lambda_K^T - \Lambda_L^T \right) \mathbf{n}_{x_K, \sigma}}{d_{x_K, \sigma} \hat{\lambda}_K^{(0)} + d_{x_L, \sigma} \hat{\lambda}_L^{(0)}},
\]

where \( \hat{\lambda}_K^{(0)} = \mathbf{n}_{x_K, \sigma} \mathbf{A}_K \mathbf{n}_{x_K, \sigma}, \hat{\lambda}_L^{(0)} = \mathbf{n}_{x_L, \sigma} \mathbf{A}_L \mathbf{n}_{x_L, \sigma} \), and \( d_{x_K, \sigma} \) (resp., \( d_{x_L, \sigma} \)) denotes the orthogonal distance from \( x_K \) (resp., \( x_L \)) to \( \sigma \).

We introduce the following assumption:

(A1) For any \( \sigma \in \mathcal{E}_K \cap \mathcal{E}_L \subset \mathcal{E} \), (i) \( K \) (resp., \( L \)) is a star-shaped polygonal or polyhedral cell with respect to \( x_K \) (resp., \( x_L \)); (ii) \( y_\sigma \in \sigma \).

When (A1) holds, \( y_\sigma \) coincides with the harmonic averaging point [4,13,30] and we have

\[
u(y_\sigma) \simeq \frac{d_{x_K, \sigma} \hat{\lambda}_K^{(0)} \mathbf{u}(\mathbf{x}_L) + d_{x_L, \sigma} \hat{\lambda}_L^{(0)} \mathbf{u}(\mathbf{x}_K)}{d_{x_K, \sigma} \hat{\lambda}_K^{(0)} + d_{x_L, \sigma} \hat{\lambda}_L^{(0)}},
\]
where ~ indicates that the corresponding relation satisfies the so-called linearity-preserving criterion, i.e., the truncation error vanishes in the linear case where the solution \( u \) is linear and the diffusion coefficient is constant on any cell \( K \in \mathcal{M} \). In order to drive our main idea clear, we have moved the derivations of (4) and (5) to Appendix A.

Finally, we can choose the set of interpolation points associated with \( K \), \( \mathcal{P}_K = \{ \mathbf{y}_\sigma, \sigma \in \mathcal{E}_K \} \). Due to this special choice of \( \mathcal{P}_K \), there exists only one interpolation point associated with each facet, hence we can always write

\[
\mathbf{u}_{K,\sigma} = \mathbf{u}_{K,\sigma} = \mathbf{u}_\sigma, \text{ if } \sigma = \mathcal{E}_K \cap \mathcal{E}_L; \quad \mathbf{u}_{K,\sigma} = \mathbf{u}_\sigma, \text{ if } \sigma \in \mathcal{E}_K \cap \mathcal{E}^{ext}.
\]

**Remark 3.1.** We point out that when (A1) is violated, (5) may not hold, however, the harmonic averaging point \( \mathbf{y}_\sigma \), defined by (4), is still used as an interpolation point in the present setting.

### 3.2. Construction of one-sided flux in 2D and 3D

For \( K \in \mathcal{M} \), once the cell center \( \mathbf{x}_K \) and the set of interpolation points \( \mathcal{P}_K \) are specified, we can establish through some approach, e.g., the linearity-preserving approach in [28], the relation below

\[
\mathbf{F}_K = \Lambda_K \left( \mathbf{u}_K - \mathbf{U}'_k \right),
\]

where \( \mathbf{F}_K = (F_{K,\sigma}, \sigma \in \mathcal{E}_K)^T \), \( \mathbf{U}'_k = (\mathbf{u}_\sigma, \sigma \in \mathcal{E}_K)^T \), \( \mathbf{U}'_k \) is an \( n_k \) sized vector with components equal to 1, \( \Lambda_K = (\alpha^e_{\sigma \rho})_{\rho \times n_k} \) is referred to as the cell matrix and \( n_k \) is the number of facets in \( \mathcal{E}_K \). Here we point out that, the entries of \( \Lambda_K \) are required to be non-negative to assure the extremum-preserving property of the resulting scheme. Now we will give the construction of \( \Lambda_K \) in 2D and 3D, respectively.

In the 2D case, we first introduce the assumption below

(A2) For any \( \sigma \in \mathcal{E}_K \) and any \( K \in \mathcal{M} \), there exist \( \mathbf{x}_{K,i(\sigma)}, \mathbf{x}_{K,j(\sigma)} \in \mathcal{P}_K \) such that

\[
\Lambda^1_{K} \mathbf{n}_{K,\sigma} = \xi_{i(\sigma)} \left( \mathbf{x}_{K,i(\sigma)} - \mathbf{x}_K \right) + \xi_{j(\sigma)} \left( \mathbf{x}_{K,j(\sigma)} - \mathbf{x}_K \right),
\]

where

\[
\xi_{i(\sigma)} = \frac{\left( \Lambda^1_{K} \mathbf{n}_{K,\sigma} \cdot \mathbf{x}_{K,i(\sigma)} - \mathbf{x}_K \cdot \mathbf{e}_x \right)}{\left( \mathbf{x}_{K,i(\sigma)} - \mathbf{x}_K \right) \cdot \left( \mathbf{x}_{K,j(\sigma)} - \mathbf{x}_K \right)},
\]

\[
\xi_{j(\sigma)} = \frac{\left( \Lambda^1_{K} \mathbf{n}_{K,\sigma} \cdot \mathbf{x}_{K,j(\sigma)} - \mathbf{x}_K \cdot \mathbf{e}_x \right)}{\left( \mathbf{x}_{K,i(\sigma)} - \mathbf{x}_K \right) \cdot \left( \mathbf{x}_{K,j(\sigma)} - \mathbf{x}_K \right)},
\]

are two non-negative scalars, \( (a, b, c) \) denotes the mixed product of vectors \( a, b \) and \( c, e_z \) is a vector parallel to \( (\mathbf{x}_{K,i(\sigma)} - \mathbf{x}_K) \times (\mathbf{x}_{K,j(\sigma)} - \mathbf{x}_K) \).

Under assumption (A2), we can construct a linearity-preserving one-sided flux as follows:

\[
F_{K,\sigma} = |\sigma| \left[ \xi_{i(\sigma)} (\mathbf{u}_K - \mathbf{u}_{K,i(\sigma)}) + \xi_{j(\sigma)} (\mathbf{u}_K - \mathbf{u}_{K,j(\sigma)}) \right], \quad \sigma \in \mathcal{E}_K.
\]

Then, a cell matrix follows immediately from (8). It is evident that the cell matrix has non-negative entries under assumption (A2).

As shown in Fig. 1, let \( \theta^1_{K,\sigma} \) (resp. \( \theta^2_{K,\sigma} \)) denote the angle between \( \mathbf{x}_{K,i(\sigma)} - \mathbf{x}_K \) (resp. \( \mathbf{x}_{K,j(\sigma)} - \mathbf{x}_K \)) and \( \Lambda^1_{K} \mathbf{n}_{K,\sigma} \). Then a sufficient condition for (7) is

\[
0 < \theta^1_{K,\sigma} + \theta^2_{K,\sigma} < \pi
\]

and moreover,

\[
\xi_{i(\sigma)} = \frac{\| \Lambda^1_{K} \mathbf{n}_{K,\sigma} \| \sin \theta^2_{K,\sigma}}{\| \mathbf{x}_{K,i(\sigma)} - \mathbf{x}_K \| \sin \left( \theta^1_{K,\sigma} + \theta^2_{K,\sigma} \right)},
\]

\[
\xi_{j(\sigma)} = \frac{\| \Lambda^1_{K} \mathbf{n}_{K,\sigma} \| \sin \theta^1_{K,\sigma}}{\| \mathbf{x}_{K,j(\sigma)} - \mathbf{x}_K \| \sin \left( \theta^1_{K,\sigma} + \theta^2_{K,\sigma} \right)}.
\]

Paralleled to (A2), the following assumption (A3) is introduced for the 3D case.

(A3) For any \( \sigma \in \mathcal{E}_K \) and any \( K \in \mathcal{M} \), there exist \( \mathbf{x}_{K,i(\sigma)}, \mathbf{x}_{K,j(\sigma)}, \mathbf{x}_{K,k(\sigma)} \in \mathcal{P}_K \) such that

\[
\Lambda^1_{K} \mathbf{n}_{K,\sigma} = \xi_{i(\sigma)} \left( \mathbf{x}_{K,i(\sigma)} - \mathbf{x}_K \right) + \xi_{j(\sigma)} \left( \mathbf{x}_{K,j(\sigma)} - \mathbf{x}_K \right) + \xi_{k(\sigma)} \left( \mathbf{x}_{K,k(\sigma)} - \mathbf{x}_K \right),
\]
where

\[
\zeta_{i|\sigma} = \frac{\left( \mathbf{x}_{K,j|\sigma} - \mathbf{x}_K \cdot \mathbf{x}_{K,k|\sigma} - \mathbf{x}_K \cdot \Lambda_n^T \mathbf{n}_{K,\sigma} \right)}{\left( \mathbf{x}_{K,k|\sigma} - \mathbf{x}_K \cdot \mathbf{x}_{K,j|\sigma} - \mathbf{x}_K \cdot \Lambda_n^T \mathbf{n}_{K,\sigma} \right)},
\]

\[
\zeta_{j|\sigma} = \frac{\left( \mathbf{x}_{K,k|\sigma} - \mathbf{x}_K \cdot \mathbf{x}_{K,j|\sigma} - \mathbf{x}_K \cdot \Lambda_n^T \mathbf{n}_{K,\sigma} \right)}{\left( \mathbf{x}_{K,j|\sigma} - \mathbf{x}_K \cdot \mathbf{x}_{K,k|\sigma} - \mathbf{x}_K \cdot \Lambda_n^T \mathbf{n}_{K,\sigma} \right)},
\]

\[
\zeta_{k|\sigma} = \frac{\left( \mathbf{x}_{K,j|\sigma} - \mathbf{x}_K \cdot \mathbf{x}_{K,k|\sigma} - \mathbf{x}_K \cdot \Lambda_n^T \mathbf{n}_{K,\sigma} \right)}{\left( \mathbf{x}_{K,k|\sigma} - \mathbf{x}_K \cdot \mathbf{x}_{K,j|\sigma} - \mathbf{x}_K \cdot \Lambda_n^T \mathbf{n}_{K,\sigma} \right)},
\]

are all non-negative scalars.

Under assumption (A3), we are able to construct a linearity-preserving one-sided flux as follows:

\[
F_{K,\sigma} = |\sigma| \left[ \zeta_{i|\sigma} \left( u_{K} - u_{i|\sigma} \right) + \zeta_{j|\sigma} \left( u_{K} - u_{j|\sigma} \right) + \zeta_{k|\sigma} \left( u_{K} - u_{k|\sigma} \right) \right],
\]

(10)

by which we obtain a desired cell matrix \(\mathcal{A}_K\) in 3D case.

3.3. A unique definition of the facet flux

For the one-sided flux defined in Section 3.2, we have the splitting

\[
F_{K,\sigma} = F_{K,\sigma}^{(1)} + F_{K,\sigma}^{(2)}, \quad \sigma \in \mathcal{E}_K.
\]

(11)

where

\[
F_{K,\sigma}^{(1)} = (1 - \gamma) a_{\sigma|\sigma}(u_K - u_\sigma)
\]

and

\[
F_{K,\sigma}^{(2)} = \gamma a_{\sigma|\sigma}(u_K - u_\sigma) + \sum_{\sigma' \in \mathcal{E}_K \setminus \sigma} a_{\sigma'|\sigma}(u_K - u_{\sigma'})
\]

with \(\gamma \in [0, 1]\). A unique definition of the facet flux is aimed to achieve local conservation. For an interior facet \(\sigma \in \mathcal{E}_K \cap \mathcal{E}_L\), we use two one-sided fluxes obtained in Section 3.2 to define

\[
\bar{F}_{K,\sigma} = \mu_{K,\sigma} F_{K,\sigma} - \mu_{L,\sigma} F_{L,\sigma}, \quad \bar{F}_{L,\sigma} = \mu_{L,\sigma} F_{L,\sigma} - \mu_{K,\sigma} F_{K,\sigma},
\]

(12)

where \(\mu_{K,\sigma}\) and \(\mu_{L,\sigma}\) are two nonlinear parameters, possibly dependent on \(u\) and given by

\[
\mu_{K,\sigma} = \frac{|F_{K,\sigma}^{(2)}| + \varepsilon}{|F_{K,\sigma}^{(2)}| + |F_{L,\sigma}^{(2)}| + 2\varepsilon}, \quad \mu_{L,\sigma} = \frac{|F_{K,\sigma}^{(2)}| + \varepsilon}{|F_{K,\sigma}^{(2)}| + |F_{L,\sigma}^{(2)}| + 2\varepsilon},
\]

(13)

where \(\varepsilon\) is a small positive number intended to assure non-zero denominators. We mention that nonlinear parameters of this type with \(\varepsilon = 0\) have been used by some authors, see e.g., [6,26]. The use of a small positive \(\varepsilon\) leads to a uniform derivation, avoiding the complicated and tedious discussions for different cases. As for a boundary facet \(\sigma \in \mathcal{E}_K \cap \mathcal{E}^{\text{ext}}\), we simply set

\[
\bar{F}_{K,\sigma} = \mu_{K,\sigma} F_{K,\sigma} \quad \text{with} \quad \mu_{K,\sigma} = 1.
\]

(14)
Obviously, for the above defined facet fluxes, we have the local conservation condition
\[ \mathit{F}_{K,\sigma} + \mathit{F}_{L,\sigma} = 0, \quad \sigma \in \mathcal{E}_K \cap \mathcal{E}_L. \]

Moreover, by substituting (13) and (11) into (12) and through some straightforward calculations, we obtain the unique definition of the facet flux
\[
\begin{align*}
\mathit{F}_{K,\sigma} &= \mu_{K,\sigma} \mathit{F}_{K,\sigma}^{(1)} - \mu_{L,\sigma} \mathit{F}_{L,\sigma}^{(1)} + \mu_{K,\sigma} \left[ 1 - \text{sign} \left( \mathit{F}_{K,\sigma}^{(2)} / \mathit{F}_{L,\sigma}^{(2)} \right) \right] \mathit{F}_{K,\sigma}^{(2)} + O(\varepsilon), \\
\mathit{F}_{L,\sigma} &= \mu_{L,\sigma} \mathit{F}_{L,\sigma}^{(1)} - \mu_{K,\sigma} \mathit{F}_{K,\sigma}^{(1)} + \mu_{L,\sigma} \left[ 1 - \text{sign} \left( \mathit{F}_{K,\sigma}^{(2)} / \mathit{F}_{L,\sigma}^{(2)} \right) \right] \mathit{F}_{L,\sigma}^{(2)} + O(\varepsilon),
\end{align*}
\]
where
\[
\text{sign}(x) = \begin{cases} 
1, & x > 0, \\
0, & x = 0, \\
-1, & x < 0.
\end{cases}
\]

Remark 3.2. Since we are going to drop the O(\varepsilon) terms in (15), the local conservation condition, i.e., formula after (14), will be spoiled so that it holds strictly only for some cases such as \( \varepsilon = 0 \) or \( \mathit{F}_{K,\sigma}^{(2)} / \mathit{F}_{L,\sigma}^{(2)} > 0 \). An alternative splitting instead of (15) can be done as follows:
\[
\begin{align*}
\mathit{F}_{K,\sigma} &= \mu_{K,\sigma} \mathit{F}_{K,\sigma}^{(1)} - \mu_{L,\sigma} \mathit{F}_{L,\sigma}^{(1)} + \tilde{\mu}_{K,\sigma} \left[ 1 - \text{sign} \left( \mathit{F}_{K,\sigma}^{(2)} / \mathit{F}_{L,\sigma}^{(2)} \right) \right] \mathit{F}_{K,\sigma}^{(2)} + O(\varepsilon), \\
\mathit{F}_{L,\sigma} &= \mu_{L,\sigma} \mathit{F}_{L,\sigma}^{(1)} - \mu_{K,\sigma} \mathit{F}_{K,\sigma}^{(1)} + \tilde{\mu}_{L,\sigma} \left[ 1 - \text{sign} \left( \mathit{F}_{K,\sigma}^{(2)} / \mathit{F}_{L,\sigma}^{(2)} \right) \right] \mathit{F}_{L,\sigma}^{(2)} + O(\varepsilon),
\end{align*}
\]
where
\[
\tilde{\mu}_{K,\sigma} = \frac{|\mathit{F}_{K,\sigma}^{(2)}|}{|\mathit{F}_{K,\sigma}^{(2)}| + |\mathit{F}_{L,\sigma}^{(2)}| + 2\varepsilon}, \quad \tilde{\mu}_{L,\sigma} = \frac{|\mathit{F}_{L,\sigma}^{(2)}|}{|\mathit{F}_{K,\sigma}^{(2)}| + |\mathit{F}_{L,\sigma}^{(2)}| + 2\varepsilon}.
\]

Since
\[
\left[ 1 - \text{sign} \left( \mathit{F}_{K,\sigma}^{(2)} / \mathit{F}_{L,\sigma}^{(2)} \right) \right] \left( \tilde{\mu}_{K,\sigma} \mathit{F}_{K,\sigma}^{(2)} + \tilde{\mu}_{L,\sigma} \mathit{F}_{L,\sigma}^{(2)} \right) = 0,
\]
the local conservation condition is still maintained if the O(\varepsilon) terms in the alternative splitting are dropped. The only price we have to pay is the introduction of \( \tilde{\mu}_{K,\sigma} \) and \( \tilde{\mu}_{L,\sigma} \). Our numerical experience indicates that the numerical performance of this alternative approach is as good as that of (15).

3.4. Interpolation of the auxiliary variables

To make the finite volume scheme a cell-centered one, we eliminate the auxiliary variables in the flux expressions by an interpolation procedure. Since we have chosen the harmonic averaging points as the interpolation ones, the interpolation procedure becomes quite simple and natural. Specifically, we choose the following interpolation formula
\[
u_\sigma = \omega_{K,\sigma} u_K + \omega_{L,\sigma} u_L, \quad \sigma \in \mathcal{E}_K \cap \mathcal{E}_L,
\]
where
\[
\omega_{K,\sigma} = \frac{d_{L,\sigma}^2 l_{K,\sigma}^2}{d_{K,\sigma}^2 l_{K,\sigma}^2 + d_{K,\sigma}^2 l_{L,\sigma}^2}, \quad \omega_{L,\sigma} = 1 - \omega_{K,\sigma}.
\]

By (5), (16) satisfies the linearity-preserving criterion if \((A1)\) holds. When \((A1)\) is violated, interpolation formula (16) is still adopted, however, it is no longer a linearity-preserving one. More important is that, since \( \omega_{K,\sigma} \) and \( \omega_{L,\sigma} \) are nonnegative, the positivity-preserving interpolation is assured, which is one of the main differences between our new scheme and other interpolation-based nonlinear monotone or extremum-preserving schemes studied before.

3.5. The finite volume scheme

By (16) and (17) and by recalling the definitions of \( \mathit{F}_{K,\sigma}^{(1)} \) and \( \mathit{F}_{L,\sigma}^{(1)} \), we have
\[
\mu_{K,\sigma} \mathit{F}_{K,\sigma}^{(1)} - \mu_{L,\sigma} \mathit{F}_{L,\sigma}^{(1)} = (1 - \gamma) \left( \mu_{K,\sigma} \omega_{K,\sigma} + \mu_{L,\sigma} \omega_{L,\sigma} \right) (u_K - u_L), \quad \sigma \in \mathcal{E}_K \cap \mathcal{E}_L.
\]
Substituting this into (15) yields
\[ \tilde{F}_{K,\sigma} = \alpha_{\sigma}(u_K - u_L) + \mu_{K,\sigma} \left[ 1 - \text{sign} \left( \frac{F_{K,\sigma}^2}{F_{L,\sigma}^2} \right) \right] F_{K,\sigma} + O(\varepsilon), \]
\[ \tilde{F}_{L,\sigma} = \alpha_{\sigma}(u_L - u_K) + \mu_{L,\sigma} \left[ 1 - \text{sign} \left( \frac{F_{K,\sigma}^2}{F_{L,\sigma}^2} \right) \right] F_{L,\sigma} + O(\varepsilon), \]
where \( \alpha_{\sigma} = (1 - \gamma) \left( \mu_{K,\sigma} \alpha_{\sigma} \alpha_{\sigma} + \mu_{L,\sigma} \alpha_{\sigma} \alpha_{\sigma} \alpha_{\sigma} \right) \) is a non-negative coefficient.

With the definition of \( F_{K,\sigma} \), we formulate the nonlinear finite volume scheme as follows: find \( \{ u_K, K \in \mathcal{M} \} \) such that
\[ \sum_{\sigma \in \mathcal{E}_K} \tilde{F}_{K,\sigma} = \int_K f(x) \, dx \quad \forall K \in \mathcal{M}. \] (19)
By (18) and (14), this finite volume equation can be rewritten as
\[ \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}^\text{ext}} \alpha_{\sigma}(u_K - u_{\sigma}) + \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}^\text{ext}} \alpha_{\sigma}(u_K - u_{\sigma}) + \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}^\text{ext}} \beta_{K,\sigma} F_{K,\sigma}^2 = \int_K f(x) \, dx + O(\varepsilon), \] (20)
where
\[ \beta_{K,\sigma} = \begin{cases} \mu_{K,\sigma} \left[ 1 - \text{sign}(F_{K,\sigma}^2/F_{L,\sigma}^2) \right], & \sigma \in \mathcal{E}_K \cap \mathcal{E}_L \subset \mathcal{E}_K \cap \mathcal{E}^\text{ext}, \\ 1, & \sigma \in \mathcal{E}_K \cap \mathcal{E}^\text{ext}. \end{cases} \]
By (11) and (16) and by neglecting the term \( O(\varepsilon) \), we arrive at the finite volume equation below
\[ \sum_{L \subset \mathcal{M}} A_{K,L}(u_K - u_L) + \sum_{\sigma \in \mathcal{E}^\text{ext}} B_{K,\sigma}(u_K - u_{\sigma}) = \int_K f(x) \, dx \quad \forall K \in \mathcal{M}, \] (21)
where
\[ A_{K,L} = \left\{ \begin{array}{ll} \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_L} \left[ \alpha_{\sigma} + \omega_{L,\sigma} \left( \gamma \beta_{K,\sigma} \alpha_{\sigma} \alpha_{\sigma} + \sum_{\sigma' \in \mathcal{E}_K \cap \mathcal{E}^\text{ext}} \beta_{K,\sigma'} \alpha_{\sigma'} \right) \right], & \mathcal{E}_K \cap \mathcal{E}_L \neq \emptyset, \\ 0, & \mathcal{E}_K \cap \mathcal{E}_L = \emptyset, \end{array} \right. \]
\[ B_{K,\sigma} = \left\{ \begin{array}{ll} \sum_{\sigma' \in \mathcal{E}^\text{ext}} \beta_{K,\sigma} \alpha_{\sigma'}, & \sigma' \in \mathcal{E}_K \cap \mathcal{E}^\text{ext}, \\ 0, & \sigma' \notin \mathcal{E}_K \cap \mathcal{E}^\text{ext}. \end{array} \right. \]
We have the following discrete extremum principle.

**Theorem 3.1.** Assume that
(i) for any \( K \in \mathcal{M} \), the entries of \( A_K \) are all non-negative;
(ii) \( f(x) = 0 \) and the boundary condition is of a pure Dirichlet type;
(iii) the solution of the finite volume equation (21) exists;
(iv) for any \( K, L \in \mathcal{M} \), there exists \( K_1, K_2, \ldots, K_p \in \mathcal{M} \) such that \( A_{K,K_1} A_{K_1,K_2} \cdots A_{K_p,L} \neq 0 \). Moreover, there exist \( K \in \mathcal{M} \) and \( \sigma' \in \mathcal{E}^\text{ext} \) such that \( B_{K,\sigma'} \neq 0 \).

Then, one of the following alternatives holds true
(a) for any \( K \in \mathcal{M} \), \( \min_{\sigma \in \mathcal{E}^\text{ext}} \{ u_\sigma \} < u_K < \max_{\sigma \in \mathcal{E}^\text{ext}} \{ u_\sigma \} \);
(b) for all \( K \in \mathcal{M} \), \( u_K = c \) and \( \min_{\sigma \in \mathcal{E}^\text{ext}} \{ u_\sigma \} \leq c \leq \max_{\sigma \in \mathcal{E}^\text{ext}} \{ u_\sigma \} \).

**Proof.** The proof is similar to, say, that of Theorem 2 in [5] so that we just give a sketch. Obviously, under the assumptions we have \( A_{K,L} \geq 0, B_{K,\sigma} \geq 0 \) for any \( K, L \in \mathcal{M}, \sigma' \in \mathcal{E}^\text{ext} \). Suppose that (a) is violated so that we have \( u_K \geq \max_{\sigma \in \mathcal{E}^\text{ext}} \{ u_\sigma \} \) for some \( K \in \mathcal{M} \). Similarly, in this case we can choose \( K \) such that \( u_K = c = \max_{\sigma \in \mathcal{E}^\text{ext}} \{ u_\sigma \} \). By (ii) and (21), we deduce that \( u_L = c \) for all \( A_{K,L} > 0 \) and \( u_{\sigma'} = c \) for \( B_{K,\sigma'} > 0 \). Repeating the argument and using (iv) we reach (b). If the violation of (a) is the case \( u_K \leq \min_{\sigma \in \mathcal{E}^\text{ext}} \{ u_\sigma \} \), the argument is similar, which completes the proof. \( \square \)

**Remark 3.3.** For orthogonal meshes and anisotropic diffusion tensor aligned with the mesh, the cell matrix \( A_K = (\alpha_{\sigma' \sigma})_{\mathcal{E}_K \times \mathcal{E}_K} \) is diagonal so that we have \( F_{K,\sigma}^2 = 0 \) in (11) when \( \gamma = 0 \), and then \( \mu_{K,\sigma} = \mu_{L,\sigma} = 1/2 \). From (18), our scheme reduces to the
classical two-point flux approximation scheme (TPFA). If $\gamma > 0$, we get a nonlinear two-point flux approximation from (18), (11) and (16) by a similar argument.

3.6. Discrete system

By (21), we get a nonlinear algebraic system

$$M(U)U = F(U),$$

where $U$ denotes the vector unknowns and $M(U)$ the coefficient matrix. The right-hand side vector $F(U)$ is generated by the source term and the boundary data.

The nonlinear system may be solved by a number of different methods. We use Picard iterations in this article: choose a small value $\epsilon_{\text{mon}} > 0$ and initial solution vector $U^0$ with positive entries, and repeat for $k = 0, 1, 2, \ldots$.

1. Assemble the global matrix $M(U^0)$, and calculate the right-hand side vector $F(U^0)$.
2. Solve $M(U^k)U^{k+1} = F(U^k)$ to obtain $U^{k+1}$.
3. Stop if $\|M(U^{k+1})U^{k+1} - F(U^{k+1})\| \leq \epsilon_{\text{mon}}\|M(U^0)U^0 - F(U^0)\|$.

The linear system in Step 2 with the non-symmetric matrix $M(U^k)$ is solved by GMRES method, and the GMRES iterations are terminated when the relative norm of the initial residual becomes smaller than a small positive parameter $\epsilon_{\text{lin}}$.

In addition, it is obvious that in each Picard iteration, we have to solve the following linearized equation which comes from (21)

$$\sum_{K \in \mathcal{M}} B_{K}^{(k)}(u_{K}^{(k+1)} - u_{K}^{(k+1)}) + \sum_{\sigma \in \mathcal{E}_d} B_{K,\sigma}^{(k)}(u_{K}^{(k+1)} - u_{\sigma}) = \int_{K} f(x) \, dx, \quad \forall K \in \mathcal{M}. \tag{22}$$

where $k$ denotes the nonlinear iteration index. Obviously, the result in Theorem 3.1 still holds for this equation. According to numerical evidence, the Picard iterations always converge provided that the linear system is solved with a very small tolerance $\epsilon_{\text{lin}}$.

Remark 3.4. We give some interesting facts for the linearized equation (22).

- (22) has a small stencil that involves only cell $K$ and the cells that share a common facet with $K$. Particularly, for the structured hexahedral meshes, (22) leads to a seven-point stencil. This nature makes the computations of the linear system, including the matrix structure, the assembling of the non-zero matrix entries and the right-hand side vector, to be essentially the same for both 2D and 3D cases and as a result, it is possible to write a uniform code.
- Computations dependent on dimensions are mainly those of the harmonic averaging point and the cell matrix $\lambda_K$. From the definitions of the harmonic averaging point and the cell matrix, one can see that computations of this kind can be done locally and only the information of a single facet or cell is involved.
- As for the data structure for the mesh topology, the bilateral relations between cells and facets are fundamental and different cells are related through their common facet, instead of their common vertices or common edges in 3D.

Due to the above facts, the implementation of our new scheme in 3D becomes as easy as that of its 2D counterpart and the handling of the complex mesh topology around vertices or 3D cell edges is avoided.

4. Stability analysis

In this section, we shall study the stability of the new scheme by following the idea proposed in our previous paper [30]. For simplicity, we just consider 2D case and diffusion equation (1) with full homogeneous Dirichlet boundary condition. Problems on 3D meshes or with other types of boundary conditions can be discussed analogously.

To begin with, we introduce some notations:

- $X(\mathcal{M})$ denotes the set of discrete functions that have constant values corresponding to each element of $\mathcal{M}$;
- $\mathcal{E}_K$ is an $n_e \times n_e$ diagonal matrix associated with cell $K$ whose $\sigma$th diagonal entry is $|\sigma|/d_{\sigma}$, where $d_{\sigma} = d_{K,\sigma} + d_{L,\sigma}$ if $\sigma = \mathcal{E}_K \cap \mathcal{E}_L$ and $d_{\sigma} = d_{K,\sigma}$ if $\sigma \in \mathcal{E}_K \cap \mathcal{E}'$;
- $\mathcal{M}_K = (m_{K,\sigma})$ is an $n_e \times n_e$ diagonal matrix with $m_{\sigma\sigma} = \mu_{K,\sigma}$;
- $\mathcal{W}_K$ is an $n_e \times n_e$ diagonal matrix whose $\sigma$th diagonal entry is $\omega_{L,\sigma}$ if $\sigma = \mathcal{E}_L \cap \mathcal{E}_K$ and 1 otherwise;
- $U_K$ denotes an $n_e$-sized vector, whose $\sigma$th component is $u_{L}$ if $\sigma = \mathcal{E}_K \cap \mathcal{E}_L$ and vanishes if $\sigma \in \mathcal{E}'$.
For \( u_M = \{u_K, K \in M\} \in X(M) \), we define the following two norms:

- **Discrete \( L_2 \) norm**
  \[
  \|u_M\|_{0,M} = \left( \sum_{K \in M} |K| u_K^2 \right)^{1/2} = \|u_M\|_{0,\Omega}.
  \] (23)

where \( \| \cdot \|_{0,\Omega} \) denotes the standard \( L_2 \) norm.

- **Discrete \( H_1 \) norm**
  \[
  \|u_M\|_{1,M} = \left( \sum_{K \in M} \| \partial_K (u_K - u_K) \|_2^2 \right)^{1/2}.
  \] (24)

In addition, we introduce the following assumptions:

**(A4)** For the finite volume discretization \( D = (M, \mathcal{E}, \mathcal{O}, \mathcal{P}) \), there exist constants \( \alpha, \beta \) and \( \gamma \), such that

\[
|K| \geq \alpha h^2, \quad n_K \leq \gamma, \quad \forall K \in M \quad \text{and} \quad |\sigma| \geq \beta h, \quad \forall \sigma \in \mathcal{E}.
\]

where \( h = \max_{K \in M} (h_K) \) denotes the mesh size and \( h_K \) is the diameter of \( K \).

**(A5)** There exist positive constants \( q \) and \( \bar{q} \), independent of \( K, \sigma \) and \( h \), such that

\[
\bar{q} \leq \frac{|\sigma|}{\alpha}, \quad \forall \sigma \in \mathcal{E}_K, \quad \forall K \in M.
\]

**(A6)** The symmetric part of \( \delta K \delta K \delta K \), is positive definite, i.e., there exists a positive constant \( \sigma_K \), such that

\[
\frac{1}{2} \nu^T (\delta K \delta K \delta K + \delta K \delta K \delta K) \nu \geq \sigma_K \| \nu \|^2, \quad \forall \nu \in \mathbb{R}^{n_K}, \quad \forall K \in M.
\] (25)

Under assumption **(A4)**, we have the following discrete Poincaré inequality

\[
\|u_M\|_{0,M} \leq C_F \|u_M\|_{1,M}, \quad \forall u_M \in X(M),
\] (26)

where \( C_F \) is a positive constant, independent of \( u_M \) and \( h \). A similar proof of (26) can be found in [10].

**Theorem 4.1 (Stability of the scheme).** Assume that \( \Gamma_D = \partial \Omega \) with \( g_D = 0 \), and assumptions **(A4)-(A6)** hold. Let \( u_M = \{u_K, K \in M\} \in X(M) \) be the solution of the finite volume equation (19), we have

\[
\|u_M\|_{1,M} \leq \frac{C_F \bar{q}}{\min_{K \in M} \|f\|_{0,\Omega}} \|f\|_{0,\Omega}.
\] (27)

**Proof.** Multiplying the finite volume equation (19) with \( u_K \), summing over all cells in \( M \), we obtain

\[
\sum_{K \in M} \sum_{x \in K} u_K \tilde{f}_{K,x} \sigma = \sum_{K \in M} u_K \int_K f(x) dx.
\] (28)

For the right-hand side of (28), we have

\[
\sum_{K \in M} u_K \int_K f(x) dx \leq \left( \sum_{K \in M} |K| u_K^2 \right)^{1/2} \left( \sum_{K \in M} \frac{1}{|K|} \left( \int_K f(x) dx \right)^2 \right)^{1/2} \leq \|u_M\|_{0,M} \|f\|_{0,\Omega} \leq C_F \|u_M\|_{1,M} \|f\|_{0,\Omega}.
\] (29)

where we have used twice the Cauchy–Schwartz inequality and discrete Poincaré inequality (26).

On the other hand, we give the estimate of the left-hand side of (28). Firstly, using the notations introduced in this section, (16) can be rewritten as

\[
U_K^c = \mathbb{W}_K U_K + u_K (I_K - \mathbb{W}_K I_K).
\] (30)

Substituting (30) into (6) yields

\[
F_K = A_K \mathbb{W}_K (u_K I_K - U_K^c).
\] (31)
Secondly, by (12) and (14), we have
\[
\sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} u_k \hat{F}_K, \sigma = \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} (u_k - u_l) \mu_{K, \sigma} F_{K, \sigma} + \sum_{\sigma \in \mathcal{E}_K} u_k \mu_{K, \sigma} F_{K, \sigma} = \sum_{K \in \mathcal{M}} (u_k I_K - U_k)^T M_K F_K.
\]

Now substituting (31) into the above equation, and employing assumptions (A5) and (A6), we obtain
\[
\sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} u_k \hat{F}_K, \sigma = \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{E}_K} (u_k I_K - U_k)^T M_K A_K W_K (u_k I_K - U_k) = \frac{1}{2} \sum_{K \in \mathcal{M}} (u_k I_K - U_k)^T (M_K A_K W_K + W_K A_K^T M_K) (u_k I_K - U_k) \\
\geq \sum_{K \in \mathcal{M}} \sigma_K \| (u_k I_K - U_k) \|^2 \geq \frac{1}{q} \min_{K \in \mathcal{M}} \sigma_K \sum_{K \in \mathcal{M}} \left\| Q_k^\frac{1}{2} (u_k I_K - U_k) \right\|^2 = \frac{1}{q} \min_{K \in \mathcal{M}} \| u_M \|^2_{1, M}.
\]

Combining this with (29), we obtain (27), which ends the proof. □

Here we must point out that the stability result given in Theorem 4.1 is preliminary and has certainly a room to be improved. The main issue is the lack of a uniform bound of the nonlinear parameter \( \sigma \). This problem can be remedied somewhat by a proper selection of \( \gamma \) in (11), see the last subsection of the next section.

5. Numerical examples

In this section, we present some numerical results to demonstrate efficiency and robustness of the small stencil and extremum-preserving scheme (BSEP5 for short) on various mesh families.

If numerical tests presented in this section have analytical solutions, absolute error in the discrete extremum principle in this test.


\[ \text{log}(E_\delta(h_2)/E_\delta(h_1)) \]
\[ \text{log}(h_2/h_1), \]

where \( h_1, h_2 \) denote the mesh sizes of the two successive meshes, and \( E_\delta(h_1), E_\delta(h_2) \) the corresponding discrete errors.

Without special mention, all tests are performed for the case \( \gamma = 0 \) (see (11)) in double precision, and we use GMRES for solving linear systems with stopping tolerance \( \varepsilon_{\text{max}} = 1.0E^{-15} \). The nonlinear iterations are terminated when the reduction of the initial residual norm becomes smaller than \( \varepsilon_{\text{max}} = 1.0E^{-07} \).

In addition, we use the following notations for the numerical tests:

- \text{nunkw}: number of unknowns;
- \text{nnmat}: number of non-zero terms in the coefficient matrix;
- \text{umin}: minimal value of the approximate solution;
- \text{umax}: maximal value of the approximate solution;
- \text{nlin}: number of nonlinear iterations;
- \text{itn}: number of linear iterations in the last nonlinear iteration;
- \text{as}: averaging stencil which equals to the ratio between \text{nnmat} and \text{nunkw}.

5.1. Test 1: mild anisotropy

We consider a test similar to test 1,2 of the benchmark on the fifth conference on discretization schemes for anisotropic diffusion problems on general grids [15] (FVCA V for short). We consider the problem (1)–(2) and \( \Omega = [0, 1]^2 \). A homogeneous anisotropic tensor and the exact solution are given below:

\[ K = \begin{pmatrix} 1.5 & 0.5 \\ 0.5 & 1.5 \end{pmatrix}, \quad u(x, y) = \frac{1}{2} \left[ \frac{\sin((1-x)(1-y))}{\sin(1)} + (1-x)^2(1-y)^2 \right], \]

where the exact solution is located in the interval \([0, 1]\).

We use a sequence of the strictly acute triangular mesh Mesh1, Shestakov mesh Mesh2, randomly perturbed mesh Mesh3 with large distortion, polygonal mesh Mesh4 and locally refined non-conforming rectangular mesh Mesh5 in this numerical test (see Fig. 2), and the mesh refinement levels are also given.

The numerical results for our scheme with two choices of cell centers (geometric center geomc and barycenter baryc) on four mesh types are given in Table 1 which shows the following:

- The minimum and maximum solutions are given on the coarsest meshes, we can see that our scheme satisfies the discrete extremum principle in this test.
The sixth column shows that on the triangle mesh \( \text{Mesh1} \), the ratio \( \text{as} \approx 4 \) and on the quadrilateral meshes \( \text{Mesh2} \) and \( \text{Mesh3} \), the ratio \( \text{as} \approx 5 \), which conform that our scheme \( \text{SSEPS} \) has a four-point stencil for the triangle mesh and five-point stencil for the structured quadrilateral mesh.

Our scheme \( \text{SSEPS} \) with two choices of cell center shows a second (resp. first) order convergence rate with respect to the discrete \( L_2 \) (resp. \( H_1 \)) norm of the solution errors on the four types of meshes.

Numbers of nonlinear iterations \( \text{nitn} \) are given in the fourth column. In the fifth column, numbers of linear iterations \( \text{itn} \) on four mesh types are nearly in the same level. Moreover, we find that the average number of linear iterations in all nonlinear iterations is very close to \( \text{itn} \) and it is still true for the rest numerical tests.

---

**Table 1**

Results for Test 1 on \( \text{Mesh1-Mesh4} \) with different choices of cell centers.

<table>
<thead>
<tr>
<th>Mesh (cell center)</th>
<th>( \text{umin} )</th>
<th>( \text{umax} )</th>
<th>( \text{nitn} )</th>
<th>( \text{itn} )</th>
<th>( \text{as} )</th>
<th>( R_u )</th>
<th>( R_q )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{Mesh1(geom)} )</td>
<td>2.80E-03</td>
<td>0.754</td>
<td>12</td>
<td>39</td>
<td>3.982</td>
<td>1.929</td>
<td>1.025</td>
</tr>
<tr>
<td>( \text{Mesh1(bary)} )</td>
<td>2.80E-03</td>
<td>0.754</td>
<td>12</td>
<td>39</td>
<td>3.982</td>
<td>1.929</td>
<td>1.025</td>
</tr>
<tr>
<td>( \text{Mesh2(geom)} )</td>
<td>1.16E-03</td>
<td>0.826</td>
<td>71</td>
<td>42</td>
<td>4.969</td>
<td>2.304</td>
<td>1.032</td>
</tr>
<tr>
<td>( \text{Mesh2(bary)} )</td>
<td>1.27E-03</td>
<td>0.821</td>
<td>68</td>
<td>44</td>
<td>4.969</td>
<td>2.337</td>
<td>1.031</td>
</tr>
<tr>
<td>( \text{Mesh3(geom)} )</td>
<td>1.84E-03</td>
<td>0.810</td>
<td>19</td>
<td>40</td>
<td>4.969</td>
<td>1.986</td>
<td>1.027</td>
</tr>
<tr>
<td>( \text{Mesh3(bary)} )</td>
<td>1.88E-03</td>
<td>0.808</td>
<td>14</td>
<td>39</td>
<td>4.969</td>
<td>1.993</td>
<td>1.022</td>
</tr>
<tr>
<td>( \text{Mesh4(geom)} )</td>
<td>4.19E-04</td>
<td>0.906</td>
<td>7</td>
<td>36</td>
<td>6.938</td>
<td>1.814</td>
<td>1.240</td>
</tr>
<tr>
<td>( \text{Mesh4(bary)} )</td>
<td>2.42E-04</td>
<td>0.906</td>
<td>6</td>
<td>35</td>
<td>6.938</td>
<td>1.746</td>
<td>1.293</td>
</tr>
</tbody>
</table>

- The sixth column shows that on the triangle mesh \( \text{Mesh1} \), the ratio \( \text{as} \approx 4 \) and on the quadrilateral meshes \( \text{Mesh2} \) and \( \text{Mesh3} \), the ratio \( \text{as} \approx 5 \), which conform that our scheme \( \text{SSEPS} \) has a four-point stencil for the triangle mesh and five-point stencil for the structured quadrilateral mesh.
- Our scheme \( \text{SSEPS} \) with two choices of cell center shows a second (resp. first) order convergence rate with respect to the discrete \( L_2 \) (resp. \( H_1 \)) norm of the solution errors on the four types of meshes.
- Numbers of nonlinear iterations \( \text{nitn} \) are given in the fourth column. In the fifth column, numbers of linear iterations \( \text{itn} \) on four mesh types are nearly in the same level. Moreover, we find that the average number of linear iterations in all nonlinear iterations is very close to \( \text{itn} \) and it is still true for the rest numerical tests.
The percentage of harmonic averaging points that are outside mesh edges on five successive mesh levels $h_1$–$h_5$ are given in Table 2. We find that even though the mesh is too distorted (such as Mesh2 and Mesh3) and some of mesh cells may be non-convex, the percentage of harmonic averaging points that are outside mesh edges is very low. In this case, we can still use formula (4) and the resulting scheme can also have expected accuracy as shown in Table 1.

The convergence rates for the discrete $L_2$-norm and $H_1$-norm of solution errors are graphically depicted in Figs. 3–5 as log–log plots of the solution errors versus the square root of the number of unknowns $n_{unkw}$, square root of the number of non-zero matrix terms $n_{nnmat}$, and the inverse of mesh size $h$. The actual convergence orders are reflected by the slopes of the experimental error curves. Three figures show a second order convergence rate with respect to the discrete $L_2$-norm.

<table>
<thead>
<tr>
<th>Mesh</th>
<th>$h_1$</th>
<th>$h_2$</th>
<th>$h_3$</th>
<th>$h_4$</th>
<th>$h_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mesh1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mesh2(geom)</td>
<td>0</td>
<td>0.18%</td>
<td>0.19%</td>
<td>0.16%</td>
<td>0.16%</td>
</tr>
<tr>
<td>Mesh2(baryc)</td>
<td>0</td>
<td>0.37%</td>
<td>0.19%</td>
<td>0.22%</td>
<td>0.26%</td>
</tr>
<tr>
<td>Mesh3(geom)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.003%</td>
</tr>
<tr>
<td>Mesh3(baryc)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.02%</td>
<td>0.03%</td>
</tr>
<tr>
<td>Mesh4(geom)</td>
<td>1.64%</td>
<td>2.07%</td>
<td>1.53%</td>
<td>1.56%</td>
<td>1.32%</td>
</tr>
<tr>
<td>Mesh4(baryc)</td>
<td>0.62%</td>
<td>0.69%</td>
<td>1.01%</td>
<td>0.45%</td>
<td>0.23%</td>
</tr>
</tbody>
</table>

Fig. 3. Test 1: solution errors in $L_2$ and $H_1$ norm versus $n_{unkw}$.

Fig. 4. Test 1: solution errors in $L_2$ and $H_1$ norm versus $n_{nnmat}$.
and first order convergence rate with respect to the discrete $H_1$-norm on four mesh types. Note that when computing the solution error with respect to $\text{nunkw}$ and $\text{nnmat}$, convergence rates remains the same as expected but the relative position varies.

In addition, from Table 3, we find that similar results are observed on the locally refined non-conforming rectangular mesh Mesh5 when using geometric center as cell center.

Finally, we find that the results for barycenter are generally a little better than those for geometric center and we suggest the use of barycenter as cell center in our scheme. However, for simplicity of presentation, only the numerical results for geometric center will be presented in the following numerical tests.

5.2. Test 2: discontinuous anisotropy

In this subsection, we investigate the convergence of our method for a problem with a discontinuous diffusion tensor. We solve the problem (1) with the full Dirichlet boundary condition (2) and $\Omega = [0, 1]^2$. Let the diffusion tensor $A$ change the eigenvalues and orientation of eigenvectors across the line $x = 0.5$,

$$ A = \begin{cases} 
  \begin{pmatrix} 10 & 2 \\ 2 & 5 \end{pmatrix}, & x \leq 0.5, \\
  \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, & x > 0.5. 
\end{cases} $$

We choose the following exact solution

$$ u(x, y) = \begin{cases} 
  [1 + (x - 0.5)(0.1 + 8\pi(y - 0.5))] \exp \left( -20\pi(y - 0.5)^2 \right), & x \leq 0.5, \\
  \exp(x - 0.5) \exp \left( -20\pi(y - 0.5)^2 \right), & x > 0.5. 
\end{cases} $$

This test is rebuilt from a numerical experiment in [32]. The numerical tests were performed on triangle mesh Mesh1, polygonal mesh Mesh4 and modified Shestakov mesh Mesh6 (see Fig. 2(f)).

The number of nonlinear iterations $\text{nitr}$, number of linear iterations $\text{itn}$ and the convergence results with respect to $L_2$-norm and $H_1$-norm of solution errors on five successive mesh levels are presented in Table 4. On the last row of the table,
Test 2: numbers of nonlinear and linear iterations, convergence results on three mesh types Mesh1, Mesh4 and Mesh6.

<table>
<thead>
<tr>
<th>$h$</th>
<th>Mesh1</th>
<th></th>
<th>Mesh4</th>
<th></th>
<th>Mesh6</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>nitn</td>
<td>itn</td>
<td>$E_u$</td>
<td></td>
<td>nitn</td>
<td>itn</td>
</tr>
<tr>
<td></td>
<td>2.50E-1</td>
<td>2.18E0</td>
<td>7</td>
<td>2</td>
<td>3.01E-1</td>
<td>3.61E0</td>
</tr>
<tr>
<td>$h_2$</td>
<td>12</td>
<td>7</td>
<td>4.22E-2</td>
<td>8.67E-1</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>$h_3$</td>
<td>16</td>
<td>10</td>
<td>1.38E-2</td>
<td>4.35E-1</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>$h_4$</td>
<td>16</td>
<td>18</td>
<td>3.92E-3</td>
<td>1.79E-1</td>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td>$h_5$</td>
<td>9</td>
<td>33</td>
<td>1.58E-3</td>
<td>8.01E-2</td>
<td>3</td>
<td>38</td>
</tr>
<tr>
<td>Rates</td>
<td>1.827</td>
<td>1.191</td>
<td>1.997</td>
<td>1.420</td>
<td>2.280</td>
<td>1.494</td>
</tr>
</tbody>
</table>

5.3. Test 3: oblique flow using boundary conditions with steep gradients

In this test, we consider test 3 described in FVCA V[15]. In this test, the main directions of an anisotropic diffusion tensor are tilted with respect to the boundary conditions and the mesh. The computational domain is $[0, 1]^2$, and the anisotropic tensor is

$$\Lambda = R_\theta \begin{pmatrix} 1 & 0 \\ 0 & 10^{-4} \end{pmatrix} R_\theta^{-1},$$

where $R_\theta$ is the rotation of angle $\theta = 40^\circ$, and the source term $f = 0$. Dirichlet boundary conditions are considered, $u = g_D$ on $\partial\Omega$, with $g_D$ a continuous and piecewise linear function defined by

$$g_D(x, y) = \begin{cases} 
1 & \text{on } (0.2) \times \{0\} \cup \{0\} \times (0.2), \\
0 & \text{on } (0.8) \times \{0\} \cup \{0\} \times (0.8), \\
0.5 & \text{on } (0.3) \times \{0\} \cup \{0\} \times (0.3), \\
0.5 & \text{on } (0.7) \times \{0\} \cup \{0\} \times (0.7).
\end{cases}$$

The solution features a Z across the $y = x$ axis. We will give the results on the uniform triangle mesh Mesh7 and small disturbed random mesh Mesh8 (see Fig. 6).

Fig. 7 shows the solution behavior of classical MPFA-O method [1] and our scheme SSEPS on Mesh7 with 128 triangles. We observe oscillations on the MPFA-O method (see Fig. 7(a)), and they disappear with SSEPS scheme. Note that if we choose a uniform rectangular mesh, no oscillations appear with MPFA-O method. In addition, solution profile with SSEPS on the fine triangle mesh Mesh7 is given in Fig. 8 with $u_{\text{min}} = 0.57E-04$ and $u_{\text{max}} = 0.996$.

Fig. 9 represents the solution profile with SSEPS scheme on the coarse and fine random meshes Mesh8. On the coarse mesh, we have $u_{\text{min}} = 6.54E-02$, $u_{\text{max}} = 0.926$, and on the fine mesh we have $u_{\text{min}} = 1.11E-03$, $u_{\text{max}} = 0.999$.

Finally, although the extremum principle is not easy to verify for such a solution and is violated on coarse and fine mesh by many schemes (Table 4 in [15]), our scheme SSEPS provides satisfactory results and satisfies the discrete extremum principle in this test.

![Fig. 6](a) Mesh7: (3.54E-01, 2.21E-02) ![Fig. 6](b) Mesh8: (4.66E-01, 3.22E-02)
5.4. Test 4: problem with point source on perturbed parallelograms

This test was given in [3,15], and is meant to test the schemes for the violation of the maximum principle within the domain. Problem (1) is considered with the diffusion tensor $K = \text{Id}$ and the full Dirichlet boundary condition (2) on the domain.

![Figure 7](image1.png)

**Fig. 7.** Test 3: solution profile with MPFA and SSEPS schemes on Mesh7: 128 triangular cells. (White color denotes position of values greater than 1 and pink color denotes position of negative values.) (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

![Figure 8](image2.png)

**Fig. 8.** Test 3: solution profile with SSEPS scheme on the fine triangle mesh Mesh7: 8192 cells.

![Figure 9](image3.png)

**Fig. 9.** Test 3: solution profile with SSEPS scheme on the coarse and fine meshes Mesh8.

5.4. Test 4: problem with point source on perturbed parallelograms

This test was given in [3,15], and is meant to test the schemes for the violation of the maximum principle within the domain. Problem (1) is considered with the diffusion tensor $K = \text{Id}$ and the full Dirichlet boundary condition (2) on the
As shown in Fig. 11, a perturbed parallelogram mesh is employed and the source term $f$ is equal to zero in all cells except cell $x^* = (6, 6)$ where

$$\int_{cell(6,6)} f(x) \, dx = 1.$$  

Note that the solution should be a function with a maximum in cell $x^* = (6, 6)$. If the solution shows internal oscillations or is negative, Hopf’s first lemma is violated.

Table 5
Test 4: minimum and maximum solutions of schemes on the parallelogram mesh (Table 12 in [15]).

<table>
<thead>
<tr>
<th>Scheme</th>
<th>$u_{min}$</th>
<th>$u_{max}$</th>
<th>Scheme</th>
<th>$u_{min}$</th>
<th>$u_{max}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fine grid</td>
<td>1.07E−24</td>
<td>4.10E−01</td>
<td>FVHYB</td>
<td>−3.38E−02</td>
<td>1.12E−01</td>
</tr>
<tr>
<td>CMPFA</td>
<td>−2.31E−02</td>
<td>1.03E−01</td>
<td>FVSYM</td>
<td>−7.21E−02</td>
<td>1.52E−01</td>
</tr>
<tr>
<td>CVFE</td>
<td>−1.23E−03</td>
<td>4.24E−02</td>
<td>FVPMMD</td>
<td>1.22E−09</td>
<td>3.99E−01</td>
</tr>
<tr>
<td>DDFV-BHU</td>
<td>−1.25E−03</td>
<td>8.22E−02</td>
<td>MFD-BLS</td>
<td>−1.03E−01</td>
<td>1.85E−01</td>
</tr>
<tr>
<td>DDFV-HER</td>
<td>−1.61E−03</td>
<td>8.99E−02</td>
<td>MFD-FHE</td>
<td>−6.54E−02</td>
<td>1.44E−01</td>
</tr>
<tr>
<td>DDFV-MNI</td>
<td>−1.46E−03</td>
<td>6.69E−02</td>
<td>MFD-MAR</td>
<td>−2.62E−02</td>
<td>9.07E−02</td>
</tr>
<tr>
<td>DDFV-OMN</td>
<td>−1.77E−03</td>
<td>8.36E−02</td>
<td>MFV</td>
<td>−8.08E−03</td>
<td>5.81E−02</td>
</tr>
<tr>
<td>DG-C</td>
<td>−7.33E−03</td>
<td>1.05E−01</td>
<td>NMFV</td>
<td>3.05E−15</td>
<td>9.42E−02</td>
</tr>
<tr>
<td>DG-W</td>
<td>−9.03E−03</td>
<td>6.57E−02</td>
<td>SUSHI-NP</td>
<td>−1.19E−03</td>
<td>5.65E−02</td>
</tr>
<tr>
<td>FEQ1</td>
<td>−4.17E−03</td>
<td>4.90E−02</td>
<td>SUSHI-P</td>
<td>3.26E−06</td>
<td>6.77E−03</td>
</tr>
<tr>
<td>SLPS [30]</td>
<td>−7.51E−04</td>
<td>7.19E−02</td>
<td>SSEPS</td>
<td>1.58E−09</td>
<td>8.03E−02</td>
</tr>
</tbody>
</table>

In Table 5, all the results of the various schemes except our schemes SLPS [30] and SSEPS are directly copied from Table 12 in [15], we omit the detailed references of them. The results on the fine grid are obtained with the scheme SUS-HI-P [12] on a 201 × 201 uniform grid which was chosen parallel to the axes. The solution profile of our new scheme is shown in Fig. 12.

There are only four schemes which remain positive, namely NMFV, FVPMMD, SUSHI-P and our scheme SSEPS. Schemes NMFV, FVPMMD and SSEPS are nonlinear schemes.

5.5. Test 5: problems with oblique drain and oblique barrier

This test represents situations encountered in underground flow engineering, and can be found in [15]. We investigate the problem (1) with the full Dirichlet boundary condition (2) and the domain $\Omega$ is composed of three subdomains...
\[ \Omega_1 = \{(x,y) \in \Omega | \phi_1(x,y) < 0\}, \]
\[ \Omega_2 = \{(x,y) \in \Omega | \phi_1(x,y) > 0, \phi_2(x,y) < 0\}, \]
\[ \Omega_3 = \{(x,y) \in \Omega | \phi_2(x,y) > 0\}, \]

with \( \phi_1(x,y) = y - \delta(x - 0.5) - 0.475 \) and \( \phi_2(x,y) = \phi_1(x,y) - 0.05 \).

We take the slope of the drain \( \delta = 0.2 \) and define the exact solution and the permeability tensor \( \Lambda \) in the following two cases:

**Case 1: oblique drain.** The exact solution \( u_1(x,y) = -\lambda - \delta y \), and the permeability tensor \( \Lambda \) is such that its principal axes are parallel and perpendicular to the drain:

\[ \Lambda_1 = R_\theta \begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix} R_\theta^{-1}, \]

where \( \theta = \arctan \delta \) and \( \alpha = 100, \beta = 10 \) on \( \Omega_1 \); \( \alpha = 1, \beta = 0.1 \) on \( \Omega_1 \cup \Omega_3 \).

**Case 2: oblique barrier.** The exact solution

\[ u_2(x,y) = \begin{cases} -\phi_1(x,y) & \text{on } \Omega_1, \\ -100\phi_1(x,y) & \text{on } \Omega_2, \\ -\phi_2(x,y) - 5 & \text{on } \Omega_3 \end{cases} \]

and the permeability tensor \( \Lambda \) is heterogeneous and isotropic

\[ \Lambda_2 = \begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix} \text{ with } \alpha = \begin{cases} 1 & \text{on } \Omega_1 \cup \Omega_2, \\ 0.01 & \text{on } \Omega_2. \end{cases} \]

We discrete this test using the oblique mesh \texttt{Mesh9} which is depicted in Fig. 13(a), and set positive number \( \varepsilon \) in (13), \( \varepsilon_{lin} \) and \( \varepsilon_{non} \) to be equal to machine precision.

The fourth–seventh columns of Table 6 show that our scheme SSEPS satisfies extremum principle in both two cases, and for problem with oblique barrier, as shown in (Table 11 in [15]), extremum values for our scheme SSEPS are equal to many participating discretization schemes in FVCA V.

The absolute discrete \( L_2 \)-norm and \( H_1 \)-norm of solution errors \( E_u \) and \( E_q \) are also given in the last two columns of Table 6, and the associated relative discrete \( L_2 \)-norm of the solution errors with respect to the \( L_2 \)-norm of the exact solution are 8.04E–16 and 7.52E–16 for cases 1 and 2 respectively. In addition, the calculated primal contours of the approximate solution for case 2 are shown in Fig. 13(b). We can conclude that in both two cases, our scheme SSEPS is linearity-preserving

**Table 6**  
Results on the oblique mesh for problem with oblique drain or oblique barrier (\( u_{\min} \) and \( u_{\max} \) denote the minimal and maximal values of exact solution, respectively).

<table>
<thead>
<tr>
<th>Case</th>
<th>nunkw</th>
<th>nmaxn</th>
<th>uemin</th>
<th>uemax</th>
<th>umin</th>
<th>umax</th>
<th>( E_u )</th>
<th>( E_q )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>210</td>
<td>988</td>
<td>-1.2</td>
<td>0.0</td>
<td>-1.14615</td>
<td>-0.05385</td>
<td>3.67E–16</td>
<td>1.16E–14</td>
</tr>
<tr>
<td>Case 2</td>
<td>210</td>
<td>988</td>
<td>-5.575</td>
<td>0.575</td>
<td>-5.53675</td>
<td>0.53675</td>
<td>2.49E–15</td>
<td>7.58E–14</td>
</tr>
</tbody>
</table>

**Fig. 13.** Test 5: oblique mesh \texttt{Mesh9} and solution contours for problem with oblique barrier.
which means SSEPS provides the exact solution if the tensor $\Lambda$ is piecewise constant and solution is linear in each mesh cell.

### 5.6. Test 6: heterogeneous diffusion tensor

This problem is given in [18]. We solve the problem (1) in the unit square domain $\Omega = [0, 1]^2$ with the full Dirichlet boundary condition (2) and the source term

$$f(x) = \begin{cases} \frac{81}{4}, & \text{if } x \in \left[\frac{7}{18}, \frac{11}{18}\right]^2, \\ 0, & \text{otherwise}. \end{cases}$$

As shown in Fig. 15(a), the domain $\Omega$ is partitioned into four square subdomains $\Omega_i, i = 1, \ldots, 4$. The diffusion tensor $\Lambda$ is given by

$$\Lambda = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} k_1 & 0 \\ 0 & k_2 \end{pmatrix} \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix}.$$

We will give the numerical results on the randomly distorted quadrilateral mesh with $72 \times 72$ cells (see Fig. 14), and some of mesh cells are non-convex.

Firstly, we set $k_1 = 1000$, $k_2 = 1$ and vary only the angle $\theta$ (see Fig. 15(a)). Secondly, we use various $\theta$ and the chess board distribution of $k_1$ and $k_2$ (see Fig. 16(a)). The solution profiles are depicted in Figs. 15(b) and 16(b) with $\text{umin} = 1.18 \times 10^{-10}$ and $\text{umin} = 6.22 \times 10^{-10}$, respectively. In addition, we also compute them on $18 \times 18$ randomly distorted quadrilateral mesh, and the minimum value is $2.16 \times 10^{-6}$ and $3.87 \times 10^{-6}$, respectively. Hence in both cases we get the positive discrete solutions.
Finally, we can conclude that our new scheme can handle strong jumps of diffusion tensor across mesh edges in this numerical test.

5.7. Test 7: nonlinear parabolic problem

In this section, we will study the numerical solution of a nonlinear parabolic problem and show the positive impact of the time derivative on the Picard convergence.

We consider the following problem in the unit square domain $\Omega = [0, 1]^2$ with the Dirichlet boundary conditions

(P1) \[
\begin{align*}
\frac{\partial u}{\partial t}(x, y, t) - \text{div}(\Lambda(x, y, t) \nabla u(x, y, t)) &= f(x, y, t), \quad (x, y, t) \in \Omega \times (0, T], \\
u(x, y, t) &= \exp(t) g(x, y), \quad (x, y, t) \in \partial \Omega \times (0, T], \\
\nu(x, y, 0) &= g(x, y), \\
\Lambda &= 1 + u^2.
\end{align*}
\]

The exact solution is $u(x, y, t) = \exp(t) g(x, y)$ with the given function $g(x, y) = (2x^3 - 3x^2 + 2)(2y^3 - 3y^2 + 3)$. We set the terminal time $T = 0.5$ and the time step $\Delta t$ equals to the inverse of numbers of mesh cells.

We also give a similar steady diffusion problem in the unit square domain

(P2) \[
\begin{align*}
-\text{div}(\Lambda \nabla u) &= f, \quad (x, y) \in \Omega, \\
u &= g(x, y), \quad (x, y) \in \partial \Omega, \\
\Lambda &= 1 + g^2(x, y).
\end{align*}
\]

The exact solution is $u = g$ and the source term $f$ is determined by the exact solution and $\Lambda$.

The numerical tests are performed on four mesh types Mesh1–Mesh4. The $L_2$-norm of the solution errors on the finest mesh, numbers of Picard iterations $\text{nitn}$ on the finest mesh, and the convergence rates with respect to the $L_2$-norm and $H_1$-norm of the solution errors are given in Table 7. It should be noted that for the nonlinear parabolic problem (P1), in each time step, we have a number of Picard iterations, $\text{nitn}$ denotes the arithmetic averages of numbers of Picard iterations in the time interval $[0, 0.5]$. Table 7 reveals the following facts:

<table>
<thead>
<tr>
<th>Mesh</th>
<th>Mesh level ($h_i$)</th>
<th>Problem</th>
<th>$E_u$</th>
<th>$\text{nitn}$</th>
<th>$R_u$</th>
<th>$R_q$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mesh1</td>
<td>1.56E-02</td>
<td>(P1)</td>
<td>1.17E-04</td>
<td>1</td>
<td>1.993</td>
<td>1.046</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(P2)</td>
<td>1.49E-04</td>
<td>5</td>
<td>1.963</td>
<td>0.993</td>
</tr>
<tr>
<td>Mesh2</td>
<td>4.25E-02</td>
<td>(P1)</td>
<td>3.05E-04</td>
<td>21</td>
<td>2.452</td>
<td>1.303</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(P2)</td>
<td>5.20E-04</td>
<td>53</td>
<td>2.335</td>
<td>1.260</td>
</tr>
<tr>
<td>Mesh3</td>
<td>1.91E-02</td>
<td>(P1)</td>
<td>1.86E-04</td>
<td>6</td>
<td>2.034</td>
<td>1.116</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(P2)</td>
<td>2.27E-04</td>
<td>9</td>
<td>2.040</td>
<td>1.066</td>
</tr>
<tr>
<td>Mesh4</td>
<td>1.51E-02</td>
<td>(P1)</td>
<td>2.60E-04</td>
<td>1</td>
<td>1.885</td>
<td>1.415</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(P2)</td>
<td>3.90E-04</td>
<td>6</td>
<td>1.786</td>
<td>1.323</td>
</tr>
</tbody>
</table>
The expected optimal convergence rates with respect to the $L_2$-norm and $H_1$-norm of the solution errors are obtained on four mesh types for the nonlinear parabolic problem and steady problem. Solutions of the nonlinear parabolic problem ($P1$) have lower numbers of Picard iterations than that of associated steady problem ($P2$). The reduction of the number of nonlinear iterations is due to a better initial guess on each time step.

5.8. Test 8: 3D study

We consider the problem (1) with full Dirichlet boundary condition (2) on the unit cube $\Omega = [0,1]^3$. We choose a constant anisotropic diffusion tensor

$$A(x,y,z) = \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0.5 \\ 0 & 0.5 & 1 \end{pmatrix}$$

and a regular solution $u_1(x,y,z) = 1 + \sin(\pi x) \sin\left(\pi \left(\frac{y}{2}\right)\right) \sin\left(\pi \left(\frac{z}{3}\right)\right)$ which implies a non-homogeneous Dirichlet condition on the boundary $\partial \Omega$. This test can be found in the sixth conference on 3D discretization schemes for anisotropic diffusion problems on general grids [11].

We will study the behavior of our scheme $SSEPS$ on the uniform hexahedral mesh, disturbed random hexahedral mesh and 3D-Shestakov mesh respectively. The uniform hexahedral mesh was composed of cubes, and the mesh size ranges from $4.33E-01$ to $2.71E-02$. The disturbed random hexahedral mesh (see Fig. 17(a)) is constructed from the uniform hexahedral mesh with the mesh size $h$ by randomly disturbing the interior vertices in three directions, and this distortion is performed on each refinement level. Since we just consider the case of planar facets, the distortion is done in such a way that all facets are planar or nearly planar ones. As seen in Fig. 17(b), the 3D-Shestakov mesh (3D version of Shestakov mesh [21]) has constant spacing in one dimension and varied spacing in the second and third dimensions.

The ratio $\alpha_s$ between the non-zero terms and the unknowns of the matrix, numbers of nonlinear and linear iterations $\text{nitn}$ and $\text{itn}$, extremum discrete solution $\text{umin}$ and $\text{umax}$, the discrete $L_2$-norm and $H_1$-norm of solution errors on three types of 3D meshes are given in Tables 8–10, respectively. From these tables, we find that.

- On the finest mesh with 262144 cells, the ratio $\alpha_s$ is approximately equal to 7, which verifies that our scheme $SSEPS$ has a seven-point stencil on the structured hexahedral meshes.

![Fig. 17. Test 8: random hexahedral mesh (left) and 3D-Shestakov mesh (right).]

<table>
<thead>
<tr>
<th>Cells</th>
<th>$\alpha_s$</th>
<th>nitn</th>
<th>itn</th>
<th>umin (E)</th>
<th>umax (E)</th>
<th>$E_u$ (E)</th>
<th>$E_q$ (E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>5.50</td>
<td>9</td>
<td>2</td>
<td>1.68E–01</td>
<td>1.855</td>
<td>3.91E–02</td>
<td>3.38E–01</td>
</tr>
<tr>
<td>512</td>
<td>6.25</td>
<td>11</td>
<td>7</td>
<td>4.66E–02</td>
<td>4.95</td>
<td>1.44E–02</td>
<td>1.51E–01</td>
</tr>
<tr>
<td>4096</td>
<td>6.63</td>
<td>12</td>
<td>14</td>
<td>1.15E–02</td>
<td>1.988</td>
<td>4.39E–03</td>
<td>5.85E–02</td>
</tr>
<tr>
<td>32,768</td>
<td>6.81</td>
<td>9</td>
<td>27</td>
<td>3.25E–03</td>
<td>1.997</td>
<td>1.17E–03</td>
<td>2.11E–02</td>
</tr>
<tr>
<td>262,144</td>
<td>6.91</td>
<td>6</td>
<td>60</td>
<td>7.68E–04</td>
<td>1.999</td>
<td>3.00E–04</td>
<td>7.43E–03</td>
</tr>
</tbody>
</table>

Rates

1.756

1.378
Table 9
Test 8: results on disturbed random hexahedral meshes.

<table>
<thead>
<tr>
<th>Mesh</th>
<th>Cells</th>
<th>as</th>
<th>nitn</th>
<th>itn</th>
<th>umin</th>
<th>umax</th>
<th>$E_u$</th>
<th>$E_q$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mesh1</td>
<td>64</td>
<td>5.50</td>
<td>9</td>
<td>2</td>
<td>1.41E-01</td>
<td>1.870</td>
<td>4.32E-02</td>
<td>3.83E-01</td>
</tr>
<tr>
<td>Mesh2</td>
<td>512</td>
<td>6.25</td>
<td>10</td>
<td>7</td>
<td>4.91E-02</td>
<td>1.959</td>
<td>1.56E-02</td>
<td>1.88E-01</td>
</tr>
<tr>
<td>Mesh3</td>
<td>4096</td>
<td>6.63</td>
<td>14</td>
<td>12</td>
<td>1.27E-02</td>
<td>1.988</td>
<td>4.88E-03</td>
<td>9.05E-02</td>
</tr>
<tr>
<td>Mesh4</td>
<td>32,768</td>
<td>6.81</td>
<td>27</td>
<td>3.36E-03</td>
<td>1.997</td>
<td>1.41E-03</td>
<td>5.46E-02</td>
<td></td>
</tr>
<tr>
<td>Mesh5</td>
<td>262,144</td>
<td>6.91</td>
<td>60</td>
<td>7.15E-04</td>
<td>1.999</td>
<td>5.01E-04</td>
<td>4.58E-02</td>
<td></td>
</tr>
<tr>
<td>Rates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.671</td>
<td>0.796</td>
</tr>
</tbody>
</table>

Table 10
Test 8: results on 3D-Shestakov meshes.

<table>
<thead>
<tr>
<th>Mesh</th>
<th>Cells</th>
<th>as</th>
<th>nitn</th>
<th>itn</th>
<th>umin</th>
<th>umax</th>
<th>$E_u$</th>
<th>$E_q$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mesh1</td>
<td>64</td>
<td>5.50</td>
<td>7</td>
<td>2</td>
<td>1.55E-01</td>
<td>1.873</td>
<td>4.66E-02</td>
<td>4.25E-01</td>
</tr>
<tr>
<td>Mesh2</td>
<td>512</td>
<td>6.25</td>
<td>10</td>
<td>7</td>
<td>4.70E-02</td>
<td>1.971</td>
<td>1.79E-02</td>
<td>1.98E-01</td>
</tr>
<tr>
<td>Mesh3</td>
<td>4096</td>
<td>6.63</td>
<td>14</td>
<td>9.65E-03</td>
<td>1.982</td>
<td>5.67E-03</td>
<td>8.36E-02</td>
<td></td>
</tr>
<tr>
<td>Mesh4</td>
<td>32,768</td>
<td>6.81</td>
<td>28</td>
<td>3.60E-03</td>
<td>1.997</td>
<td>1.65E-03</td>
<td>3.91E-02</td>
<td></td>
</tr>
<tr>
<td>Mesh5</td>
<td>262,144</td>
<td>6.91</td>
<td>59</td>
<td>8.61E-04</td>
<td>1.999</td>
<td>5.64E-04</td>
<td>1.84E-02</td>
<td></td>
</tr>
<tr>
<td>Rates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.902</td>
<td>1.347</td>
</tr>
</tbody>
</table>

- The extremum principle states that the exact solution is located in the interval $[0, 2]$. Solutions computed with our scheme $SSEPS$ on three mesh types respects the bounds everywhere in the domain $\Omega$.
- The seventh and eighth columns of three tables show that our scheme $SSEPS$ obtains nearly second order convergence rate for the discrete $L_2$-norm and first order convergence rate for $H_1$-norm of solution errors on these meshes.

In addition, if we choose a linear exact solution

$$u_2(x, y, z) = x + 2y + 3z$$

in this test, we find that on the $4 \times 4 \times 4$ uniform hexahedral mesh, the absolute discrete $L_2$-norm of solution error $E_u$ is equal to 1.20E–15, and relative discrete $L_2$-norm of solution error with respect to the $L_2$-norm of the exact solution is 3.78E–16. It is obvious that our scheme is linearity-preserving in this case.

5.9. Verification of stability condition (A6)

This section is devoted to investigate whether the stability condition (A6) in scheme $SSEPS$ holds. We also use the computational model in Test 1 and give computational results on two triangle meshes (Mesh1 and Mesh7), the largely disturbed random mesh Mesh8 and the small disturbed random mesh Mesh8 for the scheme $SSEPS$ with $\gamma = 0$ and $\gamma = 0.4$ in Tables 11 and 12, respectively. Both two tables show the following:

- The minimum and maximum solutions are given on the coarsest meshes, we find that for $\gamma = 0$ and $\gamma = 0.4$, our scheme satisfies the discrete extremum principle in this test.
- Our scheme $SSEPS$ with $\gamma = 0$ and $\gamma = 0.4$ shows a second (resp., first) order convergence rate with respect to the discrete $L_2$ (resp., $H_1$) norm of the solution errors on the four types of meshes.
- $\text{eigmin}$ and $\text{ratio}$ on three mesh levels (mesh size $h = h_1, h_1, h_3$) are given in the fourth–ninth columns, where $\text{eigmin}$ denotes the smallest eigenvalue of $M_K A_\tau W_K$ over all cells and $\text{ratio}$ is the percentage of matrices $M_K A_\tau W_K$ that have negative eigenvalues. For $\gamma = 0$, the stability condition (A6) holds only on the uniform triangle mesh Mesh7.

Table 11
Results for four meshes Mesh1, Mesh5, Mesh7 and Mesh8 with $\gamma = 0$.

<table>
<thead>
<tr>
<th>Mesh</th>
<th>$h_1$</th>
<th>$h_1$</th>
<th>$h_1$</th>
<th>$h_1$</th>
<th>$R_u$</th>
<th>$R_q$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>umin</td>
<td>max</td>
<td>eigmin</td>
<td>ratio</td>
<td>eigmin</td>
<td>ratio</td>
</tr>
<tr>
<td>Mesh1</td>
<td>2.80E-03</td>
<td>0.754</td>
<td>0.010</td>
<td>0%</td>
<td>-0.036</td>
<td>0.56%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.103</td>
<td>0.75%</td>
</tr>
<tr>
<td>Mesh2</td>
<td>1.84E-03</td>
<td>0.810</td>
<td>-0.187</td>
<td>9.38%</td>
<td>-0.789</td>
<td>5.47%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.986</td>
<td>5.18%</td>
</tr>
<tr>
<td>Mesh3</td>
<td>8.83E-03</td>
<td>0.674</td>
<td>1.160</td>
<td>0%</td>
<td>1.192</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.054</td>
<td>1.613</td>
</tr>
<tr>
<td>Mesh4</td>
<td>5.08E-03</td>
<td>0.639</td>
<td>0.311</td>
<td>0%</td>
<td>-0.100</td>
<td>0.78%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.110</td>
<td>0.27%</td>
</tr>
<tr>
<td>Mesh5</td>
<td>2.04E-03</td>
<td>0.536</td>
<td>0.222</td>
<td>0%</td>
<td>1.007</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.015</td>
<td>1.120</td>
</tr>
</tbody>
</table>
However, for $\gamma = 0.4$, condition (A6) holds on Mesh1, Mesh7 and Mesh8. We also find that on Mesh3, the largely disturbed random mesh, we cannot choose a suitable parameter $\gamma$ to keep all eigenvalues positive. The choice of $\gamma$ needs to be further studied.

6. Conclusion

We have developed and analyzed a new extremum-preserving finite volume scheme for anisotropic diffusion equations. This scheme is a further development of the 2D schemes given in [26,27,33] and has three key properties: (1) It is a scheme designed for both 2D and 3D meshes, providing a possibility of a uniform coding for both cases; (2) It has a small stencil, i.e., a four-point stencil for the triangle meshes, a five-point stencil for the structured quadrilateral meshes, and a seven-point stencil for structured hexahedral meshes, which is well adapted to domain decomposition techniques and parallelization; (3) Its interpolation algorithm is simple and positivity-preserving, and moreover, has second-order accuracy if assumption (A1) holds.

Many dedicated experiments using meshes with unstructured, severely distorted and nonmatching elements and arbitrary (continuous or discontinuous, homogeneous or heterogeneous) anisotropic diffusion tensors show the robustness of the proposed scheme (quadratic convergence rate for the approximate solution and higher than first-order accuracy for its gradient), and the new scheme satisfies discrete extremum principle in 2D and 3D numerical experiments. The price for these appealing features is nonlinearity of the scheme.

Some future works include the application to practical multi-material physical problems, convergence analysis and so on.

Appendix A. A new derivation for the harmonic averaging point

The concept of harmonic averaging point was first suggested in [4] where a sketch of derivation was also suggested. Following that sketch, we gave a detailed derivation in [30] through the linearity-preserving approach and under the following conditions:

- For $\sigma \in \mathcal{E}_K \cap \mathcal{E}_L$, (i) $K$ (resp. $L$) is a star-shaped polygonal cell with respect to $x_k$ (resp. $x_L$); (ii) $x_k$ and $x_L$ are not located on the same side of the line where $\sigma$ lies; (iii) $\{y_K, y_L\} \cap \sigma \neq \emptyset$; (iv) $y_\sigma \in \sigma$.

Actually at present, we find that condition (i) implies (ii), and condition (iii) does not need to be satisfied in the following new and simple derivation.

As shown in Fig. 18, let $y_\sigma$ be an interior point on the common edge $\sigma$ of cells $K$ and $L$, and the endpoints of $\sigma$ are denoted by $x_{\sigma,1}$ and $x_{\sigma,2}$. For simplicity, we just consider 2D case, the 3D case can be discussed analogously. On cells $K$ and $L$, we can always write

![Fig. 18. Notations for the derivation of harmonic averaging point.](image-url)
\[ \begin{align*}
\Lambda_1^T n_{K,\sigma} &= a_k(y_{\sigma} - x_k) + b_k(x_{\sigma,1} - x_{\sigma,1}), \\
\Lambda_2^T n_{K,\sigma} &= a_l(x_l - y_k) + b_l(x_{\sigma,2} - x_{\sigma,1}).
\end{align*} \tag{32} \]

Noting that \( K \) and \( L \) are star-shaped, we have
\[ a_k = \frac{\tilde{z}_{\sigma}^{(n)}(y_{\sigma})}{\tilde{d}_{\sigma}^{(n)}} , \quad a_l = \frac{\tilde{z}_{\sigma}^{(n)}(x_l)}{\tilde{d}_{\sigma}^{(n)}} \tag{33} \]
and
\[ u(y_{\sigma}) - u(x_k) \approx \nabla u \cdot (y_{\sigma} - x_k), \quad u(x_k) - u(y_{\sigma}) \approx \nabla u \cdot (x_k - y_{\sigma}) , \quad u(x_{\sigma,2}) - u(x_{\sigma,1}) \approx \nabla u \cdot (x_{\sigma,2} - x_{\sigma,1}). \]

Then, multiplying the two equations in (32) with \( -\nabla \), respectively, we have
\[
F_{K,\sigma} \approx |\sigma| [-a_k(u(y_{\sigma}) - u(x_k)) - b_k(u(x_{\sigma,2}) - u(x_{\sigma,1}))], \quad -F_{L,\sigma} \approx |\sigma| [-a_l(u(x_l) - u(y_{\sigma})) - b_l(u(x_{\sigma,2}) - u(x_{\sigma,1}))].
\]

Since \( F_{K,\sigma} + F_{L,\sigma} = 0 \), we deduce that
\[ u(y_{\sigma}) \approx \frac{a_k u(x_k) + a_l u(x_l)}{a_k + a_l} + \frac{b_k}{a_k + a_l} [u(x_{\sigma,2}) - u(x_{\sigma,1})]. \]

Thus a sufficient condition for (5) is that \( b_k = b_l \). Substituting this condition into (32) gives
\[ y_{\sigma} = \frac{a_k x_k + a_l x_l + (\Lambda_1^T - \Lambda_2^T) n_{K,\sigma}}{a_k + a_l}, \tag{34} \]
which leads to (4) by noting (33) and completes the derivation.

Let \( y_k, y_l \) (resp., \( y_{\sigma} \)) denotes the orthogonal projection point of \( x_k \) (resp., \( x_l \)) onto the hyperplane containing \( \sigma \). Obviously,
\[ y_k - x_k = d_{K,\sigma} n_{K,\sigma}, \quad y_l - x_l = d_{L,\sigma} n_{K,\sigma}. \tag{35} \]

Substituting (35) into (4) and denoting by \( I_d \) the \( d \times d \) identity matrix, we have
\[ d_{L,\sigma} \tilde{z}_{\sigma}^{(n)}(y_k) + d_{L,\sigma} \tilde{z}_{\sigma}^{(n)}(y_l) + d_{K,\sigma} d_{L,\sigma} \left[ (\Lambda_1^T - \Lambda_2^T) I_d \right] n_{K,\sigma} - \left( \tilde{z}_{\sigma}^{(n)}(x_l) - \tilde{z}_{\sigma}^{(n)}(x_k) \right) n_{K,\sigma}, \]
which is none other than the definition for harmonic averaging point given in [4,13].

References