A monotone finite volume method for advection-diffusion equations on unstructured polygonal meshes *

K. Lipnikov [a], D. Svyatskiy [a,*], Y. Vassilevski [b]

[a] *Mathematical Modeling and Analysis Group, Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545*

[b] *Institute of Numerical Mathematics, Russian Academy of Sciences, 8, Gubkina, 119333, Moscow, Russia,*

Abstract

We present a new second-order accurate monotone finite volume (FV) method for the steady-state advection-diffusion equation. The new method uses nonlinear approximation for both diffusive and advective fluxes and guarantees solution positivity. The interpolation-free approximation of the diffusive flux uses the nonlinear stencil proposed in [23]. Approximation of the advective flux is based on the second order upwind method with a specially designed minimal nonlinear correction. The second-order convergence rate and monotonicity are verified with numerical experiments.

Key words: advection-diffusion equation, finite volume method, discrete maximum principle, monotone method, unstructured mesh, polygonal mesh.

1 Introduction

Predictive numerical simulations require accurate and reliable discretization methods. The resulting discrete systems must inherit or mimic fundamental properties of continuous systems.

* This work was carried out under the auspices of the National Nuclear Security Administration of the U.S. Department of Energy at Los Alamos National Laboratory under Contract No. DE-AC52-06NA25396 and the DOE Office of Science Advanced Scientific Computing Research (ASCR) Program in Applied Mathematics Research.

* Corresponding author.
Email addresses: lipnikov@lanl.gov (K. Lipnikov), dasvyat@lanl.gov (D. Svyatskiy), vasilevs@dodo.inm.ras.ru (Y. Vassilevski).

The maximum principle and local mass conservation are the essential properties of the steady-state advection-diffusion equation. Despite a relative simplicity of this equation, an accurate discretization method that satisfies the discrete maximum principle (DMP) is hard to develop. Our focus is on a simplified version of the DMP that provides only solution positivity as is referred to as the monotonicity condition. Some physical quantities, such as a concentration and temperature, are non-negative by their nature and their approximations should be non-negative as well. We develop a nonlinear finite volume (FV) method that satisfies the monotonicity condition for both diffusion dominated and advection dominated regimes.

In advection dominated problems a solution may have internal shock and exponential or parabolic boundary layers. The thickness of these features is small compared to mesh size and hence the layers cannot be resolved properly. In the case of diffusion dominated problems diffusive fluxes may be very poor approximated in a particular direction since mesh cells are not aligned with the dominated diffusion directions. In particular, these problems may appear in the case of highly anisotropic media. This leads to unwanted spurious (nonphysical) oscillations in the numerical solution. The design of advanced discretization schemes which eliminate or significantly reduce these oscillations is the field of extensive research for more than three decades.

In the finite element (FE) framework one of the most popular approach was proposed by Brooks and Hughes in [5] and is referred to as the streamline upwind Petrov Galerkin (SUPG) method. Although the stabilization procedure proposed in this method significantly improves the robustness of the FE discretization, the spurious oscillations along sharp layers may still appear in the numerical solution. They are caused by the fact that the SUPG method is neither monotone nor monotonicity preserving method. The review of several modifications and improvements for SUPG method is presented in [13]. These modifications aimed to design discretization methods that satisfy the DMP, at least in some model cases, and the authors of [13] proposed to call them *spurious oscillations at layers diminishing* (SOLD) methods. Another approach towards a robust FE discretization method was proposed in [17,18] and is referred to as *algebraic flux correction* approach. One of the disadvantages of FE discretizations that they are not locally conservative in terms of the original computational mesh. The local mass conservation is a very desirable property if the advection-diffusion equation is nonlinear and coupled with other transport processes.

The FV type discretizations guarantee the local mass conservation by construction. Many FV methods for advection-diffusion equation have been developed for the last decades, see [32], [3], [4], [10], [24], [20] and references therein. It turns out that in the design of monotone, second order accurate discretizations the approximation of diffusive fluxes is as challenging as the approximation of the advective fluxes. The advective fluxes are usually controlled by using upwinding approach [2] along with different slope-limiting techniques [4] or introduction of artificial viscosity [24]. For a long time it was not clear how to construct a monotone discretization of a diffusive part in the case of general meshes and diffusion tensors. Many advanced *linear* methods for approximation of diffusive fluxes fail to satisfy the monotonicity condition when the media is heterogeneous and anisotropic or the computational mesh is strongly perturbed. This includes the mixed finite element (MFE), mimetic finite difference (MFD), and multi-point flux approximation (MPFA) methods that are locally conservative and second-order accurate on unstructured meshes [22]. The linear two-point flux approximation FV method, still used in modeling flows in porous media, is monotone but not even first-order accurate for anisotropic problems. Monotonicity limits of the MPFA methods are analyzed in [1,25]. The theoretical analysis of sufficient mesh conditions providing the DMP has been formulated in 70's by P.Ciarlet and P.Raviart [8] for piecewise-linear finite element approximations. Later, the DMP has been shown for weaker mesh conditions [16,27]. It was noticed in [4,13] that *nonlinear* approximations is the key ingredient and the price which has to be paid to construct monotone and at least the second order accurate discretization. To guarantee solution positivity for arbitrary meshes, a number of *nonlinear* methods have been proposed for the Poisson equation [6] and more recently for general diffusion equation [9,15,21,22,26,31,33].

The approximation of diffusive fluxes in the proposed monotone FV method is based on a nonlinear two-point flux approximation scheme. The original idea was proposed by C.LePotier [21] for the case of triangular meshes. In [22], we proved monotonicity of his method for steady-state diffusion problems and extended it to shape regular polygonal meshes and scalar diffusion coefficients. The method has been extended to tetrahedral meshes by I.Kapyrin for the diffusion equation [15] and by I.Kapyrin and Yu.Vassilevski for the unsteady advection-diffusion equations [31]. Further development of the method was made by A.Yuan and Z.Sheng [33]. Their method can be applied to a much bigger class of polygonal meshes consisting of star-shaped cells and full tensor diffusion coefficients. The common property of all these methods is that in addition to *primary* unknowns defined at mesh cells, solution values at mesh vertices are involved in the method construction. These *auxiliary* unknowns are interpolated from primary, cell-based unknowns. The interpolation problem becomes even a more challenging task when the diffusion coefficient is discontinuous. The interpolation methods studied in [33] use a piecewise linear approximation to the solution around points where the coefficient is discontinuous. However, as shown in [22,33], the choice of the interpolation method affects the accuracy of the nonlinear FV method even in the case of a constant diffusion coefficient. The choice of an interpolation method depends of the problem. The particular interpolation method may be efficient for one problem and be inaccurate for another. The three-dimensional extension of the nonlinear FV method to polyhedral meshes [9] excludes the use of nodal interpolation yet may require edge interpolation in certain pathological cases. Face interpolation [9] used at faces with jumping diffusion coefficient is based on physical relations.

In [23] we proposed the nonlinear FV method which does not use any auxiliary unknowns at mesh vertices. The numerical experiments presented in [23] demonstrate that the interpolation-free approach requires less nonlinear iterations than the methods using interpolation algorithms. The nonlinear FV method proposed in this article follows the same idea for the diffusive fluxes. The approximation of advective fluxes does not use any interpolation techniques as well. It is based on the upwinding approach along with a piecewise linear reconstruction of the FV solution. This reconstruction depends on the solution so the approximation of advective fluxes is also *nonlinear*. In order to guarantee monotonicity and robustness of the method, we propose a new slope limiting technique, see [7,11] for discussions on slope limiters. It is exact for linear solutions and thus has the second order truncation error. Our numerical experiments show the second-order convergence rate in the mesh-dependent L_2-norm. The slope limiter is a nonlinear operator designed to minimize a correction to the gradient of the least-square linear reconstruction and satisfy monotonicity conditions. This allows us to prove positivity of the discrete solution.

The two-point flux approximation methods result in schemes with a compact stencil. For square meshes and a diagonal diffusion tensor this stencil reduces to the conventional 5-point stencil. The major computational overhead in nonlinear FV methods comes from the solution of a nonlinear algebraic problem. The Picard method, used in this and the other papers, guarantees that the solution is positive on each iteration.

The paper outline is as follows. In Section 2, we state the steady advection-diffusion problem. In Section 3, we describe the nonlinear finite volume scheme. In Section 4, we prove monotonicity of the proposed scheme. In Section 5, we present numerical analysis of the scheme using triangular, quadrilateral and polygonal meshes.

2 Steady-state advection-diffusion equation

Let Ω be a two-dimensional polygonal domain with boundary $\Gamma = \Gamma_R \cup \Gamma_D$ where $\Gamma_D = \bar{\Gamma}_D$ and $\Gamma_D \neq \emptyset$. We consider a model advection-diffusion problem for unknown concentration c:

$$\begin{aligned} \text{div}\,(\mathbf{v}c - \mathbb{K}\nabla c) &= f \quad \text{in } \Omega \\ c &= g_D \text{ on } \Gamma_D \\ -\mathbb{K}\frac{\partial c}{\partial \mathbf{n}} + c\mathbf{v}\cdot\mathbf{n} &= g_R \text{ on } \Gamma_R \end{aligned} \tag{1}$$

where $\mathbb{K}(\mathbf{x}) = \mathbb{K}^T(\mathbf{x}) > 0$ is a continuous (possibly anisotropic) diffusion tensor, $\mathbf{v}(\mathbf{x}) \in C^1(\bar{\Omega})$ is a velocity field, div $\mathbf{v} \geqslant 0$, f is a source term, and $\mathbf{n}$ is the exterior normal vector. We denote by Γ_{out} the outflow part of Γ where $\mathbf{v}\cdot\mathbf{n} \geqslant 0$, and define $\Gamma_{in} = \Gamma \setminus \Gamma_{out}$. The set $\Gamma_R \subset \Gamma_{out}$ can be empty.

In order to guarantee non-negativity of the solution $c(x)$ we have to require additionally that $f(x) \geqslant 0$, $g_D \geqslant 0$ and $g_R \leqslant 0$. Under these assumption $c(x)$ will be non-negative. From a physical point of view the requirements $f(x) \geqslant 0$ and $g_R \leqslant 0$ mean that no mass or energy can be taken out of the system.

For advection-dominated problems the Dirichlet boundary conditions on Γ_{out} may result in parabolic and/or exponential boundary layers. A parabolic boundary layer can be also generated by discontinuity in boundary data g_D. An ideal discretization scheme must introduce a minimal amount of numerical diffusion to avoid excessive smearing of boundary layers but sufficient to damp non-physical oscillations.

Remark 2.1 *A monotone FV scheme for a discontinuous tensor coefficient $\mathbb{K}$ can be developed using a modified discretization of the diffusive flux as described in [9,23].*

3 Monotone nonlinear FV scheme on polygonal meshes

Let $\mathbf{q} = -\mathbb{K}\nabla c + c\mathbf{v}$ denote the total flux which satisfies the mass balance equation:

$$\operatorname{div} \mathbf{q} = f \quad \text{in} \quad \Omega. \tag{2}$$

In this section, we derive a FV scheme with a nonlinear two-point flux approximation.

Let $\mathcal{T}$ be a conformal polygonal mesh composed of shape-regular cells. Let $N_{\mathcal{T}}$ be the number of polygonal cells and $N_{\mathcal{B}}$ be the number of boundary edges. We assume that $\mathcal{T}$ is edge-connected, i.e. it cannot be split into two meshes having no common edges.

We denote by $\mathcal{E}_I$, $\mathcal{E}_B$ disjoint sets of interior and boundary edges. The set $\mathcal{E}_B$ is further split into subsets $\mathcal{E}_B^D$ and $\mathcal{E}_B^R$ where the Dirichlet and Robin boundary conditions, respectively, are imposed. Alternatively, the set $\mathcal{E}_B$ is split into subsets $\mathcal{E}_B^{out}$ and $\mathcal{E}_B^{in}$ of edges belonging to Γ_{out} and Γ_{in}, respectively. Finally, $\mathcal{E}_T$ denotes the set of edges of polygon T.

Integrating equation (2) over a polygon T and using Green's formula we get:

$$\int_{\partial T} \mathbf{q} \cdot \mathbf{n}_T \,\mathrm{d}s = \int_T f \,\mathrm{d}x, \tag{3}$$

where $\mathbf{n}_T$ denotes the outer unit normal to ∂T. Let e denote an edge of cell T and $\mathbf{n}_e$ be the corresponding normal vector. For a single cell T, we always assume that $\mathbf{n}_e$ is the outward normal vector. In all other cases, we specify orientation of $\mathbf{n}_e$. It will be convenient to assume that $|\mathbf{n}_e| = |e|$ where $|e|$ denotes the length of edge e. The equation (3) becomes

$$\sum_{e \in \partial T} \mathbf{q}_e \cdot \mathbf{n}_e = \int_T f \,\mathrm{d}x, \tag{4}$$

where $\mathbf{q}_e$ is the average flux density for edge e

$$\mathbf{q}_e = \frac{1}{|e|} \int_e \mathbf{q} \,\mathrm{d}s.$$

For each cell T, we assign one degree of freedom, C_T, for concentration c. Let C be the vector of all discrete concentrations. If two cells T_+ and T_- have a common edge e, the two-point flux approximation is as follows:

$$\mathbf{q}_e^h \cdot \mathbf{n}_e = M_e^+ C_{T_+} - M_e^- C_{T_-}, \tag{5}$$

where M_e^+ and M_e^- are some coefficients. In a linear FV method, these coefficients are equal and fixed. In the nonlinear FV method, they may be different and depend on concentrations in surrounding cells. On edge $e \in \Gamma_D$, the flux has a form similar to (5) with an explicit value for one of the concentrations. For the Dirichlet boundary value problem, $\Gamma_D = \partial\Omega$, substituting (5) into (4), we obtain a system of $N_{\mathcal{T}}$ equations with $N_{\mathcal{T}}$ unknowns C_T. Dirichlet and Robin boundary conditions are considered in Section 3.4.

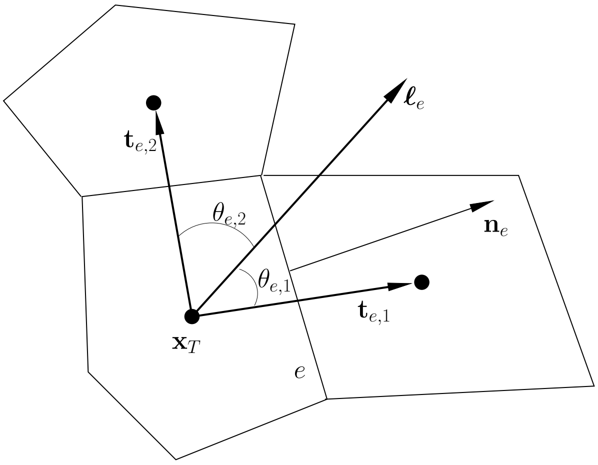


Fig. 1. Notation: vector $\boldsymbol{\ell}_e$ forms acute angles with vectors $\mathbf{t}_{e,1}$ and $\mathbf{t}_{e,2}$; the collocation points are marked by solid circles.

3.1 Notations

For every T in $\mathcal{T}$, we define the collocation interior point $\mathbf{x}_T$ at the barycenter of T. Similarly, for every edge $e \in \mathcal{E}_B$, we define the collocation point $\mathbf{x}_e$ at the barycenter of e.

For every T we define a set Σ_T of nearby collocation points as follows. First, we add to Σ_T the collocation point $\mathbf{x}_T$. Then, for every interior edge $e \in \mathcal{E}_T \cap \mathcal{E}_I$, we add the collocation point $\mathbf{x}_{T'_e}$, where T'_e is the cell, other than T, that has edge e. For every boundary edge $e \in \mathcal{E}_T \cap \mathcal{E}_B$, we add the collocation point $\mathbf{x}_e$. Let $N(\Sigma_T)$ denote the cardinality of Σ_T.

We shall refer to collocation points on edges $e \in \mathcal{E}_B$ as the *secondary* collocation points. They are introduced for mathematical convenience and will not enter the final algebraic system. In contrast, we shall refer to the other collocation points as the *primary* collocation points.

We assume that for every $e \in \mathcal{E}_T$, there exist two points $\mathbf{x}_{e,1}$ and $\mathbf{x}_{e,2}$ in set Σ_T such that the following two conditions are held [33].

(C1) If $\mathbf{t}_{e,1} = \mathbf{x}_{e,1} - \mathbf{x}_T$, $\mathbf{t}_{e,2} = \mathbf{x}_{e,2} - \mathbf{x}_T$, and $\theta_{e,i}$, $i = 1, 2$, is the angle between $\mathbf{t}_{e,i}$ and the co-normal vector $\boldsymbol{\ell}_e = \mathbb{K}(\mathbf{x}_e)\mathbf{n}_e$ (see Fig. 1), then

$$\theta_{e,1} < \pi, \qquad \theta_{e,2} < \pi \qquad \text{and} \qquad \theta_{e,1} + \theta_{e,2} < \pi. \tag{6}$$

(C2) The vectors $\mathbf{t}_{e,i}$ and $\boldsymbol{\ell}_e$ satisfy

$$\mathbf{t}_{e,1} \times \boldsymbol{\ell}_e \leqslant 0 \qquad \text{and} \qquad \mathbf{t}_{e,2} \times \boldsymbol{\ell}_e > 0. \tag{7}$$

In simple words, the co-normal vector $\boldsymbol{\ell}_e$ is assumed to lie between vectors $\mathbf{t}_{e,1}$ and $\mathbf{t}_{e,2}$, as shown in Fig. 1, and all angles are less than π. If conditions (6) and (7) are violated, we may extend the set Σ_T by adding neighbors of already included collocation points.

Lemma 3.1 *Under assumptions (6) and (7), there exist non-negative α_e and β_e such that*

$$\frac{1}{|\boldsymbol{\ell}_e|}\boldsymbol{\ell}_e = \frac{\alpha_e}{|\mathbf{t}_{e,1}|}\mathbf{t}_{e,1} + \frac{\beta_e}{|\mathbf{t}_{e,2}|}\mathbf{t}_{e,2}. \tag{8}$$

Moreover,

$$\alpha_e = \frac{\sin\theta_{e,2}}{\sin(\theta_{e,1}+\theta_{e,2})} \qquad \textit{and} \qquad \beta_e = \frac{\sin\theta_{e,1}}{\sin(\theta_{e,1}+\theta_{e,2})}.$$

Proof. The formulas for α_e and β_e follow from trigonometric observations. The non-negativity of α_e and β_e follows from assumptions (6) and (7). □

3.2 Nonlinear two-point diffusion flux approximation for an interior edge

In this section, we consider the diffusion flux on an interior edge $e \in \mathcal{E}_I$

$$\mathbf{q}_{e,d} = \frac{1}{|e|}\int_e -\mathbb{K}\nabla c \, \mathrm{d}s.$$

We denote by T_+ and T_- the cells that share e and assume that $\mathbf{n}_e$ is outward for T_+ and $T = T_+$. Let $\mathbf{x}_\pm$ (or $\mathbf{x}_{T_\pm}$) be the collocation point in $T_\pm$, $\mathbb{K}_e \equiv \mathbb{K}(\mathbf{x}_e)$ and $C_\pm$ (or $C_{T_\pm}$) be the discrete concentrations in $T_\pm$.

Using definition of the directional derivative,

$$\frac{\partial c}{\partial \boldsymbol{\ell}_e}|\boldsymbol{\ell}_e| = \nabla c \cdot (\mathbb{K}_e\, \mathbf{n}_e),$$

and Lemma 3.1, we note that

$$\mathbf{q}_{e,d}\cdot\mathbf{n}_e = -(1+O(|e|))\frac{|\boldsymbol{\ell}_e|}{|e|}\int_e \frac{\partial c}{\partial \boldsymbol{\ell}_e}\,\mathrm{d}s \quad \text{and} \quad \int_e \frac{\partial c}{\partial \boldsymbol{\ell}_e}\,\mathrm{d}s = \int_e \left(\alpha_e\frac{\partial c}{\partial \mathbf{t}_{e,1}} + \beta_e\frac{\partial c}{\partial \mathbf{t}_{e,2}}\right)\mathrm{d}s. \tag{9}$$

Replacing derivatives along directions $\mathbf{t}_{e,1}$ and $\mathbf{t}_{e,2}$ by finite differences, we get

$$\int_e \frac{\partial c}{\partial \mathbf{t}_{e,i}}\,\mathrm{d}s = |e|\left(\frac{C_{e,i}-C_T}{|\mathbf{x}_{e,i}-\mathbf{x}_T|} + O(|\mathbf{x}_{e,i}-\mathbf{x}_T|)\right), \quad i = 1,2. \tag{10}$$

Note that this formula is exact for linear concentrations. If $\mathbf{x}_{e,i}$ is the secondary collocation point, we use formula (33) or (34) for $C_{e,i}$. Using the finite difference approximations (10) in (9), we get the following discrete diffusive flux:

$$\mathbf{q}^h_{e,d}\cdot\mathbf{n}_e = -|\boldsymbol{\ell}_e|\left(\frac{\alpha_e}{|\mathbf{t}_{e,1}|}(C_{e,1}-C_T) + \frac{\beta_e}{|\mathbf{t}_{e,2}|}(C_{e,2}-C_T)\right). \tag{11}$$

At the moment, this flux involves three rather than two concentrations. To derive a two-point flux approximation, we consider polygon T_- and derive another approximation of the same flux through edge e. To distinguish between T_+ and T_-, we add subscripts ± and omit subscript e. Since $\mathbf{n}_e$ is the inward normal vector for T_-, we have to change sign of the right-hand side:

$$\mathbf{q}^h_{\pm,d} \cdot \mathbf{n}_e = \mp|\boldsymbol{\ell}_e| \left(\frac{\alpha_\pm}{|\mathbf{t}_{\pm,1}|} (C_{\pm,1} - C_\pm) + \frac{\beta_\pm}{|\mathbf{t}_{\pm,2}|} (C_{\pm,2} - C_\pm) \right), \tag{12}$$

where $\alpha_\pm$ and $\beta_\pm$ are given by Lemma 3.1 and $C_{\pm,i}$ denotes concentration at collocation point $\mathbf{x}_{\pm,i}$ from $\Sigma_{T_\pm}$.

We define a new flux as a linear combination of two fluxes (12) with non-negative weights $\mu_\pm$:

$$\begin{aligned} \mathbf{q}^h_{e,d} \cdot \mathbf{n}_e &= \mu_+ \, \mathbf{q}^h_{+,d} \cdot \mathbf{n}_e + \mu_- \, \mathbf{q}^h_{-,d} \cdot \mathbf{n}_e \\ &= \mu_+ |\boldsymbol{\ell}_e| \left(\frac{\alpha_+}{|\mathbf{t}_{+,1}|} + \frac{\beta_+}{|\mathbf{t}_{+,2}|} \right) C_+ - \mu_- |\boldsymbol{\ell}_e| \left(\frac{\alpha_-}{|\mathbf{t}_{-,1}|} + \frac{\beta_-}{|\mathbf{t}_{-,2}|} \right) C_- \\ &- \mu_+ |\boldsymbol{\ell}_e| \left(\frac{\alpha_+}{|\mathbf{t}_{+,1}|} C_{+,1} + \frac{\beta_+}{|\mathbf{t}_{+,2}|} C_{+,2} \right) + \mu_- |\boldsymbol{\ell}_e| \left(\frac{\alpha_-}{|\mathbf{t}_{-,1}|} C_{-,1} + \frac{\beta_-}{|\mathbf{t}_{-,2}|} C_{-,2} \right). \end{aligned} \tag{13}$$

The first requirement for the weights is to cancel the terms in the last row of (13) which results in a two-point flux formula. The second requirement is to approximate the true flux. These requirements lead us to the following system:

$$\begin{cases} -\mu_+ d_+ + \mu_- d_- = 0, \\ \mu_+ + \mu_- = 1, \end{cases} \tag{14}$$

where

$$d_\pm = |\boldsymbol{\ell}_e| \left(\frac{\alpha_\pm}{|\mathbf{t}_{\pm,1}|} C_{\pm,1} + \frac{\beta_\pm}{|\mathbf{t}_{\pm,2}|} C_{\pm,2} \right). \tag{15}$$

Since coefficients $d_\pm$ depend on both geometry and concentration, so do weights $\mu_\pm$. Thus, the resulting two-point flux approximation is *nonlinear*.

Remark 3.1 *Note that the concentration $C_{+,i}$ (resp., $C_{-,i}$) may be defined at the same collocation point as C_- (resp., C_+). In this case the terms to be canceled are changed. By doing so, we recover the classical linear scheme for square meshes with the 4-1-1-1-1 stencil. A similar conclusion can be drawn for centroidal Voronoi meshes. To simplify the presentation, we shall not consider this and similar special cases.*

The solution of (14) can be written explicitly. In all cases, $d_\pm \geqslant 0$ for non-negative concentrations. If $d_\pm = 0$, we set $\mu_+ = \mu_- = \frac{1}{2}$. Otherwise,

$$\mu_+ = \frac{d_-}{d_- + d_+} \qquad \text{and} \qquad \mu_- = \frac{d_+}{d_- + d_+}. \tag{16}$$

Thus, the weights $\mu\pm$ are non-negative. Substituting this into (13), we get the two-point flux

$$\mathbf{q}^h_{e,d} \cdot \mathbf{n}_e = D_e^+ C_{T_+} - D_e^- C_{T_-} \tag{17}$$

with coefficients

$$D_e^{\pm} = \mu_{\pm}|\boldsymbol{\ell}_e|(\alpha_{\pm}/|\mathbf{t}_{\pm,1}| + \beta_{\pm}/|\mathbf{t}_{\pm,2}|). \tag{18}$$

Remark 3.2 *Although formula (11) is invariant with respect to the addition of a constant function, the discrete flux (17) is defined correctly only for non-negative concentrations. Analysis below requires to extend definition of the discrete diffusive flux to negative concentrations. It can be done by adding the smallest positive constant to all concentrations in (13) that makes them non-negative.*

3.3 Nonlinear advection flux on interior edges

In this section we consider the advection flux on an interior edge $e \in \mathcal{E}_I$,

$$\mathbf{q}_{e,a} = \frac{1}{|e|}\int_e c\mathbf{v}\,\mathrm{d}s,$$

and its nonlinear upwind approximation

$$\mathbf{q}_{e,a}^h \cdot \mathbf{n}_e = v_e^+ \mathcal{R}_{T^+}(\mathbf{x}_e) + v_e^- \mathcal{R}_{T^-}(\mathbf{x}_e), \tag{19}$$

where

$$v_e^+ = \frac{1}{2}(v_e + |v_e|), \quad v_e^- = \frac{1}{2}(v_e - |v_e|), \quad v_e = \frac{1}{|e|}\int_e \mathbf{v}\cdot\mathbf{n}_e\,\mathrm{d}s,$$

$\mathcal{R}_T$ is a linear reconstruction of the concentration over cell T which depends on the concentration values from neighboring cells.

On each cell T we define the linear reconstruction

$$\mathcal{R}_T(\mathbf{x}) = \begin{cases} C_T + \mathcal{L}_T\mathbf{g}_T \cdot (\mathbf{x} - \mathbf{x}_T), & \mathbf{x} \in T, \\ 0, & \mathbf{x} \notin T, \end{cases} \tag{20}$$

where $\mathbf{g}_T$ denotes the gradient of the least-square linear reconstruction, and $\mathcal{L}_T$ is a slope limiting operator. The vector $\mathbf{g}_T$ is recovered from values C_k collocated at points $\mathbf{x}_k$ from a set $\tilde{\Sigma}_T$ which is defined as follows. First, the set $\hat{\Sigma}_T$ is defined by eliminating the secondary collocation points $\mathbf{x}_e$, $e \in \mathcal{E}_B^{out}$, from Σ_T. Second, we set $\tilde{\Sigma}_T = \hat{\Sigma}_T$ and extend it if the least-square system is degenerate or ill-conditioned. More precisely, if $\tilde{\Sigma}_T = \{\mathbf{x}_T, \mathbf{x}_{T'}\}$, we add to $\tilde{\Sigma}_T$ the elements of $\hat{\Sigma}_{T'}$ other than $\mathbf{x}_T$. If $\tilde{\Sigma}_T = \{\mathbf{x}_T, \mathbf{x}_{T'}, \mathbf{x}_{T''}\}$ and area of the triangle with vertices $\mathbf{x}_T$, $\mathbf{x}_{T'}$, $\mathbf{x}_{T''}$ is less than $10^{-3}|T|$, we add to $\tilde{\Sigma}_T$ the elements of $\hat{\Sigma}_{T'}$ and $\hat{\Sigma}_{T''}$ other than $\mathbf{x}_T$. Then, the vector $\mathbf{g}_T$ is defined as the minimizer of the least-square functional

$$\mathcal{J}_{LS}(\mathbf{g}_T) = \min_{\mathbf{g}\in\Re^2} \sum_{\mathbf{x}_k\in\tilde{\Sigma}_T} [C_T + \mathbf{g}\cdot(\mathbf{x}_k - \mathbf{x}_T) - C_k)]^2. \tag{21}$$

By construction, we ensure the following result.

Lemma 3.2 *The problem (21) has a unique solution.*

The slope limiting operator $\mathcal{L}_T$ is introduced to avoid non-physical extrema. The modified slope $\mathcal{L}_T\mathbf{g}_T$ must result in linear reconstruction that satisfies the following restrictions at collocation points $\mathbf{x}_k \in \hat{\Sigma}_T$:

$$\min\{C_T, C_1, \dots, C_{N(\hat{\Sigma}_T)}\} \leqslant C_T + \mathcal{L}_T\mathbf{g}_T \cdot (\mathbf{x}_k - \mathbf{x}_T) \leqslant \max\{C_T, C_1, \dots, C_{N(\hat{\Sigma}_T)}\}. \quad (22)$$

Additionally, vector $\mathcal{L}_T\mathbf{g}_T$ must meet the following restrictions at points $\mathbf{x}_e$ on edges $e \in \mathcal{E}_T$ where $v_e > 0$:

$$\pm v_e^{\pm}(C_T + \mathcal{L}_T\mathbf{g}_T \cdot (\mathbf{x}_e - \mathbf{x}_T)) \geqslant 0, \quad e \in \mathcal{E}_T. \quad (23)$$

This condition guarantees correct sign of the advective flux. Finally, the reconstructed solution must be bounded from below at the secondary collocation points on the outflow boundary:

$$\min\{C_T, C_1, \dots, C_{N(\hat{\Sigma}_T)}\} \leqslant C_T + \mathcal{L}_T\mathbf{g}_T \cdot (\mathbf{x}_e - \mathbf{x}_T), \qquad e \in \mathcal{E}_T \cap \mathcal{E}_B^{out}. \quad (24)$$

These restrictions were designed to bring as small as possible changes of the reconstructed least-square slope $\mathbf{g}_T$. Note also that by virtue of (20), any change of $\mathbf{g}_T$ will preserve the mass on T and by virtue of (22), $\mathcal{L}_T\mathbf{g}_T \equiv 0$ in local minima and maxima.

We define the action of the slope limiting operator, $\mathcal{L}_T\mathbf{g}_T$, as the solution of the constrained minimization problem

$$\mathcal{J}_{SL}(\mathcal{L}_T\mathbf{g}_T) = \min_{\mathbf{g}' \text{satisfies (22),(23),(24)}} \mathcal{J}_{SL}(\mathbf{g}') \quad (25)$$

where the deviation functional $\mathcal{J}_{SL}$ is

$$\mathcal{J}_{SL}(\mathbf{g}') = \frac{1}{2} \sum_{\mathbf{x}_k \in \tilde{\Sigma}_T} |(\mathbf{g}' - \mathbf{g}_T) \cdot (\mathbf{x}_k - \mathbf{x}_T)|^2.$$

Lemma 3.3 *Minimization problem (25) has a unique solution.*

Proof. A solution to problem (25) does exist, since the constant reconstruction $\mathbf{g}' = (0,0)^T$ satisfies (22), (23) and (24). However, it does not provide the minimum of $\mathcal{J}_{SL}(\mathbf{g}')$. The problem (25) reduces to a problem of linear programming [28]: given a point $\boldsymbol{\xi}_g$ on a plane and a convex polygon P, find

$$\boldsymbol{\xi} = \arg\min_{\boldsymbol{\xi}' \in \mathrm{P}} |\boldsymbol{\xi}' - \boldsymbol{\xi}_g|. \quad (26)$$

Indeed, our restrictions generate a set of strips and half-planes whose intersection is a convex polygon P. If $\boldsymbol{\xi}_g \notin \mathrm{P}$, the solution $\boldsymbol{\xi}$ is the orthogonal projection of $\boldsymbol{\xi}_g$ on $\partial\mathrm{P}$ which is unique. □

Using (19) and (20), we represent the advective flux as the sum of a linear part (the first-order approximation) and a nonlinear part (the second-order correction):

$$\mathbf{q}_{e,a}^h \cdot \mathbf{n}_e = A_e^+ C_+ - A_e^- C_-, \quad (27)$$

where

$$A_e^{\pm} = \pm v_e^{\pm}(1 + \mathcal{L}_{\pm}\mathbf{g}_{\pm} \cdot (\mathbf{x}_e - \mathbf{x}_{\pm})C_{\pm}^{-1}) \quad (28)$$

and subscript ± stands for $T_\pm$.

We note that the coefficients $A_e^\pm$ are non-negative for positive concentrations. If $C_T = 0$ in a cell T then $\mathcal{L}_T$ becomes the zero matrix and $A_e^\pm = \pm v_e^\pm$.

3.4 Fluxes on boundary edges

Let us consider a Robin boundary edge $e \in \mathcal{E}_B^R$. The total flux through this edge is

$$\mathbf{q}_e^h \cdot \mathbf{n}_e = \bar{g}_{R,e}|e|, \tag{29}$$

where $\bar{g}_{R,e}$ is the mean value of g_R on edge e. Despite that this flux is given, there may be diffusive fluxes (13) that use the concentration C_e. Thus, an independent equation for concentration C_e is needed. In the subsequent discussion, it may be convenient to think about e as the cell with zero area. Let T be the cell with edge e. Replacing C_+ and C_- with C_T and C_e, respectively, we get

$$\mathbf{q}_{e,d}^h \cdot \mathbf{n}_e = D_e^+ C_T - D_e^- C_e, \tag{30}$$

where coefficients $D_e^\pm$ are given by (18).

The approximation of the advective flux adopts formula (27):

$$\mathbf{q}_{e,a}^h \cdot \mathbf{n}_e = A_e^+ C_T. \tag{31}$$

Thus, the equation for the total flux is

$$(\mathbf{q}_{e,d}^h + \mathbf{q}_{e,a}^h) \cdot \mathbf{n}_e = \bar{g}_{R,e}|\mathbf{n}_e| \quad e \in \mathcal{E}_B^R. \tag{32}$$

Substituting (30) and (31) in (32), we get the required equation for C_e:

$$A_e^+ C_T + D_e^+ C_T - D_e^- C_e = \bar{g}_{R,e}|\mathbf{n}_e|, \tag{33}$$

where coefficients D_e^+, D_e^- and A_e^+ are non-negative for positive concentrations. Since $\bar{g}_{R,e} \leqslant 0$ then C_e is non-negative if C_T is non-negative.

Let us consider a Dirichlet boundary edge $e \in \mathcal{E}_B^D$. Let T be again the cell containing this edge. The equation for concentration is trivial,

$$C_e = \bar{g}_{D,e} = \frac{1}{|e|}\int_e g_D \,\mathrm{d}s. \tag{34}$$

The approximation of the diffusive flux is given by formula (30). The approximation of the advective flux depends on velocity direction on edge e. If $e \in \mathcal{E}_B^{out}$, the approximation adopts formulas (31) and (28). If $e \in \mathcal{E}_B^{in}$, we use

$$\mathbf{q}_{e,a}^h \cdot \mathbf{n}_e = -A_e^-, \tag{35}$$

where

$$A_e^- = -\frac{1}{|e|}\int_e g_D \mathbf{v}\cdot\mathbf{n}_e \,\mathrm{d}s \geqslant 0. \tag{36}$$

4 Discrete system and monotonicity analysis

For every T in $\mathcal{T}$, the cell equation (4) is

$$\sum_{e\in\mathcal{E}_T} \chi(T,e)\,\mathbf{q}_e^h\cdot\mathbf{n}_e = \int_T f\,\mathrm{d}x, \tag{37}$$

where $\chi(T,e)$ is either 1 or -1 depending on mutual orientation of normal vectors $\mathbf{n}_e$ and $\mathbf{n}_T$. Substituting two-point flux formula (5) with non-negative coefficients

$$M_e^\pm = D_e^\pm + A_e^\pm$$

given by (18) and (28) into (37), and using equations (33) and (34) to eliminate boundary concentrations, we get a nonlinear system of $N_\mathcal{T}$ equations

$$\mathbf{M}(\mathbf{C})\mathbf{C} = \mathbf{F}(\mathbf{C}), \tag{38}$$

where $\mathbf{C}$ is the vector of discrete concentrations at the primary collocation points. The matrix $\mathbf{M}(\mathbf{C})$ is assembled from 2×2 matrices

$$\mathbf{M}_e(\mathbf{C}) = \begin{pmatrix} M_e^+(\mathbf{C}) & -M_e^-(\mathbf{C}) \\ -M_e^+(\mathbf{C}) & M_e^-(\mathbf{C}) \end{pmatrix} \tag{39}$$

for the interior edges and 1×1 matrices $\mathbf{M}_e(\mathbf{C}) = M_e^+(\mathbf{C})$ for Dirichlet edges. The right-hand side vector $\mathbf{F}(\mathbf{C})$ is generated by the source and the boundary data:

$$F_T(\mathbf{C}) = \int_T f\,\mathrm{d}x + \sum_{e\in\mathcal{E}_B^D\cap\partial T} M_e^-(\mathbf{C})\bar{g}_{D,e} - \sum_{e\in\mathcal{E}_B^R\cap\partial T} |e|\bar{g}_{R,e}, \qquad \forall T\in\mathcal{T}. \tag{40}$$

For $f(x)\geqslant 0$, $g_D\geqslant 0$ and $g_R\leqslant 0$ the components of vector F are non-negative. We use the Picard iterations to solve the nonlinear system (38) (see Algorithm 1).

The linear system in Step 8 with the non-symmetric matrix $\mathbf{M}(\mathbf{C}^k)$ is solved by the Bi-Conjugate Gradient Stabilized (BiCGStab) method [29] with the second-order ILU preconditioner [14]. The BiCGStab iterations are terminated when the relative norm of the residual becomes smaller than ε_{lin}.

The next theorem shows that the solution to (38) is non-negative provided that it exists.

Theorem 4.1 *Let $\Gamma_R = \emptyset$ ($\mathcal{E}_B^D \equiv \mathcal{E}_B$), $f\geqslant 0$ in Ω, $g_D\geqslant 0$ on $\Gamma_D\equiv\partial\Omega$ and the solution C to (38) exist. Then $\mathbf{C}\geqslant 0$.*

Algorithm 1 Generation and solution of nonlinear system (38)

```
For each interior edge $e \in \mathcal{E}_I$ shared by elements $T_\pm$ find vectors $\mathbf{t}_{\pm,1}$, $\mathbf{t}_{\pm,2}$ satisfying conditions (6) and (7). Find similar vectors for boundary edges.
Select an initial vector $\mathbf{C}^0$ with non-negative entries and a small value $\varepsilon_{non} > 0$.
for $k = 0, \dots,$ do
    Calculate concentrations $C_e$ at the secondary collocation points on edges $e \in \mathcal{E}_B$ using (33),(34).
    Assemble the global matrix $\mathbf{M}(\mathbf{C}^k)$ from the edge-based matrices $\mathbf{M}_e(\mathbf{C}^k)$. Use formulas (18) with (15), (16), and (28) to form $\mathbf{M}_e(\mathbf{C}^k)$.
    Calculate the right-hand side vector $\mathbf{F}(\mathbf{C}^k)$ using (40).
    Stop if $\|\mathbf{M}(\mathbf{C}^k)\mathbf{C}^k - \mathbf{F}(\mathbf{C}^k)\| \leqslant \varepsilon_{non}\,\|\mathbf{M}(C^0)\mathbf{C}^0 - \mathbf{F}(\mathbf{C}^0)\|$.
    Solve $\mathbf{M}(\mathbf{C}^k)\mathbf{C}^{k+1} = \mathbf{F}(\mathbf{C}^k)$.
end for
```

Proof. The proof is by contradiction. Let us consider the cell T with the smallest concentration C_T and assume that $C_T < 0$. Let $T = T_+$ in the flux formulas. Since C_T is minimal, $\mathcal{R}_T \equiv C_T$. By adding and subtracting $v_e^- C_T$, we get

$$\sum_{e\in\mathcal{E}_T} \mathbf{q}_{e,a}^h \cdot \mathbf{n}_e = C_T \sum_{e\in\mathcal{E}_T} v_e + \sum_{e\in\mathcal{E}_T\setminus\mathcal{E}_B^{in}} v_e^- (\mathcal{R}_{T'_e}(\mathbf{x}_e) - C_T) + \sum_{e\in\mathcal{E}_T\cap\mathcal{E}_B^{in}} v_e^- (\bar{g}_{D,e} - C_T).$$

Therefore, from (37) we derive

$$-C_T \sum_{e\in\mathcal{E}_T} v_e + \int\limits_T f \,\mathrm{d}x - \sum_{e\in\mathcal{E}_T} \mathbf{q}_{e,d}^h \cdot \mathbf{n}_e - \sum_{e\in\mathcal{E}_T\setminus\mathcal{E}_B^{in}} v_e^- (\mathcal{R}_{T'_e}(\mathbf{x}_e) - C_T) - \sum_{e\in\mathcal{E}_T\cap\mathcal{E}_B^{in}} v_e^- (\bar{g}_{D,e} - C_T) = 0. \tag{41}$$

We have

$$\sum_{e\in\mathcal{E}_T} v_e = \int\limits_{\partial T} \mathbf{v}\cdot\mathbf{n}_e \,\mathrm{d}s = \int\limits_T div(\mathbf{v})\,\mathrm{d}x \geqslant 0,$$

and, by assumption,

$$C_T \sum_{e\in\mathcal{E}_T} v_e \leqslant 0.$$

Since C_T is minimal, it holds $\mathcal{R}_{T'_e}(\mathbf{x}_e) \geqslant C_T$, and since $C_T < 0$, it holds $\bar{g}_{D,e} > C_T$. Let $\widetilde{\mathbf{C}}$ be a vector with non-negative entries obtained by adding positive constant $-C_T$ to every entry of $\mathbf{C}$. For $e \in \mathcal{E}_T$, we have

$$\mathbf{q}_{e,d}^h(\widetilde{\mathbf{C}}) \cdot \mathbf{n}_e = D_e^+ \widetilde{C}_T - D_e^- \widetilde{C}_{T'_e} = -D_e^- \widetilde{C}_{T'_e} \leqslant 0.$$

As explained in Remark 3.2, the discrete diffusive flux for $\mathbf{C}$ is equal to that for $\widetilde{\mathbf{C}}$. Therefore, $\mathbf{q}_{e,d}^h \cdot \mathbf{n}_e \leqslant 0$ and therefore

$$\sum_{e\in\mathcal{E}_T} \mathbf{q}_{e,d}^h \cdot \mathbf{n}_e \leqslant 0.$$

By virtue of $v_e^- \leqslant 0$ we conclude that all the terms in (41) are non-negative and must be equal to zero. Thus, $\sum\limits_{e\in\mathcal{E}_T} \mathbf{q}_{e,d}^h \cdot \mathbf{n}_e = 0$ and $C_T = \min\limits_{e\in\mathcal{E}_T}\{C_T; C_{T'_e}\}$. This implies

$$0 = \sum_{e\in\mathcal{E}_T} \mathbf{q}_{e,d}^h(\widetilde{\mathbf{C}}) \cdot \mathbf{n}_e = -\sum_{e\in\mathcal{E}_T} D_e^- \, \widetilde{C}_{T'_e}$$

which means $C_{T'_e} = C_T$ for all $e \in \mathcal{E}_T$.

Therefore, instead of T we can consider any neighboring cell T'_e. Since $\mathcal{T}$ is edge-connected, we conclude that C is constant on $\mathcal{T}$. Considering a cell T with edge $e \in \mathcal{E}_B$, from $\bar{g}_{D,e} - C_T = 0$, we get that this constant is non-negative. This contradicts our assumption. □

Let us show that the matrix $\mathbf{M}(\mathbf{C}^k)$ is the M-matrix provided that $\mathbf{C}^k \geqslant 0$. Our derivation shows that coefficients $M_e^\pm(\mathbf{C}^k)$ are positive. Thus, all diagonal entries of matrix $\mathbf{M}(\mathbf{C}^k)$ are positive and all off-diagonal entries of $\mathbf{M}(\mathbf{C}^k)$ are non-positive. The structure of edge-based matrices (39) implies that each column sum in $\mathbf{M}_e(\mathbf{C}^k)$ is non-negative. Moreover, for elements with Dirichlet edges, the corresponding column sum is strictly positive. For a connected mesh, matrices $\mathbf{M}(\mathbf{C}^k)$ and $\mathbf{M}^T(\mathbf{C}^k)$ are irreducible since their directed graphs are strongly connected. Under the above conditions, the well known linear algebra result [30] implies that matrix $\mathbf{M}^T(\mathbf{C}^k)$ is the M-matrix and all entries of $(\mathbf{M}^T(\mathbf{C}^k))^{-1}$ are positive. Since the inverse and transpose operations commute, $(\mathbf{M}^T(\mathbf{C}^k))^{-1} = (\mathbf{M}^{-1}(\mathbf{C}^k))^T$, we conclude that $\mathbf{M}(\mathbf{C}^k)$ is monotone. Since diagonal entries of $\mathbf{M}(\mathbf{C}^k)$ are positive and off-diagonal entries are negative, it is also the M-matrix. Therefore, we proved the following theorem.

Theorem 4.2 *Let $f \geqslant 0$, $g_D \geqslant 0$, $g_R \leqslant 0$ and $\Gamma_D \neq \emptyset$ in (1). If $C^0 \geqslant 0$ and linear systems in the Picard method are solved exactly, then $C^k \geqslant 0$ for $k \geqslant 1$.*

Remark 4.1 *The theorem holds true also for linear advective fluxes:*

$$\mathbf{q}_{e,a}^h \cdot \mathbf{n}_e = A_e^+ C_+ - A_e^- C_-, \qquad A_e^\pm = \pm v_e^\pm.$$

5 Numerical experiments

5.1 Implementation issues

In all experiments, we set $\Gamma_R = \emptyset$. For advection-dominated problems, this helps to find more analytical solutions such that the right-hand side vector is non-negative, $F(C) \geqslant 0$, for any non-negative C.

5.1.1 Errors

We use the following discrete L_2-norms to evaluate discretization errors for the concentration c and the flux $\mathbf{q}$:

$$\varepsilon_2^c = \left[\frac{\sum\limits_{T\in\mathcal{T}} (c(\mathbf{x}_T) - C_T)^2\, |T|}{\sum\limits_{T\in\mathcal{T}} (c(\mathbf{x}_T))^2\, |T|}\right]^{1/2} \qquad \text{and} \qquad \varepsilon_2^q = \left[\frac{\sum\limits_{e\in\mathcal{E}_I\cup\mathcal{E}_B} \left((\mathbf{q}_e - \mathbf{q}_e^h)\cdot\mathbf{n}_e\right)^2 |S_e|}{\sum\limits_{e\in\mathcal{E}_I\cup\mathcal{E}_B} (\mathbf{q}_e\cdot\mathbf{n}_e)^2\, |S_e|}\right]^{1/2},$$

where $|S_e|$ is a representative area for edge e. More precisely, $|S_e|$ is the arithmetic average of areas of mesh cells sharing the edge. In convergence studies the nonlinear iterations are terminated when the reduction of the initial residual norm becomes smaller then $\varepsilon_{non} = 10^{-8}$. The convergence tolerance for the linear solver is set to $\varepsilon_{lin} = 10^{-12}$.

5.1.2 Meshes

The numerical tests are performed on three sequences of uniform meshes, two sequences of distorted structured meshes, and one sequence of polygonal meshes. The uniform meshes are square meshes $\{\mathbf{M1}\}$ and two types of triangular meshes produced by splitting each square cell into two triangles by the north-east $\{\mathbf{M2}\}$ or north-west diagonal $\{\mathbf{M3}\}$, as shown in Fig. 2.

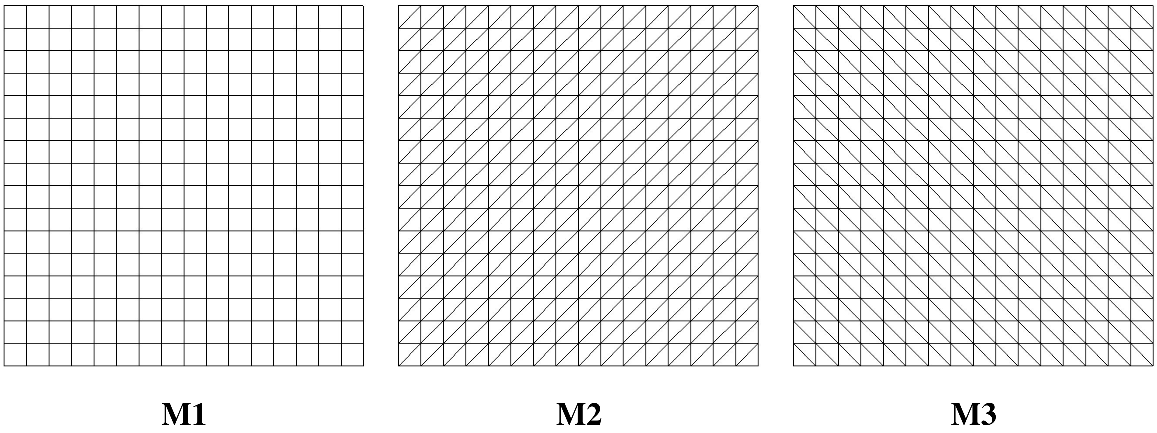


Fig. 2. Examples of three types of uniform meshes.

The distorted structured meshes include triangular $\{\mathbf{M4}\}$ and quadrilateral $\{\mathbf{M5}\}$ meshes. The distorted mesh is constructed from the uniform mesh with the mesh size h by random distortion of internal nodes (x, y):

$$x := x + \alpha\xi_x h, \qquad y := y + \alpha\xi_y h, \tag{42}$$

where ξ_x and ξ_y are random variables with values between -0.5 and 0.5 and $\alpha \in [0, 1]$ is the degree of distortion. To avoid mesh tangling, we set $\alpha = 0.6$ for both types of meshes. It is pertinent to emphasize that the distortion is performed on each refinement level. A polygonal mesh from sequence $\{\mathbf{M6}\}$ is a dual mesh for a smoothly transformed uniform triangular mesh. Examples of these meshes are shown in Fig. 3. For each space resolution, the quadrilateral and polygonal meshes have roughly the same number of cells. The corresponding triangular meshes have twice more cells.

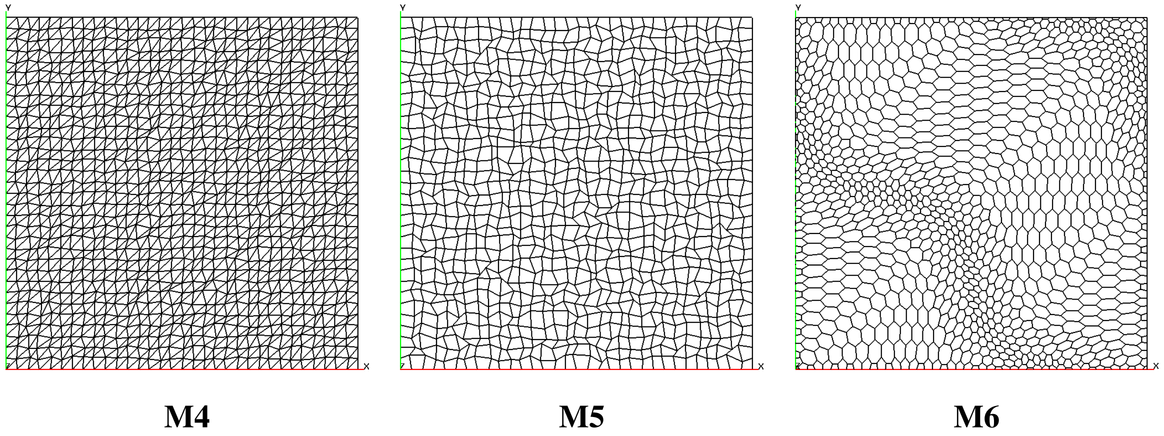


Fig. 3. Examples of two types of distorted structured meshes and a polygonal mesh.

5.2 *Anisotropic diffusion with advection*

5.2.1 *Convergence study*

The convergence study is performed for a smooth solution on mesh sequences $\{\mathbf{M4}\}$, $\{\mathbf{M5}\}$ and $\{\mathbf{M6}\}$. A sequence of distorted meshes is the most challenging test for a numerical scheme due to fix amount of random noise in position of mesh nodes. Let $\Omega = (0,1)^2$, and the exact solution, velocity field and anisotropic diffusion tensor be as follows

$$c(x,y) = x\cos(0.5\pi y), \qquad \mathbf{v} = (1,-1)^T, \qquad \mathbb{K} = \begin{pmatrix} 10 & 0 \\ 0 & 0.1 \end{pmatrix}.$$

The forcing term f and the Dirichlet boundary data g_D are set accordingly to the exact solution. Table 1 shows the relative L_2 norms of the errors. The convergence rate for the concentration is close to the second-order while the convergence rate for the flux is higher than the first-order. It is interesting to note that the order of convergence on quadrilateral and polygonal meshes is better than on triangular ones. This can be explained by the fact that a greater number of neighboring cells allows one to approximate a diffusion flux more accurately. This is one of the advantages of usage of polygonal meshes in the discretization.

5.2.2 *Monotonicity test*

The monotonicity study is performed of mesh sequences $\{\mathbf{M1}\}$, $\{\mathbf{M2}\}$ and $\{\mathbf{M3}\}$ for a problem with anisotropic solution due to highly anisotropic diffusion tensor. Such a problem is a challenging task for a wide range of discretization methods, e.g. [19,22], which may significantly violate the discrete maximum principle and produce a numerical solution with non-physical oscillations. We consider problem (1) in the unit square with a square hole, $\Omega = (0,1)^2/[4/9,5/9]^2$,

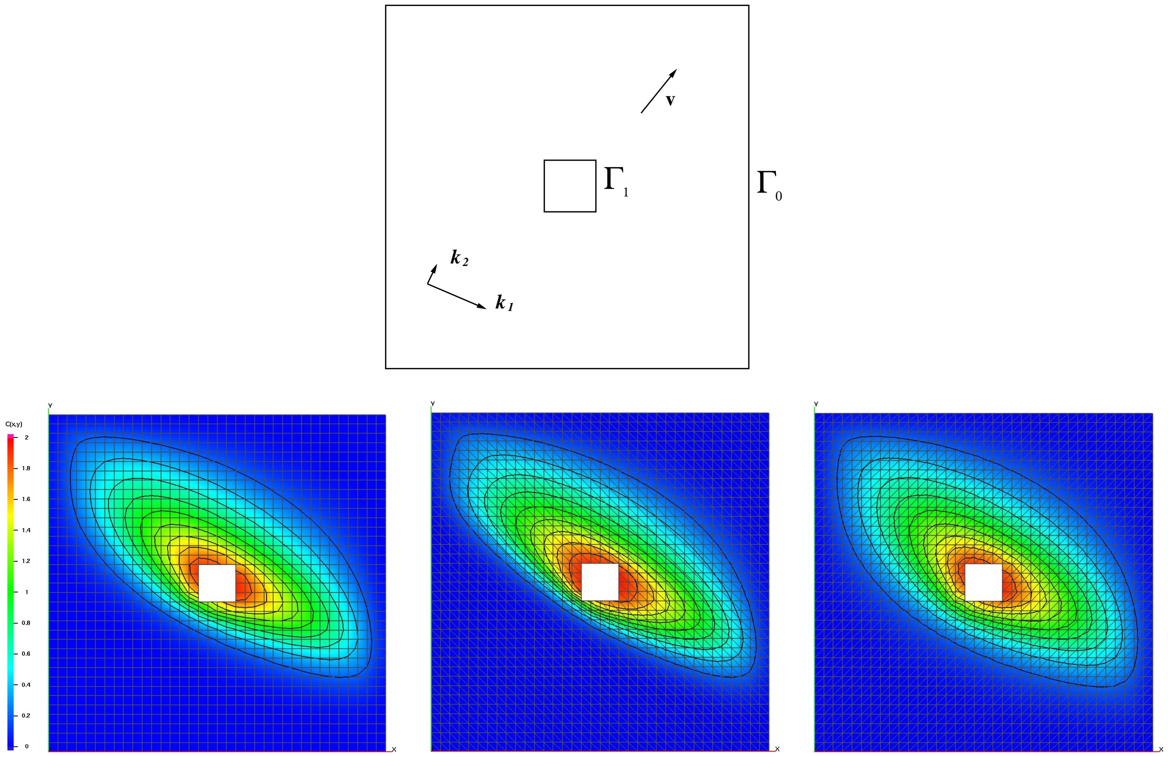
h	$\{$**M4**$\}$		$\{$**M5**$\}$		$\{$**M6**$\}$	
	ε_2^C	ε_2^q	ε_2^C	ε_2^q	ε_2^C	ε_2^q
$1/32$	7.17e-04	3.29e-03	3.28e-04	2.28e-03	8.79e-04	4.78e-03
$1/64$	2.69e-04	1.23e-03	7.10e-05	8.65e-04	2.73e-04	1.73e-03
$1/128$	9.82e-05	4.82e-04	2.01e-05	3.62e-04	7.33e-05	6.03e-04

Table 1
Convergence analysis for diffusion-dominated problems.

so that the boundary of Ω consists of two disjoint parts as shown in Fig. 4. We set $f = 0$, $g_D = 0$ on Γ_0, $g_D = 2$ on Γ_1, $\mathbf{v} = (700, 700)^T$ and take the following anisotropic diffusion tensor $\mathbb{K}$:

$$\mathbb{K} = R(-\theta) \begin{pmatrix} k_1 & 0 \\ 0 & k_2 \end{pmatrix} R(\theta), \qquad R(\theta) = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}, \tag{43}$$

where $k_1 = 1000, k_2 = 1$ and $\theta = -\pi/6$.



M1 **M2** **M3**

Fig. 4. Top panel: A sketch of the computational domain Ω with the primary directions of the diffusion tensor and the velocity field. Bottom panel: Solutions calculated with the nonlinear FV method on three different meshes.

According to the maximum principle for elliptic PDEs, the exact solution should be between 0 and 2. Solutions computed with the nonlinear FV method on triangular and square meshes are non-negative everywhere in the computational domain (see the color bar in Fig. 4). The solution profile on meshes **M1** and **M3** is wider than on mesh **M2** since the mesh size along the velocity direction is twice larger. It is pertinent to notice that our approach guarantees only the nonnegativity of the numerical solution. It means that small overshoots may occur and were observed in [23]. We note that the solution calculated with the lowest-order Raviart-Thomas MFE method is negative on both triangular meshes over large regions, even when $\mathbf{v} = (0,0)^T$ [22,23].

5.3 Advection dominated problems

5.3.1 Convergence study for smooth solutions

Firstly, we study the accuracy and the convergence order for a problem with a smooth solution. The convergence studies are performed on mesh sequences $\{$**M4**$\}$, $\{$**M5**$\}$ and $\{$**M6**$\}$. Let $\Omega = (0,1)^2$ and the exact solution, constant velocity field and anisotropic diffusion tensor be as follows:

$$c(x,y) = x\cos(0.5\pi y), \qquad \mathbf{v} = (1,-1)^T, \qquad \mathbb{K} = 10^{-5}\begin{pmatrix} 10 & 0 \\ 0 & 0.1 \end{pmatrix}.$$

The forcing term f and the Dirichlet boundary data g_D are set accordingly to the exact solution. Table 2 shows the relative L_2 norms of the errors. For all types of meshes we observe the tendency to the second-order convergence rate for the concentration and the rate higher than the first-order for the flux.

h	$\{$**M4**$\}$		$\{$**M5**$\}$		$\{$**M6**$\}$	
	ε_2^C	ε_2^q	ε_2^C	ε_2^q	ε_2^C	ε_2^q
$1/32$	3.04e-03	3.78e-03	7.49e-04	7.55e-04	1.19e-03	9.86e-04
$1/64$	9.77e-04	1.23e-03	2.89e-04	3.31e-04	3.86e-04	3.67e-04
$1/128$	3.24e-04	4.32e-04	7.09e-05	8.12e-05	1.20e-04	1.12e-04

Table 2
Convergence analysis for the advection-dominated problem and the smooth solution.

5.3.2 Convergence study for solutions with boundary layers

Secondly, we study the accuracy and the convergence order for a problem with an exponential boundary layer. We consider the problem which also was studied in [24]. The exact solution, constant velocity field and isotropic diffusion tensor are defined by

$$c(x,y) = \left(x - \exp\left(\frac{2(x-1)}{\nu}\right)\right)\left(y^2 - \exp\left(\frac{3(y-1)}{\nu}\right)\right), \qquad \mathbf{v} = (2,3)^T \qquad \mathbb{K} = \nu\,\mathbb{I},$$

where ν characterizes the thickness of the boundary layer in the top-right corner of the unit square $(0,1)^2$. For the advection-dominated problem, we set $\nu = 10^{-4}$. The goal of our numerical tests is to demonstrate that the nonlinear FV method has good convergence properties and produces the numerical solution without oscillations in a subdomain outside the boundary layer. More precisely, the errors are computed in the domain $(0,0.8)^2$. The results presented in Table 3 demonstrate the second-order convergence rate for the concentration and the first-order for the flux on all types of considered meshes. Moreover, in all numerical tests the numerical solutions vary between 0 and 1.

h	{**M4**}		{**M5**}		{**M6**}	
	ε_2^C	ε_2^q	ε_2^C	ε_2^q	ε_2^C	ε_2^q
$1/32$	1.07e-03	3.85e-04	1.91e-03	1.66e-03	4.97e-03	4.43e-03
$1/64$	2.74e-04	1.03e-04	4.94e-04	4.32e-04	1.31e-03	1.13e-03
$1/128$	6.83e-05	2.62e-05	1.27e-04	1.13e-04	3.30e-04	2.82e-04

Table 3
Convergence analysis for the advection-dominated problem and the solution with the boundary layer.

5.3.3 Monotonicity test

In this subsection we consider the advection-dominated problem with discontinuous Dirichlet boundary data. The discontinuity produces an internal shock in the solution, in addition to exponential boundary layers. This is a popular test case for the discretization schemes designed for advection-dominated regimes, see [12] and [13]. Following [13], we set

$$\mathbf{v} = \left(\cos\frac{\pi}{3}, -\sin\frac{\pi}{3}\right), \qquad \mathbb{K} = \nu\mathbb{I}, \qquad \nu = 10^{-8}.$$

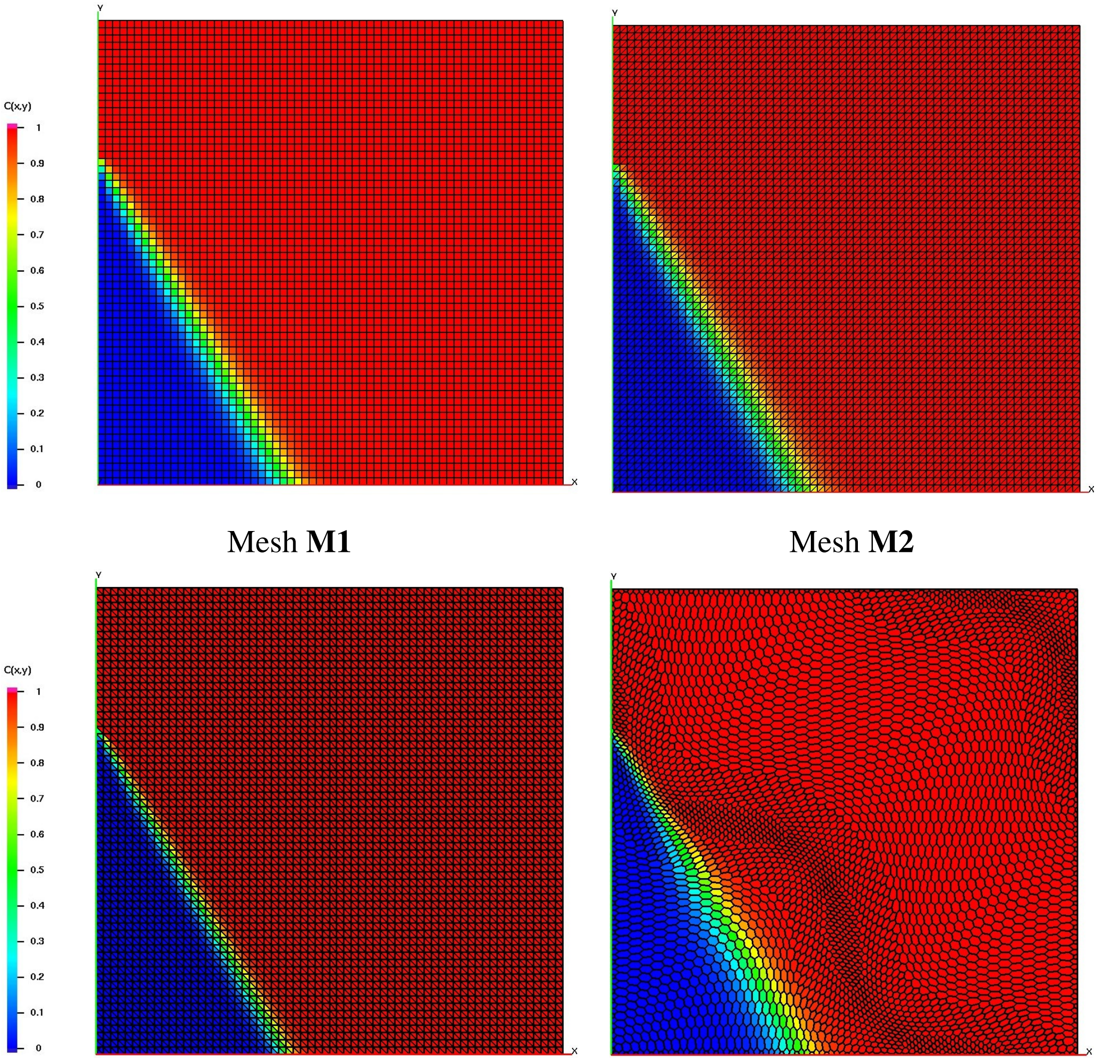
The Dirichlet boundary conditions are imposed as follows:

$$c(x,y) = \begin{cases} 0 & \text{if } x = 1 \text{ or } y \leqslant 0.7, \\ 1 & \text{otherwise} \end{cases}$$

The exact solution has a boundary layer next to the lines $y = 0$, $x = 1$ and has an internal layer along the streamline passing through the point $(0, 0.7)$.

The computations were performed on meshes **M1**, **M2**, **M3** and **M6** with the effective mesh parameter $h = 1/64$, so that the number of degrees of freedom for concentration is 4096 on the square and polygonal meshes and 8192 on the triangular meshes. The Péclet number is $Pe = 7.81^5$. According to Theorems 1 and 2, the numerical solution must be non-negative. The numerical solutions for the four meshes are shown in Fig. 5.

In order to measure quality of the numerical solution, the authors of [13] have proposed several estimates which quantify solution oscillations and smearing effects caused by a discretization



Mesh **M1**

Mesh **M2**

Mesh **M3**

Mesh **M6**

Fig. 5. Monotonicity test: the numerical solutions vary between 0 and 1.

scheme:

$$osc_{\text{int}}^{\text{min}} \equiv \left(\sum_{(x,y)\in\Omega_1} (\min\{0, c_h(x,y)\})^2 \right)^{1/2}, \tag{44}$$

$$osc_{\text{int}}^{\text{max}} \equiv \left(\sum_{(x,y)\in\Omega_1} (\max\{0, c_h(x,y)-1\})^2 \right)^{1/2}, \tag{45}$$

$$osc_{\text{exp}} \equiv \left(\sum_{(x,y)\in\Omega_2} (\max\{0, c_h(x,y)-1\})^2 \right)^{1/2}, \tag{46}$$

$$smear_{\text{exp}} \equiv \left(\sum_{(x,y)\in\Omega_2} (\min\{0, c_h(x,y) - 1\})^2 \right)^{1/2}, \tag{47}$$

$$smear_{\text{int}} \equiv x_2 - x_1, \tag{48}$$

where

$$\Omega_1 = \{(x,y) \in \Omega : x \leqslant 0.5, y \geqslant 0.1\}, \qquad \Omega_2 = \{(x,y) \in \Omega : x \geqslant 0.7\},$$
$$x_1 = \min_{\mathbf{x}_T \in \Omega_3, C(\mathbf{x}_T) \geqslant 0.1} x_T \qquad \text{and} \qquad x_2 = \max_{\mathbf{x}_T \in \Omega_3, C(\mathbf{x}_T) \leqslant 0.9} x_T,$$

with Ω_3 denoting the cell strip in the vicinity of the line $y = 0.25$, $\Omega_3 = \{T \in \mathcal{T} : \mathbf{x}_T = (x_T, y_T), |y_T - 0.25| <= |T|^{1/2}\}$. In the case of mesh **M1** the width of this strip is equal to $2h$.

The estimates (44) and (45) characterize the values of undershoots and overshoots in Ω_1, correspondingly. The estimate (46) quantifies oscillations near the boundary layer in Ω_2 whereas the estimates (47) and (48) measure the width of the boundary layer and the internal shock. In the continuous solution these estimates depend on the diffusion process only, so they are much smaller than the considered mesh size. Small values of estimates (44)-(48) characterize almost non-oscillatory and almost non-diffusive discrete solution.

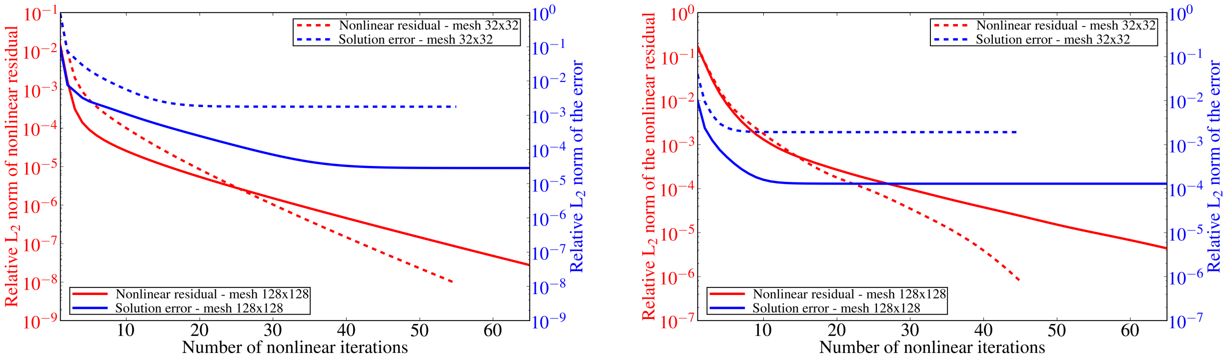
The results obtained by the nonlinear FV method are shown in Table 4. They are competitive with the best results presented in review [13]. The increase of the internal shock width on the polygonal mesh is caused by non-uniformity of mesh density. The cells near the shock are larger than the average cell size.

Mesh	$osc_{\text{int}}^{\min}$	$osc_{\text{int}}^{\max}$	osc_{exp}	$smear_{\text{int}}$	$smear_{\text{exp}}$
M1	0	2.22e-12	1.01e-11	7.81e-02	2.13e-05
M2	0	2.43e-07	7.54e-11	9.90e-02	4.49e-05
M3	0	1.29e-11	6.27e-06	4.69e-02	4.36e-05
M6	0	5.71e-08	1.64e-11	1.13e-01	8.41e-05

Table 4
The quantities that characterize the quality of the numerical solution for the problem described in subsection 5.3.3

5.4 Nonlinear iteration

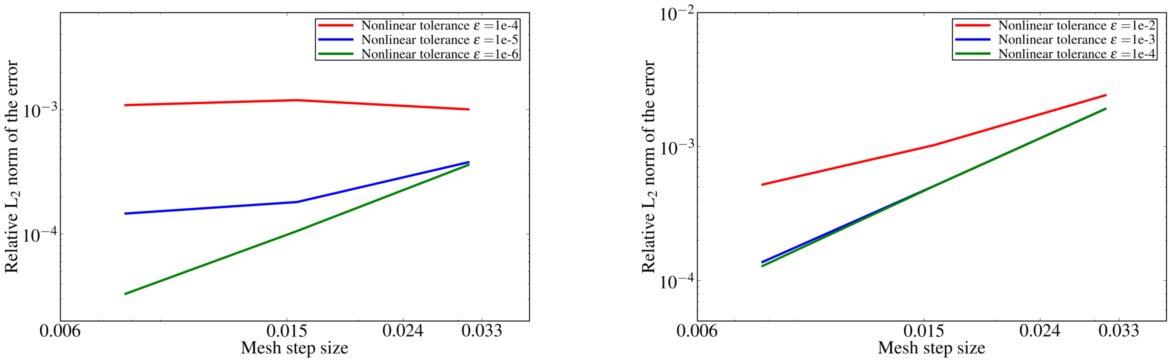
In the last group of tests we investigate the convergence of nonlinear iterations in Algorithm 1. In all numerical experiments presented above, the Picard method was terminated when the discrete L_2 norm of the nonlinear residual was reduced by factor $\varepsilon_{non} = 10^{-8}$. Each iteration of this method is computationally expensive; therefore, reduction in the number of iterations will greatly reduce the overall cost. The goal of this study is to demonstrate that the numerical solution is sufficiently accurate when the nonlinear system (38) is solved with much larger tolerance than 10^{-8}.



(a)

(b)

Fig. 6. The convergence of the Picard method:(a) diffusion-dominated problem from subsection 5.2.1, (b) advection-dominated problem from subsection 5.3.2.



(a)

(b)

Fig. 7. The convergence study for different values nonlinear tolerance ε_{non}:(a) diffusion-dominated problem from subsection 5.2.1, (b) advection-dominated problem from subsection 5.3.2.

We consider the problem with the smooth solution described in subsection 5.2.1 and the problem with the exponential boundary layer described in subsection 5.3.2. Both of these problems are solved on a sequence of distorted quadrilateral meshes $\{\mathbf{M5}\}$. In Fig. 6a,b, the relative L_2 error for concentration and the relative Euclidean norm of the nonlinear residual are plotted for each iteration. The error stabilizes much earlier than the nonlinear residual reaches the prescribed tolerance $\varepsilon_{non} = 10^{-8}$. This difference is even more distinct in the advection-dominated problem. In Fig. 7a,b, the relative L_2 error for concentration is plotted against the mesh size for three different values of the convergence tolerance ε_{non}. These results demonstrate that the second-order convergence can be achieved with much larger tolerance and, respectively, with much smaller number of nonlinear iterations. For example, only 10 nonlinear iterations are required to achieve the second-order convergence in the problem with the exponential boundary layer. For the problem with the smooth solution, gradual decrease of ε_{non} with the mesh size is required to achieve the second-order convergence. Respectively, the number of nonlinear iterations increases from 20 ($h = 1/32$) to 40 ($h = 1/128$).

We noticed that the Picard method may not converge up to the prescribed tolerance in some cases, especially on highly distorted meshes. In these cases, a relaxed version of the Picard method demonstrates much more robust behavior. The iterative process is reformulated as follows:

$$\mathbf{M}(\mathbf{C}^k)\tilde{\mathbf{C}}^{k+1} = \mathbf{F}(\mathbf{C}^k), \qquad \mathbf{C}^{k+1} = \mathbf{C}^k + \omega_k(\tilde{\mathbf{C}}^{k+1} - \mathbf{C}^k),$$

where ω_k is the damping factor, $0 < \omega_k \leqslant 1$. If $\omega_k \equiv 1$, we recover the method described in Algorithm 1. The choice of the damping factor $\{\omega_k\}$ is determined by the balance between the robustness and the convergence speed of the iterative process. Our experience shows that the choice $\omega_k = 0.75$ provides robust behavior for the considered problems. A dynamic choice of the damping factor will be analyzed in the future.

Conclusion

We developed and analyzed the new monotone finite volume method for the advection-diffusion equation with full anisotropic diffusion tensor. We proved that this method guarantees non-negativity of the numerical solution if the source term and the initial guess are non-negative. The numerical scheme does not use any interpolation to the mesh nodes. The method is applicable to polygonal meshes and full anisotropic diffusion tensors with continuous components. Generalization of the method to the case of heterogeneous diffusion coefficients can be done by following the path described in [9,23]. The numerical experiments demonstrate the second-order convergence rate for the concentration and the first-order convergence rate for the flux (a) on randomly distorted meshes, (b) for problems with highly anisotropic coefficients and (c) for advection-dominated and diffusion-dominated problems.

Acknowledgments.

This work was carried out under the auspices of the National Nuclear Security Administration of the U.S. Department of Energy at Los Alamos National Laboratory under Contract No. DE-AC52-06NA25396 and the DOE Office of Science Advanced Scientific Computing Research (ASCR) Program in Applied Mathematics.

The authors thank M.Shashkov (LANL) for fruitful discussions on the paper topic. The authors are also grateful to I.Kapyrin (INM) and K.Nikitin (INM) for useful comments.

References

[1] I. Aavatsmark, G.T. Eigestad, B.T. Mallison, and J.M. Nordbotten. A compact multipoint flux approximation method with improved robustness. *Numer. Methods Partial Differential Equations*, 24(5):1329–1360, 2008.

[2] T.J. Barth. A 3-d upwind euler solver for unstructured meshes. In *AIAA 10th Computational Fluid Dynamics Conference*, pages 228–238, Washington, 1991. Amer Inst Aeronautics & Astronautics.

[3] E. Bertolazzi and G. Manzini. A cell-centered second-order accurate finite volume method for convection-diffusion problems on unstructured meshes. *Mathematical Models &; Methods In Applied Sciences*, 14(8):1235–1260, 2004.

[4] E. Bertolazzi and G. Manzini. A second-order maximum principle preserving finite volume method for steady convection-diffusion problems. *SIAM J. Numer. Anal.*, 43(5):2172–2199, 2005.

[5] A.N. Brooks and T.J.R. Hughes. Streamline upwind/petrov-galerkin formulations for convection dominated flows with particular emphasis on the incompressible navier-stokes equations. *Computer Methods in Applied Mechanics and Engineering*, 32(1-3):199–259, 1982.

[6] E. Burman and A. Ern. Discrete maximum principle for galerkin approximations of the laplace operator on arbitrary meshes. *C. R. Math. Acad. Sci. Paris*, 338(8):641–646, 2004.

[7] G. Chavent and J. Jaffré. *Mathematical models and finite elements for reservoir simulation*. Elsevier Science Publishers, B.V., Netherlands, 1986.

[8] P. G. Ciarlet and P.-A. Raviart. Maximum principle and uniform convergence for the finite element method. *Comput. Methods Appl. Mech. Engrg.*, 2:17–31, 1973.

[9] A. Danilov and Yu. Vassilevski. A monotone nonlinear finite volume method for diffusion equations on conformal polyhedral meshes. *Russian J. Numer. Anal. Math. Modelling*, 24(3):207–227, 2009.

[10] F. Gao, Y. Yuan, and D. Yang. An upwind finite-volume element scheme and its maximum-principle-preserving property for nonlinear convection-diffusion problem. *International Journal for Numerical Methods in Fluids*, 56(12):2301–2320, 2008.

[11] R. Ghostine, G. Kesserwani, R. Mose, J. Vazquez, and A. Ghenaim. An improvement of classical slope limiters for high-order discontinuous galerkin method. *Int. J. Numer. Meth. Fl.*, 59(4):423–442, 2009.

[12] T.J.R. Hughes, M. Mallet, and A. Mizukami. A new finite element formulation for computational fluid dynamics. II. Beyond SUPG. *Computer Methods in Applied Mechanics and Engineering*, 54(3):341–55, 1986.

[13] V. John and P. Knobloch. On spurious oscillations at layers diminishing (sold) methods for convection-diffusion equations: Part I - a review. *Computer Methods In Applied Mechanics And Engineering*, 196(17-20):2197–2215, 2007.

[14] I. E. Kaporin. High quality preconditioning of a general symmetric positive definite matrix based on its $u^tu + u^tr + r^tu$-decomposition. *Numer. Linear Algebra Appl.*, 5(6):483 – 509, NOV-DEC 1998.

[15] Ivan Kapyrin. A family of monotone methods for the numerical solution of three-dimensional diffusion problems on unstructured tetrahedral meshes. *Dokl. Math.*, 76(2):734–738, 2007.

[16] Sergey Korotov, Michal Křížek, and Pekka Neittaanmäki. Weakened acute type condition for tetrahedral triangulations and the discrete maximum principle. *Math. Comp.*, 70(233):107–119 (electronic), 2001.

[17] D. Kuzmin. On the design of general-purpose flux limiters for finite element schemes. I. scalar convection. *Journal of Computational Physics*, 219(2):513–31, 2006.

[18] D. Kuzmin and M. Moller. Algebraic flux correction I. scalar conservation laws. In D Kuzmin, R Lohner, and S Turek, editors, *Flux-Corrected Transport: Principles, Algorithms, And Applications*, pages 155–206, Heidelberger Platz 3, D-14197 Berlin, Germany, 2005. Springer-Verlag Berlin.

[19] D. Kuzmin, M. J. Shashkov, and D. Svyatskiy. A constrained finite element method satisfying the discrete maximum principle for anisotropic diffusion problems. *J. Comput. Phys.*, 228(9):3448–3463, 2009.

[20] S. Lamine and M.G. Edwards. Higher-resolution convection schemes for flow in porous media on highly distorted unstructured grids. *Int. J. Numer. Meth. Eng.*, 76(8):1139–1158, 2008.

[21] C. LePotier. Schema volumes finis monotone pour des operateurs de diffusion fortement anisotropes sur des maillages de triangle non structures. *C. R. Math. Acad. Sci. Paris*, 341:787 – 792, 2005.

[22] K. Lipnikov, D. Svyatskiy, M. Shashkov, and Yu. Vassilevski. Monotone finite volume schemes for diffusion equations on unstructured trangular ans shape-regular polygonal meshes. *J. Comput. Phys.*, 227:492 – 512, 2007.

[23] K. Lipnikov, D. Svyatskiy, and Y. Vassilevski. Interpolation-free monotone finite volume method for diffusion equations on polygonal meshes. *J. Comput. Phys.*, 228(3):703–716, 2009.

[24] G. Manzini and A. Russo. A finite volume method for advection-diffusion problems in convection-dominated regimes. *Computer Methods in Applied Mechanics and Engineering*, 197(13-16):1242–61, 2008.

[25] J. M. Nordbotten, I. Aavatsmark, and G. T. Eigestad. Monotonicity of control volume methods. *Numer. Math.*, 106(2):255–288, 2007.

[26] C. Le Potier. Finite volume scheme satisfying maxcimum and minimum preinciples for anisotropic diffusion operators. In R. Eymard and J.-M. Herard, editors, *Finite Volumes for Complex Applications V*, pages 103–118, 2008.

[27] Vitoriano Ruas Santos. On the strong maximum principle for some piecewise linear finite element approximate problems of nonpositive type. *J. Fac. Sci. Univ. Tokyo Sect. IA Math.*, 29(2):473–491, 1982.

[28] P. Schneider and D. Eberly. *Geometric tools for computer graphics*. Morgan Kaufmann Publishers, 2003.

[29] H.A. van der Vorst. Bi-CGSTAB: a fast and smoothly converging variant of Bi-CG for the solution of non-symmetric linear systems. *SIAM J. Sci. Stat. Comput.*, 13(2):631 –644, MAR 1992.

[30] Richard S. Varga. *Matrix iterative analysis*. Prentice-Hall Inc., Englewood Cliffs, N.J., 1962.

[31] Yu. Vassilevski and I. Kapyrin. Two splitting schemes for nonstationary convection-diffusion problems on tetrahedral meshes. *Comput. Math. Math. Phys.*, 48(8):1349–1366, 2008.

[32] D. Wollstein, T. Linss, and R. Hans-Gorg. Uniformly accurate finite volume discretization for a convection-diffusion problem. *Electronic Transactions On Numerical Analysis*, 13:1–11, 2002.

[33] A. Yuan and Z. Sheng. Monotone finite volume schemes for diffusion equationson polygonal meshes. *J. Comput. Phys.*, 227:6288–6312, 6 2008.