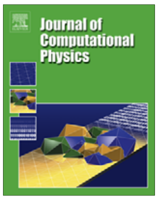
Journal of Computational Physics 229 (2010) 4017–4032

Contents lists available at ScienceDirect

Journal of Computational Physics

journal homepage: www.elsevier.com/locate/jcp



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

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

[a] Applied Mathematics and Plasma Physics Group, Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, United States
[b] Institute of Numerical Mathematics, Russian Academy of Sciences, 8, Gubkina, 119333 Moscow, Russia

ARTICLE INFO

Article history:
Received 12 May 2009
Received in revised form 22 January 2010
Accepted 26 January 2010
Available online 4 February 2010

Keywords:
Advection–diffusion equation
Finite volume method
Discrete maximum principle
Monotone method
Unstructured mesh
Polygonal mesh

ABSTRACT

We present a new second-order accurate monotone finite volume (FV) method for the steady-state advection–diffusion equation. The method uses a nonlinear approximation for both diffusive and advective fluxes and guarantees solution non-negativity. The interpolation-free approximation of the diffusive flux uses the nonlinear two-point stencil proposed in Lipnikov [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.

Published by Elsevier Inc.

1. Introduction

Accurate and reliable discretization methods inherit or mimic fundamental properties of continuous systems. 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. Therefore, our focus is on a simplified version of the DMP that provides only solution non-negativity as is referred to as the monotonicity condition. Some physical quantities, such as 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 shocks and exponential or parabolic boundary layers. The thickness of these features is usually small compared to the mesh size and hence they cannot be resolved properly. In diffusion-dominated problems and highly anisotropic media, some of the diffusive fluxes may be poorly approximated if mesh cells are not aligned with the principle directions of the diffusive tensor. In both regimes, unwanted spurious (non-physical) oscillations may appear in the numerical solution. The design of advanced discretization methods that eliminate or significantly reduce these oscillations remains the field of extensive research for the last five decades.

One of the most popular finite element (FE) methods was developed by Brooks and Hughes in [6] and is referred to as the streamline upwind Petrov Galerkin (SUPG) method. The stabilization procedure proposed in this method improves significantly robustness of the FE discretization; however, the spurious oscillations around sharp layers may still appear in the numerical solution. Indeed, the SUPG method is neither monotone nor a monotonicity preserving method. Several

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

0021-9991/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.jcp.2010.01.035

modifications and improvements of the SUPG method are reviewed in [14]. These modifications aim to design methods that satisfy the DMP, at least in some model cases, and were dubbed in [14] as the *spurious oscillations at layers diminishing* (SOLD) methods. Recently, another approach towards a robust FE method was developed in [17,18] and was dubbed as the *algebraic flux correction* method. The drawback of many FE methods is that they are formally not locally conservative on the original computational mesh, the property which is very desirable when a nonlinear advection–diffusion equation is coupled with other transport equations.

The finite volume (FV) methods guarantee the local mass conservation by construction. Recently, many new FV methods have been developed for the advection–diffusion equation (see [36,4,5,11,25,20] and references therein). It turns out that in the design of a monotone second-order accurate method, the approximation of diffusive fluxes is as challenging as that of the advective fluxes. The advective fluxes can be approximated via the upwinding approach [2] and controlled with different slope-limiting techniques [8,21,5] or introduction of an artificial viscosity [3,25]. For a long time, it was not clear how to approximate and control the diffusive fluxes in the case of general meshes and diffusion tensors. The theoretical analysis of DMP in the FE methods [9,16,35] imposes severe restrictions on the coefficients and computational mesh that are often violated in real-life simulations where the media is heterogeneous and anisotropic and the computational mesh may be strongly perturbed. In such a case, many advanced *linear* methods fail to satisfy the monotonicity condition [1,27,**?**]. 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. 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. It was noticed in [5,14] that *nonlinear* approximations is the key ingredient and the price to pay for construction of a monotone and second-order accurate discretization. In [7] a nonlinear method has been developed for the Poisson equation. For a general diffusion equation, a number of nonlinear methods have been developed [10,15,19,28,22,29,34,37]. The optimization procedures were developed in [24,26] for tensorial diffusion equation.

In this article, the approximation of diffusive fluxes is based on a nonlinear two-point flux approximation method [23]. The original idea was proposed by Le Potier in [28] for triangular meshes. It was further analyzed and extended to shape-regular polygonal meshes (but scalar diffusion coefficient) in [22] and to tetrahedral meshes in [15]. Yuan and Sheng [37] extended the method to a bigger class of polygonal meshes with star-shaped cells and full tensor coefficients. In [34], the nonlinear diffusive fluxes and the operator splitting method were used for solving the unsteady advection–diffusion equation. All the above methods, in addition to *primary* unknowns defined at mesh cells, require solution values at mesh vertices that must be interpolated from the primary unknowns. As shown in [22,37], the choice of the interpolation method affects the accuracy of the nonlinear FV method for problems with constant diffusion coefficients. The interpolation problem becomes even a more challenging task for problems with discontinuous coefficients [37]. A interpolation-free nonlinear FV method was developed in Lipnikov et al. [23]. The numerical experiments presented there demonstrate that the new method requires lesser number of nonlinear iterations compared to the methods using interpolation algorithms. The interpolation-free method was extended to polyhedral meshes in Danilov and Vassilevski [10]; however, interpolation of solution at mid-edge points may be still required in certain pathological cases. Finally, new results on the DMP were reported in the conference proceedings [29].

The approximation of advective fluxes follows ideas of the *monotonic upstream-centered scheme for conservation laws* (MUSCL) introduced in van Leer [32]. A piecewise linear discontinuous reconstruction of the FV solution on polygonal cells allows to build more accurate advective fluxes that are also *nonlinear*. In order to control monotonicity and robustness of the method, we use a new slope limiting technique. In each cell, we minimize deviation of the reconstructed linear function from given values at selected points subject to some monotonicity constraints. For each cell, majority of these points are centers of the closest neighboring cells, except a few special cases. Other limiting procedures more closely related to the proposed method are discussed in Hubbard [12]. The essential difference lies in the points where monotonicity constraints are imposed and in the norm used to measure the difference between the unlimited and limited linear functions. The methods in Hubbard [12] use the Cartesian distance between gradients of these functions. We use an analog of the discrete L^2-norm over the reconstruction area.

In this article, we prove non-negativity of the discrete solution and verify it with numerical experiments. The developed nonlinear FV method is exact for linear solutions; therefore, the second-order asymptotic convergence rate is expected for problems with smooth solutions. This rate is observed in our numerical experiments.

One of the goals of this article it to study impact of coupling of diffusive and advective fluxes on the iterative nonlinear solver which is the major computational overhead in the proposed method. To focus numerical analysis on this issue, we consider only continuous anisotropic diffusion tensors and refer for derivation of diffusive fluxes for discontinuous problems to [28,15,37,10,23]. We consider the Picard method and prove that each iterative approximation to the discrete solution is non-negative. This extends similar results for diffusion problems [28,23] to advection–diffusion problems. We found out that difference in various methods for approximation of advective fluxes, which may be subtle from viewpoint of numerical methods for hyperbolic problems, may become important for stability of the Picard method. For instance, our selection of the set of admissible gradients is driven by this stability issue.

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 $\Gamma = \Gamma_N \cup \Gamma_D$ where $\Gamma_D \cap \Gamma_N = \emptyset$ and $\Gamma_D \neq \emptyset$. We consider a model advection–diffusion problem for unknown concentration c:

$$\begin{aligned} &\mathrm{div}(\mathbf{v}c - \mathbb{K}\nabla c) = f \quad \text{in } \Omega \\ &c = g_D \quad \text{on } \Gamma_D \\ &-(\mathbb{K}\nabla c)\cdot \mathbf{n} = g_N \quad \text{on } \Gamma_N \end{aligned} \tag{1}$$

where $\mathbb{K}(\mathbf{x})$ is a symmetric positive definite continuous (possibly anisotropic) diffusion tensor, $\mathbf{v}(\mathbf{x}) \in C^1(\bar{\Omega})$ is a velocity field, $\mathrm{div}\mathbf{v} \geqslant 0, f$ is a source term, $\mathbf{n}$ is the exterior normal vector, and g_D, g_N are given data. We denote by Γ_{out} the outflow part of Γ where $\mathbf{v}\cdot\mathbf{n} \geqslant 0$, and define $\Gamma_{in} = \Gamma \setminus \Gamma_{out}$. We assume that $\Gamma_N \subset \Gamma_{out}$.

The sufficient conditions for non-negativity of the solution $c(x)$ are $f(x) \geqslant 0, g_D \geqslant 0$ and $g_N \leqslant 0$. We assume that these conditions are hold. From a physical viewpoint the requirements $f(x) \geqslant 0$ and $g_N \leqslant 0$ mean that no mass can be taken out of the system.

The Dirichlet boundary condition 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 minimal amount of a numerical diffusion to avoid excessive smearing of boundary layers but sufficient to damp non-physical oscillations.

3. Monotone nonlinear FV scheme on polygonal meshes

In this section, we derive a FV scheme with a nonlinear two-point flux approximation. Let $\mathbf{q} = -\mathbb{K}\nabla c + c\mathbf{v}$ denote the total flux which satisfies the mass balance equation:

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

Let $\mathcal{T}$ be a conformal polygonal mesh composed of $N_{\mathcal{T}}$ shape-regular cells T. We assume that $\mathcal{T}$ is edge-connected, i.e. it cannot be split into two meshes having no common edges. We denote by $\mathbf{n}_T$ the exterior unit normal vector to ∂T and by $\mathbf{n}_e$ the normal vector to edge e fixed once and for all. On a boundary edge, the vector $\mathbf{n}_e$ is exterior. We assume that $|\mathbf{n}_e| = |e|$ where $|e|$ denotes the length of edge e. Let the set $\bar{\Gamma}_N \cap \bar{\Gamma}_D$ belong to the set of nodes of $\mathcal{T}$.

Let $N_{\mathcal{B}}$ be the number of boundary 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 $\mathcal{E}_B^N$ where the Dirichlet and Neumann 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 Eq. (2) over a polygon T and using Green's formula we get:

$$\sum_{e\in\partial T} \chi_{T,e}\, \mathbf{q}_e \cdot \mathbf{n}_e = \int_T f \,\mathrm{d}x, \quad \mathbf{q}_e = \frac{1}{|e|}\int_e \mathbf{q}\,\mathrm{d}s \tag{3}$$

where $\mathbf{q}_e$ is the average flux density for edge e, and $\chi_{T,e}$ is either 1 or -1 depending on mutual orientation of normal vectors $\mathbf{n}_e$ and $\mathbf{n}_T$.

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:

$$\mathbf{q}_e^h \cdot \mathbf{n}_e = M_e^+ C_{T_+} - M_e^- C_{T_-}, \tag{4}$$

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 (4) with an explicit value for one of the concentrations. For the Dirichlet boundary value problem, $\Gamma_D = \partial\Omega$, substituting (4) into (3), we obtain a system of $N_{\mathcal{T}}$ equations with $N_{\mathcal{T}}$ unknowns C_T. Dirichlet and Neumann boundary conditions are considered in Section 3.4.

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 number of elements in the set 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 [37] (see Fig. 1 for graphical interpretation).

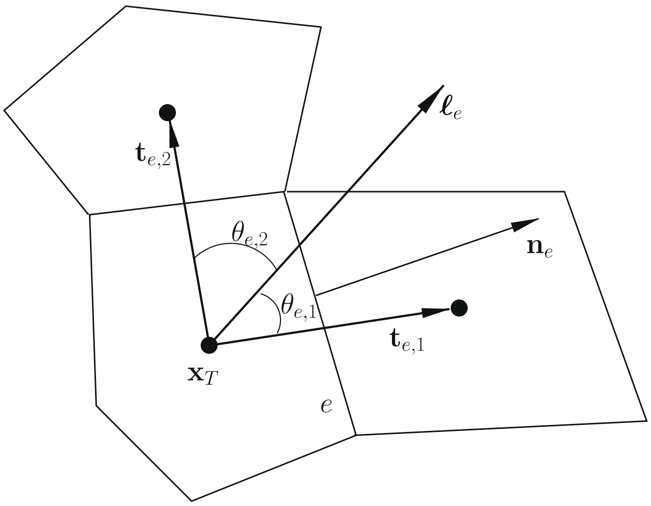


Fig. 1. The vector ℓ_e forms acute angles with vectors $\mathbf{t}_{e,1}$ and $\mathbf{t}_{e,2}$. The collocation points are marked by solid circles.

(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 $\ell_e = \mathbb{K}(\mathbf{x}_e)\mathbf{n}_e$ (see Fig. 1), then

$$\theta_{e,1} < \pi, \quad \theta_{e,2} < \pi \quad \text{and} \quad \theta_{e,1} + \theta_{e,2} < \pi. \tag{5}$$

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

$$\mathbf{t}_{e,1} \times \ell_e \leqslant 0 \quad \text{and} \quad \mathbf{t}_{e,2} \times \ell_e > 0. \tag{6}$$

If conditions (5) and (6) cannot be satisfied, we may extend the set Σ_T by adding neighbors of already included collocation points. The trigonometric observations give the following lemma (see [37] for details).

Lemma 3.1. *Under assumptions (5) and (6), there exist non-negative α_e and β_e such that*

$$\frac{1}{|\ell_e|}\ell_e = \frac{\alpha_e}{|\mathbf{t}_{e,1}|}\mathbf{t}_{e,1} + \frac{\beta_e}{|\mathbf{t}_{e,2}|}\mathbf{t}_{e,2}, \tag{7}$$

where

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

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

This section follows closely Section 3.2 in Lipnikov et al. [23]. For completeness of the presentation, we summarize the key formulas. Let us 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 $T_\pm$.

Note that $\nabla c \cdot (\mathbb{K}_e \, \mathbf{n}_e)$ is the derivative in direction ℓ_e multiplied by $|\ell_e|$. Using Lemma 3.1, we rewrite the normal component of the diffusion flux as follows:

$$\mathbf{q}_{e,d} \cdot \mathbf{n}_e = -(1 + O(|e|))\frac{|\ell_e|}{|e|}\int_e \frac{\partial c}{\partial \ell_e} \, \mathrm{d}s = -(1 + O(|e|))\frac{|\ell_e|}{|e|}\int_e \left(\alpha_e \frac{\partial c}{\partial \mathbf{t}_{e,1}} + \beta_e \frac{\partial c}{\partial \mathbf{t}_{e,2}}\right) \mathrm{d}s. \tag{8}$$

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

$$\int_e \frac{\partial c}{\partial \mathbf{t}_{e,i}} \, \mathrm{d}s = |e|\left(\frac{C_{e,i} - C_T}{|\mathbf{x}_{e,i} - \mathbf{x}_T|} + O(|\mathbf{x}_{e,i} - \mathbf{x}_T|)\right), \quad i = 1, 2. \tag{9}$$

Note that this formula is exact for linear concentrations. If $\mathbf{x}_{e,i}$ is the secondary collocation point, we use formula (29) for $C_{e,i}$. Substituting (9) in (8) we get:

$$\mathbf{q}^h_{e,d} \cdot \mathbf{n}_e = -|\ell_e| \left(\frac{\alpha_e}{|\mathbf{t}_{e,1}|}(C_{e,1} - C_T) + \frac{\beta_e}{|\mathbf{t}_{e,2}|}(C_{e,2} - C_T) \right). \tag{10}$$

At the moment, this flux involves three 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:

$$\mathbf{q}^h_{\pm,d} \cdot \mathbf{n}_e = \mp|\ell_e| \left(\frac{\alpha_\pm}{|\mathbf{t}_{\pm,1}|}\ (C_{\pm,1} - C_\pm) + \frac{\beta_\pm}{|\mathbf{t}_{\pm,2}|}\ (C_{\pm,2} - C_\pm) \right), \tag{11}$$

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 (11) with non-negative weights $\mu_\pm$:

$$\begin{aligned}\mathbf{q}^h_{e,d} \cdot \mathbf{n}_e &= \mu_+ \mathbf{q}^h_{+,d} \cdot \mathbf{n}_e + \mu_- \mathbf{q}^h_{-,d} \cdot \mathbf{n}_e \\ &= \mu_+|\ell_e| \left(\frac{\alpha_+}{|\mathbf{t}_{+,1}|} + \frac{\beta_+}{|\mathbf{t}_{+,2}|} \right) C_+ - \mu_-|\ell_e| \left(\frac{\alpha_-}{|\mathbf{t}_{-,1}|} + \frac{\beta_-}{|\mathbf{t}_{-,2}|} \right) C_- - \mu_+|\ell_e| \left(\frac{\alpha_+}{|\mathbf{t}_{+,1}|} C_{+,1} + \frac{\beta_+}{|\mathbf{t}_{+,2}|} C_{+,2} \right) \\ &\quad + \mu_-|\ell_e| \left(\frac{\alpha_-}{|\mathbf{t}_{-,1}|} C_{-,1} + \frac{\beta_-}{|\mathbf{t}_{-,2}|} C_{-,2} \right). \end{aligned} \tag{12}$$

The first requirement for the weights is to cancel the two last terms in (12). The second requirement is to approximate the true flux:

$$-\mu_+ d_+ + \mu_- d_- = 0 \quad \text{and} \quad \mu_+ + \mu_- = 1 \tag{13}$$

where $d_\pm = |\ell_e|(\alpha_\pm C_{\pm,1}/|\mathbf{t}_{\pm,1}| + \beta_\pm C_{\pm,2}/|\mathbf{t}_{\pm,2}|)$. Note that $d_\pm \geqslant 0$ for non-negative concentrations. If $d_\pm = 0$, we select the symmetric solution $\mu_+ = \mu_- = \frac{1}{2}$. Otherwise,

$$\mu_+ = \frac{d_-}{d_- + d_+} \quad \text{and} \quad \mu_- = \frac{d_+}{d_- + d_+}. \tag{14}$$

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

Substituting (14) into (12), we get

$$\mathbf{q}^h_{e,d} \cdot \mathbf{n}_e = D_e^+ C_{T_+} - D_e^- C_{T_-} \tag{15}$$

with coefficients

$$D_e^\pm = \mu_\pm |\ell_e| \left(\frac{\alpha_\pm}{|\mathbf{t}_{\pm,1}} + \frac{\beta_\pm}{|\mathbf{t}_{\pm,2}|} \right). \tag{16}$$

Remark 3.1. Although formula (10) is invariant with respect to the addition of a constant function, the discrete flux (15) 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 (12) 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

$$\mathbf{q}^h_{e,a} \cdot \mathbf{n}_e = v_e^+ \mathcal{R}_{T^+}(\mathbf{x}_e) + v_e^- \mathcal{R}_{T^-}(\mathbf{x}_e), \tag{17}$$

where

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

and $\mathcal{R}_T$ is a linear reconstruction of the concentration over cell T. We define the linear reconstruction as follows:

$$\mathcal{R}_T(\mathbf{x}) = C_T + \mathbf{g}_T \cdot (\mathbf{x} - \mathbf{x}_T), \quad \forall \mathbf{x} \in T, \tag{18}$$

where $\mathbf{g}_T$ is the gradient of the linear function. Since C_T is collocated at the barycenter of T, this reconstruction preserves the mean value of the concentration for any choice of $\mathbf{g}_T$.

It is conventional to reconstruct the gradient using approximation arguments and to constrain it using stability arguments. Let $\mathcal{G}_T$ be the set of admissible gradients $\tilde{\mathbf{g}}_T$ that satisfy the constraints (20)–(22) formulated below. We consider the following constrained minimization problem:

$$\mathbf{g}_T = \arg \min_{\tilde{\mathbf{g}}_T \in \mathcal{G}_T} \mathcal{J}_T(\tilde{\mathbf{g}}_T), \tag{19}$$

where the functional

$$\mathcal{J}_T(\tilde{\mathbf{g}}_T) = \frac{1}{2} \sum\nolimits_{\mathbf{x}_k \in \tilde{\Sigma}_T} [C_T + \tilde{\mathbf{g}}_T \cdot (\mathbf{x}_k - \mathbf{x}_T) - C_k]^2$$

measures deviation of the reconstructed function from the targeted values C_k collocated at points $\mathbf{x}_k$ from a set $\tilde{\Sigma}_T$ which is built as follows. First, the auxiliary set $\hat{\Sigma}_T$ is defined by eliminating the secondary collocation points $\mathbf{x}_e, e \in \mathcal{E}_B^{out}$, from the set Σ_T. Second, the set $\hat{\Sigma}_T$ is extended whenever it is either too small or ill-conditioned. More precisely, if $\hat{\Sigma}_T = \{\mathbf{x}_T, \mathbf{x}_{T'}\}$, we add to it the elements of $\hat{\Sigma}_{T'}$ other than $\mathbf{x}_T$. If $\hat{\Sigma}_T = \{\mathbf{x}_T, \mathbf{x}_{T'}, \mathbf{x}_{T''}\}$ and area of the triangle formed by these three points is less than $10^{-3}|T|$, we add to it the elements of $\hat{\Sigma}_{T'}$ and $\hat{\Sigma}''_T$ other than $\mathbf{x}_T$. The resulting set forms the set $\tilde{\Sigma}_T$.

The following three sets of constraints are introduced to avoid non-physical extrema. The admissible gradient $\tilde{\mathbf{g}}_T$ must result in a linear reconstruction that is bounded at the collocation points $\mathbf{x}_k \in \hat{\Sigma}_T$:

$$\min\left\{C_1, C_2, \ldots, C_{N(\hat{\Sigma}_T)}\right\} \leqslant C_T + \tilde{\mathbf{g}}_T \cdot (\mathbf{x}_k - \mathbf{x}_T) \leqslant \max\left\{C_1, C_2, \ldots, C_{N(\hat{\Sigma}_T)}\right\}. \tag{20}$$

Additionally, the reconstructed function must satisfy the following restrictions at points $\mathbf{x}_e$ on edges $e \in \mathcal{E}_T$ where $v_e > 0$:

$$C_T + \tilde{\mathbf{g}}_T \cdot (\mathbf{x}_e - \mathbf{x}_T) \geqslant 0. \tag{21}$$

This condition guarantees correct sign of the advective flux. Frequently, this condition follows from (20). However, if the edge mid-point $\mathbf{x}_e$ lies outside the convex hull of points $\mathbf{x}_k \in \hat{\Sigma}_T$ the reconstructed function may be negative at this point.

Finally, the reconstructed function must be bounded from below at the secondary collocation points on the outflow boundary:

$$\min\left\{C_1, C_2, \ldots, C_{N(\hat{\Sigma}_T)}\right\} \leqslant C_T + \tilde{\mathbf{g}}_T \cdot (\mathbf{x}_e - \mathbf{x}_T), \quad e \in \mathcal{E}_T \cap \mathcal{E}_B^{out}. \tag{22}$$

The set $\hat{\Sigma}_T$ and the above constraints were designed to be practical and at the same time as weak as possible. For instance, our attempts to use only edge mid-points $\mathbf{x}_e$ in (20) or drop out (22) resulted in numerical instabilities in the nonlinear iterative method which is introduced in Section 4. Due to (20), we get that $\tilde{\mathbf{g}}_T \equiv 0$ in local minima and maxima.

Fig. 2 illustrates the elliptic nature of the deviation functional $\mathcal{J}_T$. In this particular case, solution of the unconstrained problem is located outside the admissible set P. The solution of the constrained problem is the closest point to the boundary of P when the level sets of $\mathcal{J}_T$ are circles. In a general case, the solution may deviate significantly from the closest point.

Lemma 3.2. *Minimization problem (19) with constraints (20)–(22) has a unique solution.*

Proof. A solution to problem (19) does exist, since the constant reconstruction ($\mathbf{g}_T = (0,0)^T$) satisfies inequalities (20)–(22). Each of these inequalities represents a half-plane. The admissible set $\mathcal{G}_T$ is the intersection of half planes; therefore, it is a convex polygon. The problem (19) reduces to a problem of convex analysis [30]: given a point ξ_0 on a plane and a convex polygon P, find

$$\xi = \arg\min_{\xi' \in \mathrm{P}} (\xi' - \xi_0)^T \mathbf{R} (\xi' - \xi_0), \tag{23}$$

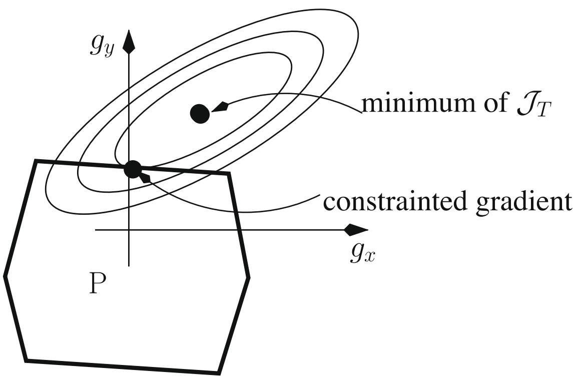


Fig. 2. Illustration of the set of admissible gradients (hexagon P), level sets of the deviation functional $\mathcal{J}_T$ (ellipses), and location of the unconstrained and constrained gradients (solid circles).

where $\mathbf{R}$ is a 2×2 symmetric positive definite matrix. The point ξ_0 and the matrix $\mathbf{R}$ can be easily derived from parameters in the deviation functional $\mathcal{J}_T$.

If $\xi_0 \in \mathrm{P}$, it is the solution of the constrained problem. Otherwise, the solution ξ is a unique point on $\partial\mathrm{P}$. This point can be found by searching for minima of quadratic functions on edges of P.

Using (17) and (18), we represent the advective flux as the sum of a linear part (the first-order approximation) and a nonlinear part (the second-order correction):

$$\mathbf{q}^h_{e,a} \cdot \mathbf{n}_e = A^+_e C_+ - A^-_e C_-, \tag{24}$$

where

$$A^\pm_e = \pm v^\pm_e (1 + \mathbf{g}_\pm \cdot (\mathbf{x}_e - \mathbf{x}_\pm) C^{-1}_\pm) \tag{25}$$

and subscript ± stands for $T_\pm$.

We note that the coefficients $A^\pm_e$ are non-negative for positive concentrations. If $C_T = 0$ in a cell T then $\mathbf{g}_T$ must be the zero and $A^\pm_e = \pm v^\pm_e$.

3.4. Fluxes on boundary edges

Let us consider a Neumann boundary edge $e \in \mathcal{E}^N_B$. The diffusive flux through this edge is

$$\mathbf{q}^h_{e,d} \cdot \mathbf{n}_e = \bar{g}_{N,e} |e|, \tag{26}$$

where $\bar{g}_{N,e}$ is the mean value of g_N on edge e. 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 from formula (24) the approximation of the advective flux:

$$\mathbf{q}^h_{e,a} \cdot \mathbf{n}_e = A^+_e C_T. \tag{27}$$

Thus, the equation for the total flux is

$$(\mathbf{q}^h_{e,d} + \mathbf{q}^h_{e,a}) \cdot \mathbf{n}_e = \bar{g}_{N,e} |\mathbf{n}_e| + A^+_e C_T, \quad e \in \mathcal{E}^N_B, \tag{28}$$

where coefficient A^+_e is non-negative for non-negative concentrations.

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

$$C_e = \bar{g}_{D,e} = \frac{1}{|e|} \int_e g_D \,\mathrm{d}s. \tag{29}$$

The approximation of the diffusive flux is given by formula

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

where coefficients $D^\pm_e$ are given by (16). The approximation of the advective flux depends on velocity direction on edge e. If $e \in \mathcal{E}^{out}_B$, the approximation adopts formulas (27) and (25). If $e \in \mathcal{E}^{in}_B$, we use

$$\mathbf{q}^h_{e,a} \cdot \mathbf{n}_e = -A^-_e, \tag{31}$$

where

$$A^-_e = -\bar{g}_{D,e} v_e \equiv -\bar{g}_{D,e} v^-_e \geqslant 0. \tag{32}$$

4. Discrete system and monotonicity analysis

Substituting two-point flux formula (4) with non-negative coefficients $M^\pm_e = D^\pm_e + A^\pm_e$ given by (16) and (25) into into the mass balance Eq. (3), and using Eq. (29) to eliminate boundary concentrations, we get a nonlinear system of $N_\mathcal{T}$ equations with $N_\mathcal{T}$ unknowns:

$$\mathbf{M}(\mathbf{C})\mathbf{C} = \mathbf{F}(\mathbf{C}), \tag{33}$$

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

$$\mathbf{M}_e(\mathbf{C}) = \begin{pmatrix} M^+_e(\mathbf{C}) & -M^-_e(\mathbf{C}) \\ -M^+_e(\mathbf{C}) & M^-_e(\mathbf{C}) \end{pmatrix} \tag{34}$$

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:

$$F_T(\mathbf{C}) = \int_T f \, \mathrm{d}x + \sum_{e \in \mathcal{E}_B^D \cap \partial T} M_e^-(\mathbf{C}) \bar{g}_{D,e} - \sum_{e \in \mathcal{E}_B^N \cap \partial T} |e| \bar{g}_{N,e}, \quad \forall T \in \mathcal{T}. \tag{35}$$

For $f(x) \geqslant 0, g_D \geqslant 0$ and $g_N \leqslant 0$ the components of vector F are non-negative. We use the Picard iterations to solve the nonlinear system (33) (see Algorithm 1).

Algorithm 1. Generation and solution of nonlinear system (33).

```
1: 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 (5) and (6). Find similar vectors for boundary edges.
2: Select an initial vector $\mathbf{C}^0$ with non-negative entries and a small value $\varepsilon_{non} > 0$.
3: for $k = 0, \ldots,$ do
4:     Calculate concentrations $C_e$ on edges $e \in \mathcal{E}_B^D$ using (29).
5:     Assemble the global matrix $\mathbf{M}(\mathbf{C}^k)$ from the edge-based matrices $\mathbf{M}_e(\mathbf{C}^k)$. Use formulas (16) and (25) to form $\mathbf{M}_e(\mathbf{C}^k)$.
6:     Calculate the right-hand side vector $\mathbf{F}(\mathbf{C}^k)$ using (35).
7:     Stop if $\|\mathbf{M}(\mathbf{C}^k)\mathbf{C}^k - \mathbf{F}(\mathbf{C}^k)\| \leqslant \varepsilon_{non} \, \|\mathbf{M}(C^0)\mathbf{C}^0 - \mathbf{F}(\mathbf{C}^0)\|$.
8:     Solve $\mathbf{M}(\mathbf{C}^k)\mathbf{C}^{k+1} = \mathbf{F}(\mathbf{C}^k)$.
9: end for
```

The linear system in Step 8 with the non-symmetric matrix $\mathbf{M}(\mathbf{C}^k)$ is solved by the ILU-preconditioned Bi-Conjugate Gradient Stabilized (BiCGStab) method [31]. 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 (33) is non-negative provided that it exists.

Theorem 4.1. *Let* $\Gamma_N = \emptyset(\mathcal{E}_B^D \equiv \mathcal{E}_B), f \geqslant 0$ *in* $\Omega, g_D \geqslant 0$ *on* $\Gamma_D \equiv \partial\Omega$ *and the solution* $\mathbf{C}$ *to* (33) *exist. Then* $\mathbf{C} \geqslant 0$.

Proof. The proof is by contradiction. Let us consider the cell T with the smallest concentration C_T and assume that $C_T < 0$. Without lose of generality, we assume that vectors $\mathbf{n}_e$ are exterior with respect to T. Let $T = T_+$ in the flux formulas. Since C_T is minimal, $\mathcal{R}_T \equiv C_T$. The definition of advective fluxes gives

$$\sum_{e \in \mathcal{E}_T} \mathbf{q}_{e,a}^h \cdot \mathbf{n}_e = \sum_{e \in \mathcal{E}_T \backslash \mathcal{E}_B^{in}} (v_e^+ C_T + v_e^- \; \mathcal{R}_{T_e^-}(\mathbf{x}_e)) + \sum_{e \in \mathcal{E}_T \cap \mathcal{E}_B^{in}} v_e^- \bar{g}_{D,e}.$$

Recall that $v_e^- = v_e$ on inflow edges and $v_e = v_e^+ + v_e^-$. By adding and subtracting $v_e^- C_T$, we get

$$\sum_{e \in \mathcal{E}_T} \mathbf{q}_{e,a}^h \cdot \mathbf{n}_e = C_T \sum_{e \in \mathcal{E}_T} v_e + \sum_{e \in \mathcal{E}_T \backslash \mathcal{E}_B^{in}} v_e^- (\mathcal{R}_{T_e^-}(\mathbf{x}_e) - C_T) + \sum_{e \in \mathcal{E}_T \cap \mathcal{E}_B^{in}} v_e^- (\bar{g}_{D,e} - C_T).$$

From the mass balance Eq. (3) we derive

$$-C_T \sum_{e \in \mathcal{E}_T} v_e + \int_T f \, \mathrm{d}x - \sum_{e \in \mathcal{E}_T} \mathbf{q}_{e,d}^h \cdot \mathbf{n}_e - \sum_{e \in \mathcal{E}_T \backslash \mathcal{E}_B^{in}} v_e^- (\mathcal{R}_{T_e^-}(\mathbf{x}_e) - C_T) - \sum_{e \in \mathcal{E}_T \cap \mathcal{E}_B^{in}} v_e^- (\bar{g}_{D,e} - C_T) = 0. \tag{36}$$

We have

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

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 3.1, 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

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

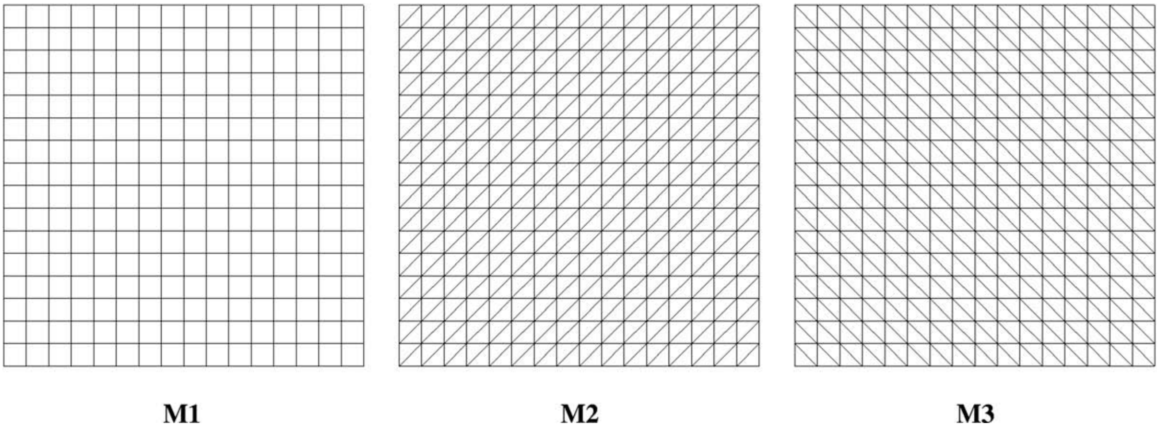


Fig. 3. Examples of three types of uniform meshes.

By virtue of $\upsilon_e^- \leqslant 0$ we conclude that all the terms in (36) are non-negative and must be equal to zero. This implies

$$0=\sum_{e\in\mathcal{E}_T}\mathbf{q}^h_{e,d}(\widetilde{\mathbf{C}})\cdot\mathbf{n}_e=-\sum_{e\in\mathcal{E}_T}D_e^-\ \widetilde{C}_{T_e^-}$$

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. □

The following result illustrates properties of the Picard iterations.

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

Proof. The proof follows closely the proof of Theorem 4.1 in [23]; therefore, it is only sketched below. First, we observe that the matrix $\mathbf{M}^T(\mathbf{C}^k)$ fulfills all conditions of Corollary 1 on p. 85 of [33] when $\mathbf{C}^k\geqslant 0$. Thus, $\mathbf{M}^T(\mathbf{C}^k)$ is the M-matrix and all entries of $(\mathbf{M}^T(\mathbf{C}^k))^{-1}$ are positive. Since the transpose and inverse are commuting operations, we get that $\mathbf{C}^{k+1}\geqslant 0$. This proves the assertion of the theorem. □

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

$$\mathbf{q}^h_{e,a}\cdot\mathbf{n}_e=A_e^+C_+-A_e^-C_-,\quad A_e^\pm=\pm\upsilon_e^\pm.$$

5. Numerical experiments

5.1. Implementation issues

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

5.1.1. Errors

We use the following discrete L^2-norms to evaluate relative discretization errors for the concentration c and the flux $\mathbf{q}$:

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

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 $\varepsilon_{non}=10^{-7}$. 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 {**M1**} and two types of triangular meshes produced by splitting each square cell into two triangles by the north-east {**M2**} or north-west diagonal {**M3**}, as shown in Fig. 3.

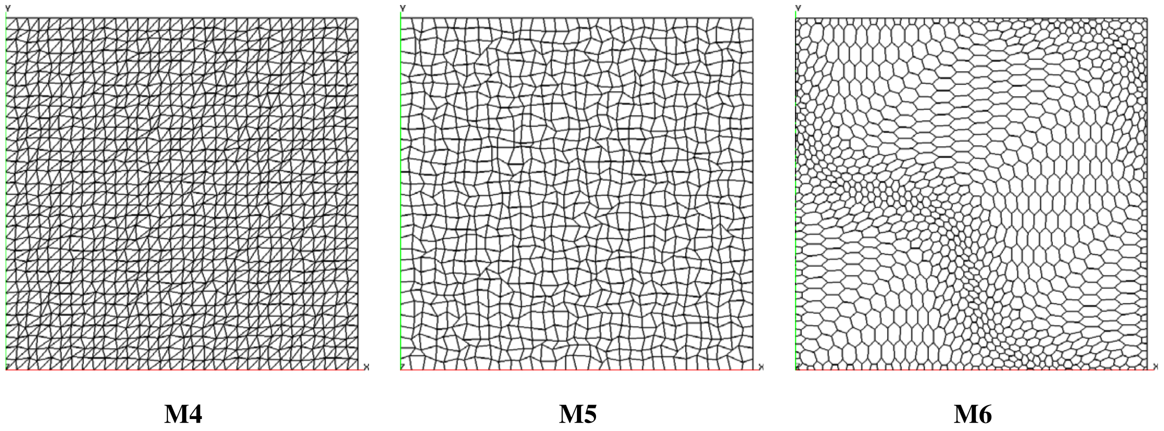


Fig. 4. Examples of two types of distorted structured meshes and a polygonal mesh.

Table 1
Convergence analysis for diffusion-dominated problems.

h	{**M4**}		{**M5**}		{**M6**}	
	ε_2^C	ε_2^q	ε_2^C	ε_2^q	ε_2^C	ε_2^q
1/32	7.13e−04	3.32e−03	3.43e−04	2.67e−03	8.76e−04	4.77e−03
1/64	2.61e−04	1.31e−03	9.49e−05	1.17e−03	2.72e−04	1.73e−03
1/128	9.80e−05	5.18e−04	2.67e−05	4.67e−04	7.35e−05	6.03e−04
1/256	3.99e−05	2.13e−04	7.24e−06	2.10e−04	1.96e−05	2.20e−04
Rate	1.39	1.32	1.85	1.23	1.83	1.48

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

$$x := x + \alpha\xi_x h, \quad y := y + \alpha\xi_y h, \tag{37}$$

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 $\alpha = 0.4$ for triangular meshes and $\alpha = 0.6$ for quadrilateral meshes. It is pertinent to emphasize that the distortion is performed on each refinement level. A polygonal mesh from sequence {**M6**} is a dual mesh for a smoothly transformed uniform triangular mesh. Examples of these meshes are shown in Fig. 4. For each space resolution, the quadrilateral and polygonal meshes have roughly the same number of cells. The corresponding triangular meshes have twice more cells.

5.2. Anisotropic diffusion with advection

5.2.1. Convergence study

The convergence study is performed for a smooth solution on mesh sequences {**M4**}, {**M5**} and {**M6**}. A sequence of distorted meshes is the most challenging test for a numerical scheme due to fixed 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:

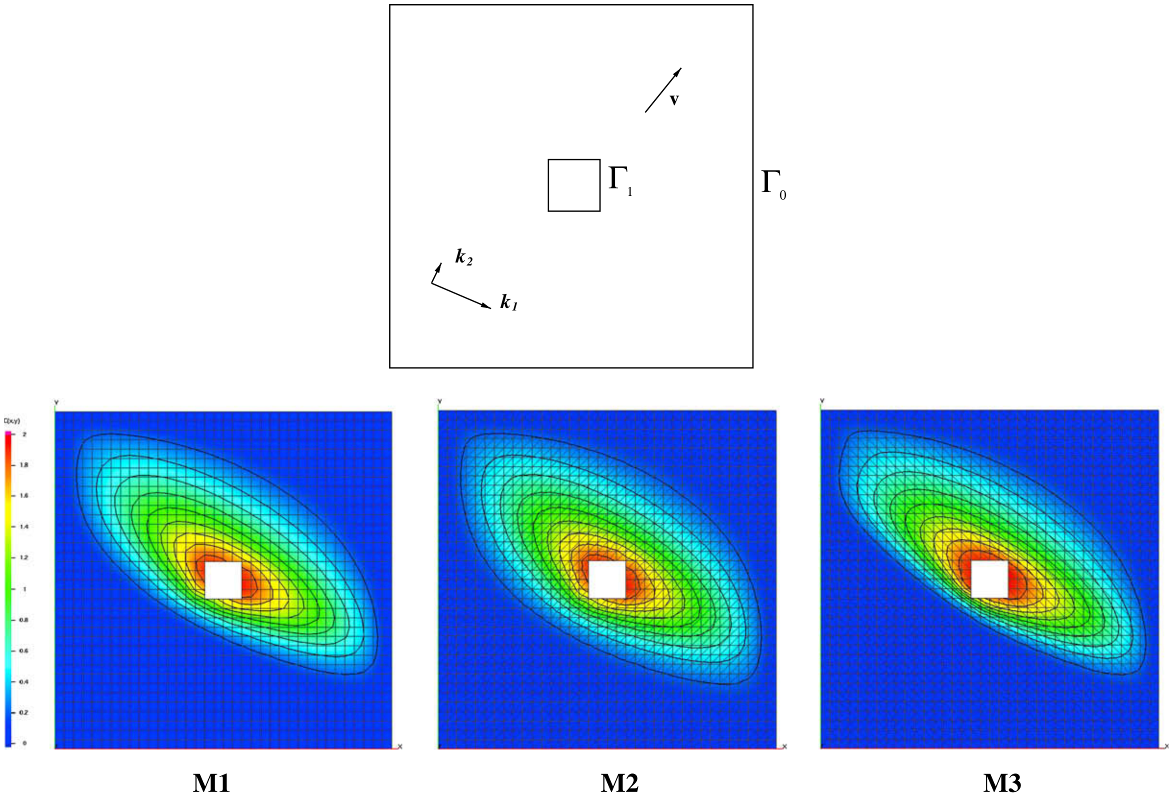
$$c(x, y) = x\cos(0.5\pi y), \quad \mathbf{v} = (1, -1)^T, \quad \mathbb{K} = \begin{pmatrix} 10 & 0 \\ 0 & 0.1 \end{pmatrix}.$$

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 linear regression algorithm has been used for calculating the convergence rates. The convergence rate for the concentration is close to the second-order for quadrilateral and polygonal meshes, while the convergence rate for the flux is higher than the first-order on all meshes. This is one of the advantages of usage of polygonal meshes in simulations.

Since our method is exact for linear functions, we may expect the asymptotic second-order convergence rate on all sequences of meshes. The triangular meshes does not show this asymptotics. This loss of convergence rate can be explained by a smaller number of neighboring cells compared to the other meshes. To verify this assumption, we modify the definition of set Σ_T for triangular meshes. More precisely, for every triangle T, we consider neighboring cells T' which share at least a vertex with triangle T. We include new collocation points $\mathbf{x}_{T'}$ into the original set Σ_T. The new set is denoted by Σ_T^{ext}. Results

Table 2
Convergence analysis for diffusion-dominated problems on triangular meshes {**M4**}.

h	{**M4**}		{**M4**}	
	Original Σ_T		Extended Σ_T^{ext}	
	ε_2^C	ε_2^q	ε_2^C	ε_2^q
1/32	7.13e−04	3.32e−03	4.61e−04	1.42e−03
1/64	2.61e−04	1.31e−03	1.10e−04	5.25e−04
1/128	9.80e−05	5.18e−04	2.83e−05	1.95e−04
1/256	3.99e−05	2.13e−04	6.87e−06	8.27e−05
Rate	1.39	1.32	2.01	1.37



M1 **M2** **M3**

Fig. 5. 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.

of calculations with Σ_T^{ext} are presented in the last two columns of Table 2. We observe essential improvement in the convergence rate for the concentration and the error reduction for both unknowns. In the remaining numerical experiments on triangular meshes, we continue to use Σ_T, since no deterioration of the convergence rate is observed.

5.2.2. Monotonicity test

The monotonicity study is performed of mesh sequences {**M1**}, {**M2**} and {**M3**} for a problem with highly anisotropic diffusion tensor. Such a problem is a challenging task for many linear discretization methods (see numerical experiments in [19,22]) that may result in significant violation of the DMP and even produce a numerical solution with non-physical oscillations. We consider problem (1) in the unit square with a square hole, $\Omega = (0,1)^2/[4/9,5/9]^2$, so that the boundary of Ω consists of two disjoint parts as shown in Fig. 5. We set $f = 0, g_D = 0$ on $\Gamma_0, g_D = 2$ on $\Gamma_1, \mathbf{v} = (700,700)^T$ and take the following anisotropic diffusion tensor $\mathbb{K}$:

$$\mathbb{K} = R(-\theta)\begin{pmatrix} k_1 & 0 \\ 0 & k_2 \end{pmatrix}R(\theta), \quad R(\theta) = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}, \tag{38}$$

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

Table 3
Convergence analysis for the advection-dominated problem and the smooth solution.

h	{**M4**}		{**M5**}		{**M6**}	
	ε_2^C	ε_2^q	ε_2^C	ε_2^q	ε_2^C	ε_2^q
1/32	1.72e−03	1.66e−03	6.91e−04	5.81e−04	9.37e−04	8.13e−04
1/64	4.77e−04	4.65e−04	1.87e−04	1.66e−04	3.29e−04	3.12e−04
1/128	1.31e−04	1.28e−04	4.67e−05	4.13e−05	9.49e−05	9.18e−05
1/256	3.34e−05	3.24e−05	1.22e−05	1.09e−05	2.61e−05	2.56e−05
Rate	1.89	1.89	1.94	1.92	1.73	1.67

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. 5). The solution profile on mesh **M2** is wider than that on mesh **M3** where the mesh size along the velocity direction is effectively twice smaller. The square mesh gives an intermediate result. It is pertinent to note that our method guarantees only non-negativity of the numerical solution. Thus, some overshoots may occur and were observed in Lipnikov et al. [23] for diffusion problems. Overshoots are more localized compared to undershoots occurring in linear methods. For instance, the solution of (38) with $\mathbf{v} = (0,0)^T$ calculated with the lowest-order Raviart-Thomas MFE method is negative on both triangular meshes over almost half of the computational domain.

5.3. Advection dominated problems

5.3.1. Convergence study for smooth solutions

Here we study the accuracy of our FV method for advection-dominated problems with smooth solutions. 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:

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

The forcing term f and the Dirichlet boundary data g_D are set accordingly to the exact solution. Table 3 shows the relative L^2-norms of the errors. For all types of meshes we observe the second-order convergence rate for the concentration and the rate higher than the first-order for the flux. Again, triangular meshes produce higher errors compared to errors on quadrilateral and polygonal meshes. Note that significant error reduction ($2^{1.86}$) on the last polygonal mesh is overpowered by slower convergence rate on coarser meshes. Thus, the linear regression algorithms gives only 1.73.

5.3.2. Convergence study for solutions with boundary layers

In this section, we study the accuracy of our FV method for problems with exponential boundary layers. Let $\Omega = (0,1)^2$. We consider the problem which also was studied in Manzini and Russo [25]. The exact solution, the constant velocity field and the isotropic diffusion tensor are defined by

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

where ν characterizes thickness of the boundary layer in the top-right corner of Ω. 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 4 demonstrate the second-order convergence rate for the concentration and the superconvergence for the flux on all types of considered meshes. Moreover, in all tests the numerical solutions vary between 0 and 1.

5.3.3. Monotonicity test

In this section 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 problems, see [13] and [14]. Following [14], we set

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

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 two lines $y = 0$ and $x = 1$. It also has an internal layer along the streamline passing through the point (0, 0.7).

Table 4
Convergence analysis for the advection-dominated problem and the solution with the boundary layer.

h	{**M4**}		{**M5**}		{**M6**}	
	ε_2^C	ε_2^q	ε_2^C	ε_2^q	ε_2^C	ε_2^q
1/32	9.71e−04	4.81e−04	1.83e−03	1.60e−03	4.90e−03	3.52e−03
1/64	2.38e−04	1.12e−04	4.92e−04	4.29e−04	1.26e−03	8.74e−04
1/128	5.94e−05	2.93e−05	1.23e−04	1.08e−04	3.15e−04	2.16e−04
1/256	1.48e−05	7.58e−06	3.07e−05	2.74e−05	7.86e−05	5.34e−05
Rate	2.01	1.99	1.97	1.96	1.99	2.01

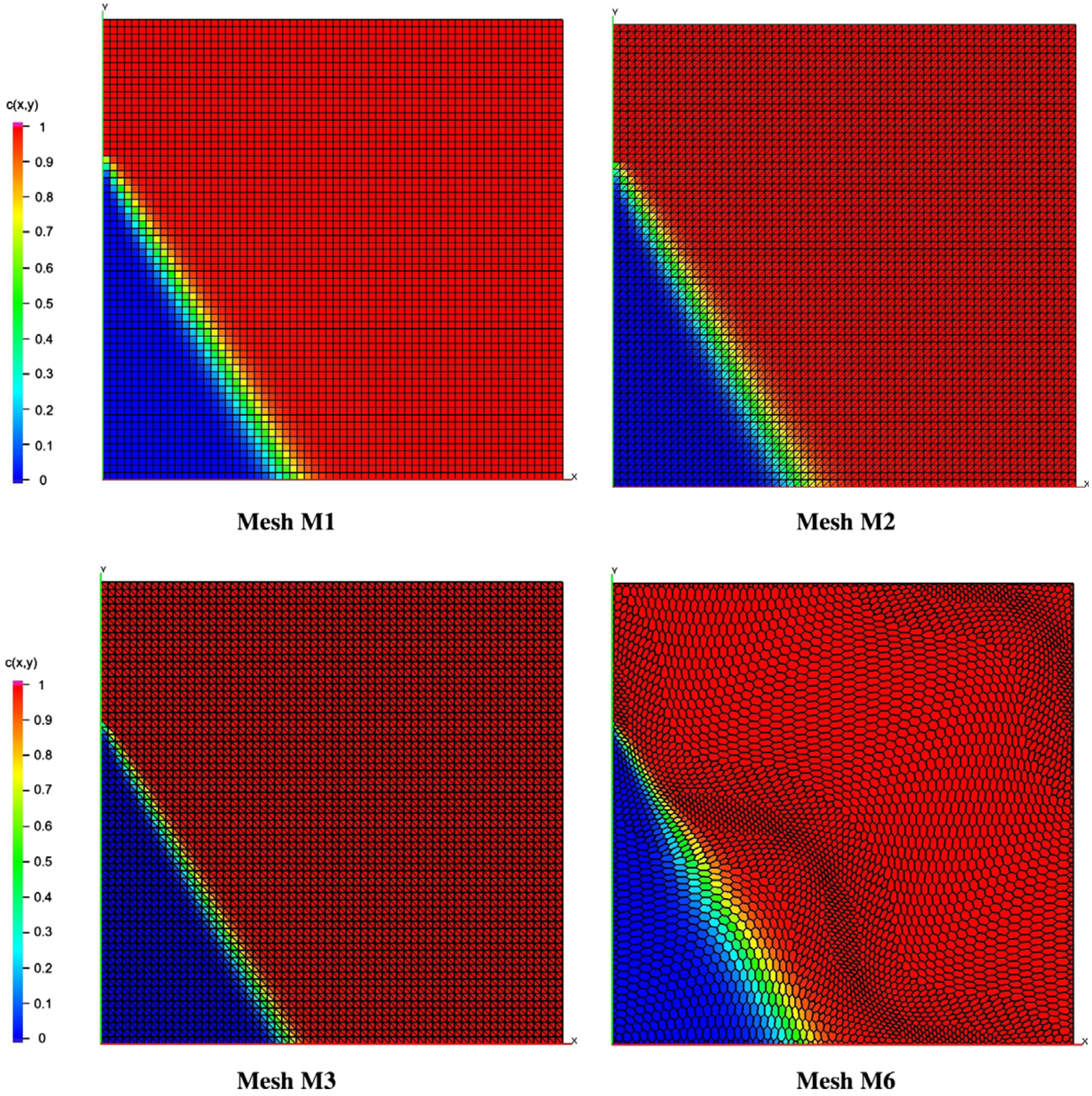


Fig. 6. Monotonicity test: the numerical solutions vary between 0 and 1.

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 $Pe_K = 781,250$. According to Theorems 4.1 and 4.2, the numerical solution must be non-negative. The numerical solutions for four meshes are shown in Fig. 6.

Table 5
The quantities that characterize the quality of the numerical solution for the problem described in Section 5.3.3.

Mesh	$osc_{\text{int}}^{\min}$	$osc_{\text{int}}^{\max}$	$osc_{\exp}$	$smear_{\text{int}}$	$smear_{\exp}$
M1	0	6.34e−12	2.19e−11	7.81e−02	2.13e−05
M2	0	2.41e−07	6.02e−14	8.33e−02	4.51e−05
M3	0	9.42e−16	6.27e−06	4.69e−02	4.33e−05
M6	0	6.96e−08	1.84e−13	1.13e−01	8.36e−05

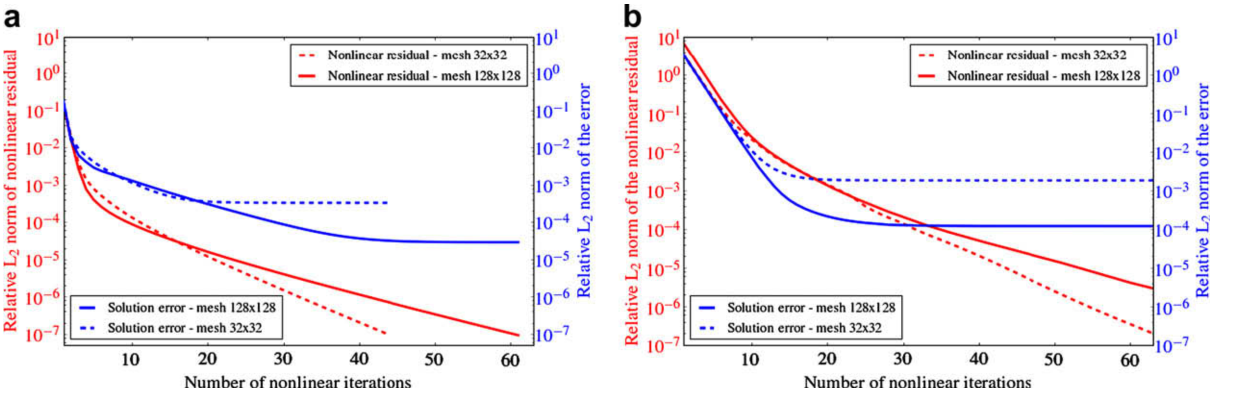


Fig. 7. The convergence of the Picard method: (a) diffusion-dominated problem from Section 5.2.1 ($\omega_k = 0.95$), and (b) advection-dominated problem from Section 5.3.2 ($\omega_k = 0.5$).

In order to measure quality of the numerical solution, the authors of [14] have proposed several estimates which quantify solution oscillations and smearing effects caused by a discretization scheme. Let $\Omega_1 = \{(x,y) \in \Omega : \; x \leqslant 0.5, y \geqslant 0.1\}$, $\Omega_2 = \{(x,y) \in \Omega : \; x \geqslant 0.7\}$, and Ω_3 denote a cell strip in the vicinity of the line $y = 0.25$,

$$\Omega_3 = \{T \in \mathcal{T} : \; \mathbf{x}_T = (x_T, y_T), |y_T - 0.25| \leqslant |T|^{1/2}\}.$$

For the square mesh **M1**, the width of this strip is equal to $2h$. First, we define two estimates (39) and (40) which characterize the values of undershoots and overshoots in Ω_1, correspondingly:

$$osc_{\text{int}}^{\min} \equiv \left(\sum_{(x,y)\in\Omega_1} (\min\{0, c_h(x,y)\})^2 \right)^{1/2}, \tag{39}$$

$$osc_{\text{int}}^{\max} \equiv \left(\sum_{(x,y)\in\Omega_1} (\max\{0, c_h(x,y) - 1\})^2 \right)^{1/2}. \tag{40}$$

Second, we define estimate (41) which quantifies oscillations near the boundary layer in Ω_2:

$$osc_{\exp} \equiv \left(\sum_{(x,y)\in\Omega_2} (\max\{0, c_h(x,y) - 1\})^2 \right)^{1/2}. \tag{41}$$

Third, we define two estimates (42) and (43) which measure thickness of the boundary layer and the internal shock, respectively:

$$smear_{\exp} \equiv \left(\sum_{(x,y)\in\Omega_2} (\min\{0, c_h(x,y) - 1\})^2 \right)^{1/2}, \tag{42}$$

$$smear_{\text{int}} \equiv x_2 - x_1, \tag{43}$$

where

$$x_1 = \min_{\mathbf{x}_T\in\Omega_3, C(\mathbf{x}_T)\geqslant 0.1} x_T \quad \text{and} \quad x_2 = \max_{\mathbf{x}_T\in\Omega_3, C(\mathbf{x}_T)\leqslant 0.9} x_T.$$

For the continuous solution these estimates depend on the diffusion process only, so they are much smaller than the considered mesh size. For the numerical solution, small values of estimates (39)–(43) characterize almost non-oscillatory and almost non-diffusive discrete solution.

The results obtained by the nonlinear FV method are shown in Table 5. They are competitive with the best results presented in review [14]. 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.

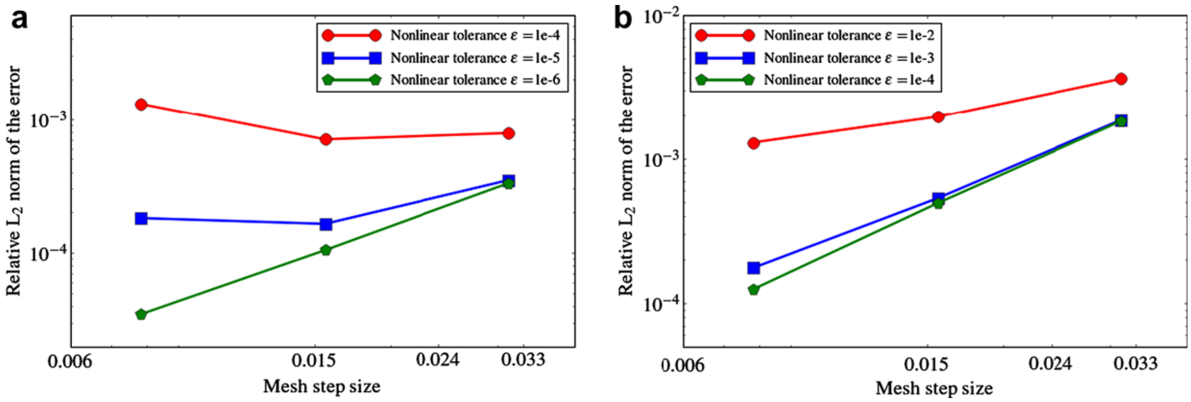


Fig. 8. The convergence study for different values of nonlinear tolerance ε_{non}: (a) diffusion-dominated problem from Section 5.2.1 ($\omega_k = 0.95$), and (b) advection-dominated problem from Section 5.3.2 ($\omega_k = 0.5$).

5.4. *Nonlinear iteration*

In the last group of tests we investigate the convergence of nonlinear iterations in Algorithm 1.

We recall that 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 $\varepsilon_{non} = 10^{-7}$. 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 (33) is solved with much cruder tolerance than 10^{-7}.

We consider the problem with the smooth solution described in Section 5.2.1 and the problem with the exponential boundary layer described in Section 5.3.2. Both of these problems are solved on a sequence of distorted quadrilateral meshes {**M5**}. In Fig. 7(a) and (b), the relative L^2-norm of error for the 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 $\varepsilon_{non} = 10^{-7}$. This difference is even more distinct in the advection-dominated problem.

In Fig. 8(a) and (b), the relative L^2-norm of error for the 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 cruder tolerance and, respectively, with much smaller number of nonlinear iterations. For example, 20 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 have found 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 more robust behavior. The iterative process is reformulated as follows:

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

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 factors $\{\omega_k\}$ is determined by the delicate balance between the robustness and the convergence speed of the iterative process. It is difficult to determine a unique damping parameter which is optimal for different types of meshes and problems considered in this section. Our experience shows that ω_k between 0.5 and 0.75 provides robust behavior for the considered problems. We noticed that the relaxation is important for the advection-dominated problems on highly distorted meshes. Also, usage of the extended set Σ_T^{ext} on triangular meshes requires additional relaxation compared to the smaller set Σ_T. A dynamic choice of the damping factor can significantly increase the efficacy of the method and it will be analyzed in the future.

6. Conclusion

We developed and analyzed the new monotone finite volume method for the advection–diffusion equation with a full anisotropic continuous diffusion tensor. We proved non-negativity of the numerical solution provided that the source term and the Dirichlet boundary data are non-negative and the flux on the Neuman boundary is non-positive. The method does not require to interpolate solution to mesh nodes and can be applied to unstructured polygonal meshes. Generalization of the method to problems with heterogeneous diffusion coefficients can be done following the path described in [10,23]. The numerical experiments demonstrated the second-order convergence rate for the concentration and the first-order convergence rate for the flux on *randomly distorted meshes* for problems with *highly anisotropic coefficients* in both advection-dominated and diffusion-dominated regimes.

Acknowledgments

This work was carried out under the auspices of the National Nuclear Security Administration of the US 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 LA-UR 09-03209.

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

References

[1] I. Aavatsmark, G.T. Eigestad, B.T. Mallison, J.M. Nordbotten, A compact multipoint flux approximation method with improved robustness, Numer. Methods Partial Differ. Equations 24 (5) (2008) 1329–1360.
[2] T.J. Barth, A 3D upwind Euler solver for unstructured meshes, in: Proceedings of the AIAA 10th Computational Fluid Dynamics Conference, 1991, pp. 228–238, in: Proceedings of the 10th Conference on Computational Fluid Dynamics, Honolulu, HI, June 24–27, 1991.
[3] D. Benson, A new two-dimensional flux-limited shock viscosity for impact calculations, Comput. Meth. Appl. Mech. Eng. 93 (1991) 39–95.
[4] E. Bertolazzi, G. Manzini, A cell-centered second-order accurate finite volume method for convection–diffusion problems on unstructured meshes, Math. Models Methods Appl. Sci. 14 (8) (2004) 1235–1260.
[5] E. Bertolazzi, G. Manzini, A second-order maximum principle preserving finite volume method for steady convection–diffusion problems, SIAM J. Numer. Anal. 43 (5) (2005) 2172–2199.
[6] A.N. Brooks, T.J.R. Hughes, Streamline upwind/Petrov–Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier–Stokes equations, Comput. Methods Appl. Mech. Eng. 32 (1–3) (1982) 199–259.
[7] E. Burman, A. Ern, Discrete maximum principle for Galerkin approximations of the Laplace operator on arbitrary meshes, C.R. Math. Acad. Sci. Paris 338 (8) (2004) 641–646.
[8] G. Chavent, J. Jaffré, Mathematical Models and Finite Elements for Reservoir Simulation, Elsevier Science Publishers, B.V., Netherlands, 1986.
[9] P.G. Ciarlet, P.-A. Raviart, Maximum principle and uniform convergence for the finite element method, Comput. Methods Appl. Mech. Eng. 2 (1973) 17–31.
[10] A. Danilov, Yu. Vassilevski, A monotone nonlinear finite volume method for diffusion equations on conformal polyhedral meshes, Russ. J. Numer. Anal. Math. Modell. 24 (3) (2009) 207–227.
[11] F. Gao, Y. Yuan, D. Yang, An upwind finite-volume element scheme and its maximum-principle-preserving property for nonlinear convection–diffusion problem, Int. J. Numer. Methods Fluids 56 (12) (2008) 2301–2320.
[12] M.E. Hubbard, Multidimensional slope limiters for muscl-type finite volume schemes on unstructured grids, J. Comput. Phys. 155 (1) (1999) 54–74.
[13] T.J.R. Hughes, M. Mallet, A. Mizukami, A new finite element formulation for computational fluid dynamics. II. Beyond SUPG, Comput. Methods Appl. Mech. Eng. 54 (3) (1986) 341–355.
[14] V. John, P. Knobloch, On spurious oscillations at layers diminishing (SOLD) methods for convection-diffusion equations: part I – a review, Comput. Methods Appl. Mech. Eng. 196 (17–20) (2007) 2197–2215.
[15] I. Kapyrin, A family of monotone methods for the numerical solution of three-dimensional diffusion problems on unstructured tetrahedral meshes, Dokl. Math. 76 (2) (2007) 734–738.
[16] S. Korotov, M. Křížek, P. Neittaanmäki, Weakened acute type condition for tetrahedral triangulations and the discrete maximum principle, Math. Comput. 70 (233) (2001) 107–119 (Electronic).
[17] D. Kuzmin, On the design of general-purpose flux limiters for finite element schemes. I. Scalar convection, J. Comput. Phys. 219 (2) (2006) 513–531.
[18] D. Kuzmin, M. Moller, Algebraic flux correction I. Scalar conservation laws, in: D. Kuzmin, R. Lohner, S. Turek (Eds.), Flux-Corrected Transport: Principles, Algorithms, and Applications, Springer-Verlag, Berlin, 2005, pp. 155–206.
[19] D. Kuzmin, M.J. Shashkov, D. Svyatskiy, A constrained finite element method satisfying the discrete maximum principle for anisotropic diffusion problems, J. Comput. Phys. 228 (9) (2009) 3448–3463.
[20] S. Lamine, M.G. Edwards, Higher-resolution convection schemes for flow in porous media on highly distorted unstructured grids, Int. J. Numer. Meth. Eng. 76 (8) (2008) 1139–1158.
[21] R.J. LeVeque, Finite Volume Methods for Hyperbolic Problems, Cambridge Texts in Applied Mathematics, 2002.
[22] K. Lipnikov, D. Svyatskiy, M. Shashkov, Yu. Vassilevski, Monotone finite volume schemes for diffusion equations on unstructured triangular ans shape-regular polygonal meshes, J. Comput. Phys. 227 (2007) 492–512.
[23] K. Lipnikov, D. Svyatskiy, Y. Vassilevski, Interpolation-free monotone finite volume method for diffusion equations on polygonal meshes, J. Comput. Phys. 228 (3) (2009) 703–716.
[24] R. Liska, M. Shashkov, Enforcing the discrete maximum principle for linear finite element solutions of second-order elliptic problems, Commun. Comput. Phys. 3 (4) (2008) 852–877.
[25] G. Manzini, A. Russo, A finite volume method for advection–diffusion problems in convection-dominated regimes, Comput. Methods Appl. Mech. Eng. 197 (13–16) (2008) 1242–1261.
[26] K.B. Nakshatrala, A.J. Valocchi, Non-negative mixed finite element formulations for a tensorial diffusion equation, J. Comput. Phys. 228 (18) (2009) 6726–6752.
[27] J.M. Nordbotten, I. Aavatsmark, G.T. Eigestad, Monotonicity of control volume methods, Numer. Math. 106 (2) (2007) 255–288.
[28] C. Le Potier, 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 (2005) 787–792.
[29] C. Le Potier, Finite volume scheme satisfying maximum and minimum principles for anisotropic diffusion operators, in: R. Eymard, J.-M. Herard (Eds.), Finite Volumes for Complex Applications V, 2008, pp. 103–118.
[30] R.T. Rockafellar, Convex Analysis. Princeton Landmarks in Mathematics, Princeton University Press, 1996.
[31] 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) (1992) 631–644.
[32] B. van Leer, Towards the ultimate conservative difference scheme. V. A second-order sequel to godunov's method, J. Comput. Phys. 32 (1) (1979) 101–136.
[33] R.S. Varga, Matrix Iterative Analysis, Prentice-Hall Inc., Englewood Cliffs, NJ, 1962.
[34] Yu. Vassilevski, I. Kapyrin, Two splitting schemes for nonstationary convection–diffusion problems on tetrahedral meshes, Comput. Math. Math. Phys. 48 (8) (2008) 1349–1366.
[35] R.S. Vitoriano, On the strong maximum principle for some piecewise linear finite element approximate problems of non-positive type, J. Fac. Sci. Univ. Tokyo Sect. IA Math. 29 (2) (1982) 473–491.
[36] D. Wollstein, T. Linss, R. Hans-Gorg, Uniformly accurate finite volume discretization for a convection–diffusion problem, Electron. Trans. Numer. Anal. 13 (2002) 1–11.
[37] A. Yuan, Z. Sheng, Monotone finite volume schemes for diffusion equations on polygonal meshes, J. Comput. Phys. 227 (2008) 6288–6312.