Journal of Computational Physics 228 (2009) 6726–6752

Contents lists available at ScienceDirect

Journal of Computational Physics

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

Non-negative mixed finite element formulations for a tensorial diffusion equation

K.B. Nakshatrala [a,*], A.J. Valocchi [b]

[a] Department of Mechanical Engineering, 216 Engineering/Physics Building, Texas A&M University, College Station, Texas 77843, United States
[b] Department of Civil and Environmental Engineering, 1110 Newmark Laboratory, University of Illinois at Urbana-Champaign, Illinois 61801, United States

ARTICLE INFO

Article history:
Received 1 October 2008
Received in revised form 25 April 2009
Accepted 14 May 2009
Available online 3 June 2009

Keywords:
Maximum–minimum principles for elliptic PDEs
Discrete maximum–minimum principle
Non-negative solutions
Active set strategy
Convex quadratic programming
Tensorial diffusion equation
Monotone methods

ABSTRACT

We consider the tensorial diffusion equation, and address the discrete maximum–minimum principle of mixed finite element formulations. In particular, we address non-negative solutions (which is a special case of the maximum–minimum principle) of mixed finite element formulations. It is well-known that the classical finite element formulations (like the single-field Galerkin formulation, and Raviart–Thomas, variational multiscale, and Galerkin/least-squares mixed formulations) do not produce non-negative solutions (that is, they do not satisfy the discrete maximum–minimum principle) on arbitrary meshes and for strongly anisotropic diffusivity coefficients.

In this paper, we present two non-negative mixed finite element formulations for tensorial diffusion equations based on constrained optimization techniques. These proposed mixed formulations produce non-negative numerical solutions on arbitrary meshes for low-order (i.e., linear, bilinear and trilinear) finite elements. The first formulation is based on the Raviart–Thomas spaces, and the second non-negative formulation is based on the variational multiscale formulation. For the former formulation we comment on the effect of adding the non-negative constraint on the local mass balance property of the Raviart–Thomas formulation.

We perform numerical convergence analysis of the proposed optimization-based non-negative mixed formulations. We also study the performance of the active set strategy for solving the resulting constrained optimization problems. The overall performance of the proposed formulation is illustrated on three canonical test problems.

Published by Elsevier Inc.

1. Introduction

Robustness of numerical methods for flow and transport problems in porous media is important for development of simulators to be used in a wide range of applications in subsurface hydrology and contaminant transport. In order to obtain robust and reliable numerical results it is imperative to preserve basic properties of solutions of mathematical models by computed approximations. In the simulation of reactive transport of contaminants one such basic properties is non-negative solutions as concentration of a chemical or biological species physically can never be negative. Since the domain of interest in subsurface flows is highly complex, one needs to employ unstructured computational grids. Therefore, obtaining non-negative solutions on unstructured meshes is an essential feature in the simulation of reactive transport of contaminants, as well as many other physical processes.

* Corresponding author. Tel.: +1 979 845 1292.
E-mail addresses: knakshatrala@tamu.edu (K.B. Nakshatrala), valocchi@uiuc.edu (A.J. Valocchi).

0021-9991/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.jcp.2009.05.039

However, obtaining non-negative solutions on unstructured grids using a numerical method (finite element, finite volume or finite difference) is not an easy task. In addition, there is another complexity arising from an anisotropic diffusion tensor. In subsurface flows, the heterogeneity in the velocity field will give rise to a non-homogeneous anisotropic diffusion tensor with non-negligible cross terms [1]. Several studies have shown that standard treatment of the cross-diffusion term will result in negative solutions on general computational grids, see [2] and references therein. Several ad-hoc procedures have been proposed in the literature. For example, a post processing step is typically employed in which one performs some sort of "smoothing." But this procedure, in many cases, is not variationally consistent. Some other methods are limited in their range of applicability (e.g., the method proposed in Ref. [2] can handle only structured grids).

Herein we consider the dispersion/diffusion process for steady single-phase flow in heterogeneous anisotropic porous media. Such a flow can be described by the Poisson's equation with a tensorial diffusion coefficient, which when written in the mixed form is similar to the governing equations of Darcy flow. In this paper we propose two optimization-based mixed methods for solving tensorial diffusion equation that gives non-negative solutions on general grids for linear finite elements. The two methods are developed by rewriting the Raviart–Thomas and variational multiscale formulations as constrained minimization problems subject to a constraint on the primary variable to be non-negative. A similar approach based on optimization techniques has been used by Liska and Shashkov [3] for a single field formulation, but herein we consider mixed finite element formulations.

The main idea behind the proposed methods is to augment a constraint on nodal values to be non-negative to the discrete (that is, after spatial finite element discretization) variational statement of the underlying formulation. Since we consider only low-order finite elements (and since the shape functions for these elements do not change their sign within an element), non-negative nodal values ensure non-negative solution every where in the element, and hence non-negative solution on the whole domain. This argument will not hold for high-order finite elements as shape functions for these finite elements (in general) change sign within an element.

Throughout this paper continuum vectors are denoted with lower case boldface normal letters, and (continuum) second-order tensors will be denoted using (LATEX) blackboard font (for example, $\mathbf{v}$ and $\mathbb{D}$, respectively). We denote finite element vectors and matrices with lower and upper case boldface italic letters, respectively. For example, vector $\boldsymbol{v}$ and matrix $\boldsymbol{K}$. The curled inequality symbols $\succeq$ and $\preceq$ are used to denote generalized inequalities between vectors, which represent component-wise inequalities. That is, given two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, $\boldsymbol{a} \succeq \boldsymbol{b}$ means $a_i \geqslant b_i \; \forall i$. A similar definition holds for the symbol $\preceq$. Other notational conventions adopted in this paper are introduced as needed.

1.1. Governing equations

Let $\Omega \subseteq \mathbb{R}^{nd}$ (where "nd" is the number of spatial dimensions) be a bounded domain with boundary $\partial\Omega = \overline{\Omega} \setminus \Omega$, where $\overline{\Omega}$ denotes the closure of Ω. Consider the diffusion of a chemical species in anisotropic heterogeneous medium, which is governed by the second-order elliptic tensorial diffusion partial differential equation. The governing equations are

$$-\nabla \cdot (\mathbb{D}(\mathbf{x})\nabla c(\mathbf{x})) = f(\mathbf{x}) \quad \text{in } \Omega \tag{1}$$

$$-\mathbf{n}(\mathbf{x}) \cdot \mathbb{D}(\mathbf{x})\nabla c = t^{\mathrm{p}}(\mathbf{x}) \quad \text{on } \Gamma^{\mathrm{N}} \tag{2}$$

$$c(\mathbf{x}) = c^{\mathrm{p}}(\mathbf{x}) \quad \text{on } \Gamma^{\mathrm{D}} \tag{3}$$

where $c(\mathbf{x})$ denotes the concentration field, $f(\mathbf{x})$ is the volumetric source, $t^{\mathrm{p}}(\mathbf{x})$ is the prescribed flux (i.e., Neumann boundary condition), $c^{\mathrm{p}}(\mathbf{x})$ is the prescribed concentration (i.e., Dirichlet boundary condition), Γ^{D} is that part of the boundary on which Dirichlet boundary condition is applied, Γ^{N} is the part of the boundary on which Neumann boundary condition is applied, $\mathbf{n}(\mathbf{x})$ is unit outward normal to the boundary, and ∇ denotes gradient operator. For well-posedness one requires $\Gamma^{\mathrm{D}} \cup \Gamma^{\mathrm{N}} = \partial\Omega$ and $\Gamma^{\mathrm{D}} \cap \Gamma^{\mathrm{N}} = \emptyset$, and for uniqueness $\Gamma^{\mathrm{D}} \neq \emptyset$. We assume that the coefficient of diffusivity $\mathbb{D}(\mathbf{x})$ is a symmetric positive definite tensor such that, for some $0 < \alpha_1 \leqslant \alpha_2 < +\infty$, we have

$$\alpha_1 \mathbf{y}^T\mathbf{y} \leqslant \mathbf{y}^T\mathbb{D}(\mathbf{x})\mathbf{y} \leqslant \alpha_2 \mathbf{y}^T\mathbf{y} \quad \forall \mathbf{x} \in \Omega, \; \forall \mathbf{y} \neq \mathbf{0} \in \mathbb{R}^{nd} \tag{4}$$

In addition, we assume that $\mathbb{D}(\mathbf{x})$ is continuously differentiable.

1.2. First-order (or mixed) form

In many situations, the primary quantity of interest is the flux. But a single field (or primal) formulation does not produce accurate solutions for the flux. One can calculate the flux by differentiating the obtained $c(\mathbf{x})$, but there will be a loss of accuracy during this process. For example, under a single field formulation, linear finite elements produce fluxes that are constant and discontinuous across elements. This means that there is no flux balance across element edges. Balance of flux along element edges is a highly desirable feature and is of physical importance in many practical engineering problems. In order to alleviate aforementioned drawbacks of single formulations, mixed formulations are often employed. Eqs. (1)–(3) in mixed (or first-order) form can be written as

$$\mathbb{D}^{-1}(\mathbf{x})\mathbf{v}(\mathbf{x}) = -\nabla c \quad \text{in } \Omega \tag{5}$$

$$\nabla \cdot \mathbf{v} = f(\mathbf{x}) \quad \text{in } \Omega \tag{6}$$

$$\mathbf{v}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) = t^{\mathrm{p}}(\mathbf{x}) \quad \text{on } \Gamma^{\mathrm{N}} \tag{7}$$

$$c(\mathbf{x}) = c^{\mathrm{p}}(\mathbf{x}) \quad \text{on } \Gamma^{\mathrm{D}} \tag{8}$$

where $\mathbf{v}(\mathbf{x})$ is an auxiliary variable, which can be interpreted as follows: given a plane defined by a normal $\mathbf{n}$, the quantity $\mathbf{v} \cdot \mathbf{n}$ will be the flux through the plane.

1.3. Maximum–minimum principle

It is well-known that some elliptic partial differential equations (under appropriate regularity assumptions) satisfy the so-called maximum–minimum principle, and the Poisson's equation is one of them [4]. We now state the classical maximum–minimum principle for second-order elliptic partial differential equations. (In Section 3 we state and prove a maximum–minimum principle under milder regularity assumptions. Note that weak solutions *may* also possess a maximum–minimum principle. For example, see Ref. [5].) Consider the boundary value problem given by Eqs. (1)–(3). Let $c(\mathbf{x}) \in C^2(\Omega) \cap C^0(\overline{\Omega})$, where $C^2(\Omega)$ denotes the set of twice continuously differentiable functions defined on Ω, and $C^0(\overline{\Omega})$ the set of uniformly continuous functions defined on Ω. If $f(\mathbf{x}) \geqslant 0$ (or if $f(\mathbf{x}) \leqslant 0$) in Ω then $c(\mathbf{x})$ attains its minimum (or its maximum) on the boundary of Ω. For a detailed discussion on maximum–minimum principles see Refs. [4,6,7,5].

One of the important consequences of maximum–minimum principles is the non-negative solution of a (tensorial) diffusion equation under non-negative forcing function with non-negative prescribed Dirichlet boundary condition. Obtaining non-negative solutions is of paramount importance in studying transport of chemical and biological species as negative concentration of a species is unphysical.

1.4. Discrete maximum–minimum principle

The discrete analogy of the maximum–minimum principle is commonly referred to as the *discrete maximum–minimum principle* (DMP). However, many numerical formulations do not *unconditionally* satisfy the discrete maximum–minimum principle. Typically, there will be restrictions on the mesh or on the magnitude of coefficients of the diffusivity tensor. For example, the single field Galerkin formulation in the case of scalar diffusion satisfies the discrete maximum–minimum principle if the mesh satisfies weak acute condition [8] (and also see Appendix). The question whether we get non-negative *numerical* solutions leads us to the discrete maximum–minimum principle.

For recent works on DMP see Refs. [9–15] and also see the discussion in Ref. [3, Section 1]. Considerable attention to DMP has also been given in the finite volume literature [2,16–18]. Optimization-based techniques have been employed in Refs. [3,19] to address DMP. For completeness, some of the classical results on discrete maximum–minimum principle are outlined in Appendix.

In this paper we concentrate on obtaining non-negative numerical solutions using mixed formulations under non-negative forcing function with non-negative Dirichlet boundary condition where ever it is prescribed. (Note that we do not assume that the Dirichlet boundary condition has to be prescribed on the whole boundary.) In all our test problems (see Section 4) we have $f(\mathbf{x}) \geqslant 0$ in Ω and $c^{\mathrm{p}}(\mathbf{x}) \geqslant 0$ on $\partial\Omega$. By using the maximum–minimum principle one can conclude that $c(\mathbf{x}) \geqslant 0$ in whole of $\overline{\Omega}$ (the closure of Ω). That is, for the chosen test problems in Section 4, we must have non-negative solutions in the whole domain.

1.5. Main contributions of this paper

Some of the main contributions of this paper are as follows:

- We numerically demonstrate that various conditions outlined in Appendix (which are sufficient for isotropic diffusion) are not sufficient for tensorial diffusion equation under the Raviart–Thomas and variational multiscale formulations to produce non-negative solutions under non-negative forcing functions with non-negative prescribed Dirichlet boundary conditions.
- We develop a non-negative formulation based on the lowest-order Raviart–Thomas spaces, and discuss the consequences of obtaining non-negative solutions on the local mass balance property.
- We extend the variational multiscale formulation to produce non-negative solutions on general grids for low-order finite elements under non-negative forcing function and prescribed non-negative Dirichlet boundary condition. We also show that the (continuous) variational multiscale formulation satisfies a continuous maximum–minimum principle (under appropriate regularity assumptions).

1.6. Organization of the paper

The remainder of this paper is organized as follows. In Section 2 we present a non-negative formulation based on the lowest-order Raviart–Thomas (RT0) finite element spaces, which is achieved by adding a non-negative constraint to the discrete

variational setting of the Raviart–Thomas formulation. For this non-negative formulation we present both primal and dual constrained optimization problems, and comment on the ease of solving these problems and also the consequences of imposing the non-negative constraint on the local mass balance. In Section 3, a non-negative formulation based on the variational multiscale formulation will be presented. We also show that the (continuous) variational multiscale formulation satisfies a continuous maximum–minimum principle. Numerical results along with a discussion on the numerical performance of both the proposed non-negative formulations will be presented in Section 4. Finally, conclusions are drawn in Section 5.

2. A non-negative mixed formulation based on Raviart–Thomas spaces

The Raviart–Thomas finite element formulation is widely used (for an example in subsurface modeling see [20]) to solve diffusion equations in mixed form, and is based on the classical mixed formulation [21]. The simplest and lowest order Raviart–Thomas space (commonly denoted as RT0) consists of fluxes evaluated on the midpoints of edges and constant pressure over elements. We first present the weak form and variational structure behind the Raviart–Thomas formulation. We then modify the variational structure by adding non-negative constraint on the concentration to build a non-negative low-order finite element formulation based on the RT0 spaces.

To this end, define function spaces as

$$\mathcal{V} := \left\{ \mathbf{v}(\mathbf{x}) \mid \mathbf{v}(\mathbf{x}) \in (L^2(\Omega))^{nd},\ \nabla \cdot \mathbf{v} \in L^2(\Omega),\ \text{trace}\ (\mathbf{v} \cdot \mathbf{n}) = t^{\mathrm{p}}(\mathbf{x})\ \text{on}\ \Gamma^{\mathrm{N}} \right\} \tag{9}$$

$$\mathcal{W} := \left\{ \mathbf{v}(\mathbf{x}) \mid \mathbf{v}(\mathbf{x}) \in (L^2(\Omega))^{nd},\ \nabla \cdot \mathbf{v} \in L^2(\Omega),\ \text{trace}\ (\mathbf{v} \cdot \mathbf{n}) = 0\ \text{on}\ \Gamma^{\mathrm{N}} \right\} \tag{10}$$

$$\mathcal{P} := L^2(\Omega) \tag{11}$$

Recall that "*nd*" denotes the number of spatial dimensions. For further details on function spaces see the monograph by Brezzi and Fortin [22]. Let $\mathbf{w}(\mathbf{x})$ and $q(\mathbf{x})$ denote the weighting functions corresponding to $\mathbf{v}(\mathbf{x})$ and $c(\mathbf{x})$, respectively. The classical mixed formulation for Eqs. (5)–(8) can be written as: Find $\mathbf{v}(\mathbf{x}) \in \mathcal{V}$ and $c(\mathbf{x}) \in \mathcal{P}$ such that

$$(\mathbf{w}; \mathbb{D}^{-1}\mathbf{v}) - (\nabla \cdot \mathbf{w}; c) + (\mathbf{w} \cdot \mathbf{n}; c^{\mathrm{p}})_{\Gamma^{\mathrm{D}}} - (q; \nabla \cdot \mathbf{v} - f) = 0 \quad \forall \mathbf{w}(\mathbf{x}) \in \mathcal{W},\ q(\mathbf{x}) \in \mathcal{P} \tag{12}$$

It is well-known that, under appropriate smoothness conditions on the domain and its boundary, the above saddle-point formulation is well-posed [22]. That is, a unique (weak) solution exists for this problem that depends continuously on the input data. However, to obtain stable results using a finite element approximation, the finite dimensional spaces $\mathcal{V}^h \subset \mathcal{V}$ and $\mathcal{P}^h \subset \mathcal{P}$ in which a numerical solution is sought have to satisfy the Ladyzhenskaya-Babuška-Brezzi (LBB) stability condition [22]. One such space that satisfies the LBB condition is the popular Raviart–Thomas (RT) finite element space. In this paper we consider only the lowest order Raviart–Thomas triangular finite element space (RT0). Let $\mathcal{T}_h$ be a triangulation on Ω. The lowest order Raviart–Thomas finite dimensional subspaces on triangles are defined as

$$\mathcal{P}_h := \{ p \mid p = \text{a constant on each triangle } K \in \mathcal{T}_h \} \tag{13}$$

$$\mathcal{V}_h := \left\{ \mathbf{v} = (v^{(1)}, v^{(2)}) \mid v_K^{(1)} = a_K + b_K x,\ v_K^{(2)} = c_K + b_K y; a_K, b_K, c_K \in \mathbb{R}; K \in \mathcal{T}_h \right\} \tag{14}$$

2.1. Discrete equations

The discretized finite element equations of the Raviart–Thomas formulation for the mixed form of tensorial diffusion equation can be written as [20]

$$\begin{bmatrix} \boldsymbol{K}_{vv} & \boldsymbol{K}_{pv}^T \\ \boldsymbol{K}_{pv} & \boldsymbol{O} \end{bmatrix} \begin{Bmatrix} \boldsymbol{v} \\ \boldsymbol{p} \end{Bmatrix} = \begin{Bmatrix} \boldsymbol{f}_v \\ \boldsymbol{f}_p \end{Bmatrix} \tag{15}$$

where $\boldsymbol{O}$ is a zero matrix of appropriate size, the matrix $\boldsymbol{K}_{vv}$ is symmetric and positive definite, $\boldsymbol{v}$ denotes the (finite element) vector of flux degrees-of-freedom, and $\boldsymbol{p}$ denotes the vector of concentration degrees-of-freedom. Comparing the weak form (12) and discrete Eq. (15), the matrices $\boldsymbol{K}_{vv}$ and $\boldsymbol{K}_{pv}$ are obtained after the finite element discretization of the terms $(\mathbf{w}; \mathbb{D}^{-1}\mathbf{v})$ and $-(q; \nabla \cdot \mathbf{v})$, respectively. Since Eq. (12) is written in symmetric form, $\boldsymbol{K}_{vp}$ (which comes from the term $-(\nabla \cdot \mathbf{w}; c)$) will be equal to $\boldsymbol{K}_{pv}^T$. The vectors $\boldsymbol{f}_v$ and $\boldsymbol{f}_p$ are, respectively, obtained from $-(\mathbf{w} \cdot \mathbf{n}; c^{\mathrm{p}})_{\Gamma^{\mathrm{D}}}$ and $-(q; f)$ after the finite element discretization. Since there is no term in Eq. (12) that contains both q in the weighting (i.e., first) slot and c in the second slot in a bilinear form $(\cdot;\cdot)$, we have the matrix $\boldsymbol{K}_{pp} = \boldsymbol{O}$ (a zero matrix).

The above system of Eq. (15) is equivalent to the following constrained minimization problem

$$\text{(P1-RT0)} \quad \begin{cases} \text{minimize}_{\boldsymbol{v}} & \frac{1}{2}\boldsymbol{v}^T \boldsymbol{K}_{vv} \boldsymbol{v} - \boldsymbol{v}^T \boldsymbol{f}_v \\ \text{subject to} & \boldsymbol{K}_{pv} \boldsymbol{v} - \boldsymbol{f}_p = \boldsymbol{0} \end{cases} \tag{16}$$

where $\boldsymbol{0}$ is a zero vector of appropriate size. Note that the constraint in Eq. (16) is the local mass balance condition for each element. We refer the above equation as the primal problem for the Raviart–Thomas formulation, and denote it as (P1-RT0).

This primal problem belongs to the class of *convex quadratic programming* problems, and from optimization theory (for example, see Ref. [23]) it can be shown that the problem has a unique global minimizer.

Remark 2.1. A quadratic program is an optimization problem in which the objective function is a quadratic function and the (equality and inequality) constraints are all linear. In a convex quadratic program the Hessian of the objective function is positive semidefinite.

Remark 2.2. It is interesting to note that, from the complexity theory, problem (16) can be solved in polynomial time (for example, using the ellipsoid and interior point methods) [23,24]. Note that the term "polynomial time" in the context of complexity theory should not be confused with the term "polynomial convergence," which is commonly used in the convergence studies using the finite element method.

Define the Lagrangian as

$$\mathcal{L}(\boldsymbol{v},\boldsymbol{p}) := \frac{1}{2}\boldsymbol{v}^T\boldsymbol{K}_{vv}\boldsymbol{v} - \boldsymbol{v}^T\boldsymbol{f}_v + \boldsymbol{p}^T(\boldsymbol{K}_{pv}\boldsymbol{v} - \boldsymbol{f}_p) \tag{17}$$

where $\boldsymbol{p}$ is the vector of Lagrange multipliers. Using the Lagrange multiplier method [23] the primal problem (16) is equivalent to

$$\underset{\boldsymbol{v},\,\boldsymbol{p}}{\text{extremize}}\ \ \mathcal{L}(\boldsymbol{v},\boldsymbol{p}) \tag{18}$$

and the first-order optimality conditions for this problem gives rise to the discretized finite element Eq. (15). We now write the dual problem corresponding to the primal problem (16). To this end, define the Lagrange dual function as

$$g(\boldsymbol{p}) := \inf_{\boldsymbol{v}}\ \mathcal{L}(\boldsymbol{v},\boldsymbol{p}) = -\frac{1}{2}\boldsymbol{p}^T\boldsymbol{K}_{pv}\boldsymbol{K}_{vv}^{-1}\boldsymbol{K}_{pv}^T\boldsymbol{p} + \boldsymbol{p}^T(\boldsymbol{K}_{pv}\boldsymbol{K}_{vv}^{-1}\boldsymbol{f}_v - \boldsymbol{f}_p) - \frac{1}{2}\boldsymbol{f}_v^T\boldsymbol{K}_{vv}^{-1}\boldsymbol{f}_v \tag{19}$$

The above expression on the right-hand side is obtained as follows. Let $\boldsymbol{v}^*$ be the minimizer that gives the infimum of $\mathcal{L}(\boldsymbol{v},\boldsymbol{p})$ with respect to $\boldsymbol{v}$. Then $\boldsymbol{v}^*$ has to satisfy

$$\boldsymbol{K}_{vv}\boldsymbol{v}^* - \boldsymbol{f}_v + \boldsymbol{K}_{pv}^T\boldsymbol{p} = \boldsymbol{0} \tag{20}$$

which is a necessary condition, and is obtained by equating the derivative of $\mathcal{L}(\boldsymbol{v},\boldsymbol{p})$ (which is defined in Eq. (17)) with respect to $\boldsymbol{v}$ to zero. Since the matrix $\boldsymbol{K}_{vv}$ is positive definite (and hence invertible) we have

$$\boldsymbol{v}^* = \boldsymbol{K}_{vv}^{-1}\left(\boldsymbol{f}_v - \boldsymbol{K}_{pv}^T\boldsymbol{p}\right) \tag{21}$$

By substituting the above expression for the minimizer $\boldsymbol{v}^*$ into the definition of $\mathcal{L}(\boldsymbol{v},\boldsymbol{p})$ (17) we obtain the expression on the right-hand side of Eq. (19).

The dual problem corresponding to the primal problem (16) can then be written as

$$\underset{\boldsymbol{p}}{\text{maximize}}\quad g(\boldsymbol{p}) \tag{22}$$

which is *equivalent* to

$$\text{(D1-RT0)}\quad \underset{\boldsymbol{p}}{\text{minimize}}\quad \frac{1}{2}\boldsymbol{p}^T\boldsymbol{K}_{pv}\boldsymbol{K}_{vv}^{-1}\boldsymbol{K}_{pv}^T\boldsymbol{p} - \boldsymbol{p}^T\left(\boldsymbol{K}_{pv}\boldsymbol{K}_{vv}^{-1}\boldsymbol{f}_v - \boldsymbol{f}_p\right) \tag{23}$$

The stationarity of the above problem implies

$$\boldsymbol{K}_{pv}\boldsymbol{K}_{vv}^{-1}\boldsymbol{K}_{pv}^T\boldsymbol{p} = \boldsymbol{K}_{pv}\boldsymbol{K}_{vv}^{-1}\boldsymbol{f}_v - \boldsymbol{f}_p \tag{24}$$

which is the Schur complement form of Eq. (15) expressed in terms of Lagrange multipliers by analytically eliminating the variable $\boldsymbol{v}$. Note that the Schur complement operator $\boldsymbol{K}_{pv}\boldsymbol{K}_{vv}^{-1}\boldsymbol{K}_{pv}^T$ is symmetric and positive definite.

A simple numerical example to be presented later (e.g., see Fig. 8) shows that RT0 triangular element does not satisfy the discrete maximum–minimum principle, and in particular, does not produce non-negative solutions for non-negative forcing functions with non-negative prescribed Dirichlet boundary conditions.

2.2. A non-negative mixed formulation

In order to get non-negative solutions under the RT0 spaces, we pose the dual problem as

$$\text{(D2-RT0)}\quad \begin{cases} \text{minimize}_{\boldsymbol{p}} & \frac{1}{2}\boldsymbol{p}^T\boldsymbol{K}_{pv}\boldsymbol{K}_{vv}^{-1}\boldsymbol{K}_{pv}^T\boldsymbol{p} - \boldsymbol{p}^T\left(\boldsymbol{K}_{pv}\boldsymbol{K}_{vv}^{-1}\boldsymbol{f}_v - \boldsymbol{f}_p\right) \\ \text{subject to} & \boldsymbol{p} \succeq \boldsymbol{0} \end{cases} \tag{25}$$

Recall that the symbol $\succeq$ denotes the generalized inequality between vectors, which represents component-wise inequality (see Introduction, just above Section 1.1, for a discussion on this notation). The primal problem corresponding to this new dual problem will then be

$$\text{(P2-RT0)} \quad \begin{cases} \text{minimize}_{\boldsymbol{v}} & \frac{1}{2}\boldsymbol{v}^T\boldsymbol{K}_{vv}\boldsymbol{v} - \boldsymbol{v}^T\boldsymbol{f}_v \\ \text{subject to} & \boldsymbol{K}_{pv}\boldsymbol{v} - \boldsymbol{f}_p \preceq \boldsymbol{0} \end{cases} \tag{26}$$

Remark 2.3. The primal problem given in Eq. (26) is obtained by inspection. That is, one can easily check (using a direct calculation) that the dual problem of this new primal problem (26) will be the same as Eq. (25). Also, it should be noted that one can write the dual problem corresponding to a given dual problem (that is, the dual of a dual). For the problem at hand, the dual of the dual problem will be the same as the primal problem, which is not the case in general [23]. Hence, one will obtain the primal problem (26) by writing the dual of the dual problem (25).

By comparing the constraints in Eqs. (16) and (26) one can conclude that under the proposed non-negative method based on the Raviart–Thomas formulation one may violate local mass balance by creating *artificial sinks*. We can infer more on local mass balance by looking at the Karush–Kuhn–Tucker (KKT) conditions (which in this case are necessary and sufficient for the optimality) for the new primal problem given by Eq. (26). The KKT optimality conditions for the (P2-RT0) problem are

$$\boldsymbol{K}_{vv}\boldsymbol{v} + \boldsymbol{K}_{pv}^T\boldsymbol{p} = \boldsymbol{f}_v \tag{27}$$

$$\boldsymbol{K}_{pv}\boldsymbol{v} - \boldsymbol{f}_p \preceq \boldsymbol{0} \tag{28}$$

$$\boldsymbol{p} \succeq \boldsymbol{0} \tag{29}$$

$$p_i(\boldsymbol{K}_{pv}\boldsymbol{v} - \boldsymbol{f}_p)_i = 0 \quad \forall i \tag{30}$$

The last condition (which is basically the complementary slackness condition in the KKT system of equations) implies that one *may* not have local mass balance in those elements for which the Lagrange multiplier vanishes (i.e., $p_i = 0$, where i denotes the element number). Note that in the RT0 formulation, the Lagrange multiplier p_i denotes the concentration in the i^{th} element.

Remark 2.4. From optimization theory [23] one can show that the primal (P2-RT0) and dual (D2-RT0) problems are equivalent. That is, there is no duality gap for the optimization-based RT0 formulation. The difference between primal and dual solutions is commonly referred to as the duality gap. In general, the solution of a dual problem gives an upper bound to its corresponding primal problem.

However, from a computational point of view the primal and dual problems can have different numerical performance (especially with respect to computational cost, and selection of numerical solvers). The primal problem (P2-RT0) has more complicated constraints than the dual problem (D2-RT0) for which the constraints are (lower) bounds on the design variable p. Compared to the primal problem (P2-RT0), the dual problem (D2-RT0) has a more complicated objective function, which is defined in terms of Schur complement operator. For all the numerical results presented in this paper, we have used the dual problem (D2-RT0). However, we have compared the numerical solutions obtained using the dual problem with the primal problem (P2-RT0), and the solutions are identical as predicted by the theory.

Special solvers are available in the literature (for example, especially designed interior point methods [25,24]) that are effective for solving problems that belong to the class of quadratic programming with constraints being just bounds on design variables. Similarly, special solvers are available for problems involving Schur complement operators. For example, the preconditioned conjugate gradient (PCG) solver is quite effective for solving large-scale problems involving Schur complement operator. A detailed analysis comparing computational costs of the primal (P2-RT0) and dual (Q2-RT0) problems is beyond the scope of this paper.

3. A non-negative variational multiscale mixed formulation

Masud and Hughes [26] have proposed a stabilized mixed formulation for the first-order form of the Poisson equation that satisfies the LBB condition. In this paper we refer to this formulation as the variational multiscale (VMS) formulation. Nakshatrala et al. [27] have shown that the variational multiscale formulation can be derived based on the multiscale framework proposed by Hughes [28]. The variational multiscale formulation possesses many favorable numerical properties and performs very well in practice. For example, the formulation passes three dimensional patch tests even for distorted elements [27]. Another feature of this formulation worth mentioning is that the equal-order interpolation for c and $\mathbf{v}$ is stable [28,27]. However, the variational multiscale formulation in general does not satisfy the discrete maximum–minimum principle, which will be illustrated below in Fig. 12 using a simple numerical example. In this section, we present a non-negative method based on the variational multiscale mixed formulation. To this end, we first present the weak form and variational structure behind the variational multiscale formulation.

Let

$$\widetilde{\mathcal{P}} \equiv H^1(\Omega) \tag{31}$$

$$\widetilde{\mathcal{Q}} \equiv H^1(\Omega) \tag{32}$$

where $H^1(\Omega)$ is a standard Sobolev space defined on domain Ω [22]. The variational multiscale formulation reads [27,26]: Find $c(\mathbf{x}) \in \widetilde{\mathcal{P}}$ and $\mathbf{v}(\mathbf{x}) \in \mathcal{V}$ such that

$$(\mathbf{w};\mathbb{D}^{-1}\mathbf{v})-(\nabla\cdot\mathbf{w};c)+(\mathbf{w}\cdot\mathbf{n};c^{\mathrm{p}})_{\Gamma^{\mathrm{D}}}-(q;\nabla\cdot\mathbf{v}-f)-\frac{1}{2}(\mathbb{D}^{-1}\mathbf{w}+\nabla q;\mathbb{D}(\mathbb{D}^{-1}\mathbf{v}+\nabla c))=0\quad\forall q(\mathbf{x})\in\widetilde{\mathcal{Q}},\ \mathbf{w}(\mathbf{x})\in\mathcal{W} \tag{33}$$

where $\mathbf{w}(\mathbf{x})$ and $q(\mathbf{x})$ are weighting functions corresponding to $\mathbf{v}(\mathbf{x})$ and $c(\mathbf{x})$, and $\mathcal{V}$ and $\mathcal{W}$ are defined in Eqs. (9) and (10), respectively. The stationarity (minimizing with respect to $\mathbf{v}$ and maximizing with respect to c) of the following (continuous) optimization problem

$$\underset{\mathbf{v}\in\mathcal{V},\ c\in\widetilde{\mathcal{P}}}{\text{extremize}}\ \frac{1}{2}(\mathbf{v};\mathbb{D}^{-1}\mathbf{v})-(c;\nabla\cdot\mathbf{v}-f)+(\mathbf{v}\cdot\mathbf{n};c^{\mathrm{p}})_{\Gamma^{\mathrm{D}}}-\frac{1}{4}(\mathbb{D}^{-1}\mathbf{v}+\nabla c;\mathbb{D}(\mathbb{D}^{-1}\mathbf{v}+\nabla c)) \tag{34}$$

is equivalent to the weak form given by Eq. (33).

Many practically important problems do not have solutions in $C^2(\Omega)\cap C^0(\overline{\Omega})$, and hence for these problem one cannot employ the classical maximum–minimum principle, which we have outlined in Section 1. For example, there exist no (classical) solutions to test problems #1 and #2 (which are defined in Section 4) that belong to $C^2(\Omega)$ as the forcing functions in both these cases are *not* continuous on Ω. (On the other hand, the solution to test problem #3 does belong to $C^2(\Omega)\cap C^0(\overline{\Omega})$.) However, weak solutions do exist for test problems #1 and #2. Using L^p regularity theory (for example, see Ref. [29]), one can show that these solutions, in fact, belong to $H^2(\Omega)\cap C^1(\Omega)\cap C^0(\overline{\Omega})$.

Remark 3.1. Note that $H^2(\Omega)$ is not a subset of $C^1(\Omega)$ or vice-versa. On the other hand, $H^2(\Omega)\subset C^0(\Omega)$.

3.1. Continuous maximum–minimum principle

In this subsection we demonstrate one of the main contributions of this paper, namely that the weak solution under the variational multiscale formulation (under appropriate regularity assumptions) satisfies a continuous maximum–minimum principle. To the authors' knowledge, this property of the VMS formulation has *not* been discussed/proved in the literature. We employ the standard notation used in mathematical analysis, for example see Ref. [6]. The standard abbreviation 'a.e.' for *almost everywhere* is often used in this subsection. We now state and prove a continuous maximum–minimum principle for the VMS formulation.

Theorem 3.2. *Assume that the Dirichlet boundary condition is prescribed on the whole of the boundary (that is, $\Gamma^{\mathrm{D}}=\partial\Omega$), and the diffusivity tensor is assumed to be continuously differentiable. Let $f(\mathbf{x})\in L^2(\Omega)$, and $f(\mathbf{x})\geqslant 0$ almost everywhere. Let the weak solution $c(\mathbf{x})$ of the variational multiscale formulation (33) belong to $H^2(\Omega)\cap C^1(\Omega)\cap C^0(\overline{\Omega})$. Then*

$$\min_{\overline{\Omega}}c(\mathbf{x})=\min_{\partial\Omega}c(\mathbf{x})$$

Proof 1. Since $c(\mathbf{x})\in H^2(\Omega)\cap C^1(\Omega)\cap C^0(\overline{\Omega})$ we have

$$\mathbf{v}:=-\mathbb{D}\nabla c\in H^1(\Omega)\cap C^0(\Omega)\subset\mathcal{V} \tag{35}$$

Define $m\in\mathbb{R}$ and a scalar field $s(\mathbf{x})$ such that

$$m=\min_{\mathbf{x}\in\Gamma^{\mathrm{D}}}c^{\mathrm{p}}(\mathbf{x}) \tag{36}$$

$$s(\mathbf{x}):=\max[m-c(\mathbf{x}),0]\quad\forall\mathbf{x}\in\overline{\Omega} \tag{37}$$

One can show that the function $s(\mathbf{x})$ is piecewise $C^1(\Omega)$, and belongs to $H^1(\Omega)\cap C^0(\overline{\Omega})$. By construction, we also have

$$s(\mathbf{x})\geqslant 0\quad\forall\mathbf{x}\in\overline{\Omega},\ \text{and}\ s(\mathbf{x})=0\quad\forall\mathbf{x}\in\Gamma^{\mathrm{D}} \tag{38}$$

Using Eq. (35) (and also employing the divergence theorem) Eq. (33) gets simplified to

$$(q;\nabla\cdot\mathbf{v}-f)=0 \tag{39}$$

Since $s(\mathbf{x})\in H^1(\Omega)$ and $s(\mathbf{x})=0$ on Γ^{D}, the scalar field $s(\mathbf{x})$ is a legitimate choice for $q(\mathbf{x})$. Substituting $s(\mathbf{x})$ in the place of $q(\mathbf{x})$, and noting that $s(\mathbf{x})\in H^1(\Omega)$ to allow the application of the divergence theorem; we get

$$(s;\mathbf{v}\cdot\mathbf{n})_{\partial\Omega}-(\nabla s;\mathbf{v})-(s;f)=0 \tag{40}$$

Since $\Gamma^{\mathrm{D}}=\partial\Omega,\ s(\mathbf{x})=0$ on $\Gamma^{\mathrm{D}}, s(\mathbf{x})\geqslant 0\ \forall\mathbf{x}\in\overline{\Omega}$, and $f(\mathbf{x})\geqslant 0$ a.e. in Ω; we conclude that

$$(\nabla s;\mathbf{v})=-(\nabla s;\mathbb{D}\nabla c)\leqslant 0 \tag{41}$$

To prove the theorem it is sufficient to show that $s(\mathbf{x})=0\ \forall\mathbf{x}\in\Omega$ (which implies that $c(\mathbf{x})\geqslant m\ \forall\mathbf{x}\in\overline{\Omega}$). Note that $s(\mathbf{x})=m-c(\mathbf{x})$ unless $s(\mathbf{x})=0$. Let

$$\mathbb{Y}:=\{\mathbf{x}\in\Omega\mid s(\mathbf{x})\neq 0\}\equiv\{\mathbf{x}\in\Omega\mid s(\mathbf{x})>0\} \tag{42}$$

The case $\mathbb{Y} = \emptyset$ is trivial. We now deal with the case when $\mathbb{Y}$ is not empty. We first note that the weak derivative of $s(\mathbf{x})$ is zero on $\Omega \setminus \mathbb{Y}$, and is $-\nabla c$ on $\mathbb{Y}$. This result along with Eq. (41) implies that

$$(\nabla s; \mathbb{D}\nabla s)_{\mathbb{Y}} \leqslant 0 \tag{43}$$

Since $\mathbb{D}(\mathbf{x})$ is a positive definite tensor, the above equation implies that $s(\mathbf{x}) = s_0 =$ constant *almost everywhere* in $\mathbb{Y}$. Since $s(\mathbf{x})$ is continuous in $\overline{\Omega}$, we conclude that $s(\mathbf{x}) = s_0$ *everywhere* in $\mathbb{Y}$. Since $\Gamma^{\mathrm{D}} = \partial\Omega \subseteq \overline{\mathbb{Y}}$, and $s(\mathbf{x}) = 0$ on Γ^{D} we conclude that $s(\mathbf{x}) = 0$ on whole of $\mathbb{Y}$ and also on $\overline{\mathbb{Y}}$. Noting the fact that the function $s(\mathbf{x})$ vanishes on the set complement of $\mathbb{Y}$, we conclude that $s(\mathbf{x}) = 0$ on whole of $\overline{\Omega}$. Hence, we have proved the desired result. □

3.2. Discrete equations

The discretized finite element equations for the variational multiscale formulation (given by Eq. (33)) can be written as

$$\begin{bmatrix} \boldsymbol{K}_{vv} & \boldsymbol{K}_{pv}^T \\ \boldsymbol{K}_{pv} & -\boldsymbol{K}_{pp} \end{bmatrix} \begin{Bmatrix} \boldsymbol{v} \\ \boldsymbol{p} \end{Bmatrix} = \begin{Bmatrix} \boldsymbol{f}_v \\ \boldsymbol{f}_p \end{Bmatrix} \tag{44}$$

where the matrices $\boldsymbol{K}_{vv}$ and $\boldsymbol{K}_{pp}$ are symmetric and positive definite, $\boldsymbol{v}$ denotes the (finite element) vector of nodal velocity (or auxiliary variable) degrees-of-freedom, and $\boldsymbol{p}$ denotes nodal vector of concentration degrees-of-freedom. Comparing the weak form (33) and discrete Eq. (44), the matrices $\boldsymbol{K}_{vv}, \boldsymbol{K}_{pv}$ and $\boldsymbol{K}_{pp}$ are obtained after the finite element discretization of the terms $\frac{1}{2}(\mathbf{w}; \mathbb{D}^{-1}\mathbf{v}), -(q, \nabla\cdot\mathbf{v}) - \frac{1}{2}(\nabla q; \mathbf{v})$ and $-\frac{1}{2}(\nabla q; \mathbb{D}\nabla c)$; respectively. Since Eq. (33) is written in symmetric form, $\boldsymbol{K}_{vp}$ (which comes from the term $-(\nabla\cdot\mathbf{w}; c) - \frac{1}{2}(\mathbf{w}; \nabla c)$) will be equal to $\boldsymbol{K}_{pv}^T$. Note that, some of the terms in Eq. (33) are combined and simplified to obtain the terms presented in the previous line. For example, we have used the symmetry of $\mathbb{D}$ in obtaining the term $-\frac{1}{2}(\mathbf{w}; \nabla c)$. The vectors $\boldsymbol{f}_v$ and $\boldsymbol{f}_p$ are, respectively, obtained from $-(\mathbf{w}\cdot\mathbf{n}; c^{\mathrm{p}})_{\Gamma^{\mathrm{D}}}$ and $-(q; f)$ after the finite element discretization.

The discrete form of Eq. (34) can be written as

$$\underset{\boldsymbol{v},\,\boldsymbol{p}}{\text{extremize}}\ \frac{1}{2}\boldsymbol{v}^T\boldsymbol{K}_{vv}\boldsymbol{v} + \boldsymbol{p}^T\boldsymbol{K}_{pv}\boldsymbol{v} - \frac{1}{2}\boldsymbol{p}^T\boldsymbol{K}_{pp}\boldsymbol{p} - \boldsymbol{v}^T\boldsymbol{f}_v - \boldsymbol{p}^T\boldsymbol{f}_p \tag{45}$$

Similar to the continuous problem (that is, Eqs. (33) and (34) are equivalent), the stationarity of the above equation (minimizing with respect to $\boldsymbol{v}$ and maximizing with respect to $\boldsymbol{p}$) is equivalent to Eq. (44). By eliminating $\boldsymbol{v}$, the Schur complement form of Eq. (44) can be written as

$$\left(\boldsymbol{K}_{pv}\boldsymbol{K}_{vv}^{-1}\boldsymbol{K}_{pv}^T + \boldsymbol{K}_{pp}\right)\boldsymbol{p} = \boldsymbol{K}_{pv}\boldsymbol{K}_{vv}^{-1}\boldsymbol{f}_v - \boldsymbol{f}_p \tag{46}$$

Clearly, the Schur complement operator $\boldsymbol{K}_{pv}\boldsymbol{K}_{vv}^{-1}\boldsymbol{K}_{pv}^T + \boldsymbol{K}_{pp}$ is symmetric and positive definite. The discrete variational statement of the variational multiscale mixed formulation can be posed solely in terms of the variable $\boldsymbol{p}$, and takes the following form:

$$\underset{\boldsymbol{p}}{\text{minimize}}\quad \frac{1}{2}\boldsymbol{p}^T\left(\boldsymbol{K}_{pv}\boldsymbol{K}_{vv}^{-1}\boldsymbol{K}_{pv}^T + \boldsymbol{K}_{pp}\right)\boldsymbol{p} - \boldsymbol{p}^T\left(\boldsymbol{K}_{pv}\boldsymbol{K}_{vv}^{-1}\boldsymbol{f}_v - \boldsymbol{f}_p\right) \tag{47}$$

As mentioned earlier the variational multiscale mixed formulation does not (always) produce non-negative solutions for the non-negative forcing function and non-negative prescribed Dirichlet boundary condition.

Remark 3.3. Unlike in the Raviart–Thomas formulation, the optimization problem (47) is not the dual problem of Eq. (45).

3.3. A non-negative formulation

A non-negative formulation based on the variational multiscale formulation can be posed as the following constrained minimization problem

$$\underset{\boldsymbol{p}}{\text{minimize}}\quad \frac{1}{2}\boldsymbol{p}^T\left(\boldsymbol{K}_{pv}\boldsymbol{K}_{vv}^{-1}\boldsymbol{K}_{pv}^T + \boldsymbol{K}_{pp}\right)\boldsymbol{p} - \boldsymbol{p}^T\left(\boldsymbol{K}_{pv}\boldsymbol{K}_{vv}^{-1}\boldsymbol{f}_v - \boldsymbol{f}_p\right) \tag{48}$$

$$\text{subject to}\quad \boldsymbol{p} \succeq \boldsymbol{0} \tag{49}$$

The above constrained optimization problem belongs to the class of convex quadratic programming, and has a unique global minimizer.

Remark 3.4. The variational multiscale formulation (given by Eq. (33)), in general, does not have the (element) local mass balance property. Specifically, one does not have the local mass balance property under linear equal-order interpolation for both $c(\mathbf{x})$ and $\mathbf{v}(\mathbf{x})$, which is employed in this paper. The corresponding non-negative formulation also does not possess the local mass balance property.

Remark 3.5. The non-negative method proposed in this section is also applicable for the mixed formulation based on the Galerkin/least-squares. As discussed in Ref. [27] the variational multiscale and Galerkin/least-squares (GLS) mixed formulations differ only in the definition of the stabilization parameter. That is, instead of the term

$$\frac{1}{2}(\mathbb{D}^{-1}\mathbf{w} + \nabla q; \mathbb{D}(\mathbb{D}^{-1}\mathbf{v} + \nabla c))$$

which is the case for the variational multiscale formulation (see Eq. (33)) we will have

$$(\mathbb{D}^{-1}\mathbf{w} + \nabla q; \tau(\mathbf{x})\mathbb{D}(\mathbb{D}^{-1}\mathbf{v} + \nabla c))$$

and $\tau(\mathbf{x}) \geqslant 0$ for the GLS mixed formulation. The discrete equations from the GLS formulation also takes the same form as given in Eq. (44).

Remark 3.6. As mentioned earlier, non-negative solution is a special case of maximum–minimum principle. Some of the formulations presented in the literature produce non-negative solutions but still may violate the (general) discrete maximum–minimum principle. For example, see the non-negative formulation presented in Ref. [30]. That is, these formulations avoid undershoots but may still produce overshoots.

Though the focus of the present paper is on non-negative solutions, the proposed two non-negative optimization-based formulations can be easily extended to satisfy the (general) discrete maximum–minimum principle. To see this, let us first define the quantities $c_{\min}$ and $c_{\max}$ to be the minimum and maximum values of $\mathbf{c}(\mathbf{x})$ based on the (continuous) maximum–minimum principle. Note that the maximum and minimum will occur on the boundary only when $f(\mathbf{x}) = 0$. In order to enforce these properties in the discrete setting modify the constraints in the corresponding optimization problem statements (i.e., Eq. $(25)_2$ and (49)) as

$$c_{\min}\boldsymbol{1} \preceq \boldsymbol{p} \preceq c_{\max}\boldsymbol{1} \tag{50}$$

where $\boldsymbol{1}$ is a vector of ones of appropriate size. The resulting problems will still belong to quadratic programming. Hence, the proposed mathematical framework and solvers are still applicable. However, some interesting questions regarding the numerical performance of the solvers (active-set strategy, interior point methods) need to be addressed in future work. For example, since in the case of general DMP we have twice the number of constraints than that in the case of the non-negative formulation, how large is the violation of the local mass balance in the RT0 formulation because of the additional constraints? How much additional computational cost will be incurred because of the additional constraints.

4. Numerical results

In this section we study the performance of the proposed formulations (with respect to non-negative solutions and local mass balance) on three canonical test problems. In our numerical experiments we have employed five different meshes – Delaunay, 45-degree, unstructured and well-centered triangular (WCT) meshes; and uniform four-node quadrilateral mesh. In two-dimensions, a well-centered triangulation means that all the triangles in the mesh are acute-angled (see Ref. [31]). We will discuss more on WCT meshes in Section 4.4.

The mesh layouts for the aforementioned meshes are shown in Figs. 2–5. For the chosen test problems, the variational multiscale and Raviart–Thomas formulations in general do not satisfy the discrete maximum–minimum principle. We now show that, for low-order finite elements, the proposed two non-negative mixed formulations produce non-negative solutions for all the three test problems and for all the chosen computational meshes. For the variational multiscale formulation, we have employed equal order interpolation for the c and $\mathbf{v}$ fields in our numerical simulations. Note that (as discussed in Introduction) WCT, 45-degree, Delaunay and square meshes are sufficient to produce non-negative solutions

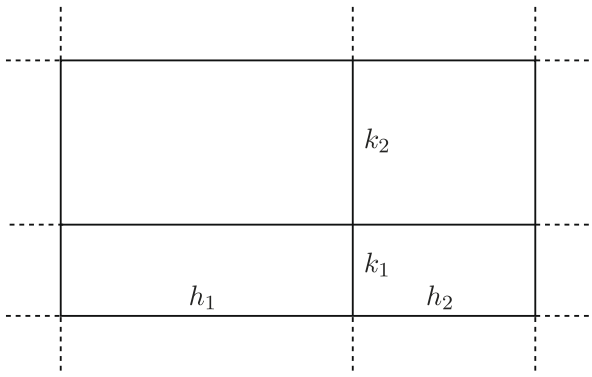


Fig. 1. Non-uniform rectangular mesh where h_1 and h_2 denote the length of horizontal edges of two neighboring elements. Similarly, k_1 and k_2 are for corresponding vertical sides.

for isotropic diffusion. However, these meshes may produce negative solutions in the case of anisotropic diffusion, which will be illustrated in this section.

4.1. Test problem #1: Anisotropic and heterogeneous medium

This test problem is taken from Ref. [16]. The computational domain is a bi-unit square with homogeneous Dirichlet boundary conditions. The forcing function is taken as

$$f = \begin{cases} 1 & \text{if } (x,y) \in [3/8, 5/8]^2 \\ 0 & \text{otherwise} \end{cases} \tag{51}$$

The diffusivity tensor is given by

$$\mathbb{D} = \begin{pmatrix} y^2 + \epsilon x^2 & -(1-\epsilon)xy \\ -(1-\epsilon)xy & \epsilon y^2 + x^2 \end{pmatrix} \tag{52}$$

In this paper we have taken the parameter $\epsilon = 0.05$. For this test problem, the numerical results for the concentration field $c(x,y)$ for various meshes using the variational multiscale and corresponding optimization-based formulations are shown in Fig. 6. The contours of the vector field $\boldsymbol{v}$ are shown in Fig. 7. The numerical results for the concentration using the RT0 and corresponding optimization-based formulation are presented in Fig. 8.

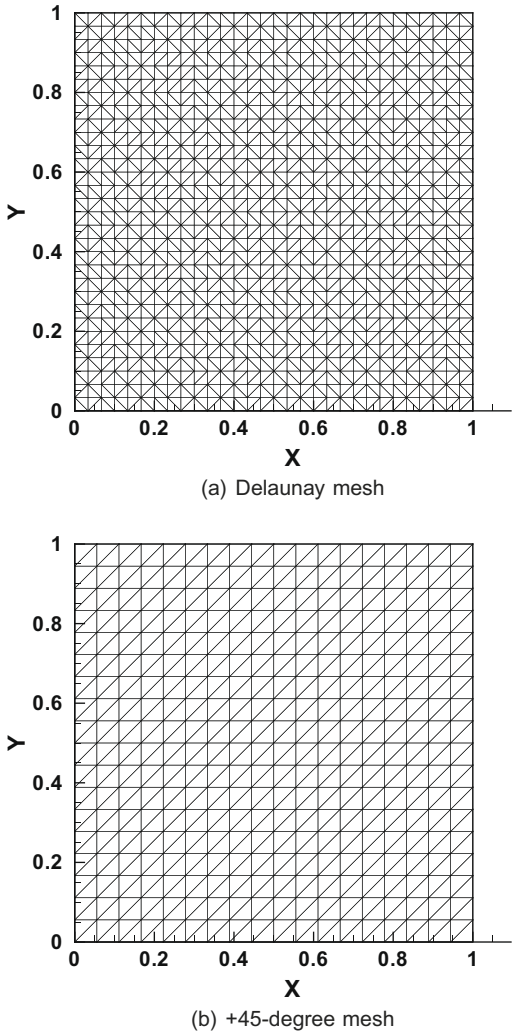


Fig. 2. Typical triangular meshes [Delaunay (top) and +45-degree (bottom)] that are used for test problems 1–3 are shown in the figure. In addition, −45-degree mesh is also used in numerical simulations, which is similar to the +45-degree except that the diagonals run along south-east to north-west direction.

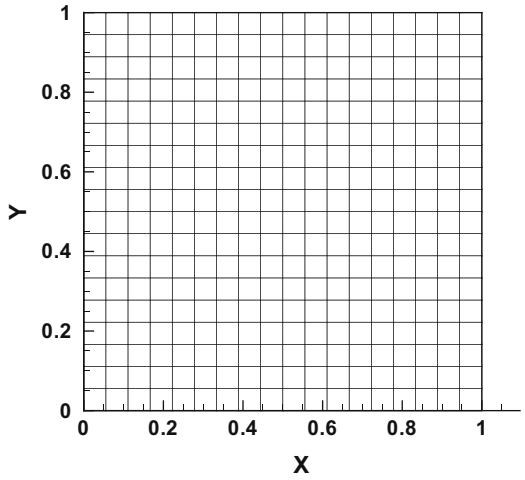


Fig. 3. Typical four-node quadrilateral mesh using in numerical simulations.

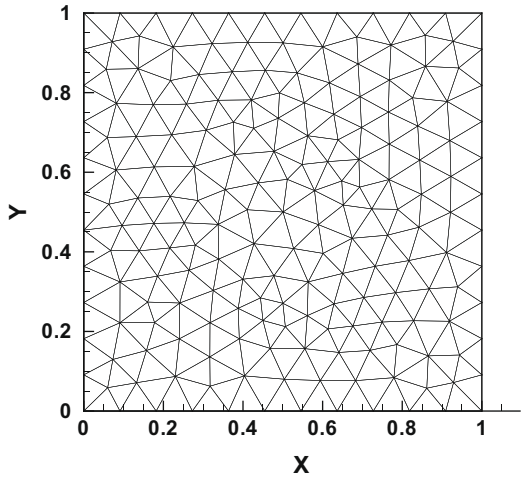


Fig. 4. Well-centered triangular (WCT) mesh.

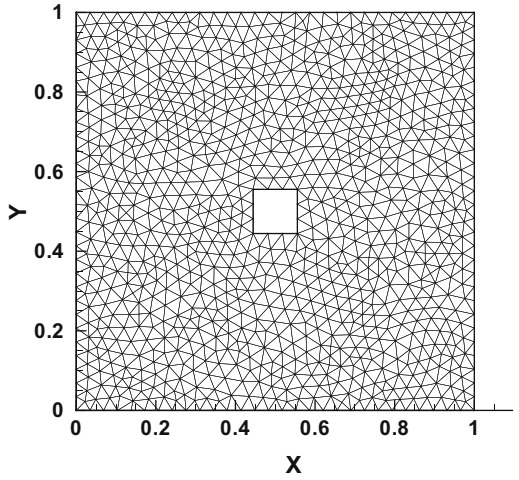
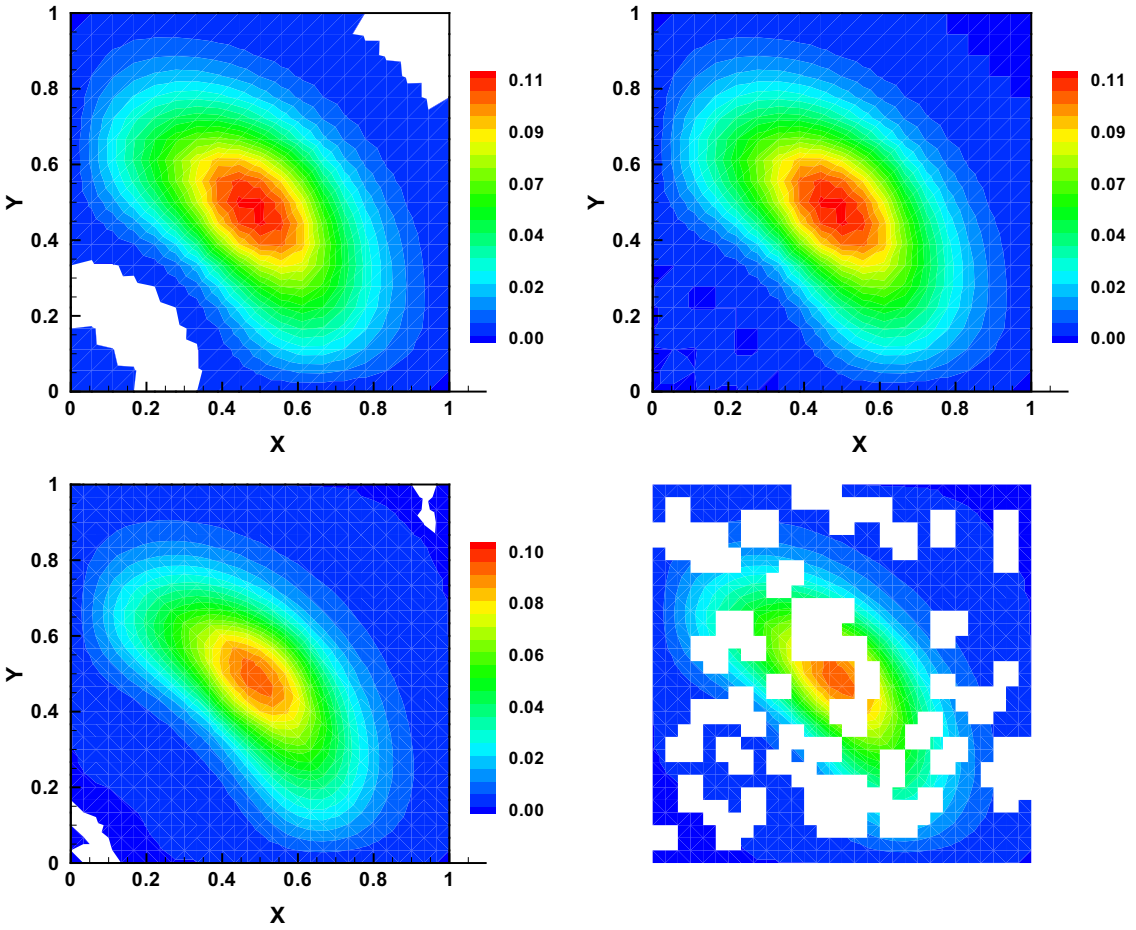


Fig. 5. Pictorial description of test problem #3: Computational domain is the bi-unit square with a square hole $[4/9, 5/9] \times [4/9, 5/9]$. On the exterior boundary $c^{\mathrm{p}}(\boldsymbol{x}) = 0$ is prescribed. On the interior boundary $c^{\mathrm{p}}(\boldsymbol{x}) = 2$ is prescribed. The computational domain is triangulated using Gmsh [42].

4.2. Test problem #2: Diffusion/dispersion tensor in subsurface flows

This problem is taken from the groundwater modeling literature (for example, see Ref. [1]). The diffusivity tensor is given by

$$\mathbb{D} = a_T \|\boldsymbol{\beta}\| \mathbb{I} + \frac{a_L - a_T}{\|\boldsymbol{\beta}\|} \boldsymbol{\beta} \otimes \boldsymbol{\beta} \tag{53}$$



where $\mathbb{I}$ denotes the second-order identity tensor, $\otimes$ the standard tensor product [32], $\boldsymbol{\beta}$ the velocity vector, and α_L and α_T are respectively the longitudinal and transverse diffusivity constants. Note that β is a eigenvector of the diffusivity tensor given in Eq. (53). In this paper we have taken $\alpha_L = 0.1$ and $\alpha_T = 0.01$, and the velocity vector to be

$$\boldsymbol{\beta} = \mathbf{e}_x + \mathbf{e}_y \tag{54}$$

where $\mathbf{e}_x$ and $\mathbf{e}_y$ are the standard unit vectors along x- and y-directions, respectively. The computational domain is again a bi-unit square with homogeneous Dirichlet boundary conditions. The forcing function is same as in test problem #1 (see Eq. (51)).

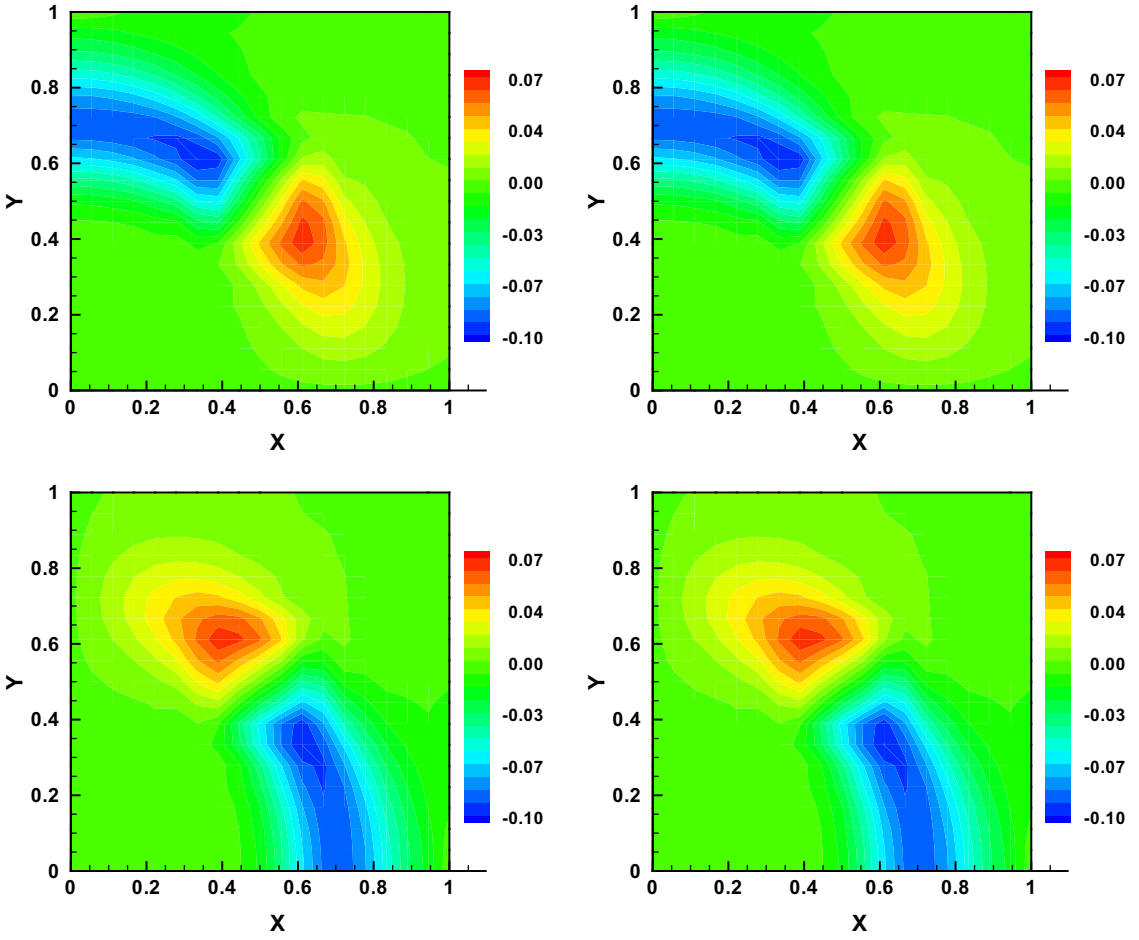


Fig. 7. Test problem #1: Contours of the components of the vector field **v** obtained using the variational multiscale (left) and corresponding optimization-based (right) formulations. The top figures show the x-component (that is, v_x), and the bottom figures the y-component of **v**. Four-node quadrilateral mesh is used in the numerical simulation.

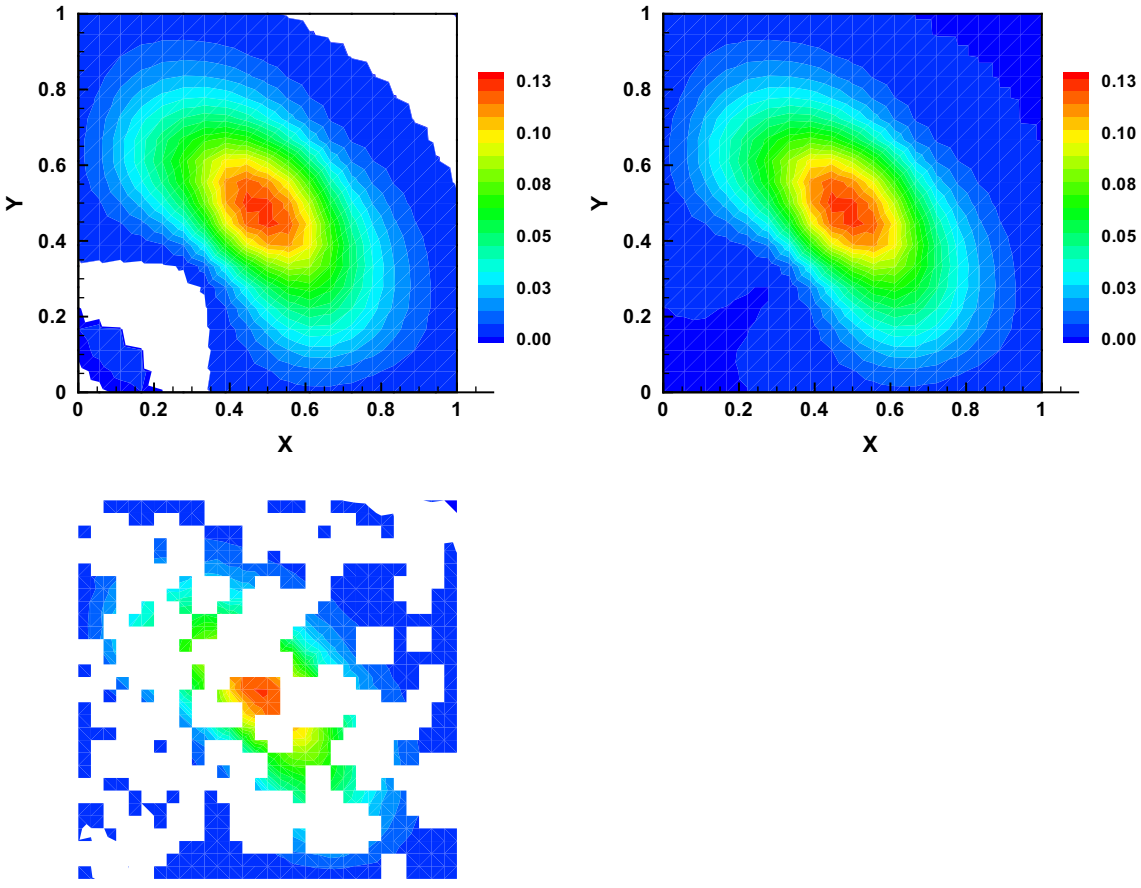
Table 1
Performance of the RT0 formulation: minimum concentration produced by the formulation, and percentage of elements that have negative concentrations (denoted as % of elements violated).

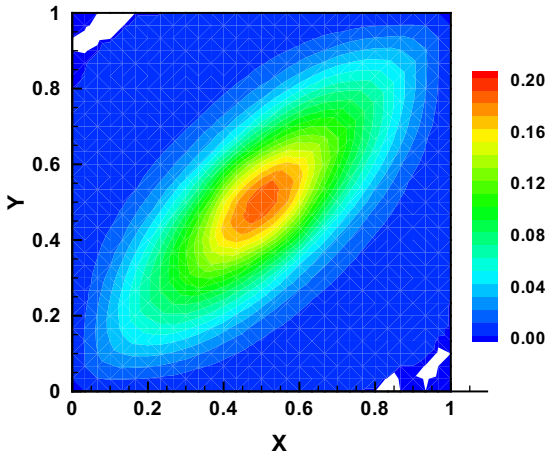
Test problem	Mesh type	Min. conc.	% of elements violated
Problem #1	+45-degree	−0.002510583	128/648 → 19.75%
	Delaunay	−0.000011158	67/1800 → 3.72%
	Well-centered	−0.000824908	45/336 → 13.39%
Problem #2	+45-degree	0.000000000	0/648 → 0.00%
	−45-degree	−0.006674991	216/648 → 33.33%
	Delaunay	−0.000468342	71/1800 → 3.94%
	Well-centered	−0.000616864	42/336 → 12.50%
Problem #3	Mesh in Fig. 5	−0.081453689	848/1868 → 45.40%

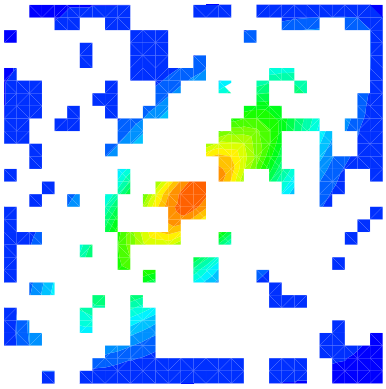
For the chosen velocity field (54) (which is aligned along south-west to north-east direction) by rotating the current coordinate system by +45 degrees (i.e., in the anticlockwise direction) the diffusivity tensor written in the transformed coordinate system will be isotropic. Therefore, a mesh aligned along +45-degree mesh should produce non-negative solutions for the chosen velocity field (which is illustrated in Tables 1 and 2). However, one will get negative solutions using a −45-degree mesh. For this test problem, the numerical results for the variational multiscale and RT0 formulations and their corresponding optimization-based methods are presented in Figs. 9 and 10, respectively.

Table 2
Performance of the variational multiscale formulation: minimum concentration produced by the formulation, and percentage of nodes that have negative concentrations (denoted as % of nodes violated).

Test problem	Mesh type	Min. conc.	% of nodes violated
Problem #1	+45-degree	−0.000986744	30/361 → 8.31%
	Delaunay	−0.000000531	11/961 → 1.14%
	Well-centered	−0.000056615	4/191 → 2.09%
	Four-node quadrilateral	−0.000000901	7/361 → 1.93%
Problem #2	+45-degree	0.000000000	0/361 → 0.00%
	−45-degree	−0.000402642	40/361 → 11.08%
	Delaunay	−0.000000941	12/961 → 1.25%
	Well-centered	−0.000007018	5/191 → 2.62%
	Four-node quadrilateral	−0.000000155	6/361 → 1.67%
Problem #3	Mesh in Fig. 5	−0.004613415	264/998 → 26.45%

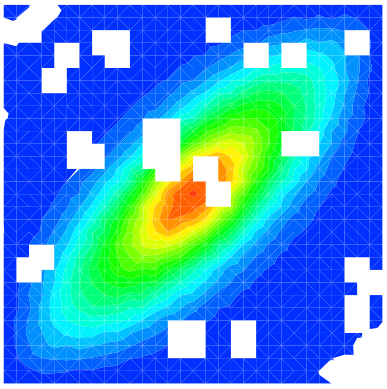






4.3. Test problem #3: Non-smooth anisotropic solution

This problem is taken from Ref. [30]. The computational domain is a bi-unit square with a square hole of dimension $[4/9, 5/9] \times [4/9, 5/9]$, which is pictorially described in Fig. 5. The forcing function is taken as $f(\boldsymbol{x}) = 0$. On the exterior boundary $c^{\mathrm{p}}(\boldsymbol{x}) = 0$ is prescribed, and on the interior boundary $c^{\mathrm{p}}(\boldsymbol{x}) = 2$ is prescribed. The diffusivity tensor is given by



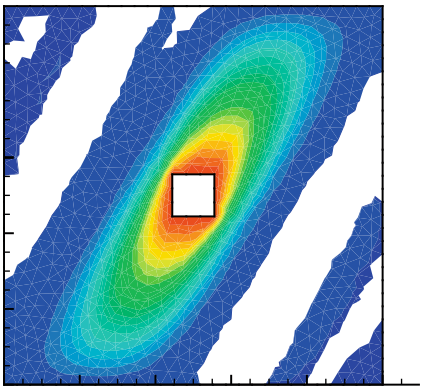
$$\mathbb{D} = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} k_1 & 0 \\ 0 & k_2 \end{pmatrix} \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \tag{55}$$

In this paper we have taken $k_1 = 1, k_2 = 100$ and $\theta = \pi/6$, which are same as the values employed in Ref. [30]. For this test problem, the performance of the variational multiscale and RT0 formulations and their corresponding optimization-based formulations are shown in Fig. 11. Also it is worth mentioning that (for this test problem) under the RT0 formulation more than 45% of the computational domain has negative concentration, which is illustrated in Table 1.

4.4. Performance on a well-centered triangular mesh

In two dimensions, a well-centered triangulation means that each element contains its circumcenter, which is equivalent to saying that all elements are acute-angled triangles. A well-centered mesh in higher dimensions can be similarly defined [31]. Note that every WCT mesh is also a Delaunay mesh but not vice-versa. For further details on how to generate WCT meshes see Ref. [31].

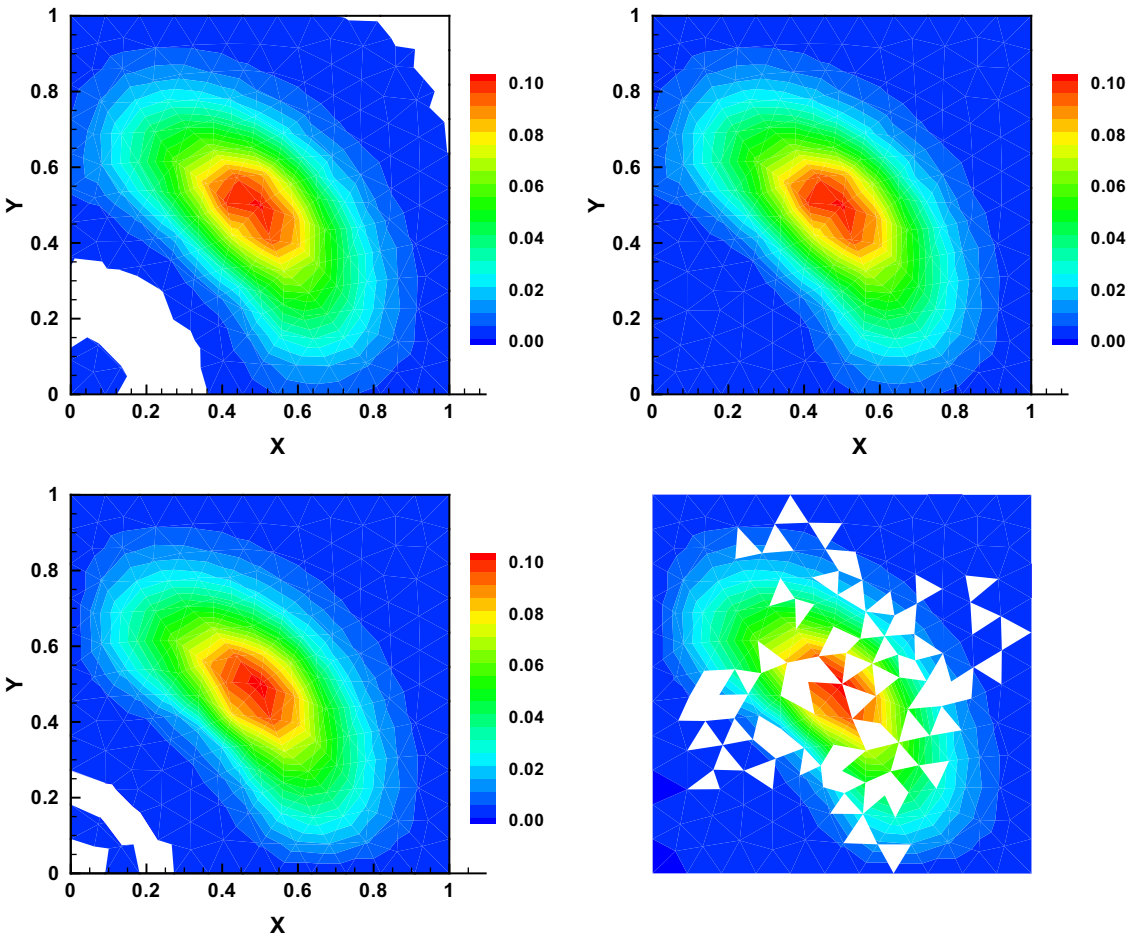
It is well known that well-centered triangular (WCT) meshes have some advantages in solving some partial differential equations as they preserve some of the underlying mathematical structure. For example, as discussed in Introduction, a WCT mesh is sufficient to produce non-negative solutions for an *isotropic* diffusion equation. In other words, a WCT mesh respects the discrete maximum–minimum principle thereby preserving this key underlying mathematical property. In addition, a WCT mesh enables construction of a compatible discretization of a Hodge star (a geometrical object in exterior calculus) [33]. In Ref. [34] this idea has been used to construct a numerical method for the mixed form of the diffusion equation that is locally and globally conservative, and also can exactly represent linear variation of concentration in a given computational domain.



However, in this subsection we numerically show that for the RT0 and variational multiscale formulations, even a well-centered triangular (WCT) mesh is not sufficient to produce non-negative solutions in the case of full diffusivity tensor. We consider test problem #1, and use the well-centered triangular mesh shown in Fig. 4. The obtained numerical results for the concentration using the Raviart–Thomas and variational multiscale formulations are shown in Fig. 12. As one can see, there are regions of negative concentration (which are indicated in white color). The obtained numerical results using the corresponding optimization-based formulations are also shown in the figure, and (as expected) we have non-negative concentration in the whole domain.

4.5. Active set strategy and its numerical performance

The two main classes of methods for solving quadratic programming problems are active set strategy and interior point methods. In this paper we employ the active set strategy, which is very effective for small to medium sized convex quadratic programming. For a detailed discussion on active set strategy (including a convergence proof) see Luenberger and Ye [35, Section 12.4], and for an algorithmic outline of the numerical method see Nocedal and Wright [24, page 462].



In Tables 3 and 4 we have studied the performance of active set strategy for the proposed non-negative formulations. We considered two cases for the initial active set. The first case is trivial, which is the empty set. In the second case, we have take the initial active set as the degrees-of-freedom that have negative concentration under the chosen formulation (either VMS or RT0). That is, initially one will solve a given problem using either the VMS or RT0 formulation, and then identify the nodes

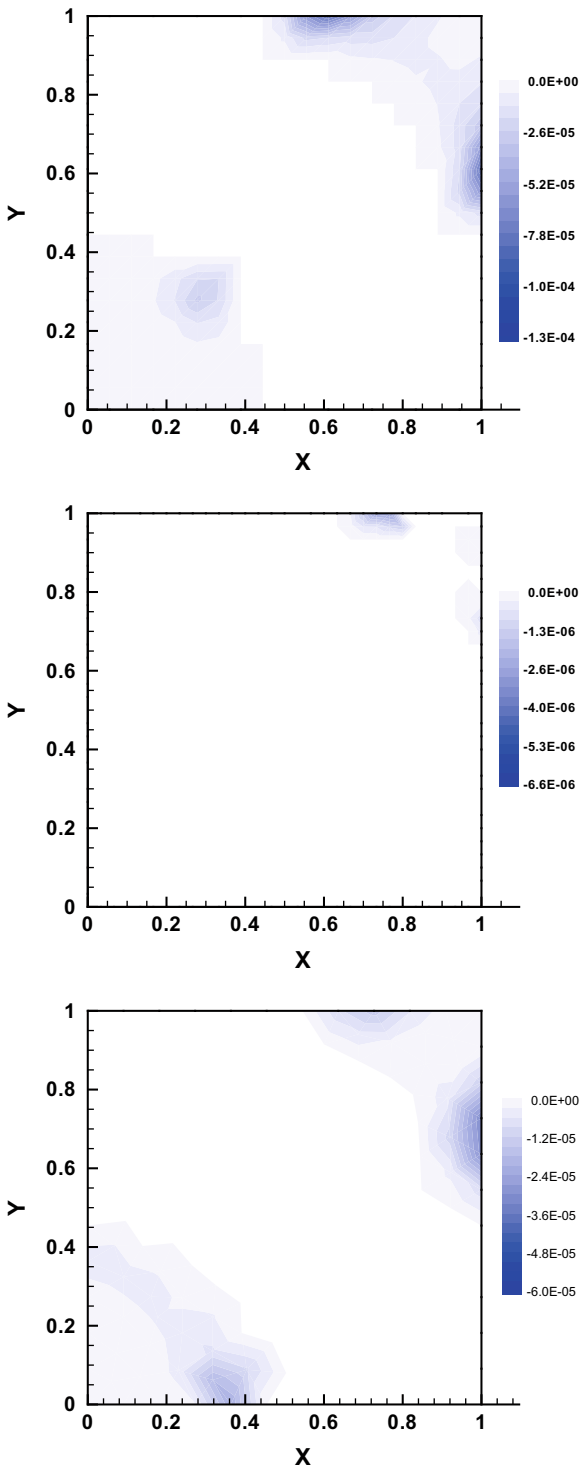


Fig. 13. Test problem #1: Contours of mass balance error under the optimization-based RT0 formulation. In the numerical simulations +45-degree (top), Delaunay (middle) and WCT (bottom) meshes are employed.

(in the case of VMS) or elements (in the case of RT0) that have negative concentration. The initial active set is taken as those degrees-of-freedom for concentration that have negative values. Based on numerical experiments we found that in many problems the second case takes fewer iterations. However, this is not the case always, which is illustrated in Tables 3 and 4. Note that in these tables, the two choices for initial active set are denoted as 'empty set' and 'initial violated set.'

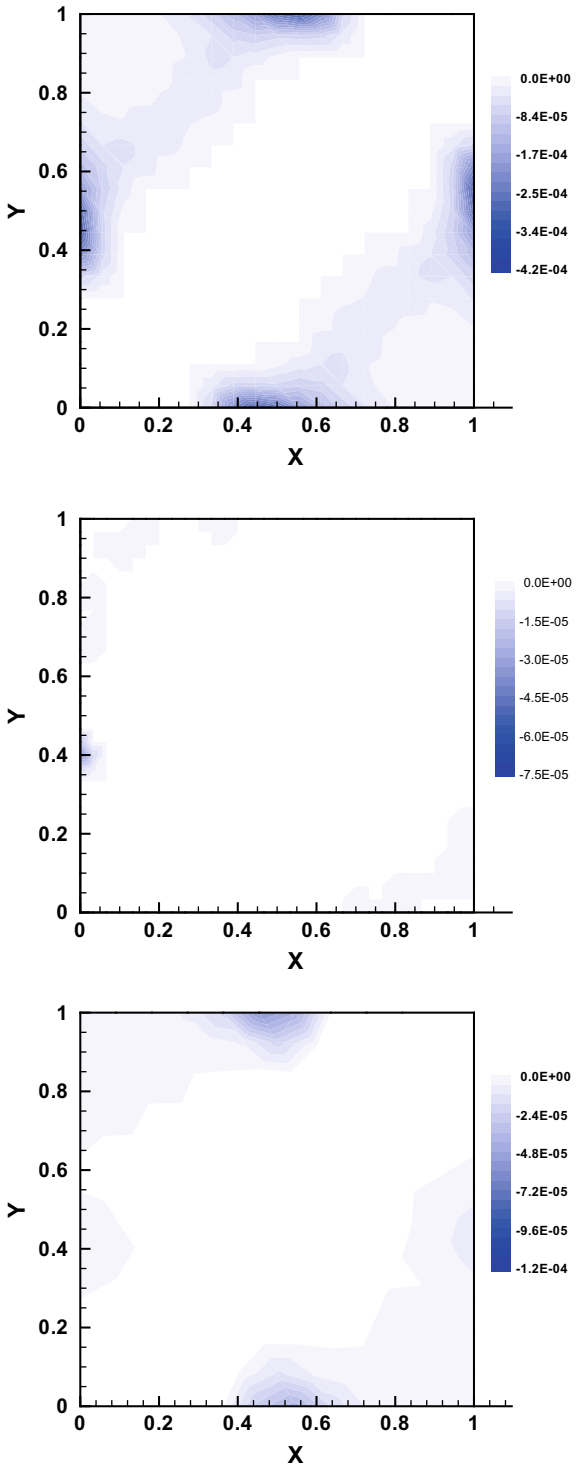


Fig. 14. Test problem #2: Contours of mass balance error under the optimization-based RT0 formulation. In the numerical simulations −45-degree (top), Delaunay (middle) and WCT (bottom) meshes are employed.

4.6. Error in local mass balance under the optimization-based RT0 formulation

In Section 2.2 we have shown mathematically that one may have violation of local mass balance under the optimization-based RT0 formulation. Based on the KKT optimal conditions we have also shown that the violation of local mass balance can occur only in the form of *artificial sinks* in some elements. These elements are those for which $(\boldsymbol{K}_{pv}\boldsymbol{v} - \boldsymbol{f}_p)_i < 0$, where i denotes the element number.

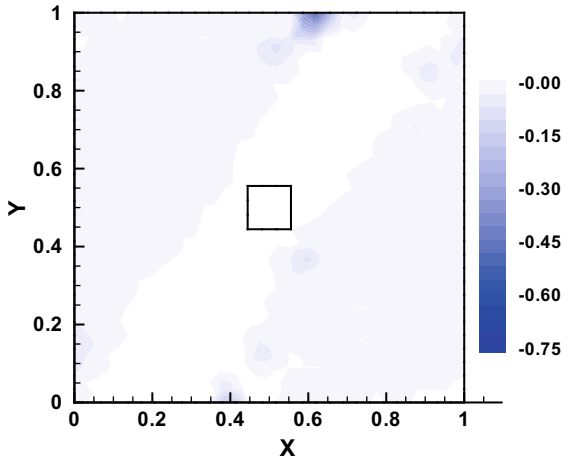


Fig. 15. Test problem #3: Contours of mass balance error under the optimization-based RT0 formulation.

Table 5
Performance of the variational multiscale formulation using three-node triangular element with respect to mesh refinement. We have employed +45-degree and −45-degree meshes for test problems #1 and #2, respectively.

Test problem	# of nodes per side	Min. Conc.	% nodes violated
Problem #1	10	−3.19E-003	9/100 → 9%
	19	−9.87E-004	30/361 → 8.31%
	28	−1.01E-004	54/784 → 6.89%
	37	−8.41E-006	64/1369 → 4.67%
	46	−1.05E-006	71/2116 → 3.36%
	55	−2.05E-007	71/3025 → 2.35%
	64	−7.38E-008	75/4096 → 1.83%
	73	−3.46E-008	77/5329 → 1.44%
Problem #2	10	−1.23E-003	6/100 → 6%
	19	−4.03E-004	40/361 → 11.08%
	28	−4.72E-005	66/784 → 8.42%
	37	−5.41E-006	66/1369 → 4.82%
	46	−9.57E-007	66/2116 → 3.12%
	55	−2.40E-007	66/3025 → 2.18%
	64	−8.50E-008	66/4096 → 1.61%
	73	−3.45E-008	66/5329 → 1.24%

Table 6
Performance of the variational multiscale formulation using four-node quadrilateral element with respect to mesh refinement.

Test problem	# of nodes per side	Min. conc.	% of nodes violated
Problem #1	11	−1.15E-004	7/121 → 5.79%
	21	−2.02E-007	8/441 → 1.81%
	31	−7.17E-009	8/961 → 0.83%
	41	−1.01E-009	9/1681 → 0.54%
	51	−2.32E-010	9/2601 → 0.35%
Problem #2	11	−1.15E-004	6/121 → 4.96%
	21	−2.02E-007	6/441 → 1.36%
	31	−7.17E-009	6/961 → 0.62%
	41	−1.01E-009	6/1681 → 0.36%
	51	−2.32E-010	6/2601 → 0.23%

In this subsection we study numerically the error in local mass balance (which is characterized by element sink strength). For test problems #1 and #2 the total source strength is 0.0625 (which is equal to $\int_\Omega f \, d\Omega$). For a given element Ω_e (with its boundary denoted by Γ_e) we calculate $\int_{\Omega_e} \nabla \cdot \boldsymbol{v} \, d\Omega \equiv \int_{\Gamma_e} \boldsymbol{v} \cdot \boldsymbol{n} \, d\Gamma$, which should be negative based on the KKT conditions. (As mentioned earlier, in the discrete finite element setting the element source/sink strength can be obtained by picking the corresponding component in the $\boldsymbol{K}_{pv}\boldsymbol{v} - \boldsymbol{f}_p$ vector.) Contours of these element sink strengths are plotted using the built-in cell-centered feature in Tecplot [36]. In Figs. 13 and 14 we have shown the contours of element sink strength for various

Table 7
Performance of the RT0 formulation with respect to mesh refinement. We have used +45-degree and −45-degree meshes for test problems #1 and #2, respectively.

Test problem	# of nodes per side	Min. conc.	% of elements violated
Problem #1	10	−0.029349220	48/162 → 29.63%
	19	−0.002510583	128/648 → 19.75%
	28	−0.000065471	158/1458 → 10.84%
	37	−0.000027412	194/2592 → 7.48%
	46	−0.000012794	224/4050 → 5.53%
	55	−0.000006529	248/5832 → 4.25%
Problem #2	10	−0.039682770	64/162 → 39.51%
	19	−0.006674991	216/648 → 33.33%
	28	−0.001901262	384/1458 → 26.34%
	37	−0.000756689	504/2592 → 19.44%
	46	−0.000337093	552/4050 → 13.63%
	55	−0.000140215	576/5832 → 9.88%

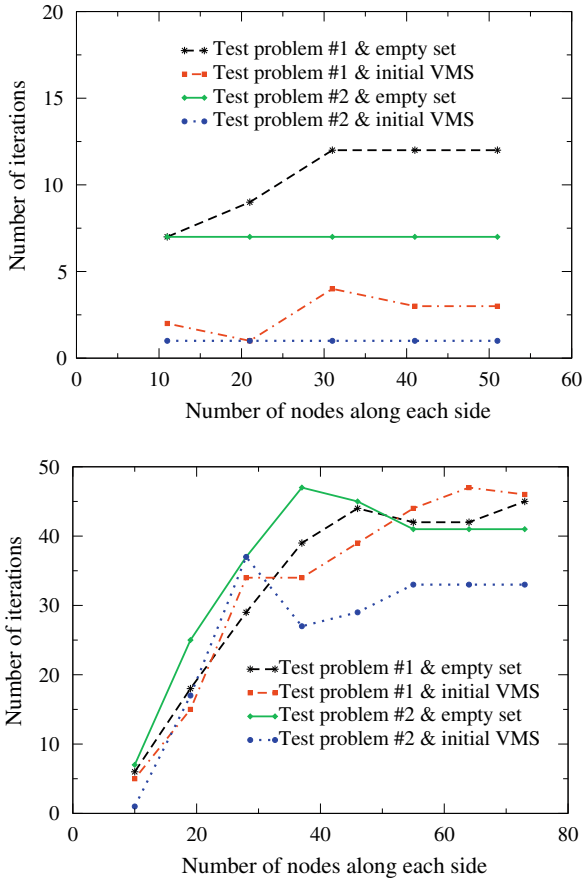


Fig. 16. This figure compares number of iterations taken by the active set strategy for the *VMS formulation* with respect to mesh refinement. We have employed four-node quadrilateral (top figure) and three-node triangular (bottom figure) meshes. Note that, in the case of three-node triangular meshes, we have used +45-degree and −45-degree meshes for test problems #1 and #2, respectively. For both the test problems two different sets are used as an initial guess in the active set strategy.

computational meshes for test problems #1 and #2, respectively. As one can see, for these representative test problems and meshes the violation of local mass balance is insignificant.

For test problem #3 the volumetric source is zero (i.e., $f(\mathbf{x}) = 0$ in Ω). (The problem is driven by non-homogeneous Dirichlet boundary conditions.) Hence, we compare the element sink strength with the total flux along the boundary. Under the

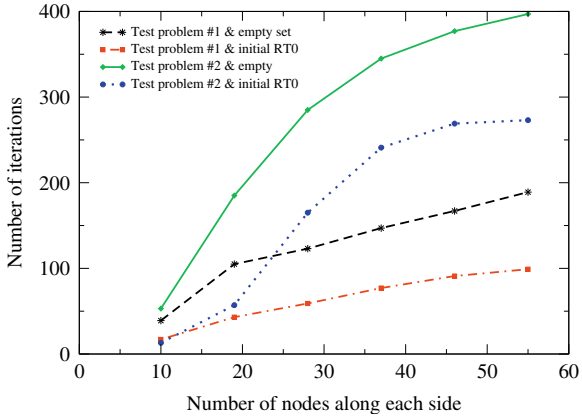


Fig. 17. This figure compares number of iterations taken by the active set strategy for the *RT0 formulation* with respect to mesh refinement. We have employed +45-degree and −45-degree triangular meshes for test problems #1 and #2, respectively. For both the test problems two different sets are used as an initial guess in the active set strategy.

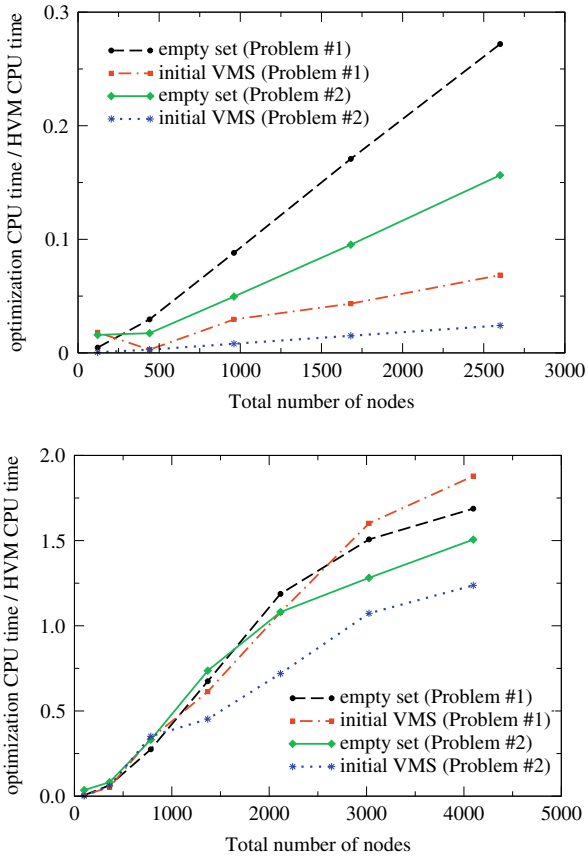


Fig. 18. This figure compares the computational effort of the *optimization-based VMS method* with respect to mesh refinement. We employed four-node quadrilateral (top figure) and 45-degree triangular (bottom figure) meshes. On the y-axis we have the ratio of the additional CPU time taken by the optimization-based VMS method to the CPU time taken by the VMS method (which produces negative solutions). We considered test problem #1 and #2. Note that each node has three degrees-of-freedom.

RT0 formulation (that is, without optimization) the total (integrated) flux along the interior and exterior boundaries are −117.3852 and +117.3852, respectively. (This is not surprising as the RT0 formulation has both local and global mass balance properties.) Under the optimization-based RT0 formulation, total integrated flux along the interior and exterior boundaries are −117.5615 and +127.5694, respectively. Maximum (in magnitude) element sink strength is 0.7477, which is 0.636% compared to the total flux along the interior boundary. The total sink strength by adding the individual element volumetric (sink) strengths is -10.0079, which matches the difference between the fluxes along interior and exterior boundaries. This means that the optimization-based RT0 formulation has the global mass balance property (but, as discussed earlier, does not possess local mass balance property). In Fig. 15 we have shown the contours of element sink strength for test problem #3.

4.7. h-Convergence analysis

In this subsection we study the convergence of the proposed optimization-based methods with respect to mesh refinement. We use test problems #1 and #2, and employ 45-degree triangular and four-node quadrilateral meshes. (As discussed earlier, we use +45-degree and −45-degree meshes for test problems #1 and #2, respectively.) Typical four-node quadrilateral and 45-degree triangular meshes are shown in Figs. 3 and 2(b), respectively. We refine the mesh by increasing the number of nodes along each side.

In Tables 5 and 6 we show the variation of minimum concentration and percentage of nodes that produce negative solutions with respect to mesh refinement under the VMS formulation for 45-degree triangular and four-node quadrilateral

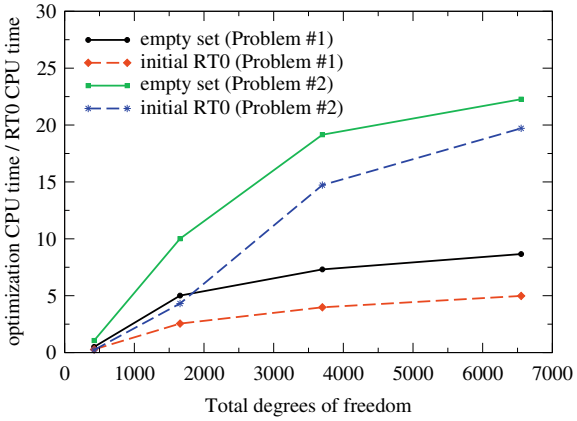


Fig. 19. This figure compares the computational effort of the *optimization-based RT0 method* with respect to mesh refinement. On the *y*-axis we have the ratio of the additional CPU time taken by the optimization-based RT0 method to the CPU time taken by the RT0 method (which produced negative solutions). We employed +45-degree and −45-degree triangular meshes for test problems #1 and #2, respectively. Note that each node has three degrees-of-freedom.

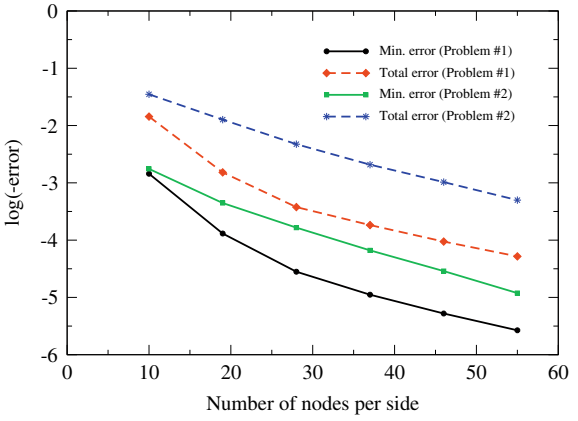


Fig. 20. This figure compares the error in the local mass balance with respect to mesh refinement using the RT0 formulation. We have used +45-degree and −45-degree triangular meshes for test problems #1 and #2, respectively. We have plotted both the maximum element sink strength, and the total error by summing the artificial sink strengths in all elements. From the above figure one can see that both these errors in the local mass balance decrease exponentially with respect to the element size.

meshes. In Table 7 we have shown the variation of minimum concentration and percentage of elements that produced non-negative solutions with respect to mesh refinement under the RT0 formulation. (Note that in the RT0 formulation, the concentration, $c(\mathbf{x})$, is a constant over each element.) As expected, the percentage of non-negative nodes and minimum concentration decrease with respect to mesh refinement, but still the persistent non-negative values and spatial extent of the violation prohibits the use of VMS and RT0 formulations in simulations for which transport is coupled with reactions.

In Figs. 16 and 17 we have compared the number of iterations taken by the active set strategy with respect to mesh refinement for the proposed optimization-based methods. As one can see from these figures, the number of iterations stabilized with respect to mesh refinement (that is, the strategy takes almost the same number of iterations as the mesh is refined). The optimization-based VMS method takes fewer active-set strategy iterations compared to the iterations taken by the optimization-based RT0 formulation.

Figs. 18 and 19, respectively, compare the CPU time taken by the optimization-based VMS and RT0 methods with respect to mesh refinement. On the y-axis we plot the ratio between the *additional* CPU time taken by the active-set strategy (or the optimization solver to obtain non-negative solution) and the CPU time taken by the corresponding underlying mixed formulation (either VMS or RT0 formulation). For the optimization-based VMS method, the additional cost to obtain non-negative solution using the four-node quadrilateral mesh is only a fraction of the computational cost of the VMS formulation. For the three-node triangular mesh, the additional cost to obtain the non-negative solution is nearly twice the cost of the VMS formulation. The optimization-based RT0 method takes relatively more additional CPU time to obtain non-negative solution, and the ratio between the additional CPU time and the CPU time taken by the RT0 formulation is nearly 5 for test problem #1 and 20 for test problem #2. This should not be surprising as in the RT0 formulation the primary variable (in our case, the concentration) is poorly approximated by its piecewise constant representation over each element. (Note that, in the VMS formulation the primary variable is C^0 continuous, that is, piecewise linear and continuous across elements.)

In Fig. 20 we compare the error in the local mass balance with respect to mesh refinement for the RT0 formulation. Since, in the non-negative version of the RT0 formulation we always have artificial sinks (that is, the error in the local mass balance will always be negative), we have taken the negative of the error before taking the logarithm. We have plotted both the maximum (artificial) element sink strength (or error in local mass balance in an element) and also the total sink strength by summing the contribution from all elements. From Fig. 20 one can see that, for the chosen problems, *the error in the local mass balance decreases exponentially with respect to the element size.*

5. Conclusions

Tensorial diffusion problems arise in a variety of important engineering and scientific applications. Although the continuous problem satisfies a maximum–minimum principle, most numerical approximations fail to satisfy this principle in a discrete sense on arbitrary meshes. For some applications, violation of the discrete maximum–minimum principle can be problematic due to physically meaningless negative values of the dependent variable.

In this paper, we proposed two non-negative low-order mixed finite element formulations for the tensorial diffusion equation. (That is, the proposed formulations provide non-negative numerical solutions for linear, bilinear and trilinear finite elements.) This is achieved by rewriting the formulations as constrained optimization problems. In both the cases, the problem belongs to convex quadratic programming, which can be effectively solved using existing numerical optimization solvers (e.g., active set strategy and interior point methods).

One of the formulations is based on the variational multiscale formulation, and the other is based on the lowest-order Raviart–Thomas spaces (that is, the RT0 element). We have demonstrated that the variational multiscale formulation satisfies a continuous maximum–minimum principle. In the case of non-negative formulation based on Raviart–Thomas spaces, two different optimization problems are presented – the primal and dual problems. From the optimization theory it has been inferred that these two problems are equivalent. In addition, from the Karush-Kuhn-Tucker optimality conditions it is inferred that one *may* have violation of (element) local mass balance in those elements that have zero concentration. These violations of local mass balance are in some limited part of the domain, and the violations are small for the test cases studies here.

We have studied the convergence properties of the proposed optimization-based methods. Through numerical experiments we have shown that the error in local mass balance under the optimization-based RT0 method decreases exponentially with respect to mesh refinement. We have also studied the performance of active-set strategy method for solving the resulting convex quadratic programming problems. The performance of the proposed non-negative formulations is illustrated on three representative problems, and the formulations performed well.

One note worthy feature is that existing solvers based on the variational multiscale and RT0 formulations can be easily extended to implement the proposed optimization-based non-negative formulations. Designing a non-negative (stabilized) mixed formulation that also possesses local mass balance property is part of our future work.

Acknowledgments

The research reported herein was supported by the Department of Energy through a SciDAC-2 project (Grant No. DOE DE-FC02-07ER64323). This support is gratefully acknowledged. The opinions expressed in this paper are those of the authors

and do not necessarily reflect that of the sponsor. We also thank Professor Anil Hirani, University of Illinois at Urbana-Champaign, for providing us with the well-centered triangular mesh that is used in this paper. The first author is grateful to Dr. Vit Pruša for valuable suggestions.

Appendix A. Some classical results on discrete maximum–minimum principle

One of the early works on DMP dating from the 1960s is by Varga [37,38], and was presented in the context of finite difference schemes. To the authors' knowledge, the initial work on DMP in the context of the finite element method was done by Ciarlet and Raviart [8]. (Another relevant work by Ciarlet is [39], which was in the context of finite difference operators.) Ref. [8] addressed linear simplicial finite elements, and considered the classical Galerkin single-field formulation for the Poisson and Helmholtz equations.

As mentioned earlier, classical finite element formulations do not satisfy the DMP on general meshes for full diffusivity tensor. The formulations which satisfy DMP impose severe restrictions on both meshes and coefficients of the diffusivity tensor. We now outline some of the classical results that are available on DMP in the context of the finite element method for a *scalar* diffusion equation.

- For linear simplicial elements, Ciarlet and Raviart [8] have shown that non-obtuseness is *sufficient* to satisfy DMP.
- Christie and Hall [40] have presented *sufficient* conditions for bilinear finite elements to satisfy the DMP under a homogeneous forcing function. The results can be summarized as follows. For a non-uniform rectangular mesh (see Fig. 1), the DMP will be satisfied provided

$$h_1 h_2 \leqslant \frac{1}{2}\max\left(k_1^2,\, k_2^2\right) \quad \text{and} \quad k_1 k_2 \leqslant \frac{1}{2}\max\left(h_1^2,\, h_2^2\right) \tag{56}$$

  This implies that a uniform rectangular mesh (that is, $h_1 = h_2$ and $k_1 = k_2$) will satisfy the DMP provided

$$\frac{1}{\sqrt{2}}k_1 \leqslant h_1 \leqslant \sqrt{2}\,k_1 \tag{57}$$

  This further implies that a mesh with squares (that is, $h_1 = h_2 = k_1 = k_2$) will always satisfy the DMP for the case of a scalar diffusion equation.
- Vanselow [41] has shown that a Delaunay triangulation along with an additional condition on boundary nodes are sufficient for the DMP under the classical single-field Galerkin formulation. We now outline how this additional condition looks for a convex domain. To this end, let P_1 and P_2 be two neighboring nodes on the boundary. Let us denote their spatial coordinates as $\mathbf{x}_1$ and $\mathbf{x}_2$, and define $\tilde{\mathbf{x}} := (\mathbf{x}_1 + \mathbf{x}_2)/2$. Then the additional condition can be written as

$$\|\mathbf{x}_i - \tilde{\mathbf{x}}\|_2 \leqslant \|\mathbf{x}_Q - \tilde{\mathbf{x}}\|_2 \text{ for all nodes Q in the triangulation with } Q \neq P_i,\ i = 1,2 \tag{58}$$

where $\mathbf{x}_Q$ denotes the spatial coordinates of the node Q. A similar condition is required for a non-convex domain. In addition, it has also been shown that under some weak additional assumptions on the triangulation, these conditions are necessary [41, Section 4].

References

[1] G.F. Pinder, M.A. Celia, Subsurface Hydrology, John Wiley & Sons. Inc., New Jersey, USA, 2006.
[2] P. Herrera, A. Valocchi, Positive solution of two-dimensional solute transport in heterogeneous aquifers, Ground Water 44 (2006) 803–813.
[3] R. Liska, M. Shashkov, Enforcing the discrete maximum principle for linear finite element solutions for elliptic problems, Communications in Computational Physics 3 (2008) 852–877.
[4] R. McOwen, Partial Differential Equations: Methods and Applications, Prentice Hall, New Jersey, USA, 1996.
[5] D. Gilbarg, N.S. Trudinger, Elliptic Partial Differential Equations of Second Order, Springer, New York, USA, 2001.
[6] L.C. Evans, Partial Differential Equations, American Mathematical Society, Providence, Rhode Island, USA, 1998.
[7] Q. Han, F. Lin, Elliptic Partial Differential Equations, American Mathematical Society, Providence, Rhode Island, USA, 2000.
[8] P.G. Ciarlet, P-A. Raviart, Maximum principle and uniform convergence for the finite element method, Computer Methods in Applied Methods and Engineering 2 (1973) 17–31.
[9] E. Burman, A. Ern, Nonlinear diffusion and discrete maximum principle for stabilized Galerkin approximations of the convection–diffusion–reaction equation, Computer Methods in Applied Mechanics and Engineering 191 (2002) 3833–3855.
[10] E. Burman, A. Ern, Stabilized Galerkin approximation of convection–diffusion–reaction equations: discrete maximum principle and convergence, Mathematics of Computation 74 (2005) 1637–1652.
[11] H. Hoteit, R. Mose, B. Philippe, Ph. Ackerer, J. Erhel, The maximum principle violations of the mixed-hybrid finite-element method applied to diffusion equations, International Journal for Numerical Methods in Engineering 55 (2002) 1373–1390.
[12] J. Karátson, S. Korotov, Discrete maximum principles for finite element solutions of nonlinear elliptic problems with mixed boundary conditions, Numerische Mathematik 99 (2005) 669–698.
[13] J. Karátson, S. Korotov, Discrete maximum principles for finite element solutions of some mixed nonlinear elliptic problems using quadratures, Journal of Computational and Applied Mathematics 192 (2006) 75–88.
[14] M. Krizek, L.P. Liu, Finite element approximation of a nonlinear heat conduction problem in anisotrpoic media, Computer Methods in Applied Mechanics and Engineering 157 (1998) 387–397.
[15] M. Krizek, L.P. Liu, On the maximum and comparison principles for a steady-state nonlinear heat conduction problem, Zeitschrift fur Angewandte Mathematik und Mechanik 83 (2003) 559–563.

[16] C. Le Potier, Finite volume monotone scheme for highly anisotropic diffusion operators on unstructured triangular meshes, Comptes Rendus Mathematique 341 (2005) 787–792.
[17] K. Lipnikov, M. Shashkov, D. Svyatskiy, The mimetic finite difference discretization of diffusion problem on unstructured polyhedral meshes, Journal of Computational Physics 211 (2006) 473–491.
[18] J.M. Nordbotten, I. Aavatsmark, G.T. Eigestad, Monotonicity of control volume methods, Numerische Mathematik 106 (2007) 255–288.
[19] M.J. Mlacnik, L.J. Durlofsky, Unstructured grid optimization for improved monotonicity of discrete solutions of elliptic equations with highly anisotropic coefficients, Journal of Computational Physics 216 (2006) 337–361.
[20] Z. Chen, G. Huan, Y. Ma, Computational Methods for Multiphase Flows in Porous Media, Society for Industrial and applied Mathematics, Philadelphia, USA, 2006.
[21] P.A. Raviart, J.M. Thomas, A mixed finite element method for 2nd order elliptic problems, in: Mathematical Aspects of the Finite Element Method, Springer Verlag, New York, 1977, pp. 292–315.
[22] F. Brezzi, M. Fortin, Mixed and Hybrid Finite Element Methods, Springer Series in Computational Mathematics, vol. 15, Springer Verlag, New York, USA, 1991.
[23] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, Cambridge, UK, 2004.
[24] J. Nocedal, S.J. Wright, Numerical Optimization, Springer Verlag, New York, USA, 1999.
[25] R. Fletcher, Practical Methods of Optimization, John Wiley & Sons. Inc., Chichester, UK, 2000.
[26] A. Masud, T.J.R. Hughes, A stabilized mixed finite element method for Darcy flow, Computer Methods in Applied Mechanics and Engineering 191 (2002) 4341–4370.
[27] K.B. Nakshatrala, D.Z. Turner, K.D. Hjelmstad, A. Masud, A stabilized mixed finite element formulation for Darcy flow based on a multiscale decomposition of the solution, Computer Methods in Applied Mechanics and Engineering 195 (2006) 4036–4049.
[28] T.J.R. Hughes, Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods, Computer Methods in Applied Mechanics and Engineering 127 (1995) 387–401.
[29] M. Borsuk, V. Kondratiev, Elliptic Boundary Value Problems of Second Order in Piecewise Smooth Domains, Elsevier Science, San Diego, USA, 2006.
[30] K. Lipnikov, M. Shashkov, D. Svyatskiy, Y. Vassilevski, Monotone finite volume schemes for diffusion equations on unstructured triangular and shape-regular polygonal meshes, Journal of Computational Physics 227 (2007) 492–512.
[31] E. Vanderzee, A.N. Hirani, D. Guoy, E. Ramos, Well-centered triangulation, Technical Report UIUCDCS-R-2008-2936, Department of Computer Science, University of Illinois at Urbana-Champaign, February 2008. Also available as a preprint at arXiv as 0802.2108v1 [cs.CG].
[32] P. Chadwick, Continuum Mechanics: Concise Theory and Problems, Dover Publications, Inc., Minealo, New York, 1999.
[33] A. Hirani, Discrete Exterior Calculus. PhD Thesis, California Institute of Technology, Pasadena, California, USA, 2003.
[34] A.N. Hirani, K.B. Nakshatrala, J.H. Chaudhry, Numerical method for Darcy flow derived using Discrete Exterior Calculus, Technical Report UIUCDCS-R-2008-2937, Department of Computer Science, University of Illinois at Urbana-Champaign, 2008. Also available as a preprint at arXiv as 0810.3434v1 [math.NA].
[35] D.G. Luenberger, Y. Ye, Linear and Nonlinear Programming, third ed., Springer Science+Business Media, Inc., New York, USA, 2008.
[36] Tecplot 360: User's Manual. URL: <http://www.tecplot.com>, Bellevue, Washington, USA, 2008.
[37] R. Varga, Matrix Iterative Analysis, Prentice-Hall, New Jersey, USA, 1962.
[38] R. Varga, On discrete maximum principle, SIAM Journal on Numerical Analysis 3 (1966) 355–359.
[39] P.G. Ciarlet, Discrete maximum principle for finite-difference operators, Aequationes Mathematicae 4 (1970) 338–352.
[40] I. Christie, C. Hall, The maximum principle for bilinear elements, International Journal for Numerical Methods in Engineering 20 (1984) 549–553.
[41] R. Vanselow, About Delaunay triangulations and discrete maximum principles for the linear conforming FEM applied to the Poisson equation, Applications of Mathematics 46 (2001) 13–28.
[42] Gmsh: A three-dimensional finite element mesh generator with pre- and post-processing facilities. URL: <http://www.geuz.org/gmsh/>.