SIAM J. NUMER. ANAL.
Vol. 57, No. 3, pp. 1238–1262

A CONSERVATIVE FLUX OPTIMIZATION FINITE ELEMENT METHOD FOR CONVECTION-DIFFUSION EQUATIONS*

YUJIE LIU†, JUNPING WANG‡, AND QINGSONG ZOU§

Abstract. This article presents a new finite element method for convection-diffusion equations by enhancing the continuous finite element space with a flux space for flux approximations that preserve the important mass conservation locally on a prescribed set of control elements. The numerical scheme is based on a constrained flux optimization approach where the constraint was given by local mass conservation equations and the flux error is minimized in a prescribed topology/metric. This new scheme provides numerical approximations for both the primal and the flux variables. It is shown that the numerical approximations for the primal and the flux variables are convergent with optimal order in some discrete Sobolev norms. Numerical experiments are conducted to confirm the convergence theory. Furthermore, the new scheme was employed in the computational simulation of a simplified two-phase flow problem in highly heterogeneous porous media. The numerical results illustrate an excellent performance of the method in scientific computing.

Key words. conservative flux, primal-dual weak Galerkin, finite element methods, finite volume method

AMS subject classifications. Primary, 65N30, 65N15, 65N12; Secondary, 35B45, 35J50, 76S05, 76T99, 76R99

DOI. 10.1137/17M1153595

1. Introduction.

This paper is concerned with the development of numerical methods for partial differential equations that maintain important conservation properties for the underlying physical variables. For simplicity, consider the model convection-diffusion equation that seeks an unknown function $u = u(x)$ satisfying

$$-\nabla \cdot (\alpha \nabla u + \boldsymbol{\beta} u) = f \qquad \text{in } \Omega, \tag{1.1}$$

$$u = 0 \qquad \text{on } \partial\Omega, \tag{1.2}$$

where $\Omega \subset \mathbf{R}^d (d = 2)$ is a bounded polygonal domain with boundary $\partial\Omega$, and $\alpha = \{\alpha_{i,j}\}_{d\times d}$ is a symmetric, positive definite tensor, i.e., there exists a positive constant α_0 such that

$$\xi^T \alpha \xi \geq \alpha_0 \xi^T \xi \qquad \forall \xi \in \Omega.$$

*Received by the editors October 25, 2017; accepted for publication (in revised form) April 15, 2019; published electronically June 4, 2019.
http://www.siam.org/journals/sinum/57-3/M115359.html
Funding: The work of the first author was partially supported by Guangdong Provincial Natural Science Foundation (2017A030310285), Shandong Provincial Natural Science Foundation (ZR2016AB15), and Youthful Teacher Foster Plan of Sun Yat-Sen University (171gpy118). The work of the second author was supported by the National Science Foundation. However, any opinion, finding, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the NSF. The work of the third author was partially supported by the special project high performance computing of the National Key Research and Development Program (2016YFB0200604), the National Natural Science Foundation of China (11571384), and Guangdong Provincial Natural Science Foundation (2017B030311001).

†School of Data and Computer Science, Sun Yat-sen University, Guangzhou, 510275, China (liuyujie5@mail.sysu.edu.cn).
‡Division of Mathematical Sciences, National Science Foundation, Alexandria, VA 22314 (jwang@nsf.gov).
§Corresponding author. School of Data and Computer Science, and Guangdong Province Key Laboratory of Computational Science, Sun Yat-sen University, Guangzhou 510006, China (mcszqs@mail.sysu.edu.cn).

In some applications, such as the flow of fluid in porous media governed by Darcy's law, the quantity of interest is often the flow velocity represented by $\mathbf{q} = -(\alpha\nabla u + \boldsymbol{\beta} u)$. With the velocity $\mathbf{q}$, (1.1) can be rewritten as $\nabla \cdot \mathbf{q} = f$, so that on any subdomain $D \subset \Omega$, from the divergence theorem one has

$$\int_D \nabla \cdot \mathbf{q} dx = \int_D f dx \iff \int_{\partial D} \mathbf{q} \cdot \mathbf{n} ds = \int_D f dx, \tag{1.3}$$

where $\mathbf{n}$ is the outward normal direction of ∂D. The equations in (1.3), especially the one on the right, characterize the mass conservation property for the porous media flow. The quantity that enters into the mass conservation equation is the flux variable $q_n = \mathbf{q} \cdot \mathbf{n}$ on the boundary of any control element D.

A numerical scheme for the model convection-diffusion equation (1.1)–(1.2) is said to be conservative if it provides a numerical solution, denoted by u_h, and an associated numerical flux $q_{n,h}$ on the boundary of a set of prescribed control elements $\mathcal{D}_h = \{D\}$ such that

$$\begin{cases} u_h \to u & \text{as } h \to 0, \\ q_{n,h} \to q_n & \text{as } h \to 0, \\ \int_{\partial D} q_{n,h} ds = \int_D f dx & \forall D \in \mathcal{D}_h, \end{cases} \tag{1.4}$$

where the convergence in (1.4) should be understood under certain prescribed topologies for the corresponding variables. The third line of (1.4) is the local mass conservation which is a highly preferable property of the algorithm in practical computing.

The classical continuous Galerkin finite element method is a popular and easy-to-implement numerical scheme for the model equation (1.1)–(1.2). In the most simple formulation, the P_1-continuous Galerkin finite element scheme seeks $u_h \in S_h$ satisfying

$$(\alpha\nabla u_h + \boldsymbol{\beta} u_h, \nabla v) = (f, v) \qquad \forall v \in S_h, \tag{1.5}$$

where $S_h \subset H_0^1(\Omega)$ consists of C^0-piecewise linear functions on a prescribed finite element triangulation $\mathcal{T}_h$. The numerical scheme (1.5) provides a direct approximation u_h to the primal variable $u = u(x)$ from which a numerical velocity can be computed as

$$\mathbf{q}_h = -(\alpha\nabla u_h + \boldsymbol{\beta} u_h). \tag{1.6}$$

On each control element $D \in \mathcal{D}_h$, the numerical velocity $\mathbf{q}_h$ obtained in (1.6) might be discontinuous across the edges of D along the normal direction $\mathbf{n}$ (i.e., $\mathbf{q}_h \cdot \mathbf{n}$, which is known as the flux). Even if $\mathbf{q}_h \cdot \mathbf{n}$ is continuous across ∂D or can be reconstructed as a continuous one through a simple average on each edge, it usually does not preserve the mass conservation property. For this reason, the continuous Galerkin finite element method is often said to be nonconservative.

Attempts at making the continuous Galerkin to be conservative have been made by various researchers in the scientific community for the last three decades. To the authors' knowledge, all the existing work in this endeavor explore the use of various postprocessing techniques for the primal variable u_h. For example, in [15], a conservative flux was obtained through a postprocessing procedure for either P_1-conforming or -nonconforming Galerkin finite element approximations, where the reconstructed numerical flux was sought in the Raviart–Thomas space of the lowest order. In [22], the authors devised a postprocessing procedure for continuous Galerkin finite element

approximations on any user-selected subdomain. Specifically, for the subdomain under consideration, the authors introduced an auxiliary boundary flux field and developed a formulation which reduces to the usual continuous Galerkin method plus a modification designed for attaining global conservation. In [23], the authors developed a different postprocessing technique for computing a numerical flux on element boundaries that is elementwise conservative for the continuous Galerkin approximation. Their technique was based on the computation of a correction of the averaged normal flux on element boundaries by using the jump of piecewise constants or linear functions to be determined as the solution of a global linear system. A modified version of [23] was employed in [31] to produce a conservative flux for both steady-state and dynamic flow models by adding a piecewise constant correction that is minimized in a weighted L^2 norm. The postprocessed flux in [31] was shown to have the same rate of convergence as the original, but nonconservative flux. In [5], the authors studied the compatible least-squares method for the Darcy flow equation, and further developed a flux-correction procedure to obtain a locally conservative numerical solution without compromising its L^2 accuracy. In [16], the authors developed a two-step postprocessing procedure for computing a conservative flux for a continuous Galerkin finite element approximation. Their first step involves the computation of a numerical flux trace defined on element interfaces. The second step is a local element-by-element postprocessing of the continuous Galerkin approximation by incorporating the result from the first step. In [28], the authors devised a computational framework for advective-diffusive-reactive systems with approximate solutions satisfying desired properties such as maximum principles, the solution nonnegative constraint, and the elementwise conservative property. This method employs a low order mixed finite element formulation based on least-squares formalism by enforcing explicit constraints of various types. In [42], the authors develop an elementwise conservative flux also by postprocessing the FEM solution element by element. Their method is valid for any order schemes and their postprocessed solution converges with optimal order both in H^1 and L^2 norms. Very recently in [44], a volumewise conservative flux field has been derived by postprocessing the finite element solution. One important feature of their method is that their derived flux field is continuous even across the internal edges of the underlying mesh.

In the literature, one can find various numerical methods for (1.1)–(1.2) that preserve the mass conservation property (1.4) locally on each element $T \in \mathcal{T}_h$. One of such methods is the finite volume method (FVM) widely used in scientific computing for problems in science and engineering, including fluid dynamics [4, 18, 19, 25, 26, 30, 34]. Most algorithms in FVM enjoy the nice feature of algorithmic simplicity and computational efficiency, and some of the low order FVMs (e.g., P_0 and P_1 schemes) have been well studied for their mathematical convergence and stability [3, 9, 13, 21, 26, 40]. It should also be noted that the high order and symmetric FVMs are generally challenging in theory and algorithmic design [10, 11, 12, 27, 43]. In the finite element context, several conservative numerical schemes have been developed. The mixed finite element method [32, 6], the discontinuous Galerkin finite element method [1], the hybridizable discontinuous Galerkin [29], and weak Galerkin finite element methods [38, 39, 37] are a few of such examples that give numerical approximations with conservative numerical flux.

In this work, we do not intend to pursue the approaches along the above-mentioned directions. Instead, we shall design a new conservative numerical scheme for (1.1)–(1.2) via a conservation-constrained optimization approach by using the continuous finite element space S_h. More precisely, let V_h be a discrete flux space defined on the

edge set of the control volume $\mathcal{D}_h$, and $r \in [1, \infty)$ be any given real number. Assume that $\mathcal{D}_h$ is nested with $\mathcal{T}_h$ in the sense that for any $D \in \mathcal{D}_h$, there exists a triangular element $T \in \mathcal{T}_h$ such that $D \subset T$. Our numerical method seeks $u_h \in S_h$ and $q_{n,h} \in V_h$ that minimizes the flux-error function

$$J_r(u_h, q_{n,h}) := \frac{1}{r} \sum_{D \in \mathcal{D}_h} \sum_{e \subset \partial D} h_D \int_e |q_{n,h} + \alpha \nabla u_h \cdot \mathbf{n} + \boldsymbol{\beta} u_h \cdot \mathbf{n}|^r ds, \tag{1.7}$$

subject to the constraint of the mass conservation equation of $\int_{\partial D} q_{n,h} ds = \int_D f dx$ on each control element $D \in \mathcal{D}_h$. We note that (1.7) is defined by first computing $\sum_{e \subset \partial D} h_D \int_e |p + \alpha^* \nabla v \cdot \mathbf{n}_e + \boldsymbol{\beta}^* v \cdot \mathbf{n}_e|^r ds$ on each control element D, and then summing over all $D \in \mathcal{D}_h$, where $\nabla u_h|_e$ is the trace of ∇u_h taken on the control element D. In the case of $r = 2$, the Euler–Lagrange form of this constrained optimization problem yields a system of linear equations for the unknown variables u_h, $q_{n,h}$, and another variable known as the Lagrange multiplier. Our new method essentially looks for a conservative flux variable that best approximates the obvious, but nonconservative, numerical velocity $\mathbf{q}_h = -(\alpha \nabla u_h + \boldsymbol{\beta} u_h)$ in a discrete metric. For this reason, the numerical scheme is named *conservative flux optimization (CFO)* finite element method in this article. It should be pointed out that the *CFO* finite element method was originally motivated by the idea of the primal-dual weak Galerkin method (namely, PDE-constraint minimization of stabilizers) presented as in [37] for the second order elliptic equation in nondivergence form.

The main contributions of this work are the following: (1) A conservative numerical scheme is devised to yield approximate solutions for the primal and the flux variables simultaneously without using any postprocessing. This flux-optimization-based discretization technique is generally applicable to other PDEs; (2) an optimal order of convergence is established for the numerical flux in the L^2 norm; (3) the numerical solution u_h for the primal variable is shown to have optimal order of convergence in $H^1(\Omega)$; (4) numerical experiments are conducted on test problems involving varying and discontinuous coefficients, and the results strongly suggest an optimal order of convergence in $L^2(\Omega)$ for the primal variable; (5) the new scheme is applied to a simplified two-phase flow problem in a highly heterogeneous porous media, and the expected fingering phenomenon is clearly shown in the corresponding computational simulation.

Our numerical scheme can also be viewed as an FVM—vertex-based for the primal variable u_h and polygonal element based for the flux variable $q_{n,h}$. In the literature, FVM algorithms are often classified into two categories: (1) vertex centered where the computational nodes are positioned at the vertices, and (2) cell centered where the computational nodes are positioned at the center of the cells/elements. Usually for vertex-centered schemes, one discretizes the underlying PDE by asking the numerical solution to satisfy the flux conservation on a *dual mesh* consisting of *control volumes* given as polygons (2-dimensional) or polyhedra (3-dimensional) surrounding the mesh vertices. The mathematical convergence for vertex-centered schemes is often established in a way that mimics the corresponding finite element formulation [3, 9, 21, 33, 40, 43]. For cell-centered finite volume schemes, their theoretical convergence is usually obtained in carefully chosen discrete norms on meshes with certain structures [20, 24, 36, 41]. The new scheme of this paper belongs to the category of element-centered FVMs in a broad sense, as the mass conservation (1.3) is imposed on an arbitrary polygonal partition $\mathcal{D}_h$ nested with $\mathcal{T}_h$. Compared with the vertex-centered FVMs, our scheme does not require a deliberate design of any dual

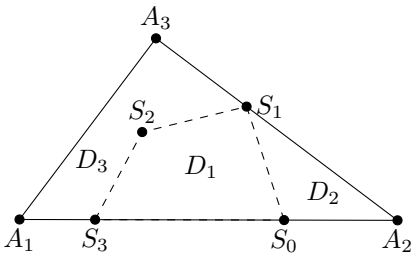


FIG. 1. *An illustrative triangular element $T = A_1A_2A_3$ with three control elements D_1, D_2, and D_3.*

mesh in the computation. Compared with the cell-centered FVMs, our numerical solution u_h belongs to the continuous finite element space so that its convergence is within the reach of currently available mathematical techniques with generalization to unstructured partitions. It is worth mentioning that our scheme is symmetric for nonsymmetric PDE problems while the classical vertex-centered FVMs are often nonsymmetric even for symmetric PDE problems.

The paper is organized as follows. In section 2, we present the CFO finite element scheme for the model problem (1.1)–(1.2). In section 3, we establish a result on the well-posedness and stability for the CFO scheme. In section 4, we derive some error estimates for the resulting numerical approximations in various Sobolev norms. Finally, in section 5, we present a few numerical results to demonstrate the efficiency and accuracy of the new scheme. In particular, we will first verify the theoretical convergence through a couple of testing examples, and then demonstrate the power of the new scheme in scientific computing through a simplified two-phase flow problem in highly heterogeneous porous media.

2. A CFO scheme. Let $\mathcal{T}_h$ be a regular triangulation of the polygonal domain $\Omega \subset \mathbb{R}^2$. Denote by $h := \max_{T\in\mathcal{T}_h} h_T$ the mesh size of $\mathcal{T}_h$, where $h_T = \mathrm{diam}(T)$ is the diameter of the element $T \in \mathcal{T}_h$. The collection of control elements forms a polygonal partition of Ω and is denoted by $\mathcal{D}_h$. For the algorithm to be presented, assume that $\mathcal{D}_h$ is nested to $\mathcal{T}_h$ in the sense that for any control element $D \in \mathcal{D}_h$, there exists a triangular element $T \in \mathcal{T}_h$ such that $D \subset T$. Figure 1 illustrates three control elements nested to the triangular element T with vertices $A_i,\ i = 1, 2, 3$. Denote by $\mathcal{E}_h$ the edge set of $\mathcal{D}_h$, and $\mathcal{E}_h^0 \subset \mathcal{E}_h$ the set of all interior edges. The set of boundary edges is denoted as $\mathcal{E}_h^B := \mathcal{E}_h \backslash \mathcal{E}_h^0$. The diameter of the edge $e \in \mathcal{E}_h$ is denoted as $h_e = \mathrm{diam}(e)$. For convenience, for each $e \in \mathcal{E}_h$ we assign a normal direction $\mathbf{n}_e$ which provides an orientation for E_h.

Recall that the classical P_1-conforming element associated with $\mathcal{T}_h$ is given by

$$S_h = \{v \in C^0(\Omega) :\ v|_T \in P_1(T), \forall T \in \mathcal{T}_h, v|_{\partial\Omega} = 0\}.$$

Here $C^0(\Omega)$ stands for the space of continuous functions on the domain Ω. Denote by W_h the space of piecewise constant functions on $\mathcal{D}_h$ and V_h the space of piecewise constant functions on the edge set $\mathcal{E}_h$.

For any $q \in L^2(\mathcal{E}_h)$, denote by $\nabla_w \cdot q$ the discrete weak divergence given as a function in W_h such that on each control element $D \in \mathcal{D}_h$

$$(\nabla_w \cdot q)|_D = \frac{1}{|D|} \int_{\partial D} q\mathbf{n} \cdot \mathbf{n}_e ds, \tag{2.1}$$

where $\mathbf{n}$ is the outward normal vector of ∂D, and $\mathbf{n}_e$ is the prescribed orientation of $e \subset \partial D$.

A flux function $p \in V_h$ is said to be an *admissible discrete flux* to the equation $\nabla \cdot \mathbf{q} = f$ if it satisfies

$$(\nabla_w \cdot p, w) = (f, w) \qquad \forall w \in W_h. \tag{2.2}$$

As W_h consists of piecewise constant functions, (2.2) is equivalent to the following elementwise identity

$$(\nabla_w \cdot p, 1)_D = (f, 1)_D \qquad \forall D \in \mathcal{D}_h. \tag{2.3}$$

From (2.2), a discrete flux $p \in V_h$ is *admissible* if and only if $\nabla_w \cdot p = Q_h f$, where Q_h is the L^2 projection onto the finite element space W_h.

The convection-diffusion equation (1.1) can be rewritten as a system of linear equations:

$$\mathbf{q} = -(\alpha \nabla u + \boldsymbol{\beta} u), \qquad \nabla \cdot \mathbf{q} = f. \tag{2.4}$$

A pair of finite element functions $(u_h; q_h) \in S_h \times V_h$ is said to be an ideal approximation of (2.4) with the boundary condition (1.2) if, on each control element $D \in \mathcal{D}_h$, one has

$$q_h + (\alpha^* \nabla u_h + \boldsymbol{\beta}^* u_h) \cdot \mathbf{n}_e = 0 \qquad \text{on } e \subset \partial D, \tag{2.5}$$

$$\nabla_w \cdot q_h = Q_h f \qquad \text{in } D, \tag{2.6}$$

where α^* and $\boldsymbol{\beta}^*$ are the trace of α and $\boldsymbol{\beta}$ on ∂D as taken from the control element D, respectively. In (2.5) on $e \subset \partial D$, the vector-valued function ∇u_h assumes the trace of ∇u_h taken on the control element D. Equation (2.6) ensures the local mass conservation by the discrete flux q_h.

It is not hard to see that the ideal approximation for (2.4) generally may not exist in the finite element space $S_h \times V_h$. A remedy to the solution nonexistence is to find a pair $(u_h; q_h) \in S_h \times V_h$ that satisfies the mass conservation equation (2.6) while the flux error $q_h + (\alpha^* \nabla u_h + \boldsymbol{\beta}^* u_h) \cdot \mathbf{n}_e$ is minimized in a metric at the user's discretion. To this end, we introduce a functional in the space $S_h \times V_h$ as follows:

$$J_r(v, p) := \frac{1}{r} \sum_{D \in \mathcal{D}_h} \sum_{e \subset \partial D} h_D \int_e |p + \alpha^* \nabla v \cdot \mathbf{n}_e + \boldsymbol{\beta}^* v \cdot \mathbf{n}_e|^r ds, \tag{2.7}$$

where $r \in [1, \infty)$ is a prescribed value. Our numerical algorithm then seeks $(u_h; q_h) \in S_h \times V_h$ which minimizes the functional J_r under the constraint (2.6). Due to the emphasis on the mass conservation and the error minimization for the flux approximation, we shall name this discretization algorithm a *CFO finite element method* or *CFO* in brief. The CFO algorithm can be mathematically stated as follows.

CFO Algorithm 2.1. *Find $u_h \in S_h$ and $q_h \in V_h$ such that*

$$(u_h; q_h) = \arg\min_{v \in S_h, p \in V_h, \nabla_w \cdot p = Q_h f} J_r(v, p). \tag{2.8}$$

By introducing a Lagrange multiplier $\lambda_h \in W_h$, the constrained minimization problem (2.8) can be reformulated into the Euler–Lagrange form that seeks $(u_h; q_h) \in$

$S_h \times V_h$ and $\lambda_h \in W_h$ satisfying

$$\langle DJ_r(u_h, q_h), (v; p)\rangle + (\nabla_w \cdot p, \lambda_h) = 0 \qquad \forall (v; p) \in S_h \times V_h, \tag{2.9}$$

$$(\nabla_w \cdot q_h, w) = (f, w) \qquad \forall w \in W_h, \tag{2.10}$$

where

$$\begin{aligned}\langle DJ_r(u_h, q_h), (v; p)\rangle := \sum_{D\in\mathcal{D}_h} h_D \sum_{e\subset\partial D} \int_e |q_h + \alpha^* \nabla u_h \cdot \mathbf{n}_e \\ + \boldsymbol{\beta}^* u_h \cdot \mathbf{n}_e|^{r-1}(p + \alpha^* \nabla v \cdot \mathbf{n}_e + \boldsymbol{\beta}^* v \cdot \mathbf{n}_e)\ Sgn\ ds\end{aligned} \tag{2.11}$$

is the Fréchet derivative of the functional J_r at $(u_h; q_h)$ along the direction of $(v; p)$. Here $Sgn = \mathrm{sgn}(q_h + \alpha^* \nabla u_h \cdot \mathbf{n}_e + \boldsymbol{\beta}^* u_h \cdot \mathbf{n}_e)$ represents the sign of the corresponding term. For the case of $r = 2$, the Fréchet derivative $\langle DJ_r(u_h, q_h), (v; p)\rangle$ defines a bilinear form given as follows:

$$\begin{aligned}&s_h(u_h; q_h), (v; p) \\ &:= \int_{D\in\mathcal{D}_h} \sum_{e\subset\partial D} h_D \int_e (q_h + \alpha^* \nabla u_h \cdot \mathbf{n}_e + \boldsymbol{\beta}^* u_h \cdot \mathbf{n}_e)(p + \alpha^* \nabla v \cdot \mathbf{n}_e + \boldsymbol{\beta}^* v \cdot \mathbf{n}_e)\ ds,\end{aligned} \tag{2.12}$$

so that the Euler–Lagrange equations (2.9)–(2.10) read as below:

$$s_h(u_h; q_h), (v; p)) + (\nabla_w \cdot p, \lambda_h) = 0 \qquad \forall (v; p) \in S_h \times V_h, \tag{2.13}$$

$$(\nabla_w \cdot q_h, w) = (f, w) \qquad \forall w \in W_h. \tag{2.14}$$

The above equations can be rewritten in their equivalent form: find $(u_h; q_h; \lambda_h) \in S_h \times V_h \times W_h$ such that

$$b_h((u_h; q_h; \lambda_h), (v; p; w)) = (f, w) \qquad \forall (v; p; w) \in S_h \times V_h \times W_h, \tag{2.15}$$

with the bilinear form $b_h(\cdot, \cdot)$ defined for all $(\tilde{u}; \tilde{q}; \tilde{\lambda}), (v; p; w) \in S_h \times V_h \times W_h$ as

$$b_h((\tilde{u}; \tilde{q}; \tilde{\lambda}), (v; p; w)) = s_h(\tilde{u}; \tilde{q}), (v; p)) + (\nabla_w \cdot p, \tilde{\lambda}) + (\nabla_w \cdot \tilde{q}, w).$$

Since the bilinear form $b_h(\cdot, \cdot)$ is symmetric, the linear system derived from numerical scheme (2.13)–(2.14) has a symmetric matrix.

Compared to the C^0-conforming finite element method, the new method (2.13)–(2.14) offers a direct approximation of the flux variable on the edge set of the control elements. Therefore, the computational cost of the CFO method is more expensive than the C^0-conforming finite element method. However, the CFO method maintains the important conservation property through the flux approximation, which is indispensable in some important applications (e.g., the two-phase fluid flow in porous media described by model (5.6)–(5.7)). To a certain extent, the cost of the CFO algorithm is comparable to the standard mixed finite element method of the lowest order. But the CFO method is advantageous over the standard mixed method in that the resulting linear system is symmetric while the mixed method is nonsymmetric. Overall, the CFO method offers more options than the standard mixed finite element method at a comparable cost of computation.

3. Well-posedness and stability. In this section, we shall study the solution existence and uniqueness of the CFO finite element scheme (2.8) with $r = 2$. Observe that the corresponding Euler–Lagrange formulation is a system of linear equations given by (2.13)–(2.14).

DEFINITION 3.1. *For any element $D \in \mathcal{D}_h$, the oscillation of a function $\sigma = \sigma(x) \in L^\infty(\Omega)$ on D is defined as*

$$\operatorname{osc}(\sigma, D) := \operatorname*{essen\,sup}_{x,y\in D} |\sigma(x) - \sigma(y)|.$$

The local oscillation of $\sigma = \sigma(x)$ with respect to the polygonal partition $\mathcal{D}_h$ is given by

$$\operatorname{osc}(\sigma, \mathcal{D}_h) := \max_{D\in\mathcal{D}_h} \operatorname{osc}(\sigma, D).$$

DEFINITION 3.2. *For a given nonnegative integer m, a function $\sigma = \sigma(x) \in L^\infty(\Omega)$ is said to be uniformly piecewise C^m with respect to the partition $\mathcal{D}_h$ if $\sigma|_D \in W^{m,\infty}(D)$ for each $D \in \mathcal{D}_h$. Furthermore, for any given $\epsilon > 0$, there exists a $\delta > 0$ such that*

$$\operatorname{osc}(\partial^s \sigma, \mathcal{D}_h) < \epsilon, \qquad 0 \le |s| = s_1 + s_2 \le m$$

whenever $h < \delta$. Here $s = (s_1, s_2)$ is a multi-index and $\partial^s = \partial_x^{s_1}\partial_y^{s_2}$ is the corresponding partial differential operator.

Denote by $C^m(\mathcal{D}_h)$ the space of functions that are uniformly piecewise C^m with respect to the finite element partition $\mathcal{D}_h$.

Next, for any given $\theta \in H^{-1}(\Omega)$ let $w = w(\theta) \in H_0^1(\Omega)$ be the solution of the following auxiliary problem:

$$(\nabla w, \alpha\nabla\phi + \boldsymbol{\beta}\phi) = (\theta, \phi) \qquad \forall \phi \in H_0^1(\Omega). \tag{3.1}$$

The problem (3.1) is given by the dual of the differential operator in the model problem (1.1). From the solution existence and uniqueness assumption for (1.1), the problem (3.1) has a unique solution $w \in H_0^1(\Omega)$ satisfying the a priori estimate

$$\|w\|_1 \le C\|\theta\|_{-1}. \tag{3.2}$$

LEMMA 3.3. *Assume that the coefficients of the model PDE (1.1) are uniformly piecewise continuous with respect to $\mathcal{D}_h$; i.e., $\alpha \in C^0(\mathcal{D}_h)$ and $\boldsymbol{\beta} \in C^0(\mathcal{D}_h)$. Given any $v \in S_h$, let $\mathbf{q}_v = \alpha\nabla v + \boldsymbol{\beta} v$ be the corresponding flux. Then for any $\theta \in H^{-1}(\Omega)$, the following identity holds true:*

$$\begin{aligned}(\theta, v) = &\sum_{D\in\mathcal{D}_h}\sum_{e\subset\partial D} \langle(\mathbf{q}_v + p\mathbf{n}_e)\cdot\mathbf{n}, w - Q_0 w\rangle_e + (\nabla_w \cdot p, w)\\ &+ \sum_{D\in\mathcal{D}_h} ((I - Q_0)\mathbf{q}_v, \nabla w)_D + \sum_{D\in\mathcal{D}_h}\sum_{e\subset\partial D} \langle(Q_0 - I)\mathbf{q}_v\cdot\mathbf{n}, w - Q_0 w\rangle_e,\end{aligned} \tag{3.3}$$

where $p \in V_h$ is arbitrary, Q_0 is the L^2 projection onto W_h, and $w \in H_0^1(\Omega)$ is the solution of the auxiliary problem (3.1).

Proof. As $v \in S_h \subset H_0^1(\Omega)$, we have from (3.1)

$$\begin{aligned}(\theta, v) &= (\alpha\nabla v + \boldsymbol{\beta} v, \nabla w)\\ &= \sum_{D\in\mathcal{D}_h} (Q_0\mathbf{q}_v, \nabla w)_D + \sum_{D\in\mathcal{D}_h} ((I-Q_0)\mathbf{q}_v, \nabla w)_D\\ &= \sum_{D\in\mathcal{D}_h}\sum_{e\subset\partial D} \langle Q_0\mathbf{q}_v\cdot\mathbf{n}, w\rangle_e + \sum_{D\in\mathcal{D}_h} ((I-Q_0)\mathbf{q}_v, \nabla w)_D\\ &= \sum_{D\in\mathcal{D}_h}\sum_{e\subset\partial D} \langle (Q_0\mathbf{q}_v + p\mathbf{n}_e)\cdot\mathbf{n}, w\rangle_e + \sum_{D\in\mathcal{D}_h} ((I-Q_0)\mathbf{q}_v, \nabla w)_D,\end{aligned} \tag{3.4}$$

where we have used the fact that $\sum_{D\in\mathcal{D}_h}\sum_{e\subset\partial D}\langle p\mathbf{n}_e\cdot\mathbf{n}, w\rangle_e = 0$ for any $p \in V_h$ in the last equality. From $\int_{\partial D} Q_0\mathbf{q}_v\cdot\mathbf{n}ds = 0$ we arrive at

$$\sum_{D\in\mathcal{D}_h}\sum_{e\subset\partial D} \langle (Q_0\mathbf{q}_v + p\mathbf{n}_e)\cdot\mathbf{n}, Q_0 w\rangle_e = (\nabla_w\cdot p, Q_0 w) = (\nabla_w\cdot p, w).$$

Using the identity in (3.4) we obtain

$$\begin{aligned}(\theta, v) &= \sum_{D\in\mathcal{D}_h}\sum_{e\subset\partial D} \langle (Q_0\mathbf{q}_v + p\mathbf{n}_e)\cdot\mathbf{n}, w - Q_0 w\rangle_e\\ &\quad + \sum_{D\in\mathcal{D}_h} ((I-Q_0)\mathbf{q}_v, \nabla w)_D + (\nabla_w\cdot p, w)\\ &= \sum_{D\in\mathcal{D}_h}\sum_{e\subset\partial D} \langle (\mathbf{q}_v + p\mathbf{n}_e)\cdot\mathbf{n}, w - Q_0 w\rangle_e + \sum_{D\in\mathcal{D}_h} ((I-Q_0)\mathbf{q}_v, \nabla w)_D\\ &\quad + \sum_{D\in\mathcal{D}_h}\sum_{e\subset\partial D} \langle (Q_0 - I)\mathbf{q}_v\cdot\mathbf{n}, w - Q_0 w\rangle_e + (\nabla_w\cdot p, w),\end{aligned} \tag{3.5}$$

which gives the identity (3.3). ∎

In the space $S_h \times V_h$, we introduce the following seminorm

$$\|(v,p)\|_h = \left(J_2(v,p) + \|\nabla_w\cdot p\|_{-1}^2\right)^{\frac{1}{2}}, \qquad (v,p)\in S_h\times V_h. \tag{3.6}$$

An application of Lemma 3.3 shows that $\|(v,p)\|_h$ indeed defines a norm in the discrete space $S_h \times V_h$. More precisely, we have the following result.

LEMMA 3.4. *Assume that the coefficients α and $\boldsymbol{\beta}$ are in $C^0(\mathcal{D}_h)$. Then for any $v \in S_h$ and $p \in V_h$, we have*

$$\|v\|_1 \le C\left((J_2(v,p)^{1/2} + \|\nabla_w\cdot p\|_{-1}\right), \tag{3.7}$$

provided that the mesh size h is sufficiently small. Consequently, the seminorm $\|\cdot\|_h$ becomes a norm in the space $S_h \times V_h$ when the mesh size is sufficiently small.

Proof. Using the identity (3.3) we have

(3.8)
$$\begin{aligned}|(\theta,v)| &\le \sum_{D\in\mathcal{D}_h}\sum_{e\subset\partial D}|\langle(\mathbf{q}_v+p\mathbf{n}_e)\cdot\mathbf{n},w-Q_0w\rangle_e| + \|\nabla_d\cdot p\|_{-1}\|w\|_1\\ &\quad+\sum_{D\in\mathcal{D}_h}|((I-Q_0)\mathbf{q}_v,\nabla w))_D| + \sum_{D\in\mathcal{D}_h}\sum_{e\subset\partial T}|\langle(Q_0-I)\mathbf{q}_v\cdot\mathbf{n},w-Q_0w\rangle_e|\\ &\le \sum_{D\in\mathcal{D}_h}\sum_{e\subset\partial D}\|(\mathbf{q}_v+p\mathbf{n}_e)\cdot\mathbf{n}\|_e\|w-Q_0w\|_e + \|\nabla_d\cdot p\|_{-1}\|w\|_1\\ &\quad+\sum_{D\in\mathcal{D}_h}\|(I-Q_0)\mathbf{q}_v\|_D\|\nabla w\|_D + \sum_{D\in\mathcal{D}_h}\sum_{e\subset\partial D}\|(Q_0-I)\mathbf{q}_v\cdot\mathbf{n}\|_e\|w-Q_0w\|_e.\end{aligned}$$

Each term on the right-hand side of (3.8) can be handled, respectively, as follows. From the Cauchy–Schwarz inequality and the trace inequality, we may estimate the first term by

(3.9)
$$\begin{aligned}&\sum_{D\in\mathcal{D}_h}\sum_{e\subset\partial D}\|(\alpha\nabla v+\boldsymbol{\beta}v+p\mathbf{n}_e)\cdot\mathbf{n}\|_e\|w-Q_0w\|_e\\ &\le C\left(\sum_{D\in\mathcal{D}_h}h_D\|(\alpha\nabla v+\boldsymbol{\beta}v+p\mathbf{n}_e)\cdot\mathbf{n}\|^2_{\partial D}\right)^{1/2}(h_D^{-1}\|w-Q_0w\|_{0,D}+\|\nabla w\|_{0,D})\\ &\le CJ_2(v,p)^{1/2}\|\nabla w\|_0.\end{aligned}$$

Next, observe that, on each control element D, $(I-Q_0)\mathbf{q}_v$ may be rewritten as

$$(I-Q_0)\mathbf{q}_v = (I-Q_0)\left((\alpha-\overline{\alpha})\nabla v+\boldsymbol{\beta}v-\overline{\boldsymbol{\beta}}\overline{v}\right), \tag{3.10}$$

where $\overline{\sigma}$ stands for the cell average of the underlying function σ over the control element D. It follows that

$$\begin{aligned}\|(I-Q_0)\mathbf{q}_v\|_{0,D} &\le \|(\alpha-\overline{\alpha})\nabla v+\boldsymbol{\beta}v-\overline{\boldsymbol{\beta}}\overline{v}\|_{0,D}\\ &\le \|(\alpha-\overline{\alpha})\nabla v\|_{0,D}+\|\overline{\boldsymbol{\beta}}(v-\overline{v})\|_{0,D}+\|(\boldsymbol{\beta}-\overline{\boldsymbol{\beta}})v\|_{0,D}\\ &\le \operatorname{osc}(\alpha,D)\|\nabla v\|_{0,D}+\operatorname{osc}(\boldsymbol{\beta},D)\|v\|_{0,D}+Ch_D\|\nabla v\|_{0,D}\\ &\le (\operatorname{osc}(\alpha,D)+\operatorname{osc}(\boldsymbol{\beta},D)+Ch_D)\left(\|\nabla v\|_{0,D}+\|v\|_{0,D}\right).\end{aligned}$$

Thus, the third term has the following estimate:

$$\begin{aligned}&\sum_{D\in\mathcal{D}_h}\|(I-Q_0)\mathbf{q}_v\|_D\|\nabla w\|_D\\ &\le \sum_{D\in\mathcal{D}_h}(\operatorname{osc}(\alpha,D)+\operatorname{osc}(\boldsymbol{\beta},D)+Ch_D)\left(\|\nabla v\|_{0,D}+\|v\|_{0,D}\right)\|\nabla w\|_D\\ &\le (\operatorname{osc}(\alpha,\mathcal{D}_h)+\operatorname{osc}(\boldsymbol{\beta},\mathcal{D}_h)+Ch)\|v\|_1\|w\|_1.\end{aligned} \tag{3.11}$$

As for the fourth term, we again use the decomposition (3.10) to arrive at

$$\begin{aligned}&\sum_{D\in\mathcal{D}_h}\sum_{e\subset\partial D}\|(Q_0-I)\mathbf{q}_v\cdot\mathbf{n}\|_e\|w-Q_0w\|_e\\ &\le \sum_{D\in\mathcal{D}_h}\sum_{e\subset\partial D}\|(I-Q_0)\left((\alpha-\overline{\alpha})\nabla v+\boldsymbol{\beta}v-\overline{\boldsymbol{\beta}}\overline{v}\right)\|_e\|w-Q_0w\|_e.\end{aligned} \tag{3.12}$$

For any function σ defined on $D \in \mathcal{D}_h$, from the trace and inverse inequality we have

$$\|Q_0\sigma\|_e \le C(h^{-1}\|Q_0\sigma\|_{0,D}^2 + h\|\nabla(Q_0\sigma)\|_{0,D}^2) \le Ch^{-1}\|Q_0\sigma\|_{0,D}^2.$$

It follows that

$$\begin{aligned}
&\|(I-Q_0)\left((\alpha-\overline{\alpha})\nabla v + \boldsymbol{\beta} v - \overline{\boldsymbol{\beta}}\overline{v}\right)\|_e \\
&\quad \le \|(\alpha-\overline{\alpha})\nabla v + \boldsymbol{\beta} v - \overline{\boldsymbol{\beta}}\overline{v}\|_e + \|Q_0\left((\alpha-\overline{\alpha})\nabla v + \boldsymbol{\beta} v - \overline{\boldsymbol{\beta}}\overline{v}\right)\|_e \\
&\quad \le \|(\alpha-\overline{\alpha})\nabla v + \boldsymbol{\beta} v - \overline{\boldsymbol{\beta}}\overline{v}\|_e + Ch^{-1/2}\|Q_0\left((\alpha-\overline{\alpha})\nabla v + \boldsymbol{\beta} v - \overline{\boldsymbol{\beta}}\overline{v}\right)\|_{0,D} \\
&\quad \le \|(\alpha-\overline{\alpha})\nabla v + \boldsymbol{\beta} v - \overline{\boldsymbol{\beta}}\overline{v}\|_e + Ch^{-1/2}\|(\alpha-\overline{\alpha})\nabla v + \boldsymbol{\beta} v - \overline{\boldsymbol{\beta}}\overline{v}\|_{0,D} \\
&\quad \le (\mathrm{osc}(\alpha, D) + \mathrm{osc}(\boldsymbol{\beta}, D) + Ch_D)(\|\nabla v\|_e + \|v\|_e) \\
&\qquad + Ch^{-1/2}\|(\alpha-\overline{\alpha})\nabla v + \boldsymbol{\beta} v - \overline{\boldsymbol{\beta}}\overline{v}\|_{0,D} \\
&\quad \le C(\mathrm{osc}(\alpha, D) + \mathrm{osc}(\boldsymbol{\beta}, D) + h_D)h^{-1/2}(\|\nabla v\|_D + \|v\|_D).
\end{aligned}$$

Substituting the above estimate into (3.12) yields

(3.13)
$$\begin{aligned}
&\sum_{D\in\mathcal{D}_h}\sum_{e\subset\partial D}\|(Q_0-I)\mathbf{q}_v\cdot\mathbf{n}\|_e\|w-Q_0w\|_e \\
&\quad \le C\sum_{D\in\mathcal{D}_h}(\mathrm{osc}(\alpha, D) + \mathrm{osc}(\boldsymbol{\beta}, D) + h_D)h^{-1/2}(\|\nabla v\|_D + \|v\|_D)\|w-Q_0w\|_{\partial D} \\
&\quad \le C(\mathrm{osc}(\alpha, \mathcal{D}_h) + \mathrm{osc}(\boldsymbol{\beta}, \mathcal{D}_h) + h)\|v\|_1\|w\|_1.
\end{aligned}$$

Now by combining (3.8) with (3.9), (3.11), and (3.13) we arrive at

(3.14)
$$\begin{aligned}
(\theta, v) &\le CJ_2(v,p)^{1/2}\|\nabla w\|_0 + \|\nabla_w\cdot p\|_{-1}\|w\|_1 \\
&\quad + C(\mathrm{osc}(\alpha, \mathcal{D}_h) + \mathrm{osc}(\boldsymbol{\beta}, \mathcal{D}_h) + h)\|v\|_1\|w\|_1 \\
&\le C(J_2(v,p)^{1/2} + \|\nabla_w\cdot p\|_{-1} + (\mathrm{osc}(\alpha, \mathcal{D}_h) + \mathrm{osc}(\boldsymbol{\beta}, \mathcal{D}_h) + h)\|v\|_1)\|\theta\|_{-1}.
\end{aligned}$$

It follows that

$$\|v\|_1 \le C\left(J_2(v,p)^{1/2} + \|\nabla_w\cdot p\|_{-1} + (\mathrm{osc}(\alpha, \mathcal{D}_h) + \mathrm{osc}(\boldsymbol{\beta}, \mathcal{D}_h) + h)\|v\|_1\right),$$

which, with the assumption of $\alpha \in C^0(\mathcal{D}_h)$ and $\boldsymbol{\beta} \in C^0(\mathcal{D}_h)$, implies the estimate (3.7) for sufficiently small mesh size h.

To show that $\|\cdot\|_h$ defines a norm in $S_h \times V_h$, for any $(v,p) \in S_h \times V_h$ satisfying $\|(v,p)\|_h = 0$ we have $J_2(v,p) = 0$ and $\nabla_w \cdot p = 0$. It then follows from (3.7) that $v = 0$. From $S(v,p) = 0$ and $v = 0$, we further have $0 = p + (\alpha\nabla v + \boldsymbol{\beta} v)\cdot\mathbf{n}_e = p$ on each edge $e \in \mathcal{E}_h$. This completes the proof of the lemma. ∎

In the space W_h, we introduce a discrete H^1 norm as follows:

$$\|\sigma\|_{1,h} = \left(\sum_{e\in\mathcal{E}_h}[\![\sigma]\!]_e^2\right)^{1/2},$$

where $[\![\sigma]\!]_e = \sigma|_{D_L} - \sigma|_{D_R}$ is the jump of the piecewise constant function σ on the edge $e \in \mathcal{E}_h$ shared by two control elements D_L and D_R. For e on the boundary $\partial\Omega$,

we assume D_R is empty so that a one-sided trace of σ will be taken as the value of $[\![\sigma]\!]_e$.

Next, we equip the space V_h with the following L^2-norm:

$$\|p\|_0 = \left(\sum_{e \in \mathcal{E}_h} h_e \int_e p^2 ds \right)^{1/2}.$$

LEMMA 3.5. *For any $\sigma \in W_h$, there exists a discrete flux $p_\sigma \in V_h$ such that*

$$(\nabla_w \cdot p_\sigma, \sigma) = \|\sigma\|_{1,h}^2, \qquad \|p_\sigma\|_0 \le C \|\sigma\|_{1,h}.$$

Proof. From the computational formula (2.1) for $\nabla_w \cdot p$, we have

$$\begin{aligned}(\nabla_w \cdot p, \sigma) &= \sum_{D \in \mathcal{D}_h} \int_{e \subset \partial D} p|_e \mathbf{n}_e \cdot \mathbf{n} \sigma ds \\ &= \sum_{e \in \mathcal{E}_h} \int_e [\![\sigma]\!]_e p|_e ds.\end{aligned}$$

By choosing $p_\sigma|_e = [\![\sigma]\!]_e / h_e$ we arrive at

$$(\nabla_w \cdot p_\sigma), \sigma) = \sum_{e \in \mathcal{E}_h} [\![\sigma]\!]_e^2 = \|\sigma\|_{1,h}^2.$$

Furthermore, it is easy to see that

$$\|p_\sigma\|_0^2 = \sum_{e \in \mathcal{E}_h} h_e \int_e p_\sigma^2 ds = \sum_{e \in \mathcal{E}_h} [\![\sigma]\!]_e^2 = \|\sigma\|_{1,h}^2.$$

This completes the proof of the lemma. ∎

THEOREM 3.6. *Assume that the coefficients α and $\boldsymbol{\beta}$ are in $C^0(\mathcal{D}_h)$. Then the numerical scheme* (2.13)–(2.14) *has one and only one solution for u_h, q_h, and λ_h, provided that the mesh size h is sufficiently small.*

Proof. It suffices to show that the homogeneous problem has only a trivial solution. To this end, let $u_h \in S_h$, $q_h \in V_h$, and $\lambda_h \in W_h$ be the solution of (2.13)–(2.14) with homogeneous data $f = 0$. From (2.14) we have $\nabla_w \cdot q_h = 0$ on each control element $D \in \mathcal{D}_h$. Next, by choosing $(v, p) = (u_h, q_h)$ in (2.13) and using $(\nabla_w \cdot q_h, \lambda_h) = 0$ we obtain

$$J_2(u_h, q_h) = 0,$$

which, together with $\nabla_w \cdot q_h = 0$, leads to $\|(u_h, q_h)\|_h = 0$ by using (3.6). It follows that $u_h = 0$ and $q_h = 0$. Thus, from (2.13) we obtain

$$(\nabla_w \cdot p, \lambda_h) = 0 \qquad \forall p \in V_h.$$

Next, from the *inf-sup* result of Lemma 3.5 there exists a discrete flux $p_{\lambda_h} \in V_h$ satisfying

$$\|\lambda_h\|_{1,h}^2 = (\nabla_w \cdot p_{\lambda_h}, \lambda_h) = 0,$$

which gives rise to $[\![\lambda_h]\!] = 0$ on each edge $e \in \mathcal{E}_h$ (including the boundary edge) and, hence, $\lambda_h \equiv 0$. This completes the proof of the theorem. ∎

4. Error estimates. We are now in a position to establish some error estimates for the approximate solutions u_h and q_h arising from the scheme (2.13)–(2.14).

THEOREM 4.1. *Assume that the coefficients of the model problem* (1.1) *are uniformly piecewise continuous with respect to* $\mathcal{D}_h$, *i.e.,* $\alpha \in C^0(\mathcal{D}_h)$ *and* $\boldsymbol{\beta} \in C^0(\mathcal{D}_h)$. *Let* u *be the exact solutions of* (1.1)–(1.2), $u_h \in S_h$, $q_h \in V_h$, *and* $\lambda_h \in W_h$ *be the approximate solution arising from the numerical scheme* (2.13)–(2.14). *If* $u \in H^2(\Omega)$, *then the following error estimate holds true:*

$$\|u-u_h\|_1 \le C(\|\alpha\|_\infty h\|u\|_2 + \|\boldsymbol{\beta}\|_\infty h\|u\|_1 + \mathrm{osc}(\alpha,\mathcal{D}_h)\|u\|_1 + \mathrm{osc}(\boldsymbol{\beta},\mathcal{D}_h)\|u\|_0). \tag{4.1}$$

In particular, if the coefficients α *and* $\boldsymbol{\beta}$ *are differentiable on each control element* $D \in \mathcal{D}_h$, *then*

$$\|u-u_h\|_1 \le Ch\|u\|_2, \tag{4.2}$$

where $C = C(\alpha,\boldsymbol{\beta})$ *is a constant depending on the elementwise* $W^{1,\infty}$ *norm of the coefficients* α *and* $\boldsymbol{\beta}$.

Proof. Let u_I be the usual nodal point interpolation of the exact solution u. Denote by $Q_b q$ the piecewise constant interpolation of the exact flux $q = -(\alpha\nabla u + \boldsymbol{\beta}u)\cdot\mathbf{n}_e$ in the finite element space V_h. Let the error functions be given by

$$e_h = u_h - u_I, \quad \eta_h = q_h - Q_b q, \quad \xi_h = \lambda_h - 0.$$

It is not hard to see that $\nabla_w \cdot \eta_h = 0$. From the estimate (3.7) we have

$$\|e_h\|_1 \le CJ_2(e_h,\eta_h)^{\frac{1}{2}}. \tag{4.3}$$

Thus, it suffices to derive an estimate for $J_2(e_h,\eta_h)$. To this end, we observe that the triplet $(u_I, Q_b q, \lambda_I = 0)$ satisfies

$$s_h((u_I,Q_bq),(v,p)) + (\nabla_w\cdot p,\lambda_I) = s_h((u_I,Q_bq),(v,p)) \quad \forall v\in S_h, p\in V_h, \tag{4.4}$$

$$(\nabla_w\cdot(Q_bq),w) = (f,w) \quad \forall w\in W_h. \tag{4.5}$$

By subtracting (4.4) from (2.13) and (4.5) from (2.14), we arrive at the following error equations,

$$s_h((e_h,\eta_h),(v,p)) + (\nabla_w\cdot p,\xi_h) = -s_h((u_I,Q_bq),(v,p)) \quad \forall v\in S_h,\ p\in V_h, \tag{4.6}$$

$$(\nabla_w\cdot\eta_h,w) = 0 \qquad \forall w\in W_h. \tag{4.7}$$

By letting $(v,p) = (e_h,\eta_h)$ in (4.6) we further obtain

$$J_2(e_h,\eta_h) = -s_h((u_I,Q_bq),(e_h,\eta_h)).$$

Using the Cauchy–Schwarz inequality on the right-hand side of the above equality gives

$$J_2(e_h,\eta_h) \le J_2(u_I,Q_bq). \tag{4.8}$$

To estimate the right-hand side of (4.8), we use the definition of $J_2(u_I, Q_b q)$ to obtain

(4.9)
$$\begin{aligned} J_2(u_I, Q_b q) &= \sum_{D\in\mathcal{D}_h}\sum_{e\subset\partial D} h_D \|(\alpha\nabla u_I + \boldsymbol{\beta} u_I)\cdot \mathbf{n}_e + Q_b q\|_e^2 \\ &= \sum_{D\in\mathcal{D}_h}\sum_{e\subset\partial D} h_D \|\alpha\nabla u_I\cdot\mathbf{n}_e - Q_b(\alpha\nabla u\cdot\mathbf{n}_e) + \boldsymbol{\beta}u_I\cdot\mathbf{n}_e - Q_b(\boldsymbol{\beta}u\cdot\mathbf{n}_e)\|_e^2 \\ &\le 2\sum_{D\in\mathcal{D}_h} h_D\left(\|\alpha\nabla u_I - Q_b(\alpha\nabla u)\|_{\partial D}^2 + \|\boldsymbol{\beta}u_I - Q_b(\boldsymbol{\beta}u)\|_{\partial D}^2\right) \\ &\le 2\sum_{D\in\mathcal{D}_h} h_D\left(\|\alpha\nabla u_I - \alpha\nabla u\|_{\partial D}^2 + \|\alpha\nabla u - Q_b(\alpha\nabla u)\|_{\partial D}^2\right) \\ &\quad + 2\sum_{D\in\mathcal{T}_h} h_D\left(\|\boldsymbol{\beta}u_I - \boldsymbol{\beta}u\|_{\partial D}^2 + \|\boldsymbol{\beta}u - Q_b(\boldsymbol{\beta}u)\|_{\partial D}^2\right). \end{aligned}$$

On each control element $D\in\mathcal{D}_h$, let $\overline{g}$ be the average of the function $g=g(x)$ on D. It follows that

$$\|\alpha\nabla u - Q_b(\alpha\nabla u)\|_{\partial D} \le \|(\alpha-\overline{\alpha})\nabla u\|_{\partial D} \tag{4.10}$$

and

$$\begin{aligned} \|\boldsymbol{\beta}u - Q_b(\boldsymbol{\beta}u)\|_{\partial D} &\le \|\boldsymbol{\beta}u - \overline{\boldsymbol{\beta}}\,\overline{u}\|_{\partial D} \\ &\le \|\boldsymbol{\beta}u - \overline{\boldsymbol{\beta}}u\|_{\partial D} + \|\overline{\boldsymbol{\beta}}u - \overline{\boldsymbol{\beta}}\,\overline{u}\|_{\partial D} \\ &= \|(\boldsymbol{\beta}-\overline{\boldsymbol{\beta}})u\|_{\partial D} + \|\overline{\boldsymbol{\beta}}(u-\overline{u})\|_{\partial D}. \end{aligned} \tag{4.11}$$

Now substituting (4.10) and (4.11) into (4.9) yields

(4.12)
$$\begin{aligned} J_2(u_I, Q_b q) &\le 2\sum_{D\in\mathcal{D}_h} h_D\left(\|\alpha\nabla(u_I-u)\|_{\partial D}^2 + \|(\alpha-\overline{\alpha})\nabla u\|_{\partial D}^2\right) \\ &\quad + 2\sum_{D\in\mathcal{D}_h} h_D\left(\|\boldsymbol{\beta}(u_I-u)\|_{\partial D}^2 + \|(\boldsymbol{\beta}-\overline{\boldsymbol{\beta}})u\|_{\partial D}^2 + \|\overline{\boldsymbol{\beta}}(u-\overline{u})\|_{\partial D}^2\right) \\ &\le C(\|\alpha\|_\infty^2 h^2\|u\|_2^2 + \|\boldsymbol{\beta}\|_\infty^2 h^2\|u\|_1^2 + \mathrm{osc}(\alpha,\mathcal{D}_h)^2\|u\|_1^2 + \mathrm{osc}(\boldsymbol{\beta},\mathcal{D}_h)^2\|u\|_0^2). \end{aligned}$$

By combining (4.8) with (4.12) we obtain

$$J_2(e_h,\eta_h)^{\frac{1}{2}} \le C(\|\alpha\|_\infty h\|u\|_2 + \|\boldsymbol{\beta}\|_\infty h\|u\|_1 + \mathrm{osc}(\alpha,\mathcal{D}_h)\|u\|_1 + \mathrm{osc}(\boldsymbol{\beta},\mathcal{D}_h)\|u\|_0).$$

Finally, substituting the above estimate into (4.3) yields

$$\|u_h - u_I\|_1 \le C(\|\alpha\|_\infty h\|u\|_2 + \|\boldsymbol{\beta}\|_\infty h\|u\|_1 + \mathrm{osc}(\alpha,\mathcal{D}_h)\|u\|_1 + \mathrm{osc}(\boldsymbol{\beta},\mathcal{D}_h)\|u\|_0),$$

which completes the proof of the theorem. ∎

We end this section by showing that q_h indeed provides a very accurate approximation of the exact flux. For any $m\in L^2(\mathcal{E}_h)$, define its L^2 norm by

$$|\!|\!| m |\!|\!|_0 = \left(\sum_{D\in\mathcal{D}_h}\sum_{e\subset\partial D} h_e\int_e m^2 ds\right)^{\frac{1}{2}}.$$

THEOREM 4.2. *Under the assumptions of Theorem* 4.1, *let* $q = -(\alpha\nabla u + \boldsymbol{\beta}u)\cdot\mathbf{n}_e$ *be the exact flux on* $e \in \mathcal{E}_h$ *along the direction* $\mathbf{n}_e$. *If the exact solution* $u \in H^2(\Omega)$, *then the following error estimate holds true:*

$$|||q - q_h|||_0 \le C(\|\alpha\|_\infty h\|u\|_2 + \|\boldsymbol{\beta}\|_\infty h\|u\|_1 + \mathrm{osc}(\alpha,\mathcal{D}_h)\|u\|_1 + \mathrm{osc}(\boldsymbol{\beta},\mathcal{D}_h)\|u\|_0). \tag{4.13}$$

In particular, if the coefficients α *and* $\boldsymbol{\beta}$ *are differentiable on each control element* $D \in \mathcal{D}_h$, *then*

$$|||q - q_h|||_0 \le Ch\|u\|_2, \tag{4.14}$$

where $C = C(\alpha,\boldsymbol{\beta})$ *is a constant depending on the elementwise* $W^{1,\infty}$ *norm of the coefficients* α *and* $\boldsymbol{\beta}$.

Proof. From $e_h = u_h - u_I$ and $\eta_h = q_h - Q_b q$ we have

$$J_2(u_h,q_h) = s_h((e_h,\eta_h),(u_h,q_h)) + s_h((u_I,Q_bq),(u_h,q_h)).$$

Using the Cauchy–Schwarz inequality and the estimate (4.8) one arrives at

$$J_2(u_h,q_h)^{\frac12} \le J_2(e_h,\eta_h)^{\frac12} + J_2(u_I,Q_bq)^{\frac12} \le 2J_2(u_I,Q_bq)^{\frac12},$$

which, by the estimate (4.12), leads to

$$J_2(u_h,q_h)^{\frac12} \le C(\|\alpha\|_\infty h\|u\|_2 + \|\boldsymbol{\beta}\|_\infty h\|u\|_1 + \mathrm{osc}(\alpha,\mathcal{D}_h)\|u\|_1 + \mathrm{osc}(\boldsymbol{\beta},\mathcal{D}_h)\|u\|_0). \tag{4.15}$$

As $q = -(\alpha\nabla u + \boldsymbol{\beta}u)\cdot\mathbf{n}_e$, we have

$$\begin{aligned} J_2(u_h,q) &= \sum_{D\in\mathcal{D}_h}\sum_{e\subset\partial D} h_e\int_e |\alpha\nabla(u_h-u)\cdot\mathbf{n}_e + \boldsymbol{\beta}(u_h-u)\cdot\mathbf{n}_e|^2 ds \\ &\le 2\sum_{D\in\mathcal{D}_h}\sum_{e\subset\partial D} h_e\int_e \left(\|\alpha\|^2_{\infty,D}|\nabla(u_h-u)|^2 ds + \|\boldsymbol{\beta}\|^2_{\infty,D}|u_h-u|^2\right) ds \\ &\le C(\|\alpha\|_\infty,\|\boldsymbol{\beta}\|_\infty)(h^2\|u\|_2^2 + \|u-u_h\|_1^2), \end{aligned} \tag{4.16}$$

where the usual trace inequality has been employed in the last line. Using the error estimate (4.2) we arrive at

$$J_2(u_h,q)^{\frac12} \le C(\|\alpha\|_\infty,\|\boldsymbol{\beta}\|_\infty)h\|u\|_2.$$

Consequently,

$$|||q - q_h|||_0 \le C(J_2(u_h,q_h)^{\frac12} + J_2(u_h,q)^{\frac12}) \le Ch\|u\|_2,$$

provided that the coefficients α and $\boldsymbol{\beta}$ are differentiable locally on each control element $D \in \mathcal{D}_h$. This proves the theorem. ∎

5. Numerical experiments. The purpose of this section is twofold. First, we shall numerically verify the theoretical error estimates developed in the previous section by applying the algorithm (2.13)–(2.14) to the elliptic problem (1.1)–(1.2). Second, we will demonstrate the significance of having a conservative numerical flux by applying the CFO algorithm to a simplified two-phase porous media flow problem with strong heterogeneity.

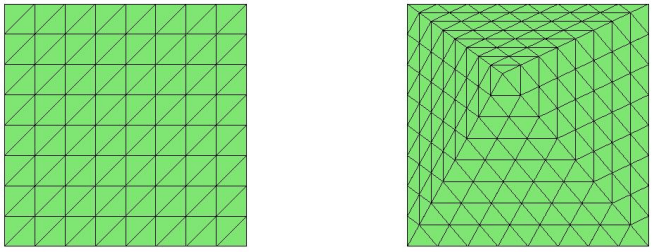


FIG. 2. *Uniform and nonuniform triangular partitions with mesh size* $h = 1/8$.

TABLE 1
Error and convergence performance of the CFO scheme for Test Case 1 *on uniform meshes.*

h^{-1}	$\|u_h - u\|_0$	Order	$\|u_h - u\|_1$	Order	$J_2(u_h, q_h)^{\frac{1}{2}}$	Order	$\|\lambda_h\|_0$	Order
2	0.234		1.54		4.49		0.246	
4	9.53e-2	1.3	0.860	0.8	2.59	0.8	0.102	1.3
8	2.92e-2	1.7	0.437	1.0	1.34	1.0	3.06e-2	1.7
16	7.80e-3	1.9	0.218	1.0	0.676	1.0	8.12e-3	1.9
32	1.99e-3	2.0	0.109	1.0	0.339	1.0	2.07e-3	2.0
64	4.99e-4	2.0	5.45e-2	1.0	0.169	1.0	5.18e-4	2.0
128	1.25e-4	2.0	2.73e-2	1.0	8.47e-2	1.0	1.30e-4	2.0

5.1. Elliptic equations. The elliptic test problem is defined in a two-dimensional square domain which seeks $u \in H^1(\Omega)$ satisfying

$$\begin{aligned} -\nabla \cdot (\alpha \nabla u) &= f \qquad \text{in } \Omega, \\ u &= g \qquad \text{on } \partial\Omega. \end{aligned} \tag{5.1}$$

The CFO algorithm (2.8) with $r = 2$ (i.e., the numerical scheme (2.13)–(2.14)) is implemented for each test case with numerical solutions denoted as u_h, q_h, and λ_h in the reporting tables. The following metrics are used to measure the magnitude of the error:

$$\begin{aligned} L^2\text{-norm: } & \|u_h - u\|_0 = \left(\sum_{T \in \mathcal{T}_h} \int_T |u_h - u|^2 dT \right)^{1/2}, \\ H^1\text{-norm: } & \|u_h - u\|_1 = \|\nabla(u_h - u)\|_0, \\ \text{residual-error: } & J_2(u_h, q_h)^{\frac{1}{2}} = \left(\sum_{D \in \mathcal{D}_h} h_D \int_{\partial D} |\alpha \nabla u_h \cdot \mathbf{n}_e + q_h|^2 ds \right)^{1/2}. \end{aligned}$$

5.1.1. Test Case 1: Smooth coefficients. In this experiment, the elliptic problem has the exact solution $u = \cos(\pi x)\cos(\pi y)$ with domain $\Omega = (0,1)^2$, the coefficient α is the identity matrix, and the right-hand side function f and the Dirichlet boundary data g are chosen to match the exact solution. The control element partition $\mathcal{D}_h$ is identical to the finite element triangulation $\mathcal{T}_h$. To numerically demonstrate the effect of the finite element partition $\mathcal{T}_h$, both uniform and nonuniform partitions are considered in our computation; see Figure 2. Table 1 shows the performance of the CFO algorithm for the test problem on a uniform mesh. It is clear that the errors

TABLE 2
Error and convergence performance of the CFO scheme for Test Case 1 *on nonuniform meshes.*

h^{-1}	$\|u_h - u\|_0$	Order	$\|u_h - u\|_1$	Order	$J_2(u_h, q_h)^{\frac{1}{2}}$	Order	$\|\lambda_h\|_0$	Order
2	0.147		1.09		2.63		0.125	
4	4.07e-2	1.9	0.571	0.9	1.59	0.7	3.51e-2	1.8
8	1.10e-2	1.9	0.288	1.0	0.836	0.9	9.23e-3	1.9
16	2.87e-3	1.9	0.144	1.0	0.424	1.0	2.36e-3	2.0
32	7.32e-4	2.0	7.21e-2	1.0	0.213	1.0	5.95e-4	2.0
64	1.84e-4	2.0	3.60e-2	1.0	0.106	1.0	1.49e-4	2.0
128	4.62e-5	2.0	1.80e-2	1.0	5.32e-2	1.0	3.73e-5	2.0

TABLE 3
The numerical performance of the CFO scheme (2.13)–(2.14) *for Test Case* 2 *with Hölder continuous coefficients.*

h^{-1}	$\|u_h - u\|_0$	Order	$\|u_h - u\|_1$	Order	$\|\|q - q_h\|\|_0$	Order
2	0.769		3.27		9.15	
4	0.265	1.5	1.74	0.9	5.19	0.8
8	7.15e-2	1.9	0.871	1.0	2.67	1.0
16	1.79 e-2	2.0	0.436	1.0	1.34	1.0
32	4.36e-3	2.0	0.218	1.0	0.67	1.0
64	1.05e-3	2.0	0.109	1.0	0.337	1.0
128	2.58e-4	2.0	5.47e-2	1.0	0.168	1.0

$\|u_h - u\|_0$ and $\|\lambda_h\|_0$ converge to zero at the rate of h^2, and that $\|\nabla(u_h - u)\|_{L^2}$ and the residual error $J_2(u_h, q_h)^{\frac{1}{2}}$ converge at the rate of h. The results are in good consistency with the theory. The numerical experiments on the irregular meshes as shown in Table 2 suggest the same convergence as on the uniform meshes.

5.1.2. Test Case 2: Hölder continuous coefficients. The coefficient matrix α in this test is given by

$$\alpha = \begin{pmatrix} 1 + |x| & 0.5|x|^{\frac{1}{3}}|y|^{\frac{1}{3}} \\ 0.5|x|^{\frac{1}{3}}|y|^{\frac{1}{3}} & 1 + |y| \end{pmatrix} \tag{5.2}$$

on the square domain $\Omega = (-1, 1)^2$. The coefficient matrix is clearly nonsmooth, but Hölder continuous in the domain. The right-hand side function and the Dirichlet boundary data are chosen to match the exact solution of $u(x, y) = \cos(\pi x)\cos(\pi y)$. Note that this example has been considered in [35, 37].

We use the CFO algorithm of $r = 2$ (i.e., (2.13)–(2.14) to approximate the solution of the above elliptic problem. The set of control elements is again identical with the finite element partition $\mathcal{T}_h$ which is obtained by first uniformly partitioning the square domain Ω into N^2, $N = h^{-1}$, small squares and then decomposing each small square into 2 similar triangles. The corresponding error and convergence information are reported in Table 3. Note that u denotes the exact solution of (5.1), and $(u_h; q_h)$ is the solution of the constrained minimization problem (2.8) with $r = 2$. It can be seen from Table 3 that the convergence of the algorithm in both the H^1 and L^2 norms has optimal order of $k = 1$ and $k = 2$, respectively. Moreover, for the flux on element edges, the convergence is also of optimal order $k = 1$. These numerical results support strongly the theoretical findings in the previous section. In fact, the numerical results outperform the theory in H^1, as the coefficient α is nonsmooth nor Lipschitz continuous on each element so that no convergence of order $k = 1$ can be deduced from the theory. We point out that no theory of optimal order of convergence has

TABLE 4
The numerical performance of the CFO algorithm (2.13)–(2.14) *for Test Case* 3 *with a discontinuous coefficient.*

h	$\|u_h - u\|_0/\|u\|_0$	Order	$\|u_h - u\|_1/\|u\|_1$	Order	$\|\|\|q - q_h\|\|\|_0/\|\|\|q\|\|\|_0$	Order
1/4	2.43e-03		6.57e-02		7.59e-02	
1/8	6.71e-04	1.9	3.28e-02	1.00	3.04e-02	1.3
1/16	2.08e-04	1.7	1.64e-02	1.00	1.32e-02	1.2
1/32	6.16e-05	1.8	8.17e-03	1.00	6.12e-03	1.1
1/64	1.79e-05	1.8	4.08e-03	1.00	2.96e-03	1.1
1/128	5.19e-06	1.8	2.04e-03	1.00	1.46e-03	1.0
1/256	1.47e-06	1.8	1.02e-03	1.00	7.30e-04	1.0

been developed for CFO in the usual L^2 norm, though the numerics strongly suggest such a result in the L^2 norm. Interested readers are encouraged to study the L^2 convergence for the CFO algorithm.

We emphasize that the CFO algorithm (2.13)–(2.14) provides not only a discrete solution u_h with optimal order of convergence, but also an elementwise conserving flux with optimal order of convergence.

5.1.3. Test Case 3: Discontinuous coefficients. In this numerical test, the domain of the elliptic problem (5.1) is chosen as the unit square $\Omega = (0,1)^2$, and the coefficient α is given by

$$\alpha = \begin{cases} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} & \text{if } x < 0.5, \\ \begin{pmatrix} 10 & 3 \\ 3 & 1 \end{pmatrix} & \text{if } x \geq 0.5, \end{cases} \tag{5.3}$$

which is clearly discontinuous along the vertical line of $x = \frac{1}{2}$. With properly chosen data on the right-hand side function and the Dirichlet boundary value, the exact solution of (5.1) is given by

$$u(x,y) = \begin{cases} 1 - 2y^2 + 4xy + 6x + 2y & \text{if } x < 0.5, \\ -2y^2 + 1.6xy - 0.6x + 3.2y + 4.3 & \text{if } x \geq 0.5. \end{cases} \tag{5.4}$$

Table 4 illustrates the performance of the CFO scheme (2.13)–(2.14) when applied to the present test case. The results suggest an optimal order of convergence for the numerical approximation u_h in the usual H^1 norm, which is in great consistency with the error estimate developed in the previous section. Likewise, the numerical approximation for the flux variable q_h also has an optimal order of convergence, as predicted by the convergence theory. On the other hand, the convergence in L^2 for u_h seems to be around $k = 1.8$.

5.1.4. Test Case 4: Discontinuous coefficients. In this test case, the domain $\Omega = (-1,1)^2$ is split into four subdomains $\Omega = \bigcup_{i=1}^4 \Omega_i$ by the x and y axes. The diffusion coefficient α is given by

$$\alpha = \begin{pmatrix} \alpha_i^x & 0 \\ 0 & \alpha_i^y \end{pmatrix} \text{ if } (x,y) \in \Omega_i,$$

and the exact solution is given by $u(x,y) = \alpha_i \sin(2\pi x)\sin(2\pi y)$. Here the values of the coefficient α_i^x, α_i^y, and α_i are specified in Table 5. It is clear that the diffusion coefficient α is discontinuous across the lines $x = 0$ and $y = 0$.

TABLE 5
Test Case 4: *Parameter values for the diffusion coefficients and the exact solution.*

$\alpha_4^x = 0.1$ $\alpha_4^y = 0.01$ $\alpha_4 = 100$	$\alpha_3^x = 1000$ $\alpha_3^y = 100$ $\alpha_3 = 0.01$
$\alpha_1^x = 100$ $\alpha_1^y = 10$ $\alpha_1 = 0.1$	$\alpha_2^x = 1$ $\alpha_2^y = 0.1$ $\alpha_2 = 10$

TABLE 6
Relative error and convergence performance of the CFO algorithm (2.13)–(2.14) *for Test Case* 4 *with discontinuous coefficients.*

h	$\|u_h - u\|_0/\|u\|_0$	Order	$\|u_h - u\|_1/\|u\|_1$	Order	$\|\|\|q - q_h\|\|\|_0/\|\|\|q\|\|\|_0$	Order
1/4	8.46e-01		8.11e-01		6.15e-01	
1/8	4.67e-01	0.9	5.13e-01	0.7	3.84e-01	0.7
1/16	1.75e-01	1.4	2.45e-01	1.1	1.78e-01	1.1
1/32	5.07e-02	1.8	1.10e-01	1.2	7.69e-02	1.2
1/64	1.34e-02	1.9	5.15e-02	1.1	3.48e-02	1.1
1/128	3.42e-03	2.0	2.49e-02	1.1	1.66e-02	1.0
1/256	8.62e-04	2.0	1.23e-02	1.0	8.15e-03	1.0

Table 6 illustrates the numerical performance of the CFO scheme (2.13)–(2.14) when applied to the Test Case 4. The results suggest an optimal order of convergence for the numerical approximation u_h in the usual H^1 norm and the flux variable q_h in L^2. The numerical results are in great consistency with the error estimate developed in the previous section. It should be noted that the numerical results strongly suggest an optimal order of convergence in L^2 for the primal variable u_h.

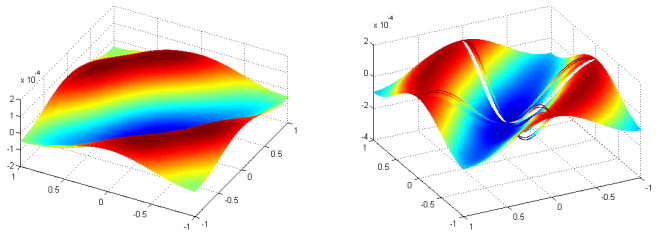
5.1.5. The Lagrange multiplier λ_h. The CFO algorithm involves two essential ideas in flux approximation: (1) the satisfaction of the mass conservation equation on the control element partition $\mathcal{D}_h$, and (2) the minimization of the object function $J_r(v, p)$ defined as in (2.7). As the value of the PDE coefficients may vary from element to element, one may modify the object function as follows by placing a weight τ_D on each element:

$$J_r^*(p, v) := \frac{1}{r} \sum_{D \in \mathcal{D}_h} \sum_{e \subset \partial D} \tau_D h_D \int_e |p + \alpha^* \nabla v \cdot \mathbf{n}_e + \boldsymbol{\beta}^* v \cdot \mathbf{n}_e|^r ds. \tag{5.5}$$

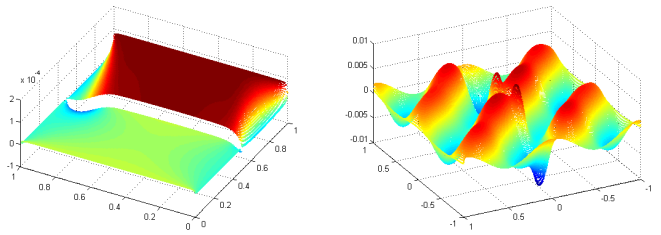
For the CFO algorithm 2.8, it can be seen from (2.9)–(2.10) that the weight parameter τ_D is automatically adjusted by the Lagrange multiplier λ_h on each element, as the weight τ_D can be easily absorbed by λ_h through the same scaling on each control element. Therefore, the CFO algorithm is quite robust in the minimization part.

Figure 3 shows the surface plot of the Lagrange multiplier λ_h for each test case. It can be seen that λ_h is quite sensitive to the continuity and smoothness of the true solution. We conjecture that λ_h should play the role of a posteriori error estimator in adaptive grid local refinements.

5.2. A two-phase flow in porous media. We consider a simplified two-phase flow problem in porous media which seeks a saturation function S and fluid pressure



(a) Smooth coefficient (left) and Hölder continuous coefficient (right)



(b) Discontinuous coefficients: Test Case 3 (left) and Test Case 4 (right)

FIG. 3. *The solution profile for the Lagrange multiplier* λ_h *on a partition of size* 128×128 *arising from the CFO scheme* (2.13)–(2.14).

p satisfying

$$-\nabla \cdot (\lambda(S)\kappa(x,y)\nabla p) = 0 \qquad \text{in } \Omega = (0,1)^2, \tag{5.6}$$

$$\frac{\partial S}{\partial t} + \operatorname{div}(\mathbf{v} f(S)) = 0, \qquad t > 0, \tag{5.7}$$

with the boundary condition

$$p(0,y) = 1,\, p(1,y) = 0, \tag{5.8}$$

$$\mathbf{v}(x,0)\cdot\mathbf{n} = 0,\, \mathbf{v}(x,1)\cdot\mathbf{n} = 0 \tag{5.9}$$

for the fluid pressure p and the following initial and boundary conditions for the saturation:

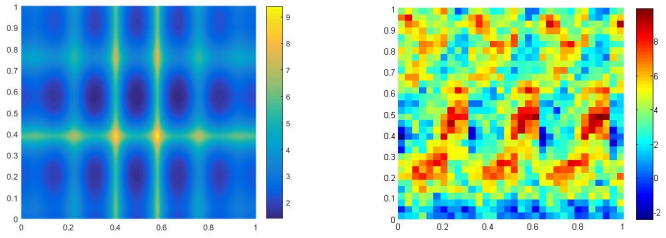
$$S(0,y,t) = 1, \qquad t \geq 0,\ y \in (0,1), \tag{5.10}$$

$$S(x,y,0) = 0, \qquad (x,y) \in (0,1)^2. \tag{5.11}$$

Here in (5.7) and the boundary condition (5.9), $\mathbf{v} = -\lambda(S)\kappa(x,y)\nabla p$ is the Darcy's velocity of the fluid, $\kappa(x,y)$ is the permeability of the porous media, $f(S)$ is the fractional flow function, and $\lambda(S)$ is the total mobility. This model problem has been studied in several existing literatures including [7, 8]. In our numerical study, we consider two examples of the permeability function $\kappa(x,y)$. The first one is a high-contrast, heterogeneous permeability function given by

$$\kappa(x,y) = \frac{1}{0.25 - 0.999(x-x^2)\sin(11.2\pi x)} \cdot \frac{1}{0.25 - 0.999(y-y^2)\sin(5.2\pi y)},$$

and the second permeability profile was generated using the experimental data from the Society of Petroleum Engineers (SPE) comparative solution project [14]. The



(a) Permeability function $\kappa(x, y)$ (b) Permeability data obtained from the SPE comparative solution project [14]

FIG. 4. *Permeability profiles plotted in the logarithmic scale.*

permeability profiles are plotted on a logarithmic scale in Figure 4. The figure shows that both types of permeability coefficients has a channelized pattern and is highly heterogenous. The flow function is given by

$$f(S) = \frac{S^2}{S^2 + (1-S)^2/5},$$

and the total mobility function is

$$\lambda(S) = S^2 + (1-S)^2/5.$$

Note that when $\lambda(S) = 1$, the equations reduce to a single-phase flow model.

We use an operator splitting technique [2] to solve the above system. That is, we substitute the saturation at the previous time step into (5.6) to compute the pressure p and the Darcy's velocity $\mathbf{v}$. Then we solve the transport equation (5.7) by an explicit time stepping scheme.

We first explain the numerical solution for (5.6) which is elliptic for any given saturation S. In existing literatures, there are many numerical methods available for approximating this elliptic equation, including the classical Galerkin finite element method (FEM). However, the straightforward numerical flux $\mathbf{v}_h = -\lambda(S)\kappa(x, y)\nabla p_h$ obtained from the classical FEM is known to be discontinuous across the element interface. Consequently, it is challenging to design conservative numerical schemes for the transport equation (5.7) based on such numerical fluxes. To overcome this difficulty, we solve the flow equation (5.6) by using the CFO algorithm proposed in section 2 where the control elements coincide with the finite element triangles of the domain.

To discretize the transport equation, we first integrate (5.7) with respect to time and then over each control element $T \in \mathcal{T}_h$ (i.e., $\mathcal{D}_h = \mathcal{T}_h$). We apply the forward Euler method in time and use integration by parts to arrive at the following scheme [7, 8]:

$$|T|(S_n^T - S_{n-1}^T) + \Delta t \int_{\partial T} \mathbf{v}_h \cdot \mathbf{n}_e f(S_{n-1}^T) ds = 0. \tag{5.12}$$

From the numerical scheme (5.12), the saturation S is defined on each element and the numerical flux $\mathbf{v}_h \cdot \mathbf{n}_e$ is needed on each edge. The continuity of this numerical flux becomes necessary for the mass conservation of the scheme. Our CFO FEM

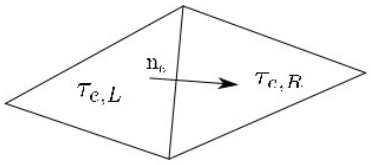


FIG. 5. *$T_{e,L}$ and $T_{e,R}$ stand for the left and right side element of edge e according to the sign of $\mathbf{v}_h \cdot \mathbf{n}_e$.*

(2.13)–(2.14) was designed to provide not only a continuous numerical flux but also one that conserves mass locally on each control element.

In our numerical computation, various methods are employed for evaluating the boundary integral in (5.12):

$$\int_{\partial T} f(S_{n-1}^T)\mathbf{v}_h \cdot \mathbf{n}_e ds \approx \sum_{e \in \partial T} |e| h_{e,T}\left(S_{n-1}^{T_{e,L}}, S_{n-1}^{T_{e,R}}\right),$$

where $h_{e,T}(S_{n-1}^{T_{e,L}}, S_{n-1}^{T_{e,R}})$ can be any admissible numerical flux, $T_{e,L}$ and $T_{e,R}$ stand for the "left" (or the upwind side) and "right" side (or the downwind side) element of edge e, respectively, according to the sign of $\mathbf{v}_h \cdot \mathbf{n}_e$ shown as in Figure 5. Note that the numerical flux $\mathbf{v}_h \cdot \mathbf{n}_e$ is constant on each edge and can be obtained directly from solving the flow equation of the system. Both the Roe type and the Lax–Friedrich numerical flux have been used in this study:

$$h_{e,T}{}^{Roe} := \begin{cases} f(S_{n-1}^{T_{e,L}})\mathbf{v}_h \cdot \mathbf{n}_e & \text{if } \mathbf{v}_h \cdot \mathbf{n}_e \geq 0, \\ f(S_{n-1}^{T_{e,R}})\mathbf{v}_h \cdot \mathbf{n}_e & \text{otherwise,} \end{cases} \tag{5.13}$$

and

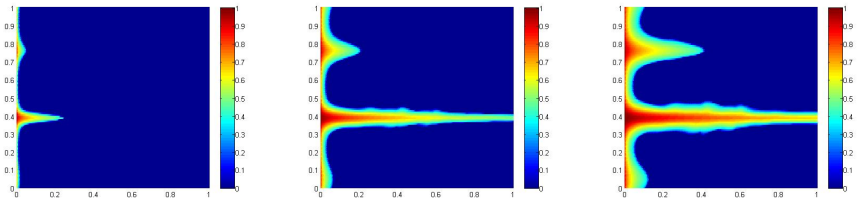
$$\begin{aligned} &h_{e,T}{}^{\text{Lax–Friedrich}} \\ &\quad := \frac{1}{2}[f(S_{n-1}^{T_{e,L}})\mathbf{v}_h \cdot \mathbf{n}_e + f(S_{n-1}^{T_{e,R}})\mathbf{v}_h \cdot \mathbf{n}_e - \lambda_{e,T}(S_{n-1}^{T_{e,R}} - S_{n-1}^{T_{e,L}})], \end{aligned} \tag{5.14}$$

where $\lambda_{e,T}$ is an estimate of the biggest eigenvalue of the Jacobian $\frac{\partial f(S,t)\mathbf{v}_h \cdot \mathbf{n}_e}{\partial S}$ for $(S;t)$ in a neighborhood of the edge e.

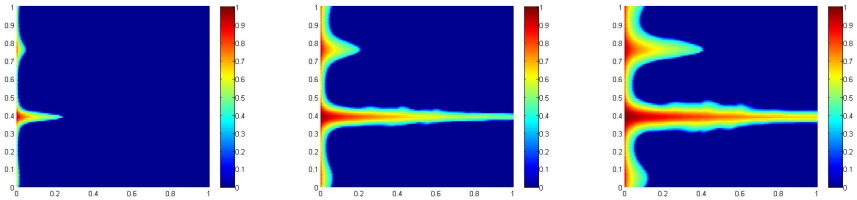
The saturation profiles of the two-phase flow for the permeability profile Figure 4(a) at time $T = 0.002$, 0.01, and 0.02 are shown in Figure 6. A mesh of 128×128 partition is employed. Both the Roe type (5.13) and the Lax–Friedrich numerical flux (5.14) have been used for the resolution of the transport equation. The following CFL condition was implemented for computing the time steps:

$$\Delta t^n \leq CFL \frac{h}{\max_{e \in \partial T \in \mathcal{T}_h} \lambda_{e,T}}. \tag{5.15}$$

For $CFL = 0.5$, we obtained time steps varying from a minimum value of 1.11e-05 to a maximum value of 2.68e-05 (Figure 7), for the two-phase flow problem on the finite element triangulation of 128×128 and the final time of $T_{final} = 0.002$. It can be seen that the numerical results corresponding to the two different methods are very close to each other, and we note that this result seems to be very close to the result shown in [7, Figure (6.10)].



(a) Saturation profiles using the Roe type numerical flux to solve the transport equation.



(b) Saturation profiles using the Lax–Friedrich numerical flux to solve the transport equation.

FIG. 6. *Saturation profiles at* $T = 0.002$, 0.01, *and* 0.02 *for the two-phase flow in porous media for the permeability profile Figure* 4(a), *obtained using the CFO scheme* (2.13)–(2.14), *with a partition of size* 128×128 *and* $CFL = 0.5$.

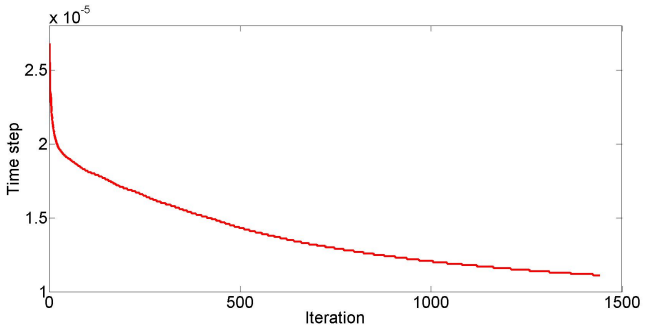


FIG. 7. *Time step for Test Case* 4(a), *with the finite element partition of size* 128×128 *and* $CFL = 0.5$.

The saturation profiles of the two-phase flow for the permeability profile Figure 4(b) at time $T = 0.002$, 0.01 and 0.02 are shown in Figure 8. A finite element partition of size 128×128 was employed with control elements being identical with the finite element partition. The computation is conducted by using the CFO method for the Darcy flow equation, the forward Euler in time and the Lax–Friedrich flux in space for the transport equation on the uniform finite element triangulation of size 128×128. The time steps are computed using the CFL condition with a value of $CFL = 0.4$, which vary from a minimum value of 1.34e-05 to a maximum value of 1.60e-05.

These computational results match the physical tendency as described in [7, 17] and, hence, confirm the effectiveness and robustness of the locally conservative fluxes provided by the CFO FEM (2.13)–(2.14).

Acknowledgments. The authors are grateful to Todd Arbogast, Malgorzata Peszynska, Ralph Showalter, and Son-Young Yi for a helpful discussion of this work during the SIAM 2017 Mathematical and Computational Geoscience conference in

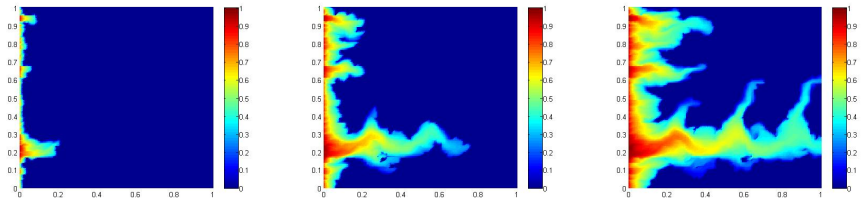


FIG. 8. *Saturation profiles at time levels* $T = 0.002$, 0.01, *and* 0.02 *for the test case with the permeability profile Figure* 4(b).

Erlangen Germany. This discussion has led to the name of "flux optimization" for the numerical scheme developed in this paper. The authors are also grateful for the valuable comments from the referees which led to an improved presentation of the results.

REFERENCES

[1] D. N. ARNOLD, F. BREZZI, B. COCKBURN, AND L. D. MARINI, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.
[2] K. AZIZ AND A. SETTARI, *Petroleum Reservoir Simulation*, Chapman & Hall, London, 1979.
[3] R. E. BANK AND D. J. ROSE, *Some error estimates for the box scheme*, SIAM Numer. Anal., 24 (1987), pp. 777–787.
[4] T. BARTH AND M. OHLBERGER, *Finite volume methods: Foundation and analysis*, in Encyclopedia of Computational Mechanics, Part 1, Wiley, Chichester, England, 2004, Chapter 15.
[5] P. B. BOCHEV AND M. D. GUNZBURGER, *A locally conservative least-squares method for Darcy flows*, Commun. Numer. Methods Engrg., (2006), pp. 00:1-6.
[6] F. BREZZI, J. DOUGLAS, AND L. MARINI, *Two families of mixed finite elements for second order elliptic problems*, Numer. Math., 47 (1985), pp. 217–235.
[7] L. BUSH AND V. GINTING, *On the application of the continuous Galerkin finite element method for conservation problems*, SIAM J. Sci. Comput., 35 (2013), pp. A2953–A2975.
[8] L. BUSH, V. GINTING, AND M. PRESHO, *Application of a conservative, generalized multiscale finite element method to flow models*, J. Comput. Appl. Math., 260 (2014), pp. 395–409.
[9] Z. CAI, *On the finite volume element method*, Numer. Math., 58 (1991), pp. 713–735.
[10] Z. CAI, J. DOUGLAS, AND M. PARK, *Development and analysis of higher order finite volume methods over rectangles for elliptic equations*, Adv. Comput. Math., 19 (2003), pp. 3–33.
[11] L. CHEN, *A new class of high order finite volume methods for second order elliptic equations*, SIAM J. Numer. Anal. 47 (2010), pp. 4021–4043.
[12] Z. CHEN, J. WU, AND Y. XU, *Higher-order finite volume methods for elliptic boundary value problems*, Adv. Comput. Math., 37 (2012), pp. 191–253.
[13] S. H. CHOU AND Q. LI, *Error estimates in* L^2, H^1 *and* L^∞ *in covolume methods for elliptic and parabolic problems: A unified approach*, Math. Comp., 69 (2000), pp. 103–120.
[14] M. A. CHRISTIE AND M. J. BLUNT, *Tenth SPE comparative solution project: A comparison of upscaling techniques*, SPE Reservoir Simulation Symposium, Society of Petroleum Engineers, Richardson, TX, 2001.
[15] S.-H. CHOU AND S. TANG, *Conservative* $P1$ *conforming and nonconforming Galerkin FEMS: Effective flux evaluation via a nonmixed method approach*, SIAM J. Numer. Anal., 38 (2000), pp. 660–680.
[16] B. COCKBURN, J. GOPALAKRISHNAN, AND H. WANG, *Locally conservative fluxes for the continuous Galerkin method*, SIAM J. Numer. Anal., 45 (2007), pp. 1742–1776.
[17] Y. EFENDIEV, V. GINTING, T. HOU, AND R. EWING, *Accurate multiscale finite element methods for two-phase flow simulations*, J. Comput. Phys., 220 (2006), pp. 155–174.
[18] PH. EMONOT, *Methods de Volume Elements Finis: Applications aux Equations de Navier-Stokes et Resultats de Convergence*, Technical report, FRCEA-TH–411, CEA, Lyon, 1992.
[19] R. EYMARD, T. GALLOUET, AND R. HERBIN, *Finite volume methods*, in Handbook of Numerical Analysis, P. G. Ciarlet and J. L. Lions, eds., North-Holland, Amsterdam, Vol. 7, 2000, pp. 713–1018.

[20] R. Eymard, T. Gallouët, and R. Herbin, *A cell-centred finite-volume approximation for anisotropic diffusion operators on unstructured meshes in any space dimension*, IMA J. Numer. Anal., 26 (2006), pp. 326–353.
[21] W. Hackbusch, *On first and second order box methods*, Computing, 41 (1989), pp. 277–296.
[22] T. J. R. Hughes, G. Engel, L. Mazzei, and M. G. Larson, *The continuous Galerkin method is locally conservative*, J. Comput. Phys., 163 (2000), pp. 467–488.
[23] M. G. Larson and A. J. Niklasson, *A conservative flux for the continuous Galerkin method based on discontinuous enrichment*, CALCOLO 41 (2004), pp. 65–76.
[24] R. D. Lazarov, I. D. Michev, and P. S. Vassilevski, *Finite volume methods for convection-diffusion problems*, SIAM J. Numer. Anal., 33 (1996), pp. 31–55.
[25] R. J. LeVeque, *Finite Volume Methods for Hyperbolic Problems*, Cambridge University Press, Cambridge, 2002.
[26] R. Li, Z. Chen, and W. Wu, *The Generalized Difference Methods for Partial Differential Equations*, Marcel Dekker, New York, 2000.
[27] Y. Lin, M. Yang, and Q. Zou, *L^2 error estimates for a class of any order finite volume schemes over quadrilateral meshes*, SIAM J. Numer. Anal., 53 (2015), pp. 2010–2050.
[28] M. K. Mudunuru and K. B. Nakshatrala, *On enforcing maximum principles and achieving element-wise species balance for advection-diffusion-reaction equations under the finite element method*, J. Comput. Phys., 305 (2016), pp. 448–493.
[29] N. C. Nguyen, J. Peraire, and B. Cockburn, *An implicit high-order hybridizable discontinuous Galerkin method for linear convection-diffusion equations*, J. Comput. Phys., 228 (2009), pp. 3232–3254.
[30] R. A. Nicolaides, T. A. Porsching, and C. A. Hall, *Covolume methods in computational fluid dynamics*, in Computational Fluid Dynamics Review, M. Hafez and K. Oshima, eds., Wiley, New York, 1995, pp. 279–299.
[31] L. H. Odsater, M. F. Wheeler, T. Kvamsdala, and M. G. Larson, *Postprocessing of non-conservative flux for compatibility with transport in heterogeneous media*, Comput. Methods Appl. Mech. Engrg., 315 (2017), pp. 799–830.
[32] P. A. Raviart and J. M. Thomas, *A mixed finite element method for second order elliptic problems*, in Mathematical Aspects of Finite Element Methods, I. Galligani and E. Magenes, eds., Springer, Berlin, 1977, pp. 292–315.
[33] T. Schmidt, *Box schemes on quadrilateral meshes*, Computing, 51 (1993), pp. 271–292.
[34] C. Shu, *High order finite difference and finite volume WENO schemes and discontinuous Galerkin methods for CFD*, J. Comput. Fluid Dyn., 17 (2003), pp. 107–118.
[35] I. Smears and E. Süli, *Discontinuous Galerkin finite element approximation of nondivergence form elliptic equations with Cordès coefficients*, SIAM J. Numer. Anal., 51 (2013), pp. 2088–2106.
[36] E. Süli, *The accuracy of cell vertex finite volume methods on quadrilateral meshes*, Math. Comp., 59 (1992), pp. 359–382.
[37] C. Wang and J. Wang, *A primal-dual weak Galerkin finite element method for second elliptic equations in non-divergence form*, Math. Comp., 87 (2018), pp. 515–545.
[38] J. Wang and X. Ye, *A weak Galerkin mixed finite element method for second-order ellliptic problems*, J. Comput. Appl. Math., 241 (2013), pp. 103-115.
[39] J. Wang and X. Ye, *A weak Galerkin mixed finite element method for second-order elliptic problems*, Math. Comp., 83 (2014), pp. 2101–2126.
[40] J. Xu and Q. Zou, *Analysis of linear and quadratic simplitical finite volume methods for elliptic equations*, Numer. Math., 111 (2009), pp. 469–492.
[41] G. Yu, B. Yu, Y. Zhao, and J. Wei, *Comparative studies on accuracy and convergence rate between the cell-centered scheme and the cell-vertex scheme for triangular grids*, Internat. J. Heat Mass Transf., 55 (2012), pp. 8051–8060.
[42] S. Zhang, Z. Zhang, and Q. Zou, *A post-process finite element solution*, Numer. Methods Partial Differential Equations, 33 (2017), pp. 1859–1883, https://doi.org/10.1002/num.22163.
[43] Z. Zhang and Q. Zou, *Vertex-centered finite volume schemes of any order over quadrilateral meshes for elliptic boundary value problems*, Numer. Math., 130 (2015), pp. 363–393.
[44] Q. Zou, L. Guo, and Q. Deng, *High order continuous local-conserving fluxes and finite-volume-like finite element solutions for elliptic equations*, SIAM J. Numer. Anal., 55 (2017), pp. 2666–2686.