1999

Mixed Upwinding Covolume Methods on Rectangular Grids for Convection-diffusion Problems

So-Hsiang Chou
Bowling Green State University, chou@bgsu.edu

Do Y. Kwak

Panayot S. Vassilevski

Follow this and additional works at: https://scholarworks.bgsu.edu/math_stat_pub

Part of the Physical Sciences and Mathematics Commons

Repository Citation
https://scholarworks.bgsu.edu/math_stat_pub/5

This Article is brought to you for free and open access by the Mathematics and Statistics at ScholarWorks@BGSU. It has been accepted for inclusion in Mathematics and Statistics Faculty Publications by an authorized administrator of ScholarWorks@BGSU.
MIXED UPWINDING COVOLUME METHODS ON RECTANGULAR GRIDS FOR CONVECTION-DIFFUSION PROBLEMS∗

SO-HSIANG CHOU†, DO Y. KWAK‡, AND PANAYOT S. VASSILEVSKI§

Abstract. We consider an upwinding covolume or control-volume method for a system of first order PDEs resulting from the mixed formulation of a convection-diffusion equation with a variable anisotropic diffusion tensor. The system can be used to model the steady state of the transport of a contaminant carried by a flow. We use the lowest order Raviart–Thomas space and show that the concentration and concentration flux both converge at one-half order provided that the exact flux is in $H^1(\Omega)^2$ and the exact concentration is in $H^1(\Omega)$. Some numerical experiments illustrating the error behavior of the scheme are provided.

Key words. stationary convection-diffusion problems, mixed finite elements, covolume methods, finite volume methods, upwind discretization, error estimates, preconditioning, M-matrices

AMS subject classifications. 65F10, 65N20, 65N30

PII. S1064827597321052

1. Introduction. The purpose of this paper is to extend the previous results or techniques of covolume conservative schemes developed and analyzed for Stokes and diffusion equations by the authors (see Chou [7]; Chou and Kwak [9], [10]; and Chou, Kwak, and Vassilevski [11]) to convection-diffusion problems, including the practically interesting convection-dominated limit case. The ensuing new scheme combines the covolume methodology from the previous papers and the upwinding strategy to handle the convective term. The present paper deals with the derivation of the scheme, provides error analysis with limiting conditions on the smoothness of the solution, and illustrates the method with numerical experiments. The error analysis is provided economically in the sense that we could have done it the long and direct way (i.e., purely within the finite volume framework); instead, we adapt an existing error analysis in a general mixed finite element setting given in Liu, Wang, and Yan [20].

The bridge of the error analysis lies in the construction of a certain transfer operator (further denoted by $\gamma_h$) between the standard Raviart–Thomas (trial) spaces and the piecewise constant test spaces commonly used in the finite volume literature. The covolume scheme simultaneously treats, as in the mixed method, the vector quantity (flux) and the original scalar one (concentration), and provides $O(h^{1/2})$ order approximation to both of them in the $L^2$ norm. This seems a reasonable approximation for piecewise constant test functions. Also, as for the vector unknown, a one-half order accuracy (at worst) may be acceptable since it is a gradient of the scalar one. In the case of no convection a first order approximation is possible, and even a second
order is possible if the mesh subdivision allows for superconvergence. If one departs from the finite volume (covolume) methodology, i.e., allowing higher-order piecewise polynomial spaces similar to the mixed method, then a higher order of convergence is also possible. Those topics, however, are not central to the present paper. The focus here is the extension of the covolume methodology to the convection-diffusion problems that works in a convection-dominated limit. Another topic that we are not concerned with is a local refinement of meshes, an important task we shall pursue in a later paper.

1.1. Problem formulation. Consider the convection-diffusion problem on an axiparallel rectangular (bounded) domain $\Omega \subset \mathbb{R}^2$,

\[
\begin{cases}
-\nabla \cdot \mathcal{K}\nabla p + \nabla \cdot (bp) + \alpha p = f \text{ in } \Omega, \\
\mathcal{K}\nabla p \cdot n = 0 \text{ on } \Gamma_+ = \{ x \in \partial \Omega : b \cdot n \geq 0 \}, \\
p = 0 \text{ on } \Gamma_- = \{ x \in \partial \Omega : b \cdot n < 0 \}.
\end{cases}
\]

One can instead have only Dirichlet boundary condition, i.e., $p = 0$ on $\partial \Omega$. The latter boundary condition simplifies the exposition; hence in most of the technical part of the analysis, only the Dirichlet boundary condition will be assumed. Here $\mathcal{K} = \mathcal{K}(x) = \text{diag} (\tau_1^{-1}(x), \tau_2^{-1}(x))$ is a positive definite diagonal matrix function whose entries are bounded from below and above by positive constants. Furthermore, we shall assume that $\tau_1, \tau_2$ are locally Lipschitz. The vector function $b$ is in the Sobolev space $[W^{1,\infty}(\Omega)]^2$, $f \in L^2(\Omega)$ and $\alpha \in L^\infty(\Omega)$. We impose the following two conditions so that (1.1) is convection-dominated and uniquely solvable.

(H1) There exist two positive constants $\epsilon_1$ and $\epsilon_2$ such that the two constants are small and proportional:

\[
\epsilon_2 \ll ||b||_{\infty}, \quad \epsilon_2/\epsilon_1 = \mathcal{O}(1),
\]

with

\[
\epsilon_1 \leq \tau_1^{-1}, \quad \tau_2^{-1} \leq \epsilon_2 \quad \text{in } \Omega.
\]

(H2) There exists a positive constant $\gamma_0$ such that

\[
\alpha + \frac{1}{2} \nabla \cdot b \geq \gamma_0 \quad \text{in } \Omega.
\]

Problem (1.1) can be used to model the steady state of the transport of a contaminant in a porous or anisotropic medium flow. The variable $p$ stands for the concentration of the contaminant, and $b$ stands for the velocity field of the flow carrying it. In this context, the molecular diffusion tensor $\mathcal{K}$ (assumed to be diagonal for simplicity) has less effect on the physics than the convection term $b \cdot \nabla p$, but the diffusion term contributes to the smoothness of the solution.

Let us introduce a new variable $u = -\mathcal{K}\nabla p$ and write (1.1) as the system of first order partial differential equations (PDEs),

\[
\begin{cases}
\mathcal{K}^{-1}u = -\nabla p, \\
\text{div} u + \text{div}(bp) + \alpha p = f,
\end{cases}
\]

together with the boundary conditions,

\[
\begin{align*}
\mathbf{u} \cdot n &= 0 \text{ on } \Gamma_+ = \{ x \in \partial \Omega : b \cdot n \geq 0 \}, \\
p &= 0 \text{ on } \Gamma_- = \{ x \in \partial \Omega : b \cdot n < 0 \}.
\end{align*}
\]
For ease of reference, we shall also refer to the variable $u$ as the flux or velocity variable and $p$ as the concentration variable. The main purpose of this new system is twofold. In the context of the mixed method we can then approximate both variables to the same order. While in the context of finite volume or difference methods, we can derive conservative schemes from these first order equations because one of them represents a conservation law and the other a constitutive law.

We shall need three domain partitions for the approximation problem. More specifically, let the domain $\Omega$ be partitioned (cf. Figure 1) into a union of rectangles $Q_{i,j}$ with centers $c_{i,j}$. This is the primal partition which we shall call $R_h$. The subindices $\{i+1,j\}, \{i-1,j\}, \{i,j+1\}$, and $\{i,j-1\}$ are assigned to the eastern, western, northern, and southern adjacent rectangles, respectively, if they exist. Given $Q_{i,j}$, the two midpoints of its vertical edges are denoted as $c_{i,j+1/2}$, and the two midpoints of horizontal edges as $c_{i+1/2,j}$. Let $c_{i,j} = (x_i, y_j)$ and $c_{i+1/2,j} = (x_{i+1/2}, y_j)$ etc., define

$$Q_{i+1/2,j} := [x_i, x_{i+1}] \times [y_{j-1/2}, y_{j+1/2}] \cap \bar{\Omega},$$
$$Q_{i,j+1/2} := [x_{i-1/2}, x_{i+1/2}] \times [y_j, y_{j+1}] \cap \bar{\Omega}$$

and

$$Q_{i,j} := [x_{i-1/2}, x_{i+1/2}] \times [y_{j-1/2}, y_{j+1/2}].$$

**Remark 1.1.** Note that at the boundary the $Q_{i+1/2,j}$ or $Q_{i,j+1/2}$ is half the size of a typical interior volume.

Let

$$H_0 := H(\text{div}; \Omega) \cap \{u \cdot n = 0 \text{ on } \Gamma_+\},$$

(1.6)
where \( \mathbf{H}(\text{div}; \Omega) \) is the space of all vector-valued functions \( \mathbf{w} \in L^2(\Omega)^2 \) such that \( \text{div} \mathbf{w} \in L^2(\Omega) \), and also define

\[
(1.7) \quad \mathcal{L} := \{ q \in L^2 : q|_Q \in H^1(Q) \quad \forall \ Q \in \mathcal{R}_h \}.
\]

The associated weak formulation of our first order system is as follows. Find \( (\mathbf{u}, p) \in \mathbf{H}_0 \times \mathcal{L} \) such that

\[
(1.8) \quad \begin{align*}
(K^{-1} \mathbf{u}, \mathbf{v}) &= (p, \text{div} \mathbf{v}) \quad \forall \ \mathbf{v} \text{ in } \mathbf{H}_0, \\
(\text{div} \mathbf{u}, q) + \tilde{d}_h(p, q) &= (f, q) \quad \forall \ q \text{ in } \mathcal{L},
\end{align*}
\]

where the bilinear form

\[
(1.9) \quad \tilde{d}_h(p, q) := \sum_{Q \in \mathcal{R}_h} \left[ -\int_{Q} p \mathbf{b} \cdot \nabla q \, dx \, dy + \int_{\partial Q} \mathbf{b} \cdot \mathbf{n} q \, d\sigma \right] + (\alpha p, q).
\]

Note that the first two terms on the right are a modification of the expression \((\text{div}(\mathbf{b} p), q)\), which does not make sense for nonsmooth \( p \).

Define the trial space as the lowest order Raviart–Thomas space:

\[
\mathbf{H}_h := \{ (\mathbf{u}_h, v_h) \in \mathbf{H}_0 : \mathbf{u}_h(x, y) = a + bx, \quad v_h(x, y) = c + dy \text{ on } Q_{i,j} \}
\]

and

\[
\mathcal{L}_h := \{ q_h \in \mathcal{L} : q \text{ is constant over } Q_{ij} \}.
\]

Then the natural (upwinding) mixed finite element method [20] corresponding to (1.8) deals with the primal grid only. Find \( (\tilde{\mathbf{u}}_h, \tilde{p}_h) \in \mathbf{H}_h \times \mathcal{L}_h \) such that

\[
(1.10) \quad \begin{align*}
(K^{-1} \tilde{\mathbf{u}}_h, \mathbf{v}_h) - (\text{div} \mathbf{v}_h, \tilde{p}_h) &= 0 \quad \forall \ \mathbf{v}_h \text{ in } \mathbf{H}_h, \\
(\text{div} \tilde{\mathbf{u}}_h, q_h) + d_h(\tilde{p}_h, q_h) &= (f, q_h) \quad \forall \ q_h \text{ in } \mathcal{L}_h,
\end{align*}
\]

where \( d_h(\tilde{p}_h, q_h) \) is the discretization of the bilinear form \( \tilde{d}_h(p, q) \) of (1.9) involving the upwinding concept

\[
(1.11) \quad d_h(p_h, q_h) = \sum_{Q \in \mathcal{R}_h} \left[ \int_{\partial Q} (\mathbf{b} \cdot \mathbf{n})_+ p_h^i q_h^i d\sigma + \int_{\partial Q} (\mathbf{b} \cdot \mathbf{n})_- p_h^e q_h^e d\sigma \right] + (\alpha \tilde{p}_h, q_h),
\]

where

\[
(\mathbf{b} \cdot \mathbf{n})_+ := \max(\mathbf{b} \cdot \mathbf{n}, 0), \quad (\mathbf{b} \cdot \mathbf{n})_- := \min(\mathbf{b} \cdot \mathbf{n}, 0),
\]

and \( p_h^i \) is the trace of \( p_h \) on \( \partial Q \) taken from the interior of \( Q \) and \( p_h^e \) is that from the exterior of \( Q \). Therefore, at the inflow boundary \( \Gamma_- \), \( p_h^e = 0 \) and hence in (1.11) integration over edges of \( \partial Q \) that intersect \( \Gamma_- \) is not actually performed.

We will not use the finite element method (1.10) except in the error analysis.
1.2. The construction of the covolume scheme. Since there are many existing hydrodynamic codes in either finite difference or finite volume methods, we are motivated to design and analyze a finite volume mixed method using the upwinding concept. We will adapt a covolume methodology for the generalized Stokes problem [7] to approximate this system. The basic idea of creating a covolume method is to find a good combination of the finite volume method and the marker and cell (MAC) [17] placements of flow variables. (A balanced survey of the covolume method literature up to 1995 is given by Nicolaides, Porsching, and Hall [21].) For an analysis of the MAC scheme for Stokes problem (see Girault and Lopez [15]). In the MAC scheme, the concentration variable is assigned to the centers of the rectangular volumes and the normal components of the velocity or fluxes are assigned to the edges of the rectangular volumes. The normal approximate velocity is assumed to be constant and constant in $y$ and constant in $x$. Thus we have four degrees of freedom assigned at midpoints of edges. For example, on the eastern vertical edge of $Q_{i,j}$, we have only one unknown; the accompanying equation is taken by integrating the first component of the vector equation (1.5) over $Q_{i+1/2,j}$. Similarly, to determine the unknown at the northern edge, we integrate (1.5) over $Q_{i,j+1/2}$. In other words, if we write the velocity field as $\mathbf{u}_h = (u_h, v_h)$, then $Q_{i,j+1/2}$ is for the determination of $v_h$ and $Q_{i+1/2,j}$ for $u_h$. We will sometimes call $Q_{i+1/2,j}$ ($Q_{i,j+1/2}$) as a $u$-volume ($v$-volume). These volumes are also called the covolumes of $Q_{i,j}$ in the literature.

Throughout this paper the primal partition $\{Q_{i,j}\}$ is quasi-regular; i.e., there exists a positive constant $C_1$ independent of $h$ such that

\[ C_1 h^2 \leq \text{area}\{Q_{i,j}\} \leq h^2 \quad \forall Q_{i,j} \in \mathcal{R}_h, \]

where $h := \max_{i,j} \{h_x^i, j h_y^i\}$, $h_x^i, j$ $h_y^i, j$ are, respectively, the width and height of $Q_{i,j}$.

We now describe the present covolume method. Let $\mathbf{u}_h = (u_h, v_h) \in \mathbf{H}_h$ and $p_h \in \mathcal{L}_h$ be the approximate solution obtained as follows. Integrate the $x$-component of the first equation of (1.5) over the volume $Q_{i+1/2,j}$ to get

\begin{equation}
(1.12) \quad \int_{Q_{i+1/2,j}} \tau_1 u_h \, dx \, dy = -(p_{i+1,j} - p_{i,j}) \int_{y_{j-1/2}}^{y_{j+1/2}} dy,
\end{equation}

where $p_{i,j} = p_h(x_i, y_j)$, $p_{i+1,j} = p_h(x_{i+1}, y_j)$. Integrate the $y$-component of the first equation of (1.5) over the (control) volume $Q_{i,j+1/2}$ to get

\begin{equation}
(1.13) \quad \int_{Q_{i,j+1/2}} \tau_2 v_h \, dx \, dy = -(p_{i+1,j+1} - p_{i,j}) \int_{x_{j-1/2}}^{x_{j+1/2}} dx,
\end{equation}

where $p_{i,j+1} = p_h(x_i, y_{j+1})$. Near the boundary of $\Omega$ one has to keep in mind that the covolumes are half of the interior covolumes (see Remark 1.1). That is, we then use the Dirichlet boundary value of $p$ at the inflow boundary. Also, if we have boundary conditions of Neumann type (at the outflow boundary) we do not integrate over boundary covolumes; then the Neumann conditions are essential conditions for the flux variable.
Denote the four edges of the boundary of $Q_{i,j}$ as $e_{i+1/2,j}, e_{i,j+1/2}, e_{i-1/2,j}$, and $e_{i,j-1/2}$. Now integrate the second equation of (1.5) over the volume $Q_{i,j}$ and use the upwinding to get

\[
\int_{Q_{i,j}} \nabla \cdot \mathbf{u}_h \, dx \, dy + \int_{Q_{i,j}} \alpha p \, dx \, dy \\
+ p_{i,j} \int_{e_{i+1/2,j}} (\mathbf{b} \cdot \mathbf{n})_+ \, d\sigma + p_{i+1,j} \int_{e_{i+1/2,j}} (\mathbf{b} \cdot \mathbf{n})_- \, d\sigma \\
+ p_{i,j} \int_{e_{i-1/2,j}} (\mathbf{b} \cdot \mathbf{n})_+ \, d\sigma + p_{i,j-1} \int_{e_{i-1/2,j}} (\mathbf{b} \cdot \mathbf{n})_- \, d\sigma \\
+ p_{i,j} \int_{e_{i,j+1/2}} (\mathbf{b} \cdot \mathbf{n})_+ \, d\sigma + p_{i,j+1} \int_{e_{i,j+1/2}} (\mathbf{b} \cdot \mathbf{n})_- \, d\sigma \\
+ p_{i,j} \int_{e_{i,j-1/2}} (\mathbf{b} \cdot \mathbf{n})_+ \, d\sigma + p_{i,j-1} \int_{e_{i,j-1/2}} (\mathbf{b} \cdot \mathbf{n})_- \, d\sigma \\
= \int_{Q_{i,j}} f \, dx \, dy.
\]

Temporarily, we shall assume all integrals are evaluated exactly. See the last section for final discretization of them using quadratures.

Equations (1.12)–(1.14) are quite intuitive from a physical point of view, but for the error analysis it is more convenient if we state them in terms of bilinear forms. Define the test space

\[
\mathbf{Y}_h := \{ \mathbf{w} = (u_h, v_h) : u_h \text{ piecewise constant on } u\text{-volumes}, \\
v_h \text{ piecewise constant on } v\text{-volumes }, \\
\mathbf{w} \cdot \mathbf{n} = 0 \text{ on covaolmes near the outflow boundary } \Gamma_+ \}
\]

which is used to pick out the control volumes. Here, $\mathbf{n}$ is a normal vector to $\Gamma_+$. Furthermore, define the following bilinear forms:

\[
a(\mathbf{u}_h, \mathbf{v}_h) = \int_{\Omega} K^{-1} \mathbf{u}_h \cdot \mathbf{v}_h \, dx, \quad \mathbf{u}_h \in \mathbf{H}_h, \mathbf{v}_h \in \mathbf{Y}_h;
\]

\[
b(\mathbf{v}_h, p_h) := \sum (v_{i,j} (e_{i+1/2,j}).0)^t \cdot \int_{\partial Q_{i+1/2,j}} p_h \mathbf{n} \, d\sigma \\
+ \sum (0, v_{i,j} (e_{i,j+1/2}).0)^t \cdot \int_{\partial Q_{i,j+1/2}} p_h \mathbf{n} \, d\sigma, \quad \mathbf{v}_h \in \mathbf{Y}_h, p_h \in \mathcal{L}_h;
\]

\[
c(\mathbf{u}_h, q_h) := \sum q_h (e_{i,j}) \int_{Q_{i,j}} \text{div} \mathbf{u}_h \, dx, \quad \mathbf{u}_h \in \mathbf{H}_h, q_h \in \mathcal{L}_h;
\]

\[
d_h(p_h, q_h) := \sum q_h (e_{i,j}) \int_{Q_{i,j}} \frac{\mathbf{b} \cdot \mathbf{n}}{\partial Q_{i,j}} + p_h^b q_h^b \, d\sigma + \int_{\partial Q_{i,j}} (\mathbf{b} \cdot \mathbf{n})_- p_h^b q_h^b \, d\sigma \\
+ (\alpha p_h, q_h), \quad p_h, q_h \in \mathcal{L}_h,
\]

where $p_h^b$ is the trace of $p_h$ on $\partial Q_{i,j}$ taken from the interior of $Q_{i,j}$ and $p_h^b$ is that from the exterior of $Q_{i,j}$. Note that there is no integration over edges of $Q_{i,j}$ near the inflow boundary (or we may formally let $p_h^b = 0$) and at the outflow boundary $(\mathbf{b} \cdot \mathbf{n})_- = 0$, hence we do not have to specify $p_h^b$ there.
Define the transfer operator \( \gamma_h : H_h \to Y_h \) connecting the trial space to the test function space in the following equation:

\[
\gamma_h w_h := (\gamma_h u_h, \gamma_h v_h) \quad w_h = (u_h, v_h) := \left(\sum_i u_h(c_{i+1/2,j})\chi_{i+1/2,j} + \sum_j v_h(c_{i,j+1/2})\chi_{i,j+1/2}\right),
\]

where \( \chi_{i+1/2,j} \) and \( \chi_{i,j+1/2} \) are the characteristic function of \( Q_{i+1/2,j} \) and \( Q_{i,j+1/2} \), respectively. Note that we used the same notation \( \gamma_h \) in the component-wise definition and that \( \gamma_h \) is one-to-one and onto.

### 1.3. Relation to the mixed finite element discretization.

The covolume method (1.12)–(1.14) is equivalent to the problem of finding \( \{u_h, p_h\} \in H_h \times L_h \) such that

\[
\begin{align*}
a(u_h, \gamma_h v_h) + b(\gamma_h v_h, p_h) &= 0 \quad \forall v_h \in H_h, \\
c(u_h, q_h) + d_h(p_h, q_h) &= (f, q_h) \quad \forall q_h \in L_h.
\end{align*}
\]

Here the substitution of \( \gamma_h v_h \) for a test function \( w_h \in Y_h \) is due to the surjectivity of the operator \( \gamma_h \). This simple observation turns the original Petrov–Galerkin statement into a standard Galerkin one.

We can reformulate (1.20) into a saddle-point problem by further introducing

\[
\begin{align*}
A(u, v) := a(u, \gamma_h v) = (K^{-1}u, \gamma_h v), \quad u, v \in H_h, \\
B(w_h, q_h) := b(\gamma_h w_h, q_h), \quad w_h \in H_h, \quad q_h \in L_h
\end{align*}
\]

and noting that \( B = -c \) (cf. Lemma 2.3) so that problem (1.20) becomes

\[
\begin{align*}
A(u_h, v_h) + B(v_h, p_h) &= 0 \quad \forall v_h \in H_h, \\
B(u_h, q_h) - d_h(p_h, q_h) &= -(f, q_h) \quad \forall q_h \in L_h.
\end{align*}
\]  

It is interesting to note that based on the transfer operator \( \gamma_h \), the forms \( A \) and \( B \) are bilinear, and hence the above system is in standard form. Nevertheless, the standard (conforming in \( H(\text{div}; \Omega) \)) mixed method analysis cannot be used here. This is so because the original PDE cannot be put into the same form: the transfer operator \( \gamma_h \) in the definition of the bilinear form \( A \) cannot be extended to the space \( H(\text{div}; \Omega) \). However, upon closer examination we see that the standard mixed method (1.10) for the convection-dominated problem (1.5) differs from the mixed covolume method (1.21) only in the bilinear form \( A \). Thus we can treat the covolume method as one resulting from a “variational crime” of the standard mixed method. A careful analysis of the transfer operator \( \gamma_h \) in connection to this deviation then leads to our error estimate in Theorem 3.1 which demonstrates the one-half order error estimate in the flux variable, as well as in the concentration variable, under the minimal regularity assumption that \( u \in H^1(\Omega)^2 \) and \( p \in H^1(\Omega) \). The starting point of the proof is a good error equation (3.10) that plays the role of Cea’s lemma in the standard finite element analysis. This methodology was initiated in Chou [7] for the generalized Stokes problem on triangular grids, in Chou and Kwak [8, 9] for the same problem on rectangular grids, and in Chou and Li [12] for the “point-centered” or vertex-centered schemes for the variable-coefficient Poisson equation. The diffusion problem (i.e., \( b = 0 \) and \( \alpha = 0 \) in (1.5)) was treated in Chou and Kwak [10] by a mixed covolume method.
1.4. Summary of results in the remaining part of the paper. We summarize now the results of the present paper. In section 2, we study the properties of our discretization scheme. In particular, we show that the saddle-point formulation (1.20) is uniquely solvable for sufficiently small $h$. The constraint on $h$ is determined by the local Lipschitz constants of the coefficient matrix $K$. In section 3 we modify the analysis from Liu, Wang, and Yan [20] in the present context to derive our main error estimates. Finally, in section 4 we present extensive numerical experiments that illustrate the error behavior of our discrete solutions. We comment that for the particular case of diagonal diffusion coefficient $K$, the saddle-point problem can be reduced (implicitly) to a problem only for the concentration variable $p_h$, which can be solved by a preconditioned iterative method. In fact, due to the upwinding, the reduced problem can be well approximated by an $M$-matrix, for which good approximate ILU factorization preconditioners are available. Details are found in the last section.

2. Properties of the saddle point formulation. In this section we prove some preliminary lemmas. Let $\| \cdot \|_j$, $j = 0, 1$ denote the usual $L^2$ and $H^1$ norms, respectively, and we also use $\| \cdot \|$ for the $L^2$ norm when there is no confusion. The symbol $C$ will be used as a generic positive constant independent of $h$ and the $\epsilon_1$ of (1.3), and may have different values at different places.

The following two lemmas are slightly more general than the original ones in Liu, Wang, and Yan [20]. To clearly indicate the influence of the boundary conditions we include our proof. Also, the two lemmas below hold in a general setting (2-D and 3-D rectangular or triangular elements; in 3-D, edges should be accordingly replaced by faces of volumes). That is why we have purposely rotated the elements in Figure 2.1.

We adopt now the following notation: for each edge $e$ of a primal volume $Q$ we assign a unit normal vector $n_e$. Then, given the pair $(e, n_e)$ with $e$ an interior edge, one can uniquely define the neighboring primal volumes $Q_e^+$ and $Q_e^-$ with the common edge $e$ so that $n_e$ points towards $Q_e^+$, see, e.g., Figure 2.1. A vector $n$ (without subscript) will be considered “outward” to a underlying domain (or volume).

**Lemma 2.1.** Let $E_0$ be the collection of interior edges of elements in $R_h$. Then
the bilinear form $d_h(\cdot, \cdot)$ of (1.19) can be rewritten as follows:

$$
d_h(p, q) = \frac{1}{2} \sum_{e \in \mathcal{E}} \int_{\Gamma_e} |p| |q| |b \cdot n| \, d\sigma
$$

(2.1)

$$
+ \frac{1}{2} \sum_{e \in \mathcal{E}_0} \int_{\Gamma_e} b \cdot n_e (p_{Q_e^+} + p_{Q_e^-}) (q_{Q_e^-} - q_{Q_e^+}) \, d\sigma
$$

$$
+ (\alpha, q) + \int b \cdot n p \, d\sigma,
$$

where for each interior edge $e$, i.e., $e \in \mathcal{E}_0$, $[p]$ denotes the jump of the discontinuous function $p$.

**Proof.** For completeness we provide a proof of the lemma. We start from the definition of $d_h(p, q)$; see (1.19). The idea is to rewrite the corresponding sums over the edges $\mathcal{E}$. For a given interior edge $e$ and a pre-assigned unit normal vector $n_e$, one has two neighboring primal volumes $Q_e^+$ and $Q_e^-$ that contribute in the rearranged sum. Also, $Q_e^+$ will contribute with the vector $n = n_e$ and $Q_e^-$ will contribute with the vector $n = -n_e$. Hence, denoting by $\mathcal{E}_0$ the set of interior edges, one gets

$$
d_h(p, q) = \sum_{e \in \mathcal{E}_0} \int_{\Gamma_e} \left[ (b \cdot n_e) + p_{Q_e^-} q_{Q_e^-} + (b \cdot n_e) - p_{Q_e^+} q_{Q_e^+} \right] \, d\sigma + (\alpha, q)
$$

$$
= \sum_{e \in \mathcal{E}_0} \int_{\Gamma_e} \left[ (b \cdot n_e) + p_{Q_e^-} q_{Q_e^-} + (b \cdot n_e) - p_{Q_e^+} q_{Q_e^+} \right] \, d\sigma + (\alpha, q) + \int b \cdot n p \, d\sigma.
$$

Using now the relations $(b \cdot n_e)_+ = -(b \cdot n_e)_-$ and $(b \cdot n_e)_- = -(b \cdot n_e)_+$ and the abbreviations $w_{Q_e^-} = p_r w_{Q_e^+} = w_r$ with $w = p, q$, one arrives at

$$
d_h(p, q) = (\alpha, q) + \int b \cdot n p \, d\sigma
$$

$$
= \sum_{e \in \mathcal{E}_0} \int_{\Gamma_e} \left[ (b \cdot n_e) + p(q_i - q_r) + (b \cdot n_e) - p_r(q_i - q_r) \right] \, d\sigma
$$

$$
= \sum_{e \in \mathcal{E}_0} \int_{\Gamma_e} \left[ (b \cdot n_e) + (b \cdot n_e) - p_r(q_i - q_r) \right] \, d\sigma
$$

$$
= \sum_{e \in \mathcal{E}_0} \int_{\Gamma_e} \left[ \frac{1}{2} (b \cdot n_e) + (q_i - q_r) \right] \, d\sigma
$$

$$
= \sum_{e \in \mathcal{E}_0} \int_{\Gamma_e} \left[ \frac{1}{2} (b \cdot n_e) + (q_i - q_r) \right] \, d\sigma.
$$

\[\square\]
Substituting $p = q$ in the formula (2.1), noticing that $-\mathbf{n}_e$ is an outward normal to $Q^+_e$ and $\mathbf{n}_e$ is an outward normal to $Q^-_e$, and that \[ \int_{Q} \text{div}(b \mathbf{p}^2) \, dx \, dy = \int_{\partial Q} b \cdot n p^2 \, d\sigma \] (since $p_0 = \text{const}$), one easily gets the next result.

**Lemma 2.2.** For any $q \in \mathbb{L}_h$ the bilinear form of (1.19) satisfies

\[ d_h(q, q) = \left( \alpha + \frac{1}{2} \nabla \cdot b \right) q, q + \frac{1}{2} \|q\|^2, \tag{2.2} \]

where

\[ \|q\|^2 = \sum_{e \in \mathcal{E}_h} \int_{\partial e} [q]^2 \, b \cdot n \, d\sigma. \tag{2.3} \]

The next lemma can be proved by direct evaluation [13].

**Lemma 2.3.** The following holds:

\[ B(w_h, q_h) = b(\gamma_h w_h, q_h) = -c(w_h, q_h) \quad \forall \ w_h \in \mathbb{H}_h, \quad q_h \in \mathbb{L}_h. \]

Now by Lemma 2.3 problem (1.20) becomes

\[ \begin{align*}
A(u_h, w_h) + B(w_h, p_h) &= 0 \quad \forall \ w_h \in \mathbb{H}_h, \tag{2.4} \\
B(u_h, q_h) - d_h(p_h, q_h) &= -f(q_h) \quad \forall \ q_h \in \mathbb{L}_h. \tag{2.5}
\end{align*} \]

The fact that $\gamma_h$ is a bounded self-adjoint operator with respect to the $L^2$ inner product can be found in [13].

**Lemma 2.4.** The following relations hold:

\[ (\gamma_h u_h, w_h) = (u_h, \gamma_h w_h) \quad \forall \ u_h, w_h \in \mathbb{H}_h. \tag{2.6} \]

There also exists a positive constant $C$ independent of $h$ such that

\[ \|\gamma_h u_h\|_0 \leq C\|u_h\|_0 \quad \forall \ u_h \in \mathbb{H}_h. \tag{2.7} \]

We next show that $A(w_h, w_h)$ is coercive for sufficiently small $h$.

**Lemma 2.5.** There exists a constant $\epsilon_0$ such that for all $0 < h \leq \epsilon_0$

\[ A(w_h, w_h) \geq C\epsilon_0^{-1} \|w_h\|_0^2 \quad \forall \ w_h \in \mathbb{H}_h, \]

where the constant $C$ is independent of $h$ and $\epsilon_0$.

**Proof.** Write $w_h = (u_h, v_h) \in \mathbb{H}_h$. Then

\[ a(w_h, w_h) = \sum_{i,j} u_h(c_{i+1/2,j}) \int_{Q^{i+1/2,j}_{i,j}} \tau_1(x, y) u_h(x, y) \, dxdy \tag{2.8} \]

\[ + \sum_{i,j} v_h(c_{i,j+1/2}) \int_{Q^{i,j+1/2}_{i,j}} \tau_2(x, y) v_h(x, y) \, dxdy \tag{2.9} \]

\[ = I + II. \tag{2.10} \]

It suffices to show that $I \geq C\epsilon_0^{-1} \|u_h\|_0^2$. Let $Q^-_{ij} := Q_{i-1/2,j} \cap Q_{ij}$ and $Q^+_{ij} := Q_{i+1/2,j} \cap Q_{ij}$. Then

\[ I = \sum_{i,j} u_h(c_{i-1/2,j}) \int_{Q^-_{ij}} \tau_1(x, y) u_h(x, y) \, dxdy \]

\[ + \sum_{i,j} u_h(c_{i+1/2,j}) \int_{Q^+_{ij}} \tau_1(x, y) u_h(x, y) \, dxdy \]

\[ = III + IV, \]
where

\[
III = \sum u_h(c_{i-1/2,j}) \int_{Q_{ij}^-} (\tau_1(x, y) - \tau_1(c_{ij})) u_h(x, y) dx dy
\]
\[
+ \sum u_h(c_{i-1/2,j}) \tau_1(c_{ij}) \int_{Q_{ij}^-} u_h(x, y) dx dy
\]
\[
= V + VI
\]

\[
IV = \sum u_h(c_{i+1/2,j}) \int_{Q_{ij}^+} (\tau_1(x, y) - \tau_1(c_{ij})) u_h(x, y) dx dy
\]
\[
+ \sum u_h(c_{i+1/2,j}) \tau_1(c_{ij}) \int_{Q_{ij}^+} u_h(x, y) dx dy
\]
\[
= VII + VIII.
\]

Using the linearity of \(u_h\) in \(x\) and constant in \(y\), we can easily derive by direct computation that

\[
VI + VIII \geq C \epsilon_1^{-1} \| u_h \|_0^2,
\]

while by Lipschitz continuity of \(\tau_1\) and the Simpson’s rule that

\[
V + VII \leq M \epsilon_1^{-1} h \left( \sum \int_{Q_{ij}^-} |u_h(c_{i-1/2,j}) u_h(x, y)| dx dy \right)
\]
\[
+ \int_{Q_{ij}^+} |u_h(c_{i+1/2,j}) u_h(x, y)| dx dy \right)
\]
\[
= M \epsilon_1^{-1} h \left( \sum \| u_h(c_{i-1/2,j}) \|_{Q_{ij}^-} \| u_h \|_{Q_{ij}^-} + \| u_h(c_{i+1/2,j}) \|_{Q_{ij}^+} \| u_h \|_{Q_{ij}^+} \right)
\]
\[
\leq 6 M \epsilon_1^{-1} h \sum \| u_h \|_{Q_{ij}}^2
\]
\[
= 6 M \epsilon_1^{-1} h \| u_h \|_{0}^2.
\]

Thus we have

\[
I \geq C \epsilon_1^{-1} \| u_h \|_0^2 - C \epsilon_1^{-1} h \| u_h \|_0^2
\]

and so

\[
I \geq C \epsilon_1^{-1} \| u_h \|_{0}^2
\]

for \(h\) sufficiently small. \(\Box\)

Now note that the bilinear form \(a\) of (1.16) is also well defined over \(H_h \times H_h\). With this in mind we state the following approximation property of \(\gamma_h\), whose simple proof can be found in [10] or directly verified.

There exists a constant \(C\) independent of \(h\) and \(\epsilon_1\) such that

\[
\|(I - \gamma_h) w_h\|_0 \leq C h \| w_h \|_{1,h},
\]

where

\[
\| w_h \|_{1,h}^2 = \| w_h \|_0^2 + \| w_h \|_{1,h}^2
\]
and the discrete seminorm $w_h = (w_h, x_h) \in H_h$ is defined as
\[ |w_h|_{1,h}^2 := \sum_Q |\nabla w_h|_{0,Q}^2 + |\nabla x_h|_{0,Q}^2. \]

**Lemma 2.6.** There exists a constant $C$ independent of $h$ and $\epsilon_1$ such that
\[ a(u, (I - \gamma_h)w_h) \leq C\epsilon_1^{-1}h|u|_1|w_h|_0 \quad \forall w_h \in H_h \]
and $\forall u \in H^1$.

**Proof.** Let $\mathcal{E}_h$ be the familiar interpolation operator from $H^1(\Omega)$ to $H_h$ with $\int_q \cdot \mathbf{n} \, d\sigma$, flux across edge, as its degrees of freedom ([22], pp. 550–554). Then
\[ \|q - \mathcal{E}_h q\|_0 \leq Ch|q|_1 \quad \forall q \in H^1(\Omega). \]
Denoting by $K_0$ the piecewise constant average of $K$, one has
\[ a(u, (I - \gamma_h)w_h) = ((K^{-1} - K_0^{-1})u, (I - \gamma_h)w_h) + (K_0^{-1}u, (I - \gamma_h)w_h) \]
\[ \leq C\epsilon_1^{-1}h|u|_0|w_h|_0 + (u, K_0^{-1}(I - \gamma_h)w_h) \]
\[ = C\epsilon_1^{-1}h|u|_0|w_h|_0 + (u, (I - \gamma_h)K_0^{-1}w_h) \]
\[ = C\epsilon_1^{-1}h|u|_0|w_h|_0 + (u - \mathcal{E}_h u, (I - \gamma_h)K_0^{-1}w_h) \]
\[ + ((I - \gamma_h)\mathcal{E}_h u, K_0^{-1}w_h), \]
where we have used the symmetry and boundedness of $\gamma_h$. Also, we used the fact that the coefficient $K^{-1}$ is locally Lipschitz. Now the second term on the right side of the last equation can be bounded by $C\epsilon_1^{-1}h|u|_1|w_h|_0 u$ using (2.13) and boundedness of $I - \gamma_h$, and the third term can be bounded using (2.11) and the fact that
\[ ||\mathcal{E}_h u||_{1,h} \leq C||u||_1. \]
This completes the proof. 

We next show that our covolume method (1.21) has a unique solution.

**Lemma 2.7.** For $h$ sufficiently small, there is a unique $(u_h, p_h) \in H_h \times L_h$ for the system:
\[ A(u_h, w_h) + B(w_h, p_h) = 0 \quad \forall w_h \text{ in } H_h, \]
\[ B(u_h, q_h) - d_h(p_h, q_h) = -f, q_h \quad \forall q_h \text{ in } L_h. \]

**Proof.** Define the bilinear form on $H_h \times L_h$:
\[ A(z_h, s; w_h, t) := A(z_h, w_h) + B(w_h, s) - B(z_h, t) + d_h(s, t). \]

Obviously the above system is equivalent to
\[ A(u_h, p_h; w_h, q_h) = \phi(w_h, q_h) \quad \forall (w_h, q_h) \in H_h \times L_h, \]
where
\[ \phi(w_h, q_h) := (f, q_h), \]
is a linear functional on $H_h \times L_h$. By Lemma 2.2
\begin{equation}
\tag{3.1}
A(w_h, q_h; w_h, q_h) = a(w_h, \gamma_h w_h) + \left( \left( \alpha + \frac{1}{2} \nabla \cdot b \right) q_h, q_h \right) + \frac{1}{2} ||q_h||^2.
\end{equation}

It suffices to show that $A(w_h, q_h; w_h, q_h) = 0$ admits only a zero solution, which can be inferred by the coercivity of $A(w_h, q_h; w_h, q_h)$ implied by Lemma 2.5 and (1.4).

3. Error estimates. We now prove our main convergence result.

**Theorem 3.1.** Let the rectangular partition family $\{Q_{ij}\}$ of the domain $\Omega$ be quasi-regular, and let $\{u_h, p_h\}$ be the solution of the problem (1.21) and $\{u, p\}$ of the problem (1.8). Then there exists a positive constant $C$ independent of $h$ and $\epsilon_1$,
\begin{equation}
\tag{3.1}
||u - u_h||_0 + \epsilon_1^{1/2} ||p - p_h||_0 \leq Ch(||u||_1 + h^{-1/2} \epsilon_1^{1/2} ||p||_1),
\end{equation}
provided that $u \in H^1$ and $p \in H^1$.

**Proof.** We will use, as a bridge of error analysis, the mixed method (1.10). Find $(\bar{u}_h, \bar{p}_h) \in H_h \times L_h$ such that
\begin{align}
\tag{3.2}
a(\bar{u}_h, w_h) + B(w_h, \bar{p}_h) &= 0 \quad \forall w_h \in H_h, \\
\tag{3.3}
B(\bar{u}_h, q_h) - d_h(\bar{p}_h, q_h) &= -(f, q_h) \quad \forall q_h \in L_h.
\end{align}

Let
\begin{equation}
\tag{3.4}
\bar{A}(z_h; s; w_h, t) := a(z_h, w_h) + B(w_h, s) - B(z_h, t) + d_h(s, t)
\end{equation}
be a bilinear form on $H_h \times L_h$. Then (3.2)–(3.3) is equivalent to the problem of finding $(\bar{u}_h, \bar{p}_h) \in H_h \times L_h$ such that
\begin{equation}
\tag{3.5}
\bar{A}(\bar{u}_h, \bar{p}_h; w_h, q_h) = \phi(w_h, q_h) \quad \forall (w_h, q_h) \in H_h \times L_h,
\end{equation}
where
\begin{equation}
\phi(w_h, q_h) := (f, q_h)
\end{equation}
is a linear functional on $H_h \times L_h$.

This system has the following known convergence result (unscaled version of Eq. 5.3 in [20]):
\begin{equation}
\tag{3.6}
||u - \bar{u}_h||_0 + \epsilon_1^{1/2} ||p - \bar{p}_h||_0 \leq Ch(||u||_1 + \epsilon_1^{1/2} h^{-1/2} ||p||_1),
\end{equation}
provided that $u \in H^1, p \in H^1$. (Note that the proof of this estimate depends on Lemma 2.2, which we have proved even for the mixed boundary condition case not covered by [20].)

On the other hand, our covolume method is equivalent to the problem of finding $(u_h, p_h) \in H_h \times L_h$ such that
\begin{equation}
\tag{3.7}
A(u_h, p_h; w_h, q_h) = \phi(w_h, q_h) \quad \forall (w_h, q_h) \in H_h \times L_h,
\end{equation}
where
\begin{equation}
\tag{3.8}
A(z_h; s; w_h, t) := A(z_h, w_h) + B(w_h, s) - B(z_h, t) + d_h(s, t).
\end{equation}
Using the bilinearity of $A$ and $B$, (3.7), (3.5), we have

$$
A(\tilde{u}_h - u_h, \bar{p}_h - p_h; w_h, q_h) = A(\tilde{u}_h, \bar{p}_h, w_h, q_h) - A(u_h, p_h, w_h, q_h)
= A(\tilde{u}_h, \bar{p}_h; w_h, q_h) - \tilde{A}(\tilde{u}_h, \bar{p}_h; w_h, q_h).
$$

Hence by (3.8) and (3.4) we have

$$
A(\tilde{u}_h - u_h, \bar{p}_h - p_h; w_h, q_h) = a(\tilde{u}_h, \gamma_h w_h) - a(u_h, w_h).
$$

Since the total error $e_h = (u - \tilde{u}_h) + (\tilde{u}_h - u_h)$, by the triangle inequality it suffices to estimate $\tilde{u}_h - u_h$. Now set $w_h = \tilde{e}_h := \tilde{u}_h - u_h$ and $q_h = \tilde{\tau}_h := \bar{p}_h - p_h$ in the above equation to get the error equation

$$
A(\tilde{e}_h, \tilde{\tau}_h; \tilde{e}_h, \tilde{\tau}_h) = a(u_h, (\gamma_h - I)\tilde{e}_h)
= a(u_h - u, (\gamma_h - I)\tilde{e}_h) + a(u, (\gamma_h - I)\tilde{e}_h)
\leq Ch^{-1}||u||_0||\tilde{e}_h||_0 + Ch^{-1}||u||_1||\tilde{e}_h||_0
\leq Ch^{-1}||||u||_1 + \epsilon_1^{1/2}h^{-1/2}||p||_1||\tilde{e}_h||_0,
$$

where we have used the boundedness of $\gamma_h - I$, (2.12) in deriving (3.12), and (3.6) in deriving (3.13). Applying (2.14) to the left side of (3.10), we get from (3.13) that

$$
a \cdot (\tilde{e}_h, \gamma_h \tilde{e}_h) + \left( \alpha + \frac{1}{2} \nabla \cdot b \right) \tilde{\tau}_h \tilde{\tau}_h = \frac{1}{2}||\tilde{\tau}_h||^2
\leq C\epsilon_1^{-1}h \left[ ||u||_1 + h^{-1/2}\epsilon_1^{1/2}||p||_1 \right] ||\tilde{e}_h||_0.
$$

Invoking (1.4) and Lemma 2.5 completes the estimate for $\tilde{e}_h$. As for the estimate for $p - p_h$, it suffices to estimate $\tilde{\tau}_h = \bar{p}_h - p_h$ due to the triangle inequality and (3.6). From (3.14), we have

$$
||\tilde{\tau}_h||_0^2 \leq C\epsilon_1^{-1}h \left[ ||u||_1 + \epsilon_1^{1/2}h^{-1/2}||p||_1 \right] ||\tilde{e}_h||_0.
$$

Now applying the just-obtained upper bound of $||\tilde{e}_h||_0$ completes the proof.

4. **Numerical results.** In this section we use quadratures to obtain the final discretization equations and show some numerical results to demonstrate the error behavior of the studied mixed covolume method. We shall deviate from the notation of the previous sections slightly for the convenience of reporting numerical results.

We consider in this section the error behavior of the following problem:

$$
\nabla \cdot (-\epsilon K \nabla p + b) p = c_0 p = f(x, y), \quad (x, y) \in \Omega = (0, 1)^2.
$$

The exact solution chosen is $p = x(1 - x)y(1 - y)$ and Dirichlet boundary conditions are imposed. The coefficients of the operator are $K = \text{diag}(k_1, k_2)$, $k_1 = 1 + 10x^2 + y^2$, $k_2 = 1 + x^2 + 10y^2$, $c_0 = 1$, and $b = (b_1, b_2)$, where

$$
b_1 = -\cos \alpha (1 - x \cos \alpha),
b_2 = -\sin \alpha (1 - y \sin \alpha),
$$

for angles $\alpha = -\frac{\pi}{4}, -\frac{\pi}{4}, 0, \frac{\pi}{4}, \frac{\pi}{4}$. Note that $\nabla \cdot b = 1$ so that condition (1.4) is satisfied and that $\epsilon = 1, 10^{-2}, 10^{-4}, 10^{-6}$. 

For the flux variable \( \mathbf{u} = (u_1, u_2) \) we used the lowest order Raviart–Thomas piecewise polynomial spaces on isosceles rectangles (squares) of size \( h \times h \), for \( h = 2^{-4}, 2^{-5}, 2^{-6}, 2^{-7} \). The concentration variable \( p \) corresponded to piecewise constants on the same rectangular elements.

In general, to evaluate the integrals

\[
\int_{Q_{i+1/2,j}} k_1^{-1}(x, y)u_1(x, y)dxdy = \int_{x_i}^{x_{i+1}} \int_{y_{j-1/2}}^{y_{j+1/2}} k_1^{-1}(x, y)u_1(x, y)dxdy \\
+ \int_{x_i}^{x_{i+1}} k_1^{-1}(x, y)u_1(x, y)dxdy,
\]

where \( Q_{i+1/2,j} = [x_i, x_{i+1}] \times [y_{j-1/2}, y_{j+1/2}] \), \( h_x = x_r - x_{r-1} \), we used the midpoint quadrature formula in \( y \)-direction and the trapezoidal rule in \( x \)-direction. This led, for example, for the first integral above, to the following expression:

\[
\approx \frac{h_x h_y}{4} \left[ k_1^{-1}(c_{i,j})u_1(c_{i,j}) + k_1^{-1}(c_{i+1/2,j})u_1(c_{i+1/2,j}) \right]
\]

\[
= \frac{h_x h_y}{4} \left[ k_1^{-1}(c_{i,j}) \frac{u_1(c_{i-1/2,j}) + u_1(c_{i+1/2,j})}{2} + k_1^{-1}(c_{i+1/2,j})u_1(c_{i+1/2,j}) \right]
\]

Note that the definition of \( Q_{i+1/2,j} \) at the boundary needs to be modified and we take either the left half or the right half, depending on the location of the boundary. One of the pressures in (1.12) is taken from the boundary value.

We remark that the above quadrature formula actually symmetrizes the bilinear form \( a(u_h, v_h) \), for \( u_h, v_h \in H_h, v_h \in Y_h \); namely, the resulting matrix \( A \) is symmetric. Also, we note that \( A \) is block-diagonal (diagonal blocks \( A_i \), \( i = 1, 2 \), corresponding to the variables \( u_1 \) and \( u_2 \), respectively) and each block \( A_i \) is block-diagonal itself, with tridiagonal (band) matrices if the unknowns corresponding to \( u_1 \) are ordered linewise along the \( x \)-direction (each horizontal line giving rise to a block) and the unknowns corresponding to \( u_2 \) are ordered linewise along \( y \)-directions (with blocks being the unknowns grouped on a given vertical line).

The edge integrals involving \( \mathbf{b} \cdot \mathbf{n} \) are computed exactly, which is simple for our particular choice of the field \( \mathbf{b} \).

After the discretization one ends up with the following linear system of equations to be solved,

\[
(4.3) \quad \mathbf{A} \begin{bmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \\ \mathbf{P} \end{bmatrix} = \begin{bmatrix} \mathbf{rhs}_{U_1} \\ \mathbf{rhs}_{U_2} \\ \mathbf{rhs}_P \end{bmatrix},
\]

with the saddle-point–like stiffness matrix:

\[
(4.4) \quad \mathbf{A} = \begin{bmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & -\mathbf{C} \end{bmatrix}.
\]

We use the fact that \( \mathbf{A} \) has simple structure, block-diagonal with blocks being tridiagonal matrices (assuming the linewise ordering explained above), to effectively compute the inverse actions of \( \mathbf{A} \). That is, one can eliminate the unknown \( \mathbf{U} = \begin{bmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{bmatrix} \) and solve the reduced system,

\[
(4.5) \quad (\mathbf{C} + \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T)\mathbf{P} = -\mathbf{rhs}_P + \mathbf{B}\mathbf{A}^{-1} \begin{bmatrix} \mathbf{rhs}_{U_1} \\ \mathbf{rhs}_{U_2} \end{bmatrix}.
\]
We do not compute the reduced Schur matrix $S \equiv C + BA^{-1}B^T$ explicitly, but rather take advantage of the fact that the actions of $S$ on vectors are inexpensively available through solutions of tridiagonal problems to get the inverse action of $A$. To solve the reduced problem, we used a generalized conjugate gradient least squares method (GCG-LS) as derived in Axelsson [1] (a mathematically-equivalent-to-the-GMRES method, see Saad [24]) with block-ILU preconditioning. The preconditioned method (GCG-LS) as derived in Axelsson [1] (a mathematically-equivalent-to-the-GMRES method, see Saad [24]) with block-ILU preconditioning. The preconditioned method (GCG-LS) as derived in Axelsson [1] (a mathematically-equivalent-to-the-GMRES method, see Saad [24]) with block-ILU preconditioning. The preconditioned method (GCG-LS) as derived in Axelsson [1] (a mathematically-equivalent-to-the-GMRES method, see Saad [24]) with block-ILU preconditioning.

Table 4.1

<table>
<thead>
<tr>
<th>$h$</th>
<th>$\delta_{u_1}$</th>
<th>$\delta_{u_2}$</th>
<th>$\approx \delta$</th>
<th>$\approx \delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h = 1/16$</td>
<td>3.3e-5</td>
<td>2.7e-5</td>
<td>1.0e-5</td>
<td>1.0e-5</td>
</tr>
<tr>
<td>$h = 1/32$</td>
<td>3.6e-5</td>
<td>2.9e-5</td>
<td>1.1e-5</td>
<td>1.1e-5</td>
</tr>
<tr>
<td>$h = 1/64$</td>
<td>2.0e-5</td>
<td>3.4e-4</td>
<td>1.4e-4</td>
<td>1.4e-4</td>
</tr>
<tr>
<td>$h = 1/128$</td>
<td>1.1e-5</td>
<td>4.2e-4</td>
<td>1.9e-4</td>
<td>1.9e-4</td>
</tr>
</tbody>
</table>

# unknowns | 800 | 3136 | 12416 | 49408 |
# iterations | 17 | 23 | 45 | 112 |
$\varrho$ | 0.28 | 0.40 | 0.62 | 0.83 |

We show in Tables 4.1–4.14, in addition to the error behavior of the covolume discretization method, also $\varrho$ and the number of iterations, where

$$\varrho = \left( \frac{\|M^{-1}r\|}{\|M^{-1}r_0\|} \right)^{\# \text{iterations}}$$

was an average reduction factor.

More specifically, denote $x_i = ih_x, y_j = jh_y, i = 0, 1, 2, \ldots, n_x, j = 0, 1, 2, \ldots, n_y, h_x = h_y = h, n_x = n_y = n = 1/h$, for a given $h = 2^{-4}, 2^{-5}, 2^{-6}, 2^{-7}$. In Tables 4.1–4.14, for a given $\epsilon$ and angle $\alpha$, we show the following items:
Table 4.2
Error behavior for $\alpha = \frac{\pi}{4}$, $\epsilon = 1$.

<table>
<thead>
<tr>
<th></th>
<th>$h = 1/16$</th>
<th>$h = 1/32$</th>
<th>$h = 1/64$</th>
<th>$h = 1/128$</th>
<th>$\approx$ order</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta_\rho$</td>
<td>9.86e-5</td>
<td>6.07e-5</td>
<td>3.33e-5</td>
<td>1.74e-5</td>
<td>1</td>
</tr>
<tr>
<td>$\delta_{u1}$</td>
<td>3.56e-3</td>
<td>1.33e-3</td>
<td>5.64e-4</td>
<td>2.57e-4</td>
<td>$&gt;$ 1</td>
</tr>
<tr>
<td>$\delta_{u2}$</td>
<td>3.56e-3</td>
<td>1.33e-3</td>
<td>5.64e-4</td>
<td>2.57e-4</td>
<td>$&gt;$ 1</td>
</tr>
<tr>
<td># unknowns</td>
<td>800</td>
<td>3136</td>
<td>12416</td>
<td>49408</td>
<td></td>
</tr>
<tr>
<td># iterations</td>
<td>17</td>
<td>23</td>
<td>45</td>
<td>111</td>
<td></td>
</tr>
<tr>
<td>$\varrho$</td>
<td>0.28</td>
<td>0.40</td>
<td>0.62</td>
<td>0.82</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.3
Error behavior for $\alpha = \frac{\pi}{2}$, $\epsilon = 1$.

<table>
<thead>
<tr>
<th></th>
<th>$h = 1/16$</th>
<th>$h = 1/32$</th>
<th>$h = 1/64$</th>
<th>$h = 1/128$</th>
<th>$\approx$ order</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta_\rho$</td>
<td>5.33e-5</td>
<td>3.62e-5</td>
<td>2.07e-5</td>
<td>1.10e-5</td>
<td>1</td>
</tr>
<tr>
<td>$\delta_{u1}$</td>
<td>3.04e-3</td>
<td>1.06e-3</td>
<td>4.29e-4</td>
<td>1.90e-4</td>
<td>$&gt;$ 1</td>
</tr>
<tr>
<td>$\delta_{u2}$</td>
<td>2.72e-3</td>
<td>9.09e-4</td>
<td>3.49e-4</td>
<td>1.49e-4</td>
<td>$&gt;$ 1</td>
</tr>
<tr>
<td># unknowns</td>
<td>800</td>
<td>3136</td>
<td>12416</td>
<td>49408</td>
<td></td>
</tr>
<tr>
<td># iterations</td>
<td>17</td>
<td>22</td>
<td>45</td>
<td>113</td>
<td></td>
</tr>
<tr>
<td>$\varrho$</td>
<td>0.28</td>
<td>0.39</td>
<td>0.62</td>
<td>0.83</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.4
Error behavior for $\alpha = 0$, $\epsilon = 0.01$.

<table>
<thead>
<tr>
<th></th>
<th>$h = 1/16$</th>
<th>$h = 1/32$</th>
<th>$h = 1/64$</th>
<th>$h = 1/128$</th>
<th>$\approx$ order</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta_\rho$</td>
<td>1.77e-3</td>
<td>9.09e-4</td>
<td>4.62e-4</td>
<td>2.33e-4</td>
<td>1</td>
</tr>
<tr>
<td>$\delta_{u1}$</td>
<td>4.09e-4</td>
<td>2.09e-4</td>
<td>1.03e-4</td>
<td>5.00e-5</td>
<td>1</td>
</tr>
<tr>
<td>$\delta_{u2}$</td>
<td>3.46e-4</td>
<td>1.67e-4</td>
<td>8.25e-5</td>
<td>4.11e-5</td>
<td>1</td>
</tr>
<tr>
<td># unknowns</td>
<td>800</td>
<td>3136</td>
<td>12416</td>
<td>49408</td>
<td></td>
</tr>
<tr>
<td># iterations</td>
<td>15</td>
<td>17</td>
<td>24</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>$\varrho$</td>
<td>0.24</td>
<td>0.29</td>
<td>0.40</td>
<td>0.67</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.5
Error behavior for $\alpha = \frac{\pi}{4}$, $\epsilon = 0.01$.

<table>
<thead>
<tr>
<th></th>
<th>$h = 1/16$</th>
<th>$h = 1/32$</th>
<th>$h = 1/64$</th>
<th>$h = 1/128$</th>
<th>$\approx$ order</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta_\rho$</td>
<td>2.88e-3</td>
<td>1.51e-3</td>
<td>7.78e-4</td>
<td>3.95e-4</td>
<td>1</td>
</tr>
<tr>
<td>$\delta_{u1}$</td>
<td>5.01e-4</td>
<td>2.49e-4</td>
<td>1.21e-4</td>
<td>5.91e-5</td>
<td>1</td>
</tr>
<tr>
<td>$\delta_{u2}$</td>
<td>5.01e-4</td>
<td>2.49e-4</td>
<td>1.21e-4</td>
<td>5.91e-5</td>
<td>1</td>
</tr>
<tr>
<td># unknowns</td>
<td>800</td>
<td>3136</td>
<td>12416</td>
<td>49408</td>
<td></td>
</tr>
<tr>
<td># iterations</td>
<td>15</td>
<td>17</td>
<td>25</td>
<td>49</td>
<td></td>
</tr>
<tr>
<td>$\varrho$</td>
<td>0.24</td>
<td>0.28</td>
<td>0.42</td>
<td>0.65</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.6
Error behavior for $\alpha = -\frac{\pi}{4}$, $\epsilon = 0.01$.

<table>
<thead>
<tr>
<th></th>
<th>$h = 1/16$</th>
<th>$h = 1/32$</th>
<th>$h = 1/64$</th>
<th>$h = 1/128$</th>
<th>$\approx$ order</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta_\rho$</td>
<td>3.10e-3</td>
<td>1.68e-3</td>
<td>8.90e-4</td>
<td>4.64e-4</td>
<td>1</td>
</tr>
<tr>
<td>$\delta_{u1}$</td>
<td>5.77e-4</td>
<td>2.95e-4</td>
<td>1.49e-4</td>
<td>7.50e-5</td>
<td>1</td>
</tr>
<tr>
<td>$\delta_{u2}$</td>
<td>1.70e-3</td>
<td>8.43e-4</td>
<td>4.45e-4</td>
<td>2.04e-4</td>
<td>1</td>
</tr>
<tr>
<td># unknowns</td>
<td>800</td>
<td>3136</td>
<td>12416</td>
<td>49408</td>
<td></td>
</tr>
<tr>
<td># iterations</td>
<td>15</td>
<td>17</td>
<td>24</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td>$\varrho$</td>
<td>0.24</td>
<td>0.27</td>
<td>0.41</td>
<td>0.63</td>
<td></td>
</tr>
</tbody>
</table>
Table 4.7
Error behavior for $\alpha = 0$, $\epsilon = 10^{-4}$.

<table>
<thead>
<tr>
<th>$h$</th>
<th>$\delta_p$</th>
<th>$\delta_{u_1}$</th>
<th>$\delta_{u_2}$</th>
<th># unknowns</th>
<th># iterations</th>
<th>$\varrho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1/16$</td>
<td>2.29e-3</td>
<td>1.14e-3</td>
<td>5.74e-4</td>
<td>10</td>
<td>0.10</td>
<td>0.70</td>
</tr>
<tr>
<td>$1/32$</td>
<td>4.40e-5</td>
<td>1.10e-5</td>
<td>5.22e-6</td>
<td>13</td>
<td>0.18</td>
<td>0.25</td>
</tr>
<tr>
<td>$1/64$</td>
<td>5.36e-6</td>
<td>2.50e-6</td>
<td>1.20e-6</td>
<td>15</td>
<td>0.22</td>
<td>0.30</td>
</tr>
<tr>
<td>$1/128$</td>
<td>5.95e-7</td>
<td>2.89e-6</td>
<td>5.95e-7</td>
<td>15</td>
<td>0.25</td>
<td>0.35</td>
</tr>
</tbody>
</table>

Table 4.8
Error behavior for $\alpha = \frac{\pi}{4}$, $\epsilon = 10^{-4}$.

<table>
<thead>
<tr>
<th>$h$</th>
<th>$\delta_p$</th>
<th>$\delta_{u_1}$</th>
<th>$\delta_{u_2}$</th>
<th># unknowns</th>
<th># iterations</th>
<th>$\varrho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1/16$</td>
<td>3.57e-3</td>
<td>1.84e-3</td>
<td>9.33e-4</td>
<td>10</td>
<td>0.10</td>
<td>0.70</td>
</tr>
<tr>
<td>$1/32$</td>
<td>4.20e-5</td>
<td>2.91e-5</td>
<td>1.66e-5</td>
<td>13</td>
<td>0.18</td>
<td>0.25</td>
</tr>
<tr>
<td>$1/64$</td>
<td>4.20e-5</td>
<td>2.91e-5</td>
<td>1.66e-5</td>
<td>15</td>
<td>0.22</td>
<td>0.30</td>
</tr>
<tr>
<td>$1/128$</td>
<td>4.68e-4</td>
<td>8.26e-6</td>
<td>8.26e-6</td>
<td>15</td>
<td>0.25</td>
<td>0.35</td>
</tr>
</tbody>
</table>

Table 4.9
Error behavior for $\alpha = \frac{\pi}{2}$, $\epsilon = 10^{-4}$.

<table>
<thead>
<tr>
<th>$h$</th>
<th>$\delta_p$</th>
<th>$\delta_{u_1}$</th>
<th>$\delta_{u_2}$</th>
<th># unknowns</th>
<th># iterations</th>
<th>$\varrho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1/16$</td>
<td>2.29e-3</td>
<td>1.14e-3</td>
<td>5.74e-4</td>
<td>10</td>
<td>0.10</td>
<td>0.70</td>
</tr>
<tr>
<td>$1/32$</td>
<td>4.40e-5</td>
<td>1.10e-5</td>
<td>5.22e-6</td>
<td>13</td>
<td>0.18</td>
<td>0.25</td>
</tr>
<tr>
<td>$1/64$</td>
<td>5.36e-6</td>
<td>2.50e-6</td>
<td>1.20e-6</td>
<td>15</td>
<td>0.22</td>
<td>0.30</td>
</tr>
<tr>
<td>$1/128$</td>
<td>5.95e-7</td>
<td>2.89e-6</td>
<td>5.95e-7</td>
<td>15</td>
<td>0.25</td>
<td>0.35</td>
</tr>
</tbody>
</table>

Table 4.10
Error behavior for $\alpha = -\frac{\pi}{4}$, $\epsilon = 10^{-4}$.

<table>
<thead>
<tr>
<th>$h$</th>
<th>$\delta_p$</th>
<th>$\delta_{u_1}$</th>
<th>$\delta_{u_2}$</th>
<th># unknowns</th>
<th># iterations</th>
<th>$\varrho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1/16$</td>
<td>3.93e-3</td>
<td>2.04e-3</td>
<td>1.03e-3</td>
<td>10</td>
<td>0.10</td>
<td>0.70</td>
</tr>
<tr>
<td>$1/32$</td>
<td>3.99e-5</td>
<td>2.88e-5</td>
<td>1.67e-5</td>
<td>13</td>
<td>0.19</td>
<td>0.23</td>
</tr>
<tr>
<td>$1/64$</td>
<td>7.25e-5</td>
<td>5.06e-5</td>
<td>3.36e-5</td>
<td>15</td>
<td>0.23</td>
<td>0.36</td>
</tr>
</tbody>
</table>

Table 4.11
Error behavior for $\alpha = 0$, $\epsilon = 10^{-6}$.

<table>
<thead>
<tr>
<th>$h$</th>
<th>$\delta_p$</th>
<th>$\delta_{u_1}$</th>
<th>$\delta_{u_2}$</th>
<th># unknowns</th>
<th># iterations</th>
<th>$\varrho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1/16$</td>
<td>2.36e-3</td>
<td>1.18e-3</td>
<td>5.95e-4</td>
<td>10</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>$1/32$</td>
<td>3.57e-7</td>
<td>2.54e-7</td>
<td>1.71e-7</td>
<td>5</td>
<td>0.006</td>
<td>0.02</td>
</tr>
<tr>
<td>$1/64$</td>
<td>5.61e-8</td>
<td>2.63e-8</td>
<td>1.27e-8</td>
<td>6</td>
<td>0.006</td>
<td>0.07</td>
</tr>
</tbody>
</table>
### Table 4.12
Error behavior for $\alpha = \frac{\pi}{4}$, $\epsilon = 10^{-6}$.

<table>
<thead>
<tr>
<th>$h$</th>
<th>$\delta_p$</th>
<th>$\delta_{u_1}$</th>
<th>$\delta_{u_2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h = 1/16$</td>
<td>3.64e-3</td>
<td>5.06e-7</td>
<td>5.06e-7</td>
</tr>
<tr>
<td>$h = 1/32$</td>
<td>1.91e-3</td>
<td>4.31e-7</td>
<td>4.31e-7</td>
</tr>
<tr>
<td>$h = 1/64$</td>
<td>9.86e-4</td>
<td>3.33e-7</td>
<td>3.33e-7</td>
</tr>
<tr>
<td>$h = 1/128$</td>
<td>4.99e-4</td>
<td>2.44e-7</td>
<td>2.44e-7</td>
</tr>
</tbody>
</table>

| $\#$ unknowns | 800 | 3.136 | 12.416 | 49.408 |
| $\#$ iterations | 3.001 | 0.006 | 0.001 | 0.003 |

### Table 4.13
Error behavior for $\alpha = \frac{\pi}{4}$, $\epsilon = 10^{-6}$.

<table>
<thead>
<tr>
<th>$h$</th>
<th>$\delta_p$</th>
<th>$\delta_{u_1}$</th>
<th>$\delta_{u_2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h = 1/16$</td>
<td>2.36e-3</td>
<td>5.61e-8</td>
<td>3.57e-7</td>
</tr>
<tr>
<td>$h = 1/32$</td>
<td>1.18e-3</td>
<td>2.63e-8</td>
<td>2.54e-7</td>
</tr>
<tr>
<td>$h = 1/64$</td>
<td>5.95e-4</td>
<td>1.27e-8</td>
<td>1.71e-7</td>
</tr>
<tr>
<td>$h = 1/128$</td>
<td>2.97e-4</td>
<td>6.23e-9</td>
<td>1.00e-7</td>
</tr>
</tbody>
</table>

| $\#$ unknowns | 800 | 3.136 | 12.416 | 49.408 |
| $\#$ iterations | 3.001 | 0.006 | 0.001 | 0.003 |

### Table 4.14
Error behavior for $\alpha = \frac{\pi}{4}$, $\epsilon = 10^{-6}$.

<table>
<thead>
<tr>
<th>$h$</th>
<th>$\delta_p$</th>
<th>$\delta_{u_1}$</th>
<th>$\delta_{u_2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h = 1/16$</td>
<td>3.97e-3</td>
<td>4.75e-7</td>
<td>7.49e-7</td>
</tr>
<tr>
<td>$h = 1/32$</td>
<td>2.07e-3</td>
<td>4.22e-7</td>
<td>5.58e-7</td>
</tr>
<tr>
<td>$h = 1/64$</td>
<td>1.06e-3</td>
<td>3.33e-7</td>
<td>4.06e-7</td>
</tr>
<tr>
<td>$h = 1/128$</td>
<td>5.39e-4</td>
<td>2.44e-7</td>
<td>2.90e-7</td>
</tr>
</tbody>
</table>

| $\#$ unknowns | 800 | 3.136 | 12.416 | 49.408 |
| $\#$ iterations | 3.001 | 0.006 | 0.001 | 0.003 |

(i) $\delta_p = \| I_h p - p_h \|_h \equiv \left[ \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} h_x h_y (p(x_i - \frac{1}{2} h_x, y_j - \frac{1}{2} h_y) - p_h(x_i - \frac{1}{2} h_x, y_j - \frac{1}{2} h_y))^2 \right]^{\frac{1}{2}}$, i.e., a discrete $L^2$-norm of the error $p - p_h$;

(ii) $\delta_{u_1} = \| I_h u_1 - u_{h, 1} \|_h \equiv \left[ \sum_{i=0}^{n_x} \sum_{j=1}^{n_y} h_x h_y (u_1(x_i, y_j - \frac{1}{2} h_y) - u_{h, 1}(x_i, y_j - \frac{1}{2} h_y))^2 \right]^{\frac{1}{2}}$, i.e., a discrete $L^2$-norm of the error $u_1 - u_{h, 1}$;

(iii) $\delta_{u_2} = \| I_h u_2 - u_{h, 2} \|_h \equiv \left[ \sum_{i=1}^{n_x} \sum_{j=0}^{n_y} h_x h_y (u_2(x_i, y_j - \frac{1}{2} h_y) - u_{h, 2}(x_i - \frac{1}{2} h_x, y_j))^2 \right]^{\frac{1}{2}}$, i.e., a discrete $L^2$-norm of the error $u_2 - u_{h, 2}$;

(iv) the number of iterations of the preconditioned GCG–LS method;

(v) the average reduction factors $\varrho$ (4.6);

(vi) the total number of unknowns (for both $U$ and $P$).

It is noticeable that the order $h^p$ of the error estimate proven in the present paper is attained for small $\epsilon = 10^{-6}$; see, for example, Tables 4.12 and 4.14. It turns out that our experiments suggest first order approximation of the concentration variable for all values of $\epsilon$ we have tested. The fast convergence of the preconditioned GCG-
LS method for small $\epsilon$ is explained due to the fact that, in that case, the reduced matrix $S \approx C$ is almost triangular and because our block-ILU factorization matrix turned out to be very accurate in that case. This also agrees with the experience of the performance of block-ILU methods in the case of streamline diffusion finite element discretization of convection-diffusion problems reported in Axelsson, Polman, Eijkhout, and Vassilevski [2]. See also, Lazarov, Mishev, and Vassilevski [19] and Iliev, Makarov, and Vassilevski [18] for finite difference or finite volume discretizations, as well as in the least-squares mixed finite element discretization of convection–diffusion problems; cf., Carey, Pehlivanov, and Vassilevski [5]. We finally remark that the preconditioners used in the present paper are not the main issue; rather, they are tools to solve the resulting systems.

Acknowledgment. The authors are grateful to one of the referees for her/his constructive comments on the original manuscript.

REFERENCES


