ESAIM: M2AN
Vol. 40, N° 1, 2006, pp. 123–147
DOI: 10.1051/m2an:2006001

ESAIM: Mathematical Modelling and Numerical Analysis
www.edpsciences.org/m2an

NEW MIXED FINITE VOLUME METHODS FOR SECOND ORDER ELIPTIC PROBLEMS *

Kwang Y. Kim[1]

Abstract. In this paper we introduce and analyze new mixed finite volume methods for second order elliptic problems which are based on $H(\text{div})$-*conforming* approximations for the vector variable and *discontinuous* approximations for the scalar variable. The discretization is fulfilled by combining the ideas of the traditional finite volume box method and the local discontinuous Galerkin method. We propose two different types of methods, called Methods I and II, and show that they have distinct advantages over the mixed methods used previously. In particular, a clever elimination of the vector variable leads to a primal formulation for the scalar variable which closely resembles discontinuous finite element methods. We establish error estimates for these methods that are optimal for the scalar variable in both methods and for the vector variable in Method II.

Mathematics Subject Classification. 65F10, 65N15, 65N30.

Received: August 19, 2004. Revised: August 5, 2005.

1. Introduction

Mixed methods for elliptic problems have been a subject of active research and widely used in a variety of applications because of their desirable properties such as good approximation of the vector quantity and local conservation of mass. Typical examples are approximation of fluid velocities in porous media flow problems and electric currents in semiconductor simulation.

There are many mixed methods developed so far. Mixed finite element methods are the standard finite element methods applied to the mixed formulation of the underlying problem with the approximation spaces subject to the *inf-sup* condition; see, for example [7–10, 14, 16, 27, 28, 32, 33, 35, 37] and the references therein. Finite volume techniques have been also applied to the mixed systems of elliptic problems. Mixed covolume methods are based on the idea of covolumes which provide control volumes around each unknown, and produce a discrete problem comparable to a lowest-order mixed finite element method (*cf.* [12, 17–19]). A mixed finite volume method on non-staggered triangular grids was proposed by Courbet and Croisille [23] and later extended to quadrilateral grids by Chou, Kwak and Kim [20]. It has many distinct features when compared with the two methods above, like violation of the *inf-sup* condition and no use of covolumes. See also [24, 31] for a Petrov-Galerkin mixed method and [2, 5, 38, 39] for finite difference type methods.

Keywords and phrases. Mixed method, finite volume method, discontinuous finite element method, conservative method.

* *This work was supported by the Post-doctoral Fellowship Program of Korea Science & Engineering Foundation (KOSEF). This work was completed while the author was visiting Department of Mathematics, Southern Methodist University, USA.*

[1] Department of Aerospace Engineering, Korea Advanced Institute of Science and Technology, Daejeon, 305-701 South Korea.
toheart@acoustic.kaist.ac.kr

For all the methods mentioned above the vector approximation is sought in $H(\mathrm{div})$-*conforming* spaces, requiring it to have continuous normal components across the interelement boundaries. For mixed finite element or covolume methods, this makes it very costly to solve the resulting discrete systems either directly or iteratively. One popular way to avoid this difficulty is to relax the continuity of normal components of the vector variable via the Lagrange multipliers so that it may be eliminated locally (usually called the mixed-hybrid method). By following this idea, it was established in [1, 3] that the whole discrete mixed system can be reduced to some nonconforming finite element method involving only the Lagrange multipliers. On the other hand, for mixed finite volume methods of Courbet and Croisille in which a nonconforming approximation is directly considered for the scalar variable, the vector variable can be decoupled in a clever way with no help of Lagrange multipliers, and this results in a nonconforming finite element method for the scalar variable only. However, this method relies heavily on the use of a nonconforming element for the scalar variable with the interface degrees of freedom, and so does not seem straightforward to extend it to elements of arbitrary orders, although a quadratic element was constructed in [25].

Recently, there has been a growing interest on the local discontinuous Galerkin methods (abbreviated as the LDG methods) which are based on discontinuous spaces for *both* the scalar and the vector approximations. Thus one can eliminate the vector variable locally element-by-element to express the whole system in terms of the scalar variable alone, which leads to a discontinuous finite element method for the scalar variable. In order to communicate information between neighboring elements, *numerical fluxes* are introduced on the interface integrals which are given as linear combinations of the averages and jumps of the nearby unknowns. We refer the interested reader to [4, 6, 11, 13, 15, 22, 26, 34] for more details.

The goal of this paper is to introduce and analyze new mixed finite volume methods for second order elliptic problems. We choose to use $H(\mathrm{div})$-*conforming* spaces for the vector approximation and *discontinuous* spaces for the scalar approximation. The discretization is fulfilled by combining the concept of the numerical fluxes used in LDG methods and the simple finite volume techniques used in the previous mixed finite volume methods on non-staggered grids.

We show that the new methods have combined advantages of the mixed methods previously developed. First, they have the same local mass conservation properties as mixed finite element methods, and can be defined for elements of arbitrary orders. Second, unlike LDG methods, we do not need to abandon the continuity of normal components of the vector for the purpose of eliminating it and obtaining a discrete system for the scalar alone. In fact, elimination of the vector variable can be done in a clever way, as in [20], which leads to a primal formulation for the scalar variable, and the vector variable can be recovered locally from the computed scalar variable. It is worthwhile to mention that this fact not only provides a convenient way of implementation but also a way of deriving error estimates without resort to the theory of saddle-point problems. In light of these facts our new methods can be viewed as a generalization of mixed finite volume methods of [20, 23] to elements of arbitrary orders, and at the same time, as an improvement over the standard mixed finite element methods which allows to decouple the vector variable from the scalar one without help of the Lagrange multipliers.

Two different types of methods, referred to as Methods I and II, will be proposed. For Method I equal-order approximations are used for both the vector and the scalar variables, which leads to a standard discontinuous finite element method without the usual symmetrizing (or anti-symmetrizing) term. For Method II a one-order higher space is used for the scalar approximation, resulting in a discontinuous finite element method of the same form which, however, incorporates the L^2-projection of interface averages and jumps onto the lower-order normal trace spaces. As we will see later, this projection of interface averages and jumps offers some computational convenience when compared to the standard form, and to the author's knowledge, has never been considered before. We will also establish error estimates that are optimal for the scalar variable in both methods and for the vector variable in Method II. Only a suboptimal result is obtained for the vector variable in Method I.

The remainder of the paper is organized as follows. In the next section we state the model problem and then introduce some notation and finite element spaces with their approximation properties. We also describe the main ideas of how the model problem is discretized. In Sections 3 and 4, our two methods, Methods I and II, are constructed and analyzed on triangular grids. Discontinuous finite element formulations for the scalar variable and their uniform ellipticity are derived, and some error estimates are established for the scalar and the vector variables. We extend these results to rectangular and prismatic elements in Section 5. Finally, in Section 6, numerical results are presented for two test problems in order to support the theoretical results obtained in the previous sections

2. Problem, finite element spaces and discretization

2.1. Model problem

We are concerned with the following second order elliptic boundary value problem with mixed boundary conditions

$$
\begin{aligned}
-\nabla \cdot (\kappa \nabla u) + \alpha u &= f \qquad \text{in } \Omega, && \text{(1a)}\\
u &= g_D \qquad \text{on } \Gamma_D, && \text{(1b)}\\
\kappa \nabla u \cdot \nu &= g_N \qquad \text{on } \Gamma_N && \text{(1c)}
\end{aligned}
$$

for a given data $f \in L^2(\Omega)$, $g_D \in H^{\frac{1}{2}}(\Gamma_D)$ and $g_N \in L^2(\Gamma_N)$. Here Ω is a bounded domain in $\mathbb{R}^n$ $(n = 2, 3)$ with a polygonal or polyhedral boundary $\partial\Omega$, and ν denotes the outward unit normal to $\partial\Omega$. Γ_D is a closed part of $\partial\Omega$, and $\Gamma_N = \partial\Omega \setminus \Gamma_D$. We assume that the matrix-valued coefficient $\kappa = \kappa(x)$ is symmetric and uniformly positive definite, *i.e.*, there exist two positive constants c_1 and c_2 such that

$$
c_1 \xi^T \xi \le \xi^T \kappa(x) \xi \le c_2 \xi^T \xi \qquad \forall \xi \in \mathbb{R}^n,\ x \in \overline{\Omega},
$$

and that $\alpha = \alpha(x)$ is a nonnegative scalar-valued function belonging to $L^\infty(\Omega)$. We restrict the discussion to the case of $\text{meas}(\Gamma_D) > 0$. All the results derived in this paper are equally valid for pure Neumann problems.

In many practical problems it is of more interest to obtain accurate approximation for the vector variable $\sigma = -\kappa \nabla u$. For this purpose we rewrite (1) as a first-order mixed system

$$
\begin{aligned}
\sigma + \kappa \nabla u &= 0 \qquad \text{in } \Omega, && \text{(2a)}\\
\nabla \cdot \sigma + \alpha u &= f \qquad \text{in } \Omega, && \text{(2b)}\\
u &= g_D \qquad \text{on } \Gamma_D, && \text{(2c)}\\
-\sigma \cdot \nu &= g_N \qquad \text{on } \Gamma_N, && \text{(2d)}
\end{aligned}
$$

and seek both approximations of σ and u simultaneously.

2.2. Notation and finite element spaces

To discretize the system (2), let $\{\mathcal{T}_h\}_{h>0}$ be a family of regular partitions of Ω into triangles $(n = 2)$ or tetrahedra $(n = 3)$ in the usual sense of Ciarlet [21], where $h = \max_{T \in \mathcal{T}_h} h_T$ and h_T stands for the diameter of T. The regularity property implies that, for all elements $T \in \mathcal{T}_h$ and all edges $(n = 2)$ or faces $(n = 3)$ e of T,

$$
\text{meas}(T) \simeq h_T^n, \qquad \text{meas}(e) \simeq h_T^{n-1},
$$

where the symbol $\simeq$ indicates that both sides are of the same order of magnitude, *i.e.*, the ratio of both sides are bounded above and below by positive constants which are independent of h_T.

By $\mathcal{E}_h$ we denote the collection of all edges $(n = 2)$ or faces $(n = 3)$ of $\mathcal{T}_h$ which is split into three disjoint parts $\mathcal{E}_I$, $\mathcal{E}_D$ and $\mathcal{E}_N$, according to whether it belongs to Ω, Γ_D or Γ_N, respectively. We also use the notation $\mathcal{E}_T$ to denote the set of edges or faces of an element T. With each $e \in \mathcal{E}_h$ we associate a unit normal ν_e which is directed outward to Ω for $e \in \mathcal{E}_D \cup \mathcal{E}_N$. For an interior edge or face $e \in \mathcal{E}_I$ shared by two elements T^+ and T^- with ν_e being directed from T^+ to T^-, we define the average and the jump of v on e to be

$$\{v\} = \frac{v^+ + v^-}{2}, \qquad [v] = v^+ - v^-,$$

where v^+ (resp. v^-) denotes the trace of $v|_{T^+}$ (resp. $v|_{T^-}$).

Let $\mathcal{P}_r(T)$ be the space of all polynomials on T of degree r (set $\mathcal{P}_{-1}(T) = \emptyset$) and $\widetilde{\mathcal{P}}_r(T)$ the space of all homogeneous polynomials on T of degree r. We also set

$$\mathcal{R}_r(\partial T) = \{\mu \in L^2(\partial T) : \mu|_e \in \mathcal{P}_r(e) \quad \forall e \in \mathcal{E}_T\}.$$

For the vector approximation we will use the Raviart–Thomas–Nedelec elements (*cf.* [7,32,35,37])

$$\mathcal{RT}_r(T) = (\mathcal{P}_r(T))^n \oplus x\widetilde{\mathcal{P}}_r(T),$$

where $x = (x_1, x_2)$ for $n = 2$ and $x = (x_1, x_2, x_3)$ for $n = 3$. It is now well known that

$$\nabla \cdot \xi_h \in \mathcal{P}_r(T), \quad \xi_h \cdot \nu_T \in \mathcal{R}_r(\partial T) \qquad \forall \xi_h \in \mathcal{RT}_r(T),$$

and that the degrees of freedom for ξ_h are provided by the moments of order up to r of $\xi_h \cdot \nu_T$ on ∂T

$$\Big\{ \int_{\partial T} \xi_h \cdot \nu_T \mu \,\mathrm{d}s : \mu \in \mathcal{R}_r(\partial T) \Big\} \tag{3}$$

and the moments of order up to $r - 1$ of ξ_h on T ($r \geq 1$)

$$\Big\{ \int_T \xi_h \cdot \tau \,\mathrm{d}x : \tau \in (\mathcal{P}_{r-1}(T))^n \Big\}. \tag{4}$$

The following approximation properties are well known for the space $\mathcal{P}_r(T)$: there exists a function $v_h \in \mathcal{P}_r(T)$ such that, for $\frac{3}{2} < s \leq r + 1$,

$$\|u - v_h\|_{0,T} + h_T^{1/2} \|u - v_h\|_{0,\partial T} \leq C h_T^s \|u\|_{s,T}, \tag{5}$$

$$\|\nabla(u - v_h)\|_{0,T} + h_T^{1/2} \|\nabla(u - v_h) \cdot \nu_T\|_{0,\partial T} \leq C h_T^{s-1} \|u\|_{s,T}. \tag{6}$$

Similarly, we have for $\frac{1}{2} < s \leq r + 1$

$$\|\sigma - \Pi_T \sigma\|_{0,T} + h_T^{1/2} \|(\sigma - \Pi_T \sigma) \cdot \nu_T\|_{0,\partial T} \leq C h_T^s \|\sigma\|_{s,T}, \tag{7}$$

$$\|\nabla \cdot (\sigma - \Pi_T \sigma)\|_{0,T} \leq C h_T^s \|\nabla \cdot \sigma\|_{s,T}, \tag{8}$$

where the operator $\Pi_T : (H^s(T))^n \to \mathcal{RT}_r(T)$ is defined by means of the degrees of freedom (3)–(4)

$$\int_{\partial T} (\sigma - \Pi_T \sigma) \cdot \nu_T \mu \,\mathrm{d}s = 0 \qquad \forall \mu \in \mathcal{R}_r(\partial T),$$

$$\int_T (\sigma - \Pi_T \sigma) \cdot \tau \,\mathrm{d}x = 0 \qquad \forall \tau \in (\mathcal{P}_{r-1}(T))^n \quad (r \geq 1).$$

Also, we will often use the following standard inequalities

$$\|v\|_{0,\partial T} \leq C(h_T^{-1/2} \|v\|_{0,T} + h_T^{1/2} \|\nabla v\|_{0,T}) \qquad \forall v \in H^1(T), \tag{9}$$

$$\|v\|_{0,\partial T} \leq C h_T^{-1/2} \|v\|_{0,T} \qquad \forall v \in \mathcal{P}_r(T). \tag{10}$$

The results (5)–(10) can be proved by using the standard scaling argument.

Finally, we define the finite element spaces

$$\begin{aligned}\mathcal{P}_r &= \{v \in L^2(\Omega) : v|_T \in \mathcal{P}_r(T) \quad \forall T \in \mathcal{T}_h\},\\ \mathcal{RT}_r &= \{\tau \in H(\mathrm{div};\Omega) : \tau|_T \in \mathcal{RT}_r(T) \quad \forall T \in \mathcal{T}_h\},\end{aligned}$$

where

$$H(\mathrm{div};\Omega) = \{\tau \in (L^2(\Omega))^n : \nabla\cdot\tau \in L^2(\Omega)\}.$$

We point out that while *discontinuous* approximations are used for the scalar variable, the $H(\mathrm{div})$-*conforming* approximations are used for the vector variable, requiring it to have continuous normal components across interelement boundaries.

2.3. **Discretization**

Let us briefly describe the main idea for discretizing the mixed system (2). With $\mathcal{RT}_k \times \mathcal{P}_l$ chosen as the trial spaces, the first two equations of (2) are discretized by testing them with functions from $\Sigma(T)$ and $V(T)$

$$\begin{aligned}\int_T (\sigma_h + \kappa\nabla u_h)\cdot\tau \,\mathrm{d}x = 0 \qquad &\forall \tau \in \Sigma(T),\\ \int_T (\nabla\cdot\sigma_h + \alpha u_h) w \,\mathrm{d}x = \int_T f w \,\mathrm{d}x \qquad &\forall w \in V(T).\end{aligned}$$

Next, to communicate between neighboring elements, we express $\sigma_h \cdot \nu_e|_e$ for $e \in \mathcal{E}_h$ in terms of $\{\kappa\nabla u_h \cdot \nu_e\}$ and $[u_h]|_e$, as is done for numerical fluxes in LDG methods.

Now the problem is how to choose an appropriate pair of test function spaces $\Sigma(T)$ and $V(T)$ in such a way that

- the number of equations are equal to that of unknowns to yield a square matrix system;
- σ_h can be easily eliminated from the whole system, leading to a discontinuous finite element method for u_h;
- σ_h can be recovered in a simple manner from the computed u_h.

We mention that these statements are easily seen to be true for the LDG methods because the trial and the test spaces are identical (namely, a Galerkin method), and σ_h is completely discontinuous and thus can be eliminated locally.

For our methods we notice that, since $\sigma_h \cdot \nu_T|_{\partial T}$ is *a priori* given in terms of $\{\kappa\nabla u_h\}$ and $[u_h]$, it suffices to be able to determine the interior degrees of freedom (4) of σ_h in terms of u_h. Thus we are naturally led to choose $\Sigma(T)$ such that $\Sigma(T) \supseteq (\mathcal{P}_{k-1}(T))^n$. In particular, this choice enables to eliminate and recover σ_h *locally* on each element. On the other hand, we will need to take $\tau = \nabla v$ for $v \in \mathcal{P}_l(T)$ in order to obtain a discontinuous finite element method for u_h, which requires

$$\Sigma(T) \supseteq (\mathcal{P}_{k-1}(T))^n + \nabla\mathcal{P}_l(T). \tag{\textbf{C1}}$$

By dimensional argument we also require

$$\dim\Sigma(T) + \dim V(T) = \dim(\mathcal{P}_{k-1}(T))^n + \dim\mathcal{P}_l(T). \tag{\textbf{C2}}$$

Based on these considerations we construct and analyze two methods (one with $l = k$ and the other with $l = k+1$) in the next sections.

3. Method I: $\mathcal{RT}_k \times \mathcal{P}_k$ $(k \geq 1)$

3.1. **Discrete problem**

Our first method is based on equal-order interpolation spaces for σ and u, and is given as follows: find $(\sigma_h, u_h) \in \mathcal{RT}_k \times \mathcal{P}_k$ satisfying

$$\int_T (\sigma_h + \kappa \nabla u_h) \cdot \tau \, \mathrm{d}x = 0 \qquad \forall \tau \in (\mathcal{P}_{k-1}(T))^n, \tag{11a}$$

$$\int_T (\nabla \cdot \sigma_h + \alpha u_h) w \, \mathrm{d}x = \int_T f w \, \mathrm{d}x \qquad \forall w \in \mathcal{P}_k(T), \tag{11b}$$

and

$$\sigma_h \cdot \nu_e|_e = \begin{cases} -Q_h(\{\kappa \nabla u_h \cdot \nu_e\}) + \gamma h_e^{-1}[u_h] & \text{for } e \in \mathcal{E}_I, \\ -Q_h(\kappa \nabla u_h \cdot \nu_e) + \gamma h_e^{-1}(u_h - Q_h g_D) & \text{for } e \in \mathcal{E}_D, \\ -Q_h g_N & \text{for } e \in \mathcal{E}_N, \end{cases} \tag{11c}$$

where Q_h is the L^2-projection onto $\mathcal{P}_k(e)$, $\gamma > 0$ is a stabilization parameter to be determined later, and h_e is the diameter of $e \in \mathcal{E}_h$.

Let us point out that equation (11b) represents the same local conservation law as in the standard mixed finite element method. Thus the difference between our scheme (11) and the standard mixed finite element method lies in the discretization of the constitutive relation $\sigma + \kappa \nabla u = 0$. Also, it is trivial to verify the requirements (**C1**)–(**C2**). As a result, the above scheme yields a square matrix system.

At first sight the scheme (11) appears to be very difficult to implement due to lack of any useful property such as symmetry or positive definiteness. One remarkable property of this method is, however, that it allows for an easy elimination of the vector variable σ_h which leads to a discrete problem for u_h only. The following theorem shows that u_h is, in fact, a solution of a discontinuous finite element method (whose existence and uniqueness will be shown later).

Theorem 3.1. *Let $(\sigma_h, u_h) \in \mathcal{RT}_k \times \mathcal{P}_k$ be a solution of the scheme (11). Then u_h is a solution of the variational problem*

$$B(u_h, v_h) = l(v_h) \qquad \forall v_h \in \mathcal{P}_k, \tag{12}$$

where we define the bilinear and the linear forms

$$\begin{aligned} B(u,v) := & \sum_{T \in \mathcal{T}_h} \int_T (\kappa \nabla u \cdot \nabla v + \alpha u v) \, \mathrm{d}x \\ & - \sum_{e \in \mathcal{E}_I} \int_e \{\kappa \nabla u \cdot \nu_e\}[v] \, \mathrm{d}s - \sum_{e \in \mathcal{E}_D} \int_e \kappa \nabla u \cdot \nu_e v \, \mathrm{d}s \\ & + \sum_{e \in \mathcal{E}_I} \gamma h_e^{-1} \int_e [u][v] \, \mathrm{d}s + \sum_{e \in \mathcal{E}_D} \gamma h_e^{-1} \int_e u v \, \mathrm{d}s, \\ l(v) := & \int_\Omega f v \, \mathrm{d}x + \int_{\Gamma_N} g_N v \, \mathrm{d}s + \sum_{e \in \mathcal{E}_D} \gamma h_e^{-1} \int_e g_D v \, \mathrm{d}s. \end{aligned}$$

Proof. Using integration by parts and the fact that σ_h has continuous normal components on $\mathcal{E}_I$, we obtain for all $v_h \in \mathcal{P}_k$

$$\sum_{T \in \mathcal{T}_h} \int_T (\sigma_h \cdot \nabla v_h + \nabla \cdot \sigma_h \, v_h) \, \mathrm{d}x = \sum_{e \in \mathcal{E}_I} \int_e \sigma_h \cdot \nu_e [v_h] \, \mathrm{d}s + \int_{\partial\Omega} (\sigma_h \cdot \nu) v_h \, \mathrm{d}s. \tag{13}$$

By taking $\tau = \nabla v_h$ in (11a) and $w = v_h$ in (11b), it is easy to see that the LHS of (13) can be expressed in terms of u_h only

$$\sum_{T\in\mathcal{T}_h}\int_T (\sigma_h \cdot \nabla v_h + \nabla\cdot\sigma_h\, v_h)\,\mathrm{d}x = -\sum_{T\in\mathcal{T}_h}\int_T \Big(\kappa\nabla u_h\cdot\nabla v_h + \alpha u_h v_h - f v_h\Big)\,\mathrm{d}x.$$

Now substitution of (11c) into the RHS of (13) gives the desired result. □

From Theorem 3.1 it is now obvious how to implement the mixed finite volume method (11). First, the scalar approximation u_h is computed from the discontinuous finite element method (12), and then the vector approximation σ_h can be *locally* recovered from u_h through the equations (11a) and (11c) which specify a complete set of degrees of freedom (3)–(4) for $\sigma_h|_T$. This fact not only provides a convenient way of implementation but also a way of deriving error estimates without resort to the theory of saddle-point problems, as is demonstrated in the next subsection.

Remark 3.2. Notice that the discontinuous finite element method (12) lacks the usual symmetrizing or anti-symmetrizing terms which appear in the traditional methods (see, *e.g.*, [4,11,15,26,36]). This does not affect the stability or the convergence rate of the method. To solve the discontinuous finite element method (12) in an efficient way, one may use the Krylov subspace method like GMRES with preconditioners, *e.g.*, from [29,30].

3.2. **Error estimates**

Now we turn to error estimates for the scheme (11). For this sake let $(s > 0)$

$$H^s(\mathcal{T}_h) = \{v \in L^2(\Omega) : v|_T \in H^s(T) \quad \forall T \in \mathcal{T}_h\},$$

and define the mesh-dependent norm for $v \in H^1(\mathcal{T}_h)$

$$|\!|\!| v |\!|\!| := \Bigg(\sum_{T\in\mathcal{T}_h} \|\kappa^{1/2}\nabla v\|_{0,T}^2 + \|\alpha^{1/2} v\|_{0,\Omega}^2 + \sum_{e\in\mathcal{E}_I} h_e^{-1}\|[v]\|_{0,e}^2 + \sum_{e\in\mathcal{E}_D} h_e^{-1}\|v\|_{0,e}^2\Bigg)^{1/2}.$$

In the sequel we make the following regularity assumption on u and $\sigma = -\kappa\nabla u$:

$$u \in H^s(\mathcal{T}_h), \quad \sigma \in (H^{s-1}(\mathcal{T}_h))^n, \qquad s > \frac{3}{2}\cdot$$

To begin with, we recall an abstract error estimate for the variational problem which is a variant of Strang's lemma (*cf.* [21]). For later use we present the following lemma in its full generality.

Lemma 3.3. *Suppose $B(u,\cdot)$ is well defined and the following uniform ellipticity holds:*

$$B(v_h, v_h) \ge c|\!|\!| v_h |\!|\!|^2 \qquad \forall v_h \in V_h.$$

Let $u_h \in V_h$ be the solution of

$$B(u_h, v_h) = l(v_h) \qquad \forall v_h \in V_h.$$

Then we have

$$|\!|\!| u - u_h |\!|\!| \le \inf_{v_h\in V_h}\left(|\!|\!| u - v_h |\!|\!| + \frac{1}{c}\sup_{w_h\in V_h}\frac{B(u - v_h, w_h)}{|\!|\!| w_h |\!|\!|}\right) + \frac{1}{c}\sup_{w_h\in V_h}\frac{B(u, w_h) - l(w_h)}{|\!|\!| w_h |\!|\!|}\cdot$$

Proof. Given any $v_h \in V_h$ we obtain

$$\begin{aligned} c|\!|\!|u_h - v_h|\!|\!|^2 &\le B(u_h - v_h, u_h - v_h) \\ &= B(u - v_h, u_h - v_h) + [l(u_h - v_h) - B(u, u_h - v_h)], \end{aligned}$$

which gives

$$c|\!|\!|u_h - v_h|\!|\!| \le \frac{B(u - v_h, u_h - v_h)}{|\!|\!|u_h - v_h|\!|\!|} + \frac{l(u_h - v_h) - B(u, u_h - v_h)}{|\!|\!|u_h - v_h|\!|\!|}.$$

The desired result follows easily by setting $w_h = u_h - v_h$ and taking the supremum on the right-hand side. □

In order to apply Lemma 3.3 in deriving an error estimate for $u - u_h$, we need to prove some preliminary results.

Lemma 3.4. *The scheme (12) is consistent in the sense that*

$$B(u - u_h, v_h) = B(u, v_h) - l(v_h) = 0 \qquad \forall v_h \in \mathcal{P}_k.$$

Proof. Use integration by parts on each element and then the continuity of u and $\kappa\nabla u \cdot \nu_e$ across $e \in \mathcal{E}_I$, together with (1), to complete the proof. □

Lemma 3.5. *The following uniform ellipticity holds:*

$$B(v_h, v_h) \ge C|\!|\!|v_h|\!|\!|^2 \qquad \forall v_h \in \mathcal{P}_k,$$

provided that $\gamma > 0$ is sufficiently large (independently of the mesh size).

Proof. For $e \in \mathcal{E}_I$ shared by two elements T^+ and T^-, we obtain

$$\int_e \{\kappa\nabla v_h \cdot \nu_e\}[v_h]\,\mathrm{d}s \le C\|\{\nabla v_h\}\|_{0,e}\|[v_h]\|_{0,e} \le C\|\nabla v_h\|_{0,T^+\cup T^-}\, h_e^{-1/2}\|[v_h]\|_{0,e},$$

where we used the inverse inequality (10). Summing over all $e \in \mathcal{E}_I$, we obtain

$$\sum_{e\in\mathcal{E}_I} \int_e \{\kappa\nabla v_h \cdot \nu_e\}[v_h]\,\mathrm{d}s \le C_1\bigg(\sum_{T\in\mathcal{T}_h} \|\kappa^{1/2}\nabla v_h\|_{0,T}^2\bigg)^{1/2}\bigg(\sum_{e\in\mathcal{E}_I} h_e^{-1}\|[v_h]\|_{0,e}^2\bigg)^{1/2}.$$

Similarly,

$$\sum_{e\in\mathcal{E}_D} \int_e \kappa\nabla v_h \cdot \nu_e v_h\,\mathrm{d}s \le C_2\bigg(\sum_{T\in\mathcal{T}_h} \|\kappa^{1/2}\nabla v_h\|_{0,T}^2\bigg)^{1/2}\bigg(\sum_{e\in\mathcal{E}_D} h_e^{-1}\|v_h\|_{0,e}^2\bigg)^{1/2}.$$

Consequently, it follows that

$$\begin{aligned} B(v_h, v_h) \ge{}& \sum_{T\in\mathcal{T}_h} \|\kappa^{1/2}\nabla v_h\|_{0,T}^2 + \|\alpha^{1/2}v_h\|_{0,\Omega}^2 + \sum_{e\in\mathcal{E}_I} \gamma h_e^{-1}\|[v_h]\|_{0,e}^2 + \sum_{e\in\mathcal{E}_D} \gamma h_e^{-1}\|v_h\|_{0,e}^2 \\ &- C_1\bigg(\sum_{T\in\mathcal{T}_h} \|\kappa^{1/2}\nabla v_h\|_{0,T}^2\bigg)^{1/2}\bigg(\sum_{e\in\mathcal{E}_I} h_e^{-1}\|[v_h]\|_{0,e}^2\bigg)^{1/2} \\ &- C_2\bigg(\sum_{T\in\mathcal{T}_h} \|\kappa^{1/2}\nabla v_h\|_{0,T}^2\bigg)^{1/2}\bigg(\sum_{e\in\mathcal{E}_D} h_e^{-1}\|v_h\|_{0,e}^2\bigg)^{1/2}. \end{aligned}$$

Using the inequality $ab \le \frac{1}{4}a^2 + b^2$, we then obtain

$$\begin{aligned} B(v_h, v_h) &\ge \frac{1}{2} \sum_{T\in\mathcal{T}_h} \|\kappa^{1/2}\nabla v_h\|_{0,T}^2 + \|\alpha^{1/2} v_h\|_{0,\Omega}^2 \\ &\qquad + (\gamma - C_1^2) \sum_{e\in\mathcal{E}_I} h_e^{-1}\|[v_h]\|_{0,e}^2 + (\gamma - C_2^2) \sum_{e\in\mathcal{E}_D} h_e^{-1}\|v_h\|_{0,e}^2 \\ &\ge C|\!|\!| v_h |\!|\!|^2, \end{aligned}$$

provided that $\gamma > 0$ is taken to be larger than $\max(C_1^2, C_2^2)$. □

Corollary 3.6. *The scheme (11) has a unique solution $(\sigma_h, u_h) \in \mathcal{RT}_k \times \mathcal{P}_k$.*

Proof. It suffices to show that, if $f = g_D = g_N = 0$, then we have $\sigma_h = u_h = 0$. By the uniform ellipticity of $B(\cdot,\cdot)$ the problem (12) has a unique solution, that is, $u_h = 0$. Since $\sigma_h|_T$ is completely determined by (11a) and (11c), it follows that $\sigma_h = 0$. □

Now it remains to estimate the terms in Lemma 3.3. In fact, we only need to estimate the first two terms, as our method is consistent. We see that, for all $v_h, w_h \in \mathcal{P}_k$,

$$\begin{aligned} B(u - v_h, w_h) &= \sum_{T\in\mathcal{T}_h} \left(\int_T \kappa\nabla(u - v_h)\cdot\nabla w_h \,\mathrm{d}x + \int_T \alpha(u - v_h)w_h \,\mathrm{d}x \right) \\ &\qquad - \sum_{e\in\mathcal{E}_I} \int_e \{\kappa\nabla(u - v_h)\cdot\nu_e\}[w_h]\,\mathrm{d}s - \sum_{e\in\mathcal{E}_D} \int_e \kappa\nabla(u - v_h)\cdot\nu_e w_h \,\mathrm{d}s \\ &\qquad + \sum_{e\in\mathcal{E}_I} \gamma h_e^{-1} \int_e [u - v_h][w_h]\,\mathrm{d}s + \sum_{e\in\mathcal{E}_D} \gamma h_e^{-1} \int_e (u - v_h)w_h \,\mathrm{d}s \\ &\le C|\!|\!| u - v_h |\!|\!| \, |\!|\!| w_h |\!|\!| + C\Bigg(\sum_{T\in\mathcal{T}_h} h_T \|\nabla(u - v_h)\cdot\nu_T\|_{0,\partial T}^2 \Bigg)^{1/2} |\!|\!| w_h |\!|\!|, \end{aligned}$$

which gives

$$\sup_{w_h\in\mathcal{P}_k} \frac{B(u - v_h, w_h)}{|\!|\!| w_h |\!|\!|} \le C|\!|\!| u - v_h |\!|\!| + C\Bigg(\sum_{T\in\mathcal{T}_h} h_T \|\nabla(u - v_h)\cdot\nu_T\|_{0,\partial T}^2 \Bigg)^{1/2}.$$

Therefore it follows from Lemma 3.3 that, for all $v_h \in \mathcal{P}_k$,

$$|\!|\!| u - u_h |\!|\!| \le C|\!|\!| u - v_h |\!|\!| + C\Bigg(\sum_{T\in\mathcal{T}_h} h_T \|\nabla(u - v_h)\cdot\nu_T\|_{0,\partial T}^2 \Bigg)^{1/2}.$$

Finally, using the approximation properties (5)–(6), we arrive at the following theorem.

Theorem 3.7. *Let $u_h \in \mathcal{P}_k$ be the solution of the problem (12). Then we have for $\frac{3}{2} < s \le k + 1$*

$$|\!|\!| u - u_h |\!|\!| \le C\Bigg(\sum_{T\in\mathcal{T}_h} h_T^{2(s-1)} \|u\|_{s,T}^2 \Bigg)^{1/2}. \tag{14}$$

Now we turn to error estimation for the vector variable. The following lemma provides a result which plays a crucial role in deriving an error estimate for $\sigma - \sigma_h$ in L^2-norm.

Lemma 3.8. *Given $p \in L^2(\partial T)$ and $\beta \in (L^2(T))^n$, let $\xi_h \in \mathcal{RT}_k(T)$ satisfy*

$$\int_{\partial T} \xi_h \cdot \nu_T \, q \, \mathrm{d}s = \int_{\partial T} p \, q \, \mathrm{d}s \qquad \forall q \in \mathcal{R}_k(\partial T),$$
$$\int_T \xi_h \cdot \tau \, \mathrm{d}x = \int_T \beta \cdot \tau \, \mathrm{d}x \qquad \forall \tau \in (\mathcal{P}_{k-1}(T))^n \quad (k \geq 1).$$

Then we obtain

$$\|\xi_h\|_{0,T} \leq C(\|\beta\|_{0,T} + h_T^{1/2}\|p\|_{0,\partial T}).$$

Proof. By considering the L^2-projections, we may assume that $p \in \mathcal{R}_k(\partial T)$ and $\beta \in (\mathcal{P}_{k-1}(T))^n$. Then the proof is done by using a simple scaling argument [3, 7]. □

We are now ready to prove the following theorem.

Theorem 3.9. *Let $(\sigma_h, u_h) \in \mathcal{RT}_k \times \mathcal{P}_k$ be the solution of the system (11). Then, for $\frac{3}{2} < s \leq k+1$ we have*

$$\|\sigma - \sigma_h\|_0 \leq C\Big(\sum_{T\in\mathcal{T}_h} h_T^{2(s-1)}(\|u\|_{s,T}^2 + \|\sigma\|_{s-1,T}^2)\Big)^{1/2}. \tag{15}$$

Furthermore, if $\nabla\cdot\sigma \in H^{s-1}(\mathcal{T}_h)$, we have

$$\|\nabla\cdot(\sigma - \sigma_h)\|_0 \leq C\Big(\sum_{T\in\mathcal{T}_h} h_T^{2(s-1)}(\|u\|_{s,T}^2 + \|\nabla\cdot\sigma\|_{s-1,T}^2)\Big)^{1/2}, \tag{16}$$

where the term $\|u\|_{s,T}^2$ is dropped and s can be up to $k+2$ in the case $\alpha \equiv 0$.

Proof. By (11a), (11c) and the fact that $(\Pi_T\sigma)\cdot\nu_T = Q_h(\sigma|_T\cdot\nu_T)$ and $[u] = 0$ on $e \in \mathcal{E}_I$, we obtain for $e \in \mathcal{E}_T$

$$(\sigma_h - \Pi_T\sigma)\cdot\nu_e|_e = \begin{cases} Q_h\big(\{\kappa\nabla(u-u_h)\cdot\nu_e\}\big) + \gamma h_e^{-1}Q_h[u_h - u] & \text{for } e \in \mathcal{E}_I, \\ Q_h\big(\kappa\nabla(u-u_h)\cdot\nu_e\big) + \gamma h_e^{-1}Q_h(u_h - u) & \text{for } e \in \mathcal{E}_D, \\ 0 & \text{for } e \in \mathcal{E}_N, \end{cases}$$

and for all $\tau \in (\mathcal{P}_{k-1}(T))^2$

$$\int_T (\sigma_h - \Pi_T\sigma)\cdot\tau \, \mathrm{d}x = \int_T \big[(\sigma - \Pi_T\sigma) + \kappa\nabla(u-u_h)\big]\cdot\tau \, \mathrm{d}x.$$

Now, applying Lemma 3.8 to these equations, we deduce that

$$\|\sigma - \sigma_h\|_0 \leq C\Big(|\!|\!| u - u_h |\!|\!|^2 + \sum_{T\in\mathcal{T}_h}\big(\|\sigma - \Pi_T\sigma\|_{0,T}^2 + h_T\|\nabla(u-u_h)\cdot\nu_T\|_{0,\partial T}^2\big)\Big)^{1/2}.$$

By using (6), (7) and (14), we obtain the first result.

For the second part we start with the error equation

$$\int_T \big(\nabla\cdot(\sigma - \sigma_h) + \alpha(u - u_h)\big)w_h \, \mathrm{d}x = 0 \qquad \forall w_h \in \mathcal{P}_k(T).$$

By taking $w_h = \nabla\cdot(\Pi_T\sigma - \sigma_h)$ it is easy to derive that

$$\|\nabla\cdot(\sigma - \sigma_h)\|_{0,T} \leq C(\|\nabla\cdot(\sigma - \Pi_T\sigma)\|_{0,T} + \|\alpha^{1/2}(u - u_h)\|_{0,T}),$$

from which the second result follows by (8) and (14). □

Remark 3.10. Theorems 3.7 and 3.9 indicate that the error bound is optimal for $u - u_h$ but only suboptimal for $\sigma - \sigma_h$ with respect to the polynomial degree.

4. Method II: $\mathcal{RT}_k \times \mathcal{P}_{k+1}$ $(k \geq 0)$

4.1. Discrete problem

In view of the suboptimality of the error estimate for $\sigma - \sigma_h$ with respect to the polynomial degree, we are naturally led to consider using the higher-order space $\mathcal{P}_{k+1}$ for the scalar approximation. So our second method reads as follows: find $(\sigma_h, u_h) \in \mathcal{RT}_k \times \mathcal{P}_{k+1}$ satisfying

$$\int_T (\sigma_h + \kappa \nabla u_h) \cdot \tau \,\mathrm{d}x = 0 \qquad \forall \tau \in (\mathcal{P}_{k-1}(T))^n \oplus \nabla \widetilde{\mathcal{P}}_{k+1}(T), \tag{17a}$$

$$\int_T (\nabla \cdot \sigma_h + \alpha u_h) w \,\mathrm{d}x = \int_T f w \,\mathrm{d}x \qquad \forall w \in \mathcal{P}_k(T), \tag{17b}$$

and

$$\sigma_h \cdot \nu_e|_e = \begin{cases} -Q_h(\{\kappa \nabla u_h \cdot \nu_e\}) + \gamma h_e^{-1} Q_h([u_h]) & \text{for } e \in \mathcal{E}_I, \\ -Q_h(\kappa \nabla u_h \cdot \nu_e) + \gamma h_e^{-1} Q_h(u_h - g_D) & \text{for } e \in \mathcal{E}_D, \\ -Q_h g_N & \text{for } e \in \mathcal{E}_N. \end{cases} \tag{17c}$$

Recall that Q_h is the L^2-projection onto $\mathcal{P}_k(e)$.

It is not difficult to check that many nice properties of Method I are also valid for Method II. In particular, note that a square matrix system is produced by introducing additional test functions from $\nabla \widetilde{\mathcal{P}}_{k+1}(T)$, since we have

$$\dim \nabla \widetilde{\mathcal{P}}_{k+1}(T) + \dim \mathcal{P}_k(T) = \dim \mathcal{P}_{k+1}(T).$$

Moreover, we can derive the following theorem similar to Theorem 3.1 which states that u_h is the solution of a slightly modified discontinuous finite element method. The proof proceeds exactly in the same way as in Theorem 3.1 and is thus omitted.

Theorem 4.1. *Let $(\sigma_h, u_h) \in \mathcal{RT}_k \times \mathcal{P}_{k+1}$ be a solution of the scheme (17). Then $u_h \in \mathcal{P}_{k+1}$ is a solution of the variational problem*

$$B(u_h, v_h) = l(v_h) \qquad \forall v_h \in \mathcal{P}_{k+1}, \tag{18}$$

where we define the bilinear and the linear forms

$$\begin{aligned} B(u, v) := & \sum_{T \in \mathcal{T}_h} \int_T (\kappa \nabla u \cdot \nabla v + \overline{\alpha u}\, v) \,\mathrm{d}x \\ & - \sum_{e \in \mathcal{E}_I} \int_e Q_h(\{\kappa \nabla u \cdot \nu_e\})[v] \,\mathrm{d}s - \sum_{e \in \mathcal{E}_D} \int_e Q_h(\kappa \nabla u \cdot \nu_e) v \,\mathrm{d}s \\ & + \sum_{e \in \mathcal{E}_I} \gamma h_e^{-1} \int_e Q_h([u])[v] \,\mathrm{d}s + \sum_{e \in \mathcal{E}_D} \gamma h_e^{-1} \int_e Q_h(u) v \,\mathrm{d}s, \\ l(v) := & \int_\Omega \bar{f} v \,\mathrm{d}x + \int_{\Gamma_N} Q_h(g_N) v \,\mathrm{d}s + \sum_{e \in \mathcal{E}_D} \gamma h_e^{-1} \int_e Q_h(g_D) v \,\mathrm{d}s, \end{aligned}$$

and $\overline{w}$ means the L^2-projection of w onto $\mathcal{P}_k(T)$.

Remark 4.2. The inclusion of the L^2-projection Q_h offers some computational advantages when calculating the stiffness matrices. Indeed, in two space dimensions, all the line integrals on $e \in \mathcal{E}_h$ involve polynomials of degree $2k+1$ as the integrands and thus can be calculated by the Gaussian quadrature

$$\int_e \phi \,\mathrm{d}s = \sum_{l=1}^{k+1} w_{l,e}\, \phi(b_{l,e}) \qquad \forall \phi \in \mathcal{P}_{2k+1}(e),$$

where $\{b_{l,e}\}_{l=1}^{k+1}$ denote the $k+1$ Gauss points and $\{w_{l,e}\}_{l=1}^{k+1}$ the corresponding weights. In particular, since ϕ and $Q_h(\phi)$ have the same values at the points $\{b_{l,e}\}_{l=1}^{k+1}$ if $\phi \in \mathcal{P}_{k+1}(e)$, it follows that

$$\int_e Q_h([u_h])[v_h]\,\mathrm{d}s = \sum_{l=1}^{k+1} w_{l,e}\,[u_h(b_{l,e})]\,[v_h(b_{l,e})].$$

4.2. **Error estimates**

To derive error estimates for the scheme (17), we assume for simplicity that α is piecewise constant on $\mathcal{T}_h$. Proceeding as in the proof of Lemma 3.5, we easily obtain for all $v_h \in \mathcal{P}_{k+1}$

$$B(v_h, v_h) \geq C\biggl(\sum_{T\in\mathcal{T}_h} \|\kappa^{1/2}\nabla v_h\|_{0,T}^2 + \|\alpha^{1/2}\bar{v}_h\|_{0,\Omega}^2 + \sum_{e\in\mathcal{E}_I} h_e^{-1}\|Q_h[v_h]\|_{0,e}^2 + \sum_{e\in\mathcal{E}_D} h_e^{-1}\|Q_h v_h\|_{0,e}^2 \biggr) := C|\!|\!| v_h |\!|\!|_*^2.$$

It is easy to see that $|\!|\!| \cdot |\!|\!|_*$ also defines a norm for the space $H^1(\mathcal{T}_h)$. The uniform ellipticity for the new bilinear form $B(\cdot,\cdot)$ is an immediate consequence of the following proposition.

Proposition 4.3. *The two norms $|\!|\!| \cdot |\!|\!|$ and $|\!|\!| \cdot |\!|\!|_*$ are uniformly equivalent on $\mathcal{P}_{k+1}$ (i.e. with constants independent of the mesh size).*

Proof. Fix $v_h \in \mathcal{P}_{k+1}$. From the property of L^2 projections, it is obvious that $|\!|\!| v_h |\!|\!| \geq |\!|\!| v_h |\!|\!|_*$. To prove the other inequality, we note that

$$|\!|\!| v_h |\!|\!| \leq C(I_1 + I_2 + |\!|\!| v_h |\!|\!|_*),$$

where

$$I_1 := \|\alpha^{1/2}(v_h - \bar{v}_h)\|_{0,\Omega}, \quad I_2 := \biggl(\sum_{e\in\mathcal{E}_I} h_e^{-1}\|(I-Q_h)[v_h]\|_{0,e}^2 + \sum_{e\in\mathcal{E}_D} h_e^{-1}\|(I-Q_h)v_h\|_{0,e}^2 \biggr)^{1/2}.$$

By using the approximation property of L^2 projections, it is not difficult to show that

$$I_1 \leq C\biggl(\sum_{T\in\mathcal{T}_h} \alpha\|\nabla v_h\|_{0,T}^2 \biggr)^{1/2} \leq C|\!|\!| v_h |\!|\!|_*,$$

$$I_2 \leq C\biggl(\sum_{T\in\mathcal{T}_h}\sum_{e\in\mathcal{E}_T} h_e^{-1}\|(I-Q_h)v_h|_T\|_{0,e}^2 \biggr)^{1/2} \leq C\biggl(\sum_{T\in\mathcal{T}_h} \|\nabla v_h\|_{0,T}^2 \biggr)^{1/2} \leq C|\!|\!| v_h |\!|\!|_*.$$

This completes the proof. □

Now we apply Lemma 3.3 to perform the error analysis for the new problem (18). It is sufficient to consider the consistency error $B(u, w_h) - l(w_h)$ as the remaining terms can be treated as previously. Using integration by parts on each $T \in \mathcal{T}_h$ and the continuity of u and $\kappa \nabla u \cdot \nu_e$ across $e \in \mathcal{E}_I$, together with (1), one can write

$$\begin{aligned}
B(u, w_h) - l(w_h) &= \sum_{T\in\mathcal{T}_h} \int_T (\kappa\nabla u \cdot \nabla w_h + \alpha \bar{u} w_h)\,\mathrm{d}x - \sum_{e\in\mathcal{E}_I} \int_e Q_h(\kappa\nabla u\cdot\nu_e)[w_h]\,\mathrm{d}s \\
&\qquad - \sum_{e\in\mathcal{E}_D}\int_e Q_h(\kappa\nabla u\cdot\nu_e)w_h\,\mathrm{d}s - \int_\Omega \bar{f} w_h\,\mathrm{d}x - \int_{\Gamma_N} Q_h(g_N)w_h\,\mathrm{d}s \\
&= \sum_{T\in\mathcal{T}_h}\Big(\int_T \kappa\nabla u\cdot\nabla w_h\,\mathrm{d}x - \int_{\partial T} Q_h(\kappa\nabla u\cdot\nu_T)w_h\,\mathrm{d}s\Big) \\
&\qquad + \sum_{T\in\mathcal{T}_h}\Big(\int_T \alpha\bar{u}w_h\,\mathrm{d}x - \int_T \bar{f}w_h\,\mathrm{d}x\Big) \\
&= \sum_{T\in\mathcal{T}_h}\int_{\partial T}(I-Q_h)(\kappa\nabla u\cdot\nu_T)w_h\,\mathrm{d}s \\
&\qquad + \sum_{T\in\mathcal{T}_h}\Big(\int_T \alpha(\bar{u}-u)w_h\,\mathrm{d}x + \int_T (f-\bar{f})w_h\,\mathrm{d}x\Big) \\
&:= E_1 + E_2.
\end{aligned}$$

Let c be a piecewise constant approximation of w_h such that

$$\|w_h - c\|_{0,T} + h_T^{1/2}\|w_h - c\|_{0,\partial T} \le Ch_T\|\nabla w_h\|_{0,T}.$$

Then we obtain by using the fact that $(\Pi_T\sigma)\cdot\nu_T = Q_h(\sigma|_T\cdot\nu_T)$

$$\begin{aligned}
E_1 &= \sum_{T\in\mathcal{T}_h}\int_{\partial T}(I-Q_h)(\kappa\nabla u\cdot\nu_T)(w_h-c)\,\mathrm{d}s \\
&\le C\Big(\sum_{T\in\mathcal{T}_h} h_T\|(\sigma-\Pi_T\sigma)\cdot\nu_T\|_{0,\partial T}^2\Big)^{1/2}\Big(\sum_{T\in\mathcal{T}_h}\|\nabla w_h\|_{0,T}^2\Big)^{1/2},
\end{aligned}$$

and

$$\begin{aligned}
E_2 &= \sum_{T\in\mathcal{T}_h}\Big(\int_T \alpha(\bar{u}-u)(w_h-c)\,\mathrm{d}x + \int_T (f-\bar{f})(w_h-c)\,\mathrm{d}x\Big) \\
&\le C\Big(\sum_{T\in\mathcal{T}_h} h_T^2(\|u-\bar{u}\|_{0,T}^2 + \|f-\bar{f}\|_{0,T}^2)\Big)^{1/2}\Big(\sum_{T\in\mathcal{T}_h}\|\nabla w_h\|_{0,T}^2\Big)^{1/2}.
\end{aligned}$$

From these results and Lemma 3.3 it follows that

$$\begin{aligned}
|\!|\!| u - u_h |\!|\!| &\le C|\!|\!| u - v_h |\!|\!| + C\Big(\sum_{T\in\mathcal{T}_h} h_T\|\nabla(u-v_h)\cdot\nu_e\|_{0,\partial T}^2\Big)^{1/2} \\
&\qquad + C\Big(\sum_{T\in\mathcal{T}_h}\big(h_T\|(\sigma-\Pi_T\sigma)\cdot\nu_T\|_{0,\partial T}^2 + h_T^2(\|u-\bar{u}\|_{0,T}^2 + \|f-\bar{f}\|_{0,T}^2)\big)\Big)^{1/2}.
\end{aligned}$$

By (5)–(7) and the standard results for the L^2 projection, we obtain the following theorem.

Theorem 4.4. *Let $u_h \in \mathcal{P}_{k+1}$ be the solution of the problem (18). Then we have for $\frac{3}{2} < s \leq k+2$*

$$\|\!|u - u_h|\!\| \leq C\bigg(\sum_{T \in \mathcal{T}_h} h_T^{2(s-1)} (\|u\|_{s,T}^2 + \|\sigma\|_{s-1,T}^2 + \|f\|_{\max(s-2,0),T}^2) \bigg)^{1/2}. \tag{19}$$

The error bound for $\sigma - \sigma_h$ can be derived in the same way as Method I and is stated in the following theorem.

Theorem 4.5. *Let $(\sigma_h, u_h) \in \mathcal{RT}_k \times \mathcal{P}_{k+1}$ be the solution of the system (17). Then, for $\frac{3}{2} < s \leq k+2$ we have*

$$\|\sigma - \sigma_h\|_0 \leq C\bigg(\sum_{T \in \mathcal{T}_h} h_T^{2(s-1)} (\|u\|_{s,T}^2 + \|\sigma\|_{s-1,T}^2 + \|f\|_{\max(s-2,0),T}^2) \bigg)^{1/2}. \tag{20}$$

Furthermore, if $\nabla \cdot \sigma \in H^{s-1}(\mathcal{T}_h)$, we have

$$\|\nabla \cdot (\sigma - \sigma_h)\|_0 \leq C\bigg(\sum_{T \in \mathcal{T}_h} h_T^{2(s-1)} (\|u\|_{s,T}^2 + \|\sigma\|_{s-1,T}^2 + \|\nabla \cdot \sigma\|_{s-1,T}^2 + \|f\|_{\max(s-2,0),T}^2) \bigg)^{1/2}, \tag{21}$$

where the term $\|u\|_{s,T}^2 + \|\sigma\|_{s-1,T}^2 + \|f\|_{\max(s-2,0),T}^2$ is dropped in the case $\alpha \equiv 0$.

Remark 4.6. From the error estimates for Methods I and II established above, we can conclude that Method II is more efficient than Method I in the sense that (1) with the same vector space Method II produces an optimal vector approximation and a one-order higher scalar approximation than Method I, and (2) with the same scalar space Method II uses a one-order lower vector space to achieve the same order of accuracy as Method I.

Remark 4.7. It is possible to make use of other $H(\text{div})$-conforming spaces for the vector approximation, for example, the Brezzi–Douglas–Marini space defined by

$$\mathcal{BDM}_r(T) = (\mathcal{P}_r(T))^n \qquad (r \geq 1).$$

The degrees of freedom for $\xi_h \in \mathcal{BDM}_r(T)$ are given by the moments of order up to r of $\xi_h \cdot \nu_T$ on ∂T

$$\bigg\{ \int_{\partial T} \xi_h \cdot \nu_T \mu \, \mathrm{d}s : \mu \in \mathcal{R}_r(\partial T) \bigg\}$$

and the moments of ξ_h on T

$$\bigg\{ \int_T \xi_h \cdot \tau \, \mathrm{d}x : \tau \in \nabla \mathcal{P}_{r-1}(T) \oplus \Psi_r(T) \bigg\},$$

where we set

$$\Psi_r(T) = \{\tau \in (\mathcal{P}_r(T))^n : \nabla \cdot \tau = 0, \ \tau \cdot \nu_T = 0\}.$$

Moreover, an operator $\Pi_T : (H^s(T))^n \to \mathcal{BDM}_r(T)$ can be defined analogously to the RT space so that it satisfies the approximation properties (7)–(8). We refer to [7–9] for a detailed discussion on this space.

For $k \geq 1$, Method II based on the BDM space is formulated as follows: find $(\sigma_h, u_h) \in \mathcal{BDM}_k \times \mathcal{P}_{k+1}$ such that

$$\begin{aligned}
&\int_T (\sigma_h + \kappa \nabla u_h) \cdot \tau \, \mathrm{d}x = 0 && \forall \tau \in \Psi_k(T) \oplus \nabla \mathcal{P}_{k+1}(T), \\
&\int_T (\nabla \cdot \sigma_h + \alpha u_h) w \, \mathrm{d}x = \int_T f w \, \mathrm{d}x && \forall w \in \mathcal{P}_{k-1}(T),
\end{aligned}$$

and

$$\sigma_h \cdot \nu_e|_e = \begin{cases} -Q_h(\{\kappa\nabla u_h \cdot \nu_e\}) + \gamma h_e^{-1} Q_h([u_h]) & \text{for } e \in \mathcal{E}_I, \\ -Q_h(\kappa\nabla u_h \cdot \nu_e) + \gamma h_e^{-1} Q_h(u_h - g_D) & \text{for } e \in \mathcal{E}_D, \\ -Q_h g_N & \text{for } e \in \mathcal{E}_N, \end{cases}$$

where Q_h is the L^2-projection onto $\mathcal{P}_k(e)$, as before.

It can be easily checked that all of the previous results hold for this case as well. The conditions (**C1**)–(**C2**) (modified in a suitable way) are satisfied, and thus (1) a square matrix system is produced, and (2) σ_h can be eliminated to yield the same discontinuous finite element method (18) for u_h only, with one minor modification:

$$\overline{\alpha u_h} \text{ and } \bar{f} \text{ now means the } L^2\text{-projections onto the lower-degree space } \mathcal{P}_{k-1}.$$

We should say that this modification does not deteriorate the order of convergence given in Theorems 4.4 and 4.5, since the term $h_T^2(\|u - \bar{u}\|_{0,T}^2 + \|f - \bar{f}\|_{0,T}^2)$ still gives optimal estimates.

4.3. Choice of stabilization parameter

In this subsection we demonstrate how the stabilization parameter $\gamma > 0$ can be calculated in *a priori* way independently of the problem at hand and the triangulation of the domain. This can be done by some judicious choice of the mesh parameter h_e, as illustrated below.

To simplify the discussion we assume that κ is piecewise constant on $\mathcal{T}_h$, and define the mesh parameter h_e for $e \in \mathcal{E}_I$ with $e = \partial T^+ \cap \partial T^-$ by (set $\kappa^+ = \kappa|_{T^+}$ and $\kappa^- = \kappa|_{T^-}$)

$$h_e = \left(\kappa^+ \frac{\text{meas}(e)}{\text{meas}(T^+)} + \kappa^- \frac{\text{meas}(e)}{\text{meas}(T^-)}\right)^{-1}, \tag{22}$$

and for $e \in \mathcal{E}_D$ with $e \subset \partial T$,

$$h_e = \left(\kappa_T \frac{\text{meas}(e)}{\text{meas}(T)}\right)^{-1}. \tag{23}$$

In establishing the uniform ellipticity for the bilinear form $B(\cdot,\cdot)$, we will use the following refined form of the inverse inequality for $e \in \mathcal{E}_T$:

$$\frac{1}{\text{meas}(e)} \|\nabla v_h \cdot \nu_T\|_{0,e}^2 \le C \frac{1}{\text{meas}(T)} \|\nabla v_h\|_{0,T}^2 \qquad \forall v_h \in \mathcal{P}_{k+1}(T). \tag{24}$$

Here it is important to observe that the constant $C > 0$ depends only on the reference triangle $\hat{T}$ and k, the polynomial degree of ∇v_h. For example, for $k = 0$, we obtain

$$\frac{1}{\text{meas}(e)} \|\nabla v_h \cdot \nu_T\|_{0,e}^2 = |\nabla v_h \cdot \nu_T|^2 \le \frac{1}{\text{meas}(T)} \|\nabla v_h\|_{0,T}^2,$$

which yields $C = 1$. For $k \ge 1$, by using the scaling argument, we obtain for $\phi \in \mathcal{P}_k(T)$

$$\frac{1}{\text{meas}(e)} \|\phi\|_{0,e}^2 = \frac{1}{\text{meas}(\hat{e})} \|\hat{\phi}\|_{0,\hat{e}}^2 \le C_{\hat{e}} \frac{1}{\text{meas}(\hat{T})} \|\hat{\phi}\|_{0,\hat{T}}^2 = C_{\hat{e}} \frac{1}{\text{meas}(T)} \|\phi\|_{0,T}^2,$$

where $\hat{e}$ is the corresponding edge of e on $\partial\hat{T}$, and

$$C_{\hat{e}} = \max_{\hat{\phi} \in \mathcal{P}_k(\hat{T})} \frac{\text{meas}(\hat{e})^{-1} \|\hat{\phi}\|_{0,\hat{e}}^2}{\text{meas}(\hat{T})^{-1} \|\hat{\phi}\|_{0,\hat{T}}^2}.$$

Thus it suffices to take $C = \max_{\hat{e} \in \mathcal{E}(\hat{T})} C_{\hat{e}}$.

Proposition 4.8. *Let C be the constant given in the inverse inequality (24). Then the uniform ellipticity for $B(\cdot,\cdot)$ with the mesh parameters (22)–(23) holds true if γ is chosen such that*

$$\gamma > \frac{C(n+1)}{4}\cdot$$

Proof. Fix $v_h \in \mathcal{P}_{k+1}$. For $e \in \mathcal{E}_I$ with $e = \partial T^+ \cap \partial T^-$,

$$\begin{aligned}
\int_e Q_h(\{\kappa \nabla v_h \cdot \nu_e\})[v_h]\,\mathrm{d}s &\le \|\{\kappa\nabla v_h\cdot\nu_e\}\|_{0,e}\|Q_h[v_h]\|_{0,e} \\
&\le \frac{1}{2}\big(\kappa^+\|\nabla v_h\cdot\nu|_{T^+}\|_{0,e} + \kappa^-\|\nabla v_h\cdot\nu|_{T^-}\|_{0,e}\big)\|Q_h[v_h]\|_{0,e} \\
&\le \frac{C^{1/2}}{2}\left(\kappa^+\left(\frac{\text{meas}(e)}{\text{meas}(T^+)}\right)^{1/2}\|\nabla v_h\|_{0,T^+} + \kappa^-\left(\frac{\text{meas}(e)}{\text{meas}(T^-)}\right)^{1/2}\|\nabla v_h\|_{0,T^-}\right) \\
&\qquad \times \|Q_h[v_h]\|_{0,e} \\
&\le \frac{C^{1/2}}{2}\big(\|\kappa^{1/2}\nabla v_h\|^2_{0,T^+} + \|\kappa^{1/2}\nabla v_h\|^2_{0,T^-}\big)^{1/2} h_e^{-1/2}\|Q_h[v_h]\|_{0,e},
\end{aligned}$$

where we used (24) and (22). For $e \in \mathcal{E}_D$ with $e \subset \partial T$,

$$\int_e Q_h(\kappa\nabla v_h\cdot\nu_e)v_h\,\mathrm{d}s \le C^{1/2}\|\kappa^{1/2}\nabla v_h\|_{0,T}\, h_e^{-1/2}\|Q_h v_h\|_{0,e},$$

where we used (24) and (23). Summing over $e \in \mathcal{E}_I \cup \mathcal{E}_D$, we obtain

$$\begin{aligned}
\sum_{e\in\mathcal{E}_I}\int_e Q_h(\{\kappa\nabla v_h\cdot\nu_e\})[v_h]\,\mathrm{d}s + &S\sum_{e\in\mathcal{E}_D}\int_e Q_h(\kappa\nabla v_h\cdot\nu_e)v_h\,\mathrm{d}s \\
&\le C^{1/2}(n+1)^{1/2}\left(\sum_{T\in\mathcal{T}_h}\|\kappa^{1/2}\nabla v_h\|^2_{0,T}\right)^{1/2} \\
&\qquad \times\left(\sum_{e\in\mathcal{E}_I} h_e^{-1}\|Q_h[v_h]\|^2_{0,e} + \sum_{e\in\mathcal{E}_D} h_e^{-1}\|Q_h v_h\|^2_{0,e}\right)^{1/2} \\
&\le \varepsilon\left(\sum_{T\in\mathcal{T}_h}\|\kappa^{1/2}\nabla v_h\|^2_{0,T}\right) \\
&\qquad + \frac{C(n+1)}{4\varepsilon}\left(\sum_{e\in\mathcal{E}_I} h_e^{-1}\|Q_h[v_h]\|^2_{0,e} + \sum_{e\in\mathcal{E}_D} h_e^{-1}\|Q_h v_h\|^2_{0,e}\right),
\end{aligned}$$

since there are $(n+1)$ faces for each element. Hence it follows that

$$\begin{aligned}
B(v_h, v_h) \ge (1-\varepsilon)&\left(\sum_{T\in\mathcal{T}_h}\|\kappa^{1/2}\nabla v_h\|^2_{0,T}\right) + \|\alpha^{1/2}\bar{v}_h\|^2_{0,\Omega} \\
&+ \left(\gamma - \frac{C(n+1)}{4\varepsilon}\right)\left(\sum_{e\in\mathcal{E}_I} h_e^{-1}\|Q_h[v_h]\|^2_{0,e} + \sum_{e\in\mathcal{E}_D} h_e^{-1}\|Q_h v_h\|^2_{0,e}\right).
\end{aligned}$$

Thus we can take $0 < \varepsilon < 1$ to ensure the uniform ellipticity, if $\gamma > \frac{C(n+1)}{4}$. □

5. Extension to other elements

In this section we will discuss some extension of Method II to rectangular and prismatic elements. Since the arguments in Section 4 directly carry over to these cases, we concentrate on describing how the methods are constructed.

5.1. **Rectangular elements**

We will consider the RT and BDFM spaces for the vector approximation. We begin with a brief review of the definitions of RT and BDFM spaces and their local degrees of freedom which are necessary to define our methods. More details on these spaces can be found in [7, 10, 35].

Let $\mathcal{Q}_{r_1,r_2}(T)$ be the space of all polynomials on a rectangle T whose degrees are less than or equal to r_1 in x_1 and r_2 in x_2, and $\widetilde{\mathcal{Q}}_{r_1,r_2}(T)$ its subspace consisting of polynomials whose degree is exactly equal to r_1 in x_1 *or* r_2 in x_2. It is easy to see that

$$\mathcal{Q}_{r_1,r_2}(T) = \mathcal{Q}_{r_1-1,r_2-1}(T) \oplus \widetilde{\mathcal{Q}}_{r_1,r_2}(T).$$

We set $\mathcal{Q}_r := \mathcal{Q}_{r,r}$ and $\widetilde{\mathcal{Q}}_r(T) := \widetilde{\mathcal{Q}}_{r,r}$. Similar definitions can be made for $n = 3$.

The rectangular RT space is defined to be

$$\mathcal{RT}_r(T) = \begin{cases} \mathcal{Q}_{r+1,r}(T) \times \mathcal{Q}_{r,r+1}(T) & \text{for } n = 2, \\ \mathcal{Q}_{r+1,r,r}(T) \times \mathcal{Q}_{r,r+1,r}(T) \times \mathcal{Q}_{r,r,r+1}(T) & \text{for } n = 3. \end{cases}$$

The degrees of freedom for $\xi_h \in \mathcal{RT}_r(T)$ are given by the moments which completely determine $\xi_h \cdot \nu_e \in \mathcal{Q}_r(e)$ on each $e \in \mathcal{E}_T$ and the moments of ξ_h on T

$$\left\{ \int_T \xi_h \cdot \tau \,\mathrm{d}x : \tau \in \Phi_r(T) \right\},$$

where

$$\Phi_r(T) = \begin{cases} \mathcal{Q}_{r-1,r}(T) \times \mathcal{Q}_{r,r-1}(T) & \text{for } n = 2, \\ \mathcal{Q}_{r-1,r,r}(T) \times \mathcal{Q}_{r,r-1,r}(T) \times \mathcal{Q}_{r,r,r-1}(T) & \text{for } n = 3. \end{cases}$$

The BDFM space is defined as follows: for $n = 2$

$$\mathcal{BDFM}_r(T) = (\mathcal{P}_{r+1}(T) \setminus \{x_2^{r+1}\}) \times (\mathcal{P}_{r+1}(T) \setminus \{x_1^{r+1}\}),$$

and for $n = 3$

$$\mathcal{BDFM}_r(T) = (\mathcal{P}_{r+1}(T) \setminus \mathrm{Hom}_{r+1}(x_2, x_3)) \times (\mathcal{P}_{r+1}(T) \setminus \mathrm{Hom}_{r+1}(x_1, x_3)) \times (\mathcal{P}_{r+1}(T) \setminus \mathrm{Hom}_{r+1}(x_1, x_2)),$$

where $\mathrm{Hom}_{r+1}(\cdot, \cdot)$ denotes the set of homogeneous polynomials of degree $r + 1$. The degrees of freedom for $\xi_h \in \mathcal{BDFM}_r(T)$ are given by the moments which completely determine $\xi_h \cdot \nu_e \in \mathcal{P}_r(e)$ on each $e \in \mathcal{E}_T$ and the moments of ξ_h on T

$$\left\{ \int_T \xi_h \cdot \tau \,\mathrm{d}x : \tau \in (\mathcal{P}_{r-1}(T))^n \right\}.$$

Now we present the following methods which are the analogues of Method II for the RT and BDFM spaces over rectangular grids:

Find $(\sigma_h, u_h) \in \mathcal{RT}_k \times \mathcal{Q}_{k+1}$ such that

$$\int_T (\sigma_h + \kappa \nabla u_h) \cdot \tau \,\mathrm{d}x = 0 \qquad \forall \tau \in \Phi_k(T) \oplus \nabla \widetilde{\mathcal{Q}}_{k+1}(T), \tag{25a}$$

$$\int_T (\nabla \cdot \sigma_h + \alpha u_h) w \,\mathrm{d}x = \int_T f w \,\mathrm{d}x \qquad \forall w \in \mathcal{Q}_k(T) \tag{25b}$$

and

$$\sigma_h \cdot \nu_e|_e = \begin{cases} -Q_h(\{\kappa \nabla u_h\} \cdot \nu_e) + \gamma h_e^{-1} Q_h([u_h]) & \text{for } e \in \mathcal{E}_I, \\ -Q_h(\kappa \nabla u_h \cdot \nu_e) + \gamma h_e^{-1} Q_h(u_h - g_D) & \text{for } e \in \mathcal{E}_D, \\ -Q_h g_N & \text{for } e \in \mathcal{E}_N, \end{cases} \tag{25c}$$

where Q_h is the L^2-projection onto $Q_k(e)$.

Find $(\sigma_h, u_h) \in \mathcal{BDFM}_k \times \mathcal{P}_{k+1}$ such that

$$\int_T (\sigma_h + \kappa \nabla u_h) \cdot \tau \,\mathrm{d}x = 0 \qquad \forall \tau \in (\mathcal{P}_{k-1}(T))^n \oplus \nabla \widetilde{\mathcal{P}}_{k+1}(T), \tag{26a}$$

$$\int_T (\nabla \cdot \sigma_h + \alpha u_h) w \,\mathrm{d}x = \int_T f w \,\mathrm{d}x \qquad \forall w \in \mathcal{P}_k(T) \tag{26b}$$

and

$$\sigma_h \cdot \nu_e|_e = \begin{cases} -Q_h(\{\kappa \nabla u_h\} \cdot \nu_e) + \gamma h_e^{-1} Q_h([u_h]) & \text{for } e \in \mathcal{E}_I, \\ -Q_h(\kappa \nabla u_h \cdot \nu_e) + \gamma h_e^{-1} Q_h(u_h - g_D) & \text{for } e \in \mathcal{E}_D, \\ -Q_h g_N & \text{for } e \in \mathcal{E}_N, \end{cases} \tag{26c}$$

where Q_h is the L^2-projection onto $P_k(e)$.

Remark 5.1. We note that u_h is sought in $\mathcal{Q}_{k+1}$ for the RT space, whereas $\mathcal{P}_{k+1}$ is used for the BDFM space.

Remark 5.2. The formulation (25), based on the RT space, can be applied to quadrilateral or hexahedral elements as well. On these elements, we define

$$\mathcal{Q}_{r_1,r_2}(T) = \{\hat{v} \circ F_T^{-1} : \hat{v} \in \mathcal{Q}_{r_1,r_2}(\hat{T})\},$$
$$\mathcal{RT}_r(T) = \{(\det B_T)^{-1} B_T \hat{\tau} \circ F_T^{-1} : \hat{\tau} \in \mathcal{RT}_r(\hat{T})\},$$

where $\hat{T}$ is the unit cube in $\mathbb{R}^n$ and $F_T : \hat{T} \to T$ is the invertible bilinear or trilinear mapping with $B_T = DF_T$. The interior degrees of freedom for $\mathcal{RT}_r(T)$ are provided by the moments with respect to the set

$$\Phi_r(T) = \{B_T^{-t} \hat{\tau} \circ F_T^{-1} : \hat{\tau} \in \Phi_r(\hat{T})\}.$$

In particular, since

$$\nabla v = B_T^{-t} \hat{\nabla} \hat{v} \circ F_T^{-1} \qquad \text{where} \quad v = \hat{v} \circ F_T^{-1},$$

we have

$$\Phi_k(T) \supseteq \nabla \mathcal{Q}_k(T),$$

which implies that the condition (**C1**) (modified in a suitable way) is satisfied.

5.2. Prismatic elements

We consider the space introduced by Nedelec [33]. The spaces developed by Chen and Douglas [16] can be similarly treated.

Let T be a prism whose base is a triangle in the (x_1, x_2)-plane, with three vertical edges parallel to the x_3 axis. By $\mathcal{P}_{l,m}(T)$ we denote the space of all polynomials whose degrees are at most l in (x_1, x_2) and at most m in x_3, and by $\widetilde{\mathcal{P}}_{l,m}(T)$ its subspace consisting of polynomials whose degrees are exactly equal to l in (x_1, x_2) *or* to m in x_3. It is easy to see that

$$\mathcal{P}_{l,m}(T) = \mathcal{P}_{l-1,m-1}(T) \oplus \widetilde{\mathcal{P}}_{l,m}(T).$$

Let $\mathcal{RT}_{l,m}(T)$ denote the space of pairs of polynomials which belong to the triangular RT space of order l, $(\mathcal{P}_l(T))^2 \oplus (x_1, x_2)\widetilde{\mathcal{P}}_l(T)$, for x_3 fixed, and are of degree at most m in x_3.

We define the Nedelec space $\mathcal{NE}_r(T)$ to be

$$\mathcal{NE}_r(T) = \{\tau = (\tau_1, \tau_2, \tau_3) : \tau_1, \tau_2 \in \mathcal{RT}_{r,r}(T),\ \tau_3 \in \mathcal{P}_{r,r+1}\}.$$

As for the degrees of freedom, we have

$$\begin{aligned} &\xi_h \cdot \nu_e \in \mathcal{P}_r(e) && \text{for the two horizontal faces;}\\ &\xi_h \cdot \nu_e \in \mathcal{Q}_r(e) && \text{for the three vertical edges,} \end{aligned}$$

and the interior degrees of freedom are provided by the moments

$$\left\{ \int_T \xi_h \cdot \tau \,\mathrm{d}x : \tau \in \Phi_r(T) \right\},$$

where $\Phi_r(T) = (\mathcal{P}_{r-1,r}(T))^2 \times \mathcal{P}_{r,r-1}(T)$.

Now we propose the following method: find $(\sigma_h, u_h) \in \mathcal{NE}_k \times \mathcal{P}_{k+1,k+1}$ satisfying

$$\int_T (\sigma_h + \kappa\nabla u_h) \cdot \tau \,\mathrm{d}x = 0 \qquad \forall \tau \in \Phi_k(T) \oplus \nabla\widetilde{\mathcal{P}}_{k+1,k+1}(T), \tag{27a}$$

$$\int_T (\nabla \cdot \sigma_h + \alpha u_h) w \,\mathrm{d}x = \int_T f w \,\mathrm{d}x \qquad \forall w \in \mathcal{P}_{k,k}(T), \tag{27b}$$

and

$$\sigma_h \cdot \nu_e|_e = \begin{cases} -Q_h(\{\kappa\nabla u_h\} \cdot \nu_e) + \gamma h_e^{-1} Q_h([u_h]) & \text{for } e \in \mathcal{E}_I,\\ -Q_h(\kappa\nabla u_h \cdot \nu_e) + \gamma h_e^{-1} Q_h(u_h - g_D) & \text{for } e \in \mathcal{E}_D,\\ -Q_h g_N & \text{for } e \in \mathcal{E}_N, \end{cases} \tag{27c}$$

where Q_h is the L^2-projection onto $\mathcal{P}_k(e)$ for the two horizontal faces and onto $\mathcal{Q}_k(e)$ for the three vertical edges.

6. Numerical results

In this section we present numerical results for two test problems on the unit square $\Omega = (0, 1)^2$ in order to demonstrate the performance of our mixed finite volume methods. For each test problem the true solution is specified together with the corresponding problem data and compared with the computed numerical solutions. Numerical experiments are carried out with polynomials of degree $k = 1, 2, 3$ for the scalar variable, that is,

$$\begin{aligned} &\text{Method I} && \text{with} \quad \mathcal{RT}_1 \times \mathcal{P}_1, \quad \mathcal{RT}_2 \times \mathcal{P}_2, \quad \mathcal{RT}_3 \times \mathcal{P}_3\\ &\text{Method II} && \text{with} \quad \mathcal{RT}_0 \times \mathcal{P}_1, \quad \mathcal{RT}_1 \times \mathcal{P}_2, \quad \mathcal{RT}_2 \times \mathcal{P}_3. \end{aligned}$$

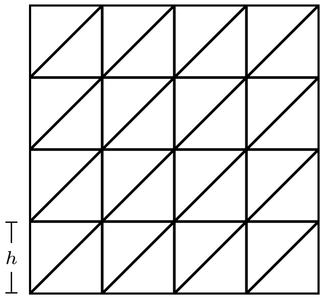

FIGURE 1. Uniform triangular grid of size h.

The stabilization parameter γ is chosen to be $5, 10$ and 15 for $k = 1, 2$ and 3, respectively. We only consider a sequence of uniform triangular grids generated by partitioning Ω into the squares of equal size h and then dividing each square into two triangles by the diagonal from the lower left corner to the upper right corner, as shown in Figure 1.

The stiffness matrices and the load vectors for the variational formulation (12) or (18) are calculated exactly by means of high order quadrature rules. For the solution method of the resulting linear systems, we use the generalized minimum residual method (GMRES) with the variable V-cycle multigrid preconditioner (*cf.* [30]) to compute the scalar solution u_h. The stopping criterion is that the residual norm should be less than 10^{-15}. The vector solution σ_h is then recovered in a local manner from (11a), (11c) for Method I and from (17a), (17c) for Method II. We found that this postprocessing is virtually cost-free.

Below we report the scalar errors measured in the broken H^1 and L^2 norms and the vector errors and their divergence measured in the L^2 norm, where the broken H^1 norm is obviously defined by

$$|u - u_h|_{1,h} = \left(\sum_{T \in \mathcal{T}_h} \|\nabla(u - u_h)\|_{0,T}^2 \right)^{1/2}.$$

Assuming that the errors are of the form Ch^β, we also estimate the values of β, the order of convergence, by using the least-squares fit to the computed errors.

6.1. Simple Poisson problem

Our first test problem is the simple Poisson problem with the homogeneous Dirichlet boundary condition

$$\begin{aligned} -\Delta u &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \partial\Omega, \end{aligned}$$

where we chose the smooth solution

$$u(x, y) = \sin(\pi x)\sin(\pi y)$$

and the corresponding right-hand side $f(x, y) = 2\pi^2 \sin(\pi x) \sin(\pi y)$.

The numerical results for Method I are reported in Tables 1–3. We first observe the optimal orders of convergence in the broken H^1 error of the scalar and in the L^2 error of the divergence of the vector for all $k = 1, 2, 3$, as predicted by the theory. Note that the latter error is, in fact, *a priori* determined by the right-hand side f, since $\alpha \equiv 0$. For the L^2 error of the scalar (not handled in this work), we see that the convergence is optimal for the odd degrees $k = 1, 3$, whereas it is one order less for the even degree $k = 2$. On the other hand, for the L^2 error of the vector, one order less convergence is observed for all degrees $k = 1, 2, 3$, which indicates that our theoretical prediction (*cf.* Thm. 3.9) cannot be improved.

Table 1. Method I: $\mathcal{RT}_1 \times \mathcal{P}_1$ for simple Poisson problem.

h	$\lvert u-u_h\rvert_{1,h}$	$\lVert u-u_h\rVert_0$	$\lVert \sigma-\sigma_h\rVert_0$	$\lVert \nabla\cdot(\sigma-\sigma_h)\rVert_0$
1/8	3.225e–1	7.263e–3	2.521e–1	9.772e–2
1/16	1.635e–1	1.939e–3	1.314e–1	2.453e–2
1/32	8.216e–2	5.005e–4	6.679e–2	6.139e–3
1/64	4.118e–2	1.271e–4	3.364e–2	1.535e–3
1/128	2.062e–2	3.203e–5	1.688e–2	3.838e–4
order	0.992	1.958	0.976	1.998

Table 2. Method I: $\mathcal{RT}_2 \times \mathcal{P}_2$ for simple Poisson problem.

h	$\lvert u-u_h\rvert_{1,h}$	$\lVert u-u_h\rVert_0$	$\lVert \sigma-\sigma_h\rVert_0$	$\lVert \nabla\cdot(\sigma-\sigma_h)\rVert_0$
1/8	2.583e–2	1.359e–3	2.437e–2	5.422e–3
1/16	6.520e–3	3.128e–4	6.309e–3	6.804e–4
1/32	1.636e–3	7.533e–5	1.601e–3	8.513e–5
1/64	4.097e–4	1.852e–5	4.027e–4	1.065e–5
1/128	1.025e–4	4.593e–6	1.010e–4	1.331e–6
order	1.994	2.049	1.979	2.998

Table 3. Method I: $\mathcal{RT}_3 \times \mathcal{P}_3$ for simple Poisson problem.

h	$\lvert u-u_h\rvert_{1,h}$	$\lVert u-u_h\rVert_0$	$\lVert \sigma-\sigma_h\rVert_0$	$\lVert \nabla\cdot(\sigma-\sigma_h)\rVert_0$
1/8	1.471e–3	2.204e–5	1.295e–3	2.369e–4
1/16	1.845e–4	1.348e–6	1.620e–4	1.486e–5
1/32	2.309e–5	8.322e–8	2.023e–5	9.293e–7
1/64	2.886e–6	5.170e–9	2.527e–6	5.810e–8
1/128	3.605e–7	3.205e–10	3.158e–7	3.631e–9
order	2.998	4.016	3.000	3.998

Tables 4–6 contain the corresponding numerical results for Method II. By comparing them with those of Method I, we observe that both methods display the same orders of convergence for the scalar error and for the L^2 error of the vector, when applied to the same-degree polynomials for the scalar variable. More specifically, Method II achieves the optimal order of convergence for the L^2 error of the vector with one lower degree for the vector variable than Method I, although Method I gains one higher order of convergence in the divergence of the vector.

6.2. **Discontinuous coefficient**

In the second test problem we deal with the discontinuous coefficient given by

$$\kappa(x,y) = \begin{cases} 1.0 & \text{if } (x-1/2)(y-1/2) > 0, \\ 100.0 & \text{if } (x-1/2)(y-1/2) < 0. \end{cases}$$

TABLE 4. Method II: $\mathcal{RT}_0 \times \mathcal{P}_1$ for simple Poisson problem.

h	$\lvert u-u_h\rvert_{1,h}$	$\lVert u-u_h\rVert_0$	$\lVert \sigma-\sigma_h\rVert_0$	$\lVert \nabla\cdot(\sigma-\sigma_h)\rVert_0$
1/8	3.236e–1	8.215e–3	2.517e–1	1.286e+0
1/16	1.624e–1	2.087e–3	1.259e–1	6.452e–1
1/32	8.126e–2	5.252e–4	6.296e–2	3.229e–1
1/64	4.064e–2	1.317e–4	3.148e–2	1.615e–1
1/128	2.032e–2	3.298e–5	1.574e–2	8.075e–2
order	0.998	1.990	0.999	0.998

TABLE 5. Method II: $\mathcal{RT}_1 \times \mathcal{P}_2$ for simple Poisson problem.

h	$\lvert u-u_h\rvert_{1,h}$	$\lVert u-u_h\rVert_0$	$\lVert \sigma-\sigma_h\rVert_0$	$\lVert \nabla\cdot(\sigma-\sigma_h)\rVert_0$
1/8	2.448e–2	1.012e–3	1.475e–2	9.772e–2
1/16	6.144e–3	2.168e–4	3.599e–3	2.453e–2
1/32	1.538e–3	5.072e–5	8.903e–4	6.139e–3
1/64	3.846e–4	1.234e–5	2.215e–4	1.535e–3
1/128	9.614e–5	3.046e–6	5.526e–5	3.838e–4
order	1.998	2.088	2.014	1.998

TABLE 6. Method II: $\mathcal{RT}_2 \times \mathcal{P}_3$ for simple Poisson problem.

h	$\lvert u-u_h\rvert_{1,h}$	$\lVert u-u_h\rVert_0$	$\lVert \sigma-\sigma_h\rVert_0$	$\lVert \nabla\cdot(\sigma-\sigma_h)\rVert_0$
1/8	1.440e–3	2.040e–5	7.101e–4	5.422e–3
1/16	1.799e–4	1.242e–6	8.425e–5	6.804e–4
1/32	2.246e–5	7.669e–8	1.025e–5	8.513e–5
1/64	2.806e–6	4.766e–9	1.264e–6	1.065e–5
1/128	3.506e–7	2.964e–10	1.570e–7	1.331e–6
order	3.001	4.016	3.034	2.998

The true solution is continuous and piecewise smooth which is chosen as follows:

$$u(x,y) = \frac{1}{\kappa(x,y)} \sin(2\pi x)\sin(2\pi y).$$

We again set $\alpha \equiv 0$ and impose the homogeneous Dirichlet boundary condition on $\partial\Omega$. Note that $\sigma = -\kappa\nabla u$ is smooth on the whole domain, in spite of the discontinuity of κ.

To take account of the discontinuity of κ, the penalty parameter γh_e^{-1} is multiplied by the additional factor

$$\frac{\kappa^+ + \kappa^-}{2} \quad \text{for} \quad e = \partial T^+ \cap \partial T^- \qquad \text{and} \qquad \kappa_T \quad \text{for} \quad e = \partial T \cap \partial\Omega$$

which is similar to the one defined in Section 4.3.

TABLE 7. Method I: $\mathcal{RT}_1 \times \mathcal{P}_1$ for discontinuous coefficient.

h	$\lvert u-u_h\rvert_{1,h}$	$\lVert u-u_h\rVert_0$	$\lVert\sigma-\sigma_h\rVert_0$	$\lVert\nabla\cdot(\sigma-\sigma_h)\rVert_0$
1/8	9.647e–1	2.385e–2	1.390e+0	1.539e+0
1/16	4.795e–1	6.032e–3	6.587e–1	3.909e–1
1/32	2.373e–1	1.497e–3	3.067e–1	9.812e–2
1/64	1.178e–1	3.715e–4	1.453e–1	2.456e–2
1/128	5.864e–2	9.246e–5	7.029e–2	6.140e–3
order	1.010	2.004	1.079	1.993

TABLE 8. Method I: $\mathcal{RT}_2 \times \mathcal{P}_2$ for discontinuous coefficient.

h	$\lvert u-u_h\rvert_{1,h}$	$\lVert u-u_h\rVert_0$	$\lVert\sigma-\sigma_h\rVert_0$	$\lVert\nabla\cdot(\sigma-\sigma_h)\rVert_0$
1/8	1.511e–1	4.234e–3	2.367e–1	1.709e–1
1/16	3.773e–2	9.006e–4	5.649e–2	2.169e–2
1/32	9.374e–3	2.124e–4	1.361e–2	2.722e–3
1/64	2.333e–3	5.210e–5	3.325e–3	3.406e–4
1/128	5.818e–4	1.294e–5	8.210e–4	4.258e–5
order	2.005	2.081	2.042	2.993

Numerical results with $k = 1, 2$ are presented in Tables 7–8 for Method I and in Tables 9–10 for Method II which show the same convergence behavior obtained for the simple Poisson problem. Finally, we note that Method II yields slightly better results in the L^2 error of the vector.

7. CONCLUSIONS

In the present work we have introduced and analyzed new mixed finite volume methods on non-staggered grids in which $H(\text{div})$-*conforming* approximations are used for the vector variable and *discontinuous* approximations are used for the scalar variable. The construction is general enough to cover all the existing vector approximation spaces.

Our new methods have some distinct advantages over other mixed methods such as the same local mass conservation property of mixed finite element methods and local elimination and recovery of the vector variable. In particular, one can obtain discontinuous finite element methods for the scalar variable only in which the usual symmetrizing or anti-symmetrizing terms are missing. In doing so, it has turned out that the L^2 projection of the interface averages and jumps onto the lower-order normal trace space plays a crucial role in recovering optimal vector approximations.

Although our method gives rise to a nonsymmetric matrix system even for the symmetric problem, this is no drawback when applied to the convection-diffusion problem, and we can solve it in an efficient way by means of the Krylov subspace method like GMRES with preconditioners, *e.g.*, from [29, 30]. This will be the subject of a forthcoming paper.

Table 9. Method II: $\mathcal{RT}_0 \times \mathcal{P}_1$ for discontinuous coefficient.

h	$\lvert u - u_h\rvert_{1,h}$	$\lVert u - u_h\rVert_0$	$\lVert\sigma - \sigma_h\rVert_0$	$\lVert\nabla\cdot(\sigma - \sigma_h)\rVert_0$
1/8	9.025e–1	2.323e–2	1.013e+0	1.014e+1
1/16	4.577e–1	5.928e–3	5.043e–1	5.143e+0
1/32	2.297e–1	1.490e–3	2.519e–1	2.581e+0
1/64	1.150e–1	3.730e–4	1.260e–1	1.292e+0
1/128	5.747e–2	9.327e–5	6.296e–2	6.460e–1
order	0.993	1.991	1.001	0.993

Table 10. Method II: $\mathcal{RT}_1 \times \mathcal{P}_2$ for discontinuous coefficient.

h	$\lvert u - u_h\rvert_{1,h}$	$\lVert u - u_h\rVert_0$	$\lVert\sigma - \sigma_h\rVert_0$	$\lVert\nabla\cdot(\sigma - \sigma_h)\rVert_0$
1/8	1.379e–1	3.480e–3	1.346e–1	1.539e+0
1/16	3.479e–2	6.404e–4	3.142e–2	3.909e–1
1/32	8.710e–3	1.427e–4	7.471e–3	9.812e–2
1/64	2.177e–3	3.450e–5	1.817e–3	2.456e–2
1/128	5.442e–4	8.548e–6	4.476e–4	6.140e–3
order	1.996	2.155	2.057	1.993

Acknowledgements. The author would like to thank Prof. Zhangxin Chen and Do Y. Kwak for their careful reading of this paper and many helpful comments.

References

[1] T. Arbogast and Z. Chen, On the implementation of mixed methods as nonconforming methods for second order elliptic problems. *Math. Comp.* **64** (1995) 943–972.

[2] T. Arbogast, M. Wheeler and I. Yotov, Mixed finite elements for elliptic problems with tensor coefficients as cell-centered finite differences. *SIAM J. Numer. Anal.* **34** (1997) 828–852.

[3] D.N. Arnold and F. Brezzi, Mixed and nonconforming finite element methods: implementation, postprocessing and error estimates. *RAIRO Modél. Math. Anal. Numér.* **19** (1985) 7–32.

[4] 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) 1749–1779.

[5] J. Baranger, J.F. Maître and F. Oudin, Connection between finite volume and mixed finite element methods. *RAIRO Modél. Math. Anal. Numér.* **30** (1996) 445–465.

[6] F. Bassi and S. Rebay, A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations. *J. Comput. Phys.* **131** (1997) 267–279.

[7] F. Brezzi and M. Fortin, *Mixed and hybrid finite element methods*. Springer-Verlag (1991).

[8] F. Brezzi, J. Douglas and L.D. Marini, Two families of mixed finite elements for second order elliptic problems. *Numer. Math.* **47** (1985) 217–235.

[9] F. Brezzi, J. Douglas, R. Durán and M. Fortin, Mixed finite elements for second order elliptic problems in three variables. *Numer. Math.* **51** (1987) 237–250.

[10] F. Brezzi, J. Douglas, M. Fortin and L.D. Marini, Efficient rectangular mixed finite elements in two and three variables. *RAIRO Modél. Math. Anal. Numér.* **21** (1987) 581–604.

[11] F. Brezzi, G. Manzini, L.D. Marini, P. Pietra and A. Russo, Discontinuous Galerkin approximations for elliptic problems. *Numer. Methods Partial Differential Equations* **16** (2000) 365–378.

[12] Z. Cai, J.E. Jones, S.F. McCormick and T.F. Russell, Control-volume mixed finite element Methods. *Comput. Geosci.* **1** (1997) 289–315.

[13] P. Castillo, B. Cockburn, I. Perugia and D. Schötzau, An *a priori* error analysis of the local discontinuous Galerkin method for elliptic problems. *SIAM J. Numer. Anal.* **38** (2000) 1676–1706.

[14] Z. Chen, Expanded mixed finite element methods for linear second-order elliptic problems I. *RAIRO Modél. Math. Anal. Numér.* **32** (1998) 479–499.

[15] Z. Chen, On the relationship of various discontinuous finite element methods for second-order elliptic equations. *East-West J. Numer. Math.* **9** (2001) 99–122.

[16] Z. Chen and J. Douglas, Prismatic mixed finite elements for second order elliptic problems. *Calcolo* **26** (1989) 135–148.

[17] S.H. Chou and P.S. Vassilevski, A general mixed covolume framework for constructing conservative schemes for elliptic problems. *Math. Comp.* **68** (1999) 991–1011.

[18] S.H. Chou, D.Y. Kwak and P. Vassilevski, Mixed covolume methods for elliptic problems on triangular grids. *SIAM J. Numer. Anal.* **35** (1998) 1850–1861.

[19] S.H. Chou, D.Y. Kwak and K.Y. Kim, A general framework for constructing and analyzing mixed finite volume methods on quadrilateral grids: the overlapping covolume case. *SIAM J. Numer. Anal.* **39** (2001) 1170–1196

[20] S.H. Chou, D.Y. Kwak and K.Y. Kim, Mixed finite volume methods on non-staggered quadrilateral grids for elliptic problems. *Math. Comp.* **72** (2003) 525–539.

[21] P. Ciarlet, *The Finite Element Method for Elliptic Problems.* North-Holland (1978).

[22] B. Cockburn and C.W. Shu, The local discontinuous Galerkin method for time-dependent convection-diffusion system. *SIAM J. Numer. Anal.* **35** (1998) 2440–2463.

[23] B. Courbet and J.P. Croisille, Finite volume box schemes on triangular meshes. *RAIRO Modél. Math. Anal. Numér.* **32** (1998) 631–649.

[24] J.P. Croisille, Finite volume box schemes and mixed methods *ESAIM: M2AN* **34** (2000) 1087–1106.

[25] J.P. Croisille and I. Greff, Some nonconforming mixed box schemes for elliptic problems. *Numer. Methods Partial Differential Equations* **18** (2002) 355–373.

[26] C. Dawson, The $\mathcal{P}^{K+1} - \mathcal{S}^K$ local discontinuous Galerkin method for elliptic equations. *SIAM J. Numer. Anal.* **40** (2002) 2151–2170.

[27] R.G. Durán, Error analysis in $L^p, 1 \leq p \leq \infty$, for mixed finite element methods for linear and quasi-linear elliptic problems. *RAIRO Modél. Math. Anal. Numér.* **22** (1988) 371–387.

[28] R.S. Falk and J.E. Osborn, Error estimates for mixed methods. *RAIRO Anal. Numér.* **14** (1980) 249–277.

[29] X. Feng and O.A. Karakashian, Two-level additive Schwarz methods for a discontinuous Galerkin approximation of second order elliptic problems. *SIAM J. Numer. Anal.* **39** (2001) 1343–1365.

[30] J. Gopalakrishnan and G. Kanschat, A multilevel discontinuous Galerkin method. *Numer. Math.* **95** (2003) 527–550.

[31] S. Micheletti and R. Sacco, Dual-primal mixed finite elements for elliptic problems. *Numer. Methods Partial Differential Equations* **17** (2001) 137–151.

[32] J.C. Nedelec, Mixed finite elements in $\mathbb{R}^3$. *Numer. Math.* **35** (1980) 315–341.

[33] J.C. Nedelec, A new family of mixed finite elements in $\mathbb{R}^3$. *Numer. Math.* **50** (1986) 57–81.

[34] I. Perugia and D. Schötzau, An *hp*-analysis of the local discontinuous Galerkin method for diffusion problems. *J. Sci. Comput.* **17** (2002) 561–571.

[35] P.A. Raviart and J.M. Thomas, A mixed finite element method for 2nd order elliptic problems, in *Proc. Conference on Mathematical Aspects of Finite Element Methods*, Springer-Verlag. *Lect. Notes Math.* **606** (1977) 292–315.

[36] B. Riviere, M.F. Wheeler and V. Girault, *A priori* error estimates for finite element methods based on discontinuous approximation spaces for elliptic problems. *SIAM J. Numer. Anal.* **39** (2001) 902–931.

[37] J.E. Roberts and J.M. Thomas, Mixed and hybrid methods, in *Handbook of Numerical Analysis*, Vol. II, North-Holland (1991) 523–639.

[38] R. Sacco and F. Saleri, Mixed finite volume methods for semiconductor device simulation. *Numer. Methods Partial Differential Equations* **13** (1997) 215–236.

[39] A. Weiser and M.F. Wheeler, On convergence of block-centered finite differences for elliptic problems. *SIAM J. Numer. Anal.* **25** (1988) 351–375.