# THE CELL FUNCTIONAL MINIMIZATION SCHEME FOR THE ANISOTROPIC DIFFUSION PROBLEMS ON ARBITRARY POLYGONAL GRIDS\*

Li  $\mathrm{Yin}^1, \ \mathrm{Jiming} \ \mathrm{Wu}^1$  and  $\mathrm{Zhiming} \ \mathrm{Gao}^1$ 

Abstract. A finite volume scheme based on minimization of a certain cell functional is constructed for unstructured polygonal meshes. This new scheme has a local stencil, allows arbitrary diffusion tensors, leads to a symmetric positive definite diffusion matrix in case that edge unknowns are defined at the midpoints of edges, and is linearity-preserving, *i.e.*, preserves linear solutions. Under a very weak geometry condition, the stability result and discrete  $H_1$  error estimate of the scheme is obtained through a discrete functional approach. Finally, numerical results on various mesh types (including a particular jigsaw puzzle mesh) demonstrate the good performance of the scheme and validate the theoretical analysis.

Mathematics Subject Classification. 65N12, 65N08, 35J25.

Received September 12, 2012. Revised March 25, 2014. Published online January 16, 2015.

### 1. INTRODUCTION

This paper will focus on the investigation of a new discretization scheme for solving the anisotropic steadystate diffusion equation

$$-\operatorname{div}(\Lambda(\mathbf{x})\nabla u) = f(\mathbf{x}) \quad \text{in } \Omega, \tag{1.1a}$$

 $u = u_D \quad \text{on} \ \Gamma_D, \tag{1.1b}$ 

$$-\Lambda(\mathbf{x})\nabla u \cdot \mathbf{n} = f_N \quad \text{on } \Gamma_N, \tag{1.1c}$$

where u denotes the scalar unknown function or the intensity,  $\Omega$  is a bounded connected polygonal domain in  $\mathbb{R}^2$ , **n** denotes the unit outward normal along the domain boundary  $\partial \Omega = \Gamma_D \cup \Gamma_N$  ( $\Gamma_D$  has a nonzero one-dimensional measure),  $\Lambda(\mathbf{x})$  is a 2 × 2 symmetric tensor, and uniformly bounded above and below in  $\Omega$ ,

yinli\_kitty122@sina.com; jiming\_wu@iapcm.ac.cn; dtgaozm@gmail.com

Keywords and phrases. Cell functional minimization, finite volume scheme, diffusion problem, polygonal mesh, convergence, stability, error estimate.

<sup>\*</sup> This work of the first author is supported by the National Natural Science Foundation of China under contract No. 91118001 and the Science Foundation of China Academy of Engineering Physics (2013B0202034), the second author is sponsored by the National Natural Science Foundation of China under contract Nos. 91330205 and 11271053, and the corresponding author is under the auspices of the National Natural Science Foundation of China (91330107) and Foundation of President of China Academy of Engineering Physics (2014-1-042).

<sup>&</sup>lt;sup>1</sup> Institute of Applied Physics and Computational Mathematics, P.O. Box 8009-9, Beijing 100088, P.R. China.

*i.e.*, there exist positive constants  $\underline{\kappa}$  and  $\overline{\kappa}$ , such that

$$\underline{\kappa} \|\mathbf{v}\|^2 \le \|\mathbf{v}^T \Lambda(\mathbf{x})\mathbf{v}\| \le \overline{\kappa} \|\mathbf{v}\|^2, \quad \forall \mathbf{v} \in \mathbb{R}^2, \quad \forall \mathbf{x} \in \Omega,$$
(1.2)

 $f, u_D$  and  $f_N$  represent the source function, the Dirichlet and flux boundary data, respectively.

Equations of this type arise in a wide range of scientific fields such as oil reservoir simulations, plasma physics, Navier–Stokes equations, biology, and so on. For example, in reservoir simulations u represents the pressure and the diffusion tensor  $\Lambda$  may be highly anisotropic and heterogeneous, which imposes a great challenge for discretization schemes.

There are several numerical methods arising from the solution of (1.1) through a functional approach, such as the mixed finite element method [11]. Here we are more interested in a finite volume scheme introduced in [9] where the flux variable  $\mathbf{F} = -\Lambda \nabla u$  is obtained by minimizing the following energy functional of a vector field  $\mathbf{G}$  on the whole domain  $\Omega$ ,

$$\widehat{W}(\mathbf{G}) := \int_{\Omega} \frac{|\mathbf{G}|^2}{\kappa_0} \,\mathrm{d}\mathbf{x} - 2 \int_{\Omega} u \,\mathrm{div}\mathbf{G} \,\mathrm{d}\mathbf{x},\tag{1.3}$$

where  $k_0$  denotes the scalar diffusion coefficient. Although the scheme avoids the errors coming from vertex values and edge diffusion coefficients calculated through certain interpolation procedure, it loses accuracy on distorted meshes and can only deal with the cases where pure flux boundary condition, scalar diffusion coefficient and structured quadrilateral mesh are involved. Later, based on minimization of a certain cell functional modified from (1.3), a new finite volume scheme together with its parallel algorithm is suggested in [14], where the relation with the local support operator method [10] is also discussed. This scheme can deal with boundary conditions of all types and much improves the accuracy on distorted meshes. However, it is still confined to the two-dimensional structured quadrilateral meshes and the diffusion coefficient is a scalar. Besides, the relevant theoretical results were not given. Recently, this same scheme has been extended to construct a nonoverlapping domain decomposition algorithm on multiblock quadrilateral meshes with non-matching interfaces [15].

The object of this paper is to present a new finite volume scheme based on minimization of a certain cell functional on general polygonal meshes. The key of our scheme is to construct a cell matrix  $A_K$  satisfying the linearity-preserving criterion (condition (3.3) holds). Of course that the form of  $A_K$  is not unique. So the presented scheme can be regarded as a family of cell functional minimization (CFM) schemes depending on different  $A_K$ . With some special forms of  $A_K$ , our schemes algebraically identify with some existing schemes [5]. However, the cell center and points on cell edges in our scheme are more flexible. Under a very weak geometry condition, the stability results and the discrete  $H_1$  error estimate of a family of CFM schemes are obtained. Finally in the numerical part, we show the good performance of our new scheme on various mesh families. In addition, we also present a so-called jigsaw puzzle mesh, most of whose cells are not star-shaped ones, in this case many geometry assumptions adopted in some previous papers, *e.g.*, [3,4], fail but our geometry condition holds fairly well. Numerical results on the jigsaw puzzle mesh validate the theoretical analysis of this paper.

In summary, our new scheme has the following characteristics:

- it has a local stencil;
- it is locally conservative;
- it leads to a symmetric positive definite linear system in case that edge unknowns are located at the midpoints of edges;
- it allows heterogeneous diffusion tensors;
- it is reliable on unstructured anisotropic meshes that may be severely distorted;
- it has second order convergence rate for the approximate solution on general meshes;
- it is linearity-preserving, *i.e.*, provides the exact solution if the diffusion tensor is piecewise constant in polygonal mesh cells and the solution is linear in each of these mesh cells;
- it has theoretical foundations: stability and discrete  $H_1$  error estimates.

The remainder of this paper is organized as follows. In Section 2, the cell functional minimization scheme is suggested for unstructured polygonal meshes. The construction and the spectral study of the cell matrix associated with the new scheme are presented in Section 3. In the fourth and fifth sections, we obtain the stability result and discrete  $H_1$  error estimate. Numerical results are reported in Section 6 to validate the theoretical analysis and conclusions are given in the last section.

#### 2. A finite volume scheme based on minimizing a certain cell functional

Throughout, we shall confine ourselves to the two dimensional case. However, most of the discussion in this paper holds for the three-dimensional case. For simplicity of exposition, we introduce some notations. Let  $\Omega$  be an open bounded connected polygonal domain in  $\mathbb{R}^2$ , and  $\partial \Omega$  its boundary. In this paper, a finite volume discretization of  $\Omega$ , denoted by  $\mathcal{D}$ , is defined as the triplex  $\mathcal{D} = (\mathcal{M}, \mathcal{E}, \mathcal{O})$ , where

- $\mathcal{M} = \{K\}$  is a finite family of disjoint open polygonal cells in  $\Omega$  such that  $\overline{\Omega} = \bigcup_{K \in \mathcal{M}} \overline{K}$ . For  $K \in \mathcal{M}$ , let  $\partial K$ , |K| and  $h_K$  denote the cell boundary, measure and diameter, respectively.
- $\mathcal{E} = \{\sigma\}$  is a finite family of disjoint edges in  $\overline{\Omega}$ . For  $\sigma \in \mathcal{E}$ ,  $\sigma$  is an open line segment whose one-dimensional measure is strictly positive and denoted as  $|\sigma|$ . Let  $\mathcal{E}^{int} = \mathcal{E} \cap \Omega$  and  $\mathcal{E}^{ext} = \mathcal{E} \cap \partial \Omega$ . For  $K \in \mathcal{M}$ , there exists a subset  $\mathcal{E}_K$  of  $\mathcal{E}$  such that  $\partial K = \bigcup_{\sigma \in \mathcal{E}_K} \overline{\sigma}$  and the number of edges in  $\mathcal{E}_K$  is  $n_K$ . For  $\sigma \in \mathcal{E}_K$ , notation  $\sigma$  may denote either an edge on  $\partial K$  or the local number of this same edge in cell K, depending on the context. This slight abuse of notation  $\sigma$  should not be a source of confusion.  $\mathbf{x}_{K,\sigma}$ , without special statement, denotes the midpoint of  $\sigma$ . In addition,  $\mathbf{n}_{K,\sigma}$  is the unit vector normal to  $\sigma$  outward to K.
- $\mathcal{O} = \{\mathbf{x}_K, K \in \mathcal{M}\}$  is a set of points, known as cell centers, where  $\mathbf{x}_K \in K$ .

In the following discussion, we shall assume that  $\Lambda(\mathbf{x})$  is constant on each cell  $K \in \mathcal{M}$  with  $\Lambda_K$  denoting the restriction of  $\Lambda(\mathbf{x})$  on K. Throughout, the hollow letters  $\mathbb{A}, \mathbb{F}, \mathbb{X}, \ldots$  will be used to denote matrices with column numbers greater than one, while the black ones  $\mathbf{F}, \mathbf{U}, \mathbf{I}, \mathbf{n}, \mathbf{x}$ , *etc.*, will be employed to denote column vectors or matrices with only one column.

By introducing the flux vector  $\mathbf{F} = -\Lambda(\mathbf{x})\nabla u$ , (1.1) can be rewritten as a system of first-order partial differential equations

$$\operatorname{div}\mathbf{F} = f(\mathbf{x}),\tag{2.1}$$

$$\mathbf{F} = -\Lambda(\mathbf{x})\nabla u. \tag{2.2}$$

Integrating (2.1) over a mesh cell K and using the divergence theorem, we obtain

$$\oint_{\partial K} \mathbf{F} \cdot \mathbf{n}_K \, \mathrm{d}s = \int_K f(\mathbf{x}) \mathrm{d}\mathbf{x},\tag{2.3}$$

where  $\mathbf{n}_K$  denotes the unit outward normal along the cell boundary  $\partial K$ . In physical terms, (2.3) represents the energy or particle conservation over the cell K. The main part of a finite volume discretization is now down to find certain approximation of the contour integration in (2.3).

### 2.1. Cell functional minimization algorithm

We begin the discussion by introducing a cell functional of a vector function  $\mathbf{G}$ , given by

$$W_K(\mathbf{G}) := \int_K \mathbf{G}^T \Lambda_K^{-1} \mathbf{G} \, \mathrm{d}\mathbf{x} - 2 \int_K u \, \mathrm{div} \mathbf{G} \, \mathrm{d}\mathbf{x} + 2 \oint_{\partial K} u \mathbf{G} \cdot \mathbf{n} \mathrm{d}s.$$
(2.4)

Obviously, (2.4) can be viewed as a modified version of (1.3), and coincides with the cell functional in [14] when  $\Lambda(\mathbf{x})$  reduces to a scalar. The finite volume scheme in this paper is based on the following theorem.



FIGURE 1. Locations of various unknowns.

**Theorem 2.1.** The flux  $\mathbf{F} = -\Lambda_K \nabla u$  minimizes the cell functional (2.4), i.e.,

$$W_K(\mathbf{G}) = W_K(\mathbf{F}) + \int_K (\mathbf{G} - \mathbf{F})^T \Lambda_K^{-1}(\mathbf{G} - \mathbf{F}) \,\mathrm{d}\mathbf{x} \ge W_K(\mathbf{F}).$$
(2.5)

*Proof.* Note that  $\Lambda(\mathbf{x})$  is symmetric. By definitions of  $W_K(\mathbf{G})$  and  $\mathbf{F}$ , we have

$$W_{K}(\mathbf{G}) = W_{K}(\mathbf{F}) + \int_{K} (\mathbf{G} - \mathbf{F})^{T} \Lambda_{K}^{-1} (\mathbf{G} - \mathbf{F}) \, \mathrm{d}\mathbf{x} - 2 \int_{K} \nabla u \cdot (\mathbf{G} - \mathbf{F}) \, \mathrm{d}\mathbf{x}$$
$$- 2 \int_{K} u \, \mathrm{div}(\mathbf{G} - \mathbf{F}) \, \mathrm{d}\mathbf{x} + 2 \oint_{\partial K} u (\mathbf{G} - \mathbf{F}) \cdot \mathbf{n} \, \mathrm{d}s.$$

Using the Green's formula, we obtain the identity in (2.5) and the last inequality follows from the positive definiteness of the diffusion tensor  $\Lambda_K$ , which completes the proof.

In the following, we shall discretize the cell functional (2.4) in a special manner. Suppose that  $\Omega$  is partitioned into a set of unstructured nonoverlapping polygonal cells. Assume also that each cell edge is on either  $\Gamma_D$  or  $\Gamma_N$ , *i.e.*, Dirichlet and flux boundary conditions are not allowed to appear simultaneously on a single edge. In addition, we introduce some more notations:

- $\mathbf{I}_K = (1, 1, \cdots, 1)^T$  is an  $n_K$ -sized vector;
- $\mathbf{I}_{K,\sigma}$  an  $n_K$ -sized vector whose  $\sigma$ th entry is 1 and the rest ones are all zero;
- $\mathbb{I}_K$  an  $n_K \times n_K$  identity matrix;
- $\mathbb{D}_K = (d_{\sigma\sigma'})$  an  $n_K \times n_K$  diagonal matrix in which  $d_{\sigma\sigma} = 0$  if the  $\sigma$ th edge of K is on  $\Gamma_D$  and  $d_{\sigma\sigma} = 1$  otherwise.

Now we are ready to describe our algorithm that consists of four steps.

Step 1. Introduction of the discrete unknowns.

As depicted in Figure 1, approximation of the solution u at the cell center  $\mathbf{x}_K$  and the edge midpoint  $\mathbf{x}_{K,\sigma}$  are denoted as  $u_K$  and  $u_{K,\sigma}$ , respectively. Since there exists a unique midpoint for an edge and the discrete solution is required to be continuous at this point, we can thus always write

$$u_{K,\sigma} = u_{L,\sigma} = u_{\sigma}$$
, if  $\sigma \in \mathcal{E}_K \cap \mathcal{E}_L$ ;  $u_{K,\sigma} = u_{\sigma}$ , if  $\sigma \in \mathcal{E}_K \cap \mathcal{E}^{ext}$ .

As for the flux, we use the notation

$$F_{K,\sigma} \approx \int_{\sigma} \mathbf{F} \cdot \mathbf{n}_{K,\sigma} \,\mathrm{d}s, \quad \sigma \in \mathcal{E}_K.$$

Step 2. Approximation of the cell functional.

We first approximate the integrals in the cell functional to get

$$\int_{K} u \nabla \cdot \mathbf{F} \, \mathrm{d}\mathbf{x} \approx u_{K} \oint_{\partial K} \mathbf{F} \cdot \mathbf{n} \, \mathrm{d}s \approx u_{K} \sum_{\sigma \in \mathcal{E}_{K}} F_{K,\sigma}, \tag{2.6}$$

$$\oint_{\partial K} u \mathbf{F} \cdot \mathbf{n} \, \mathrm{d}s \approx \sum_{\sigma \in \mathcal{E}_K} u_{K,\sigma} F_{K,\sigma} \tag{2.7}$$

and

$$\int_{K} \mathbf{F}^{T} \Lambda_{K}^{-1} \mathbf{F} \mathrm{d} \mathbf{x} \approx \mathbf{F}_{K}^{T} \mathbb{A}_{K} \mathbf{F}_{K}, \qquad (2.8)$$

where  $\mathbf{F}_K = (F_{K,\sigma}, \sigma \in \mathcal{E}_K)^T$  and  $\mathbb{A}_K$ , called as cell matrix, is an  $n_K \times n_K$  matrix whose entries will be specified in the third section. Then, by putting (2.6), (2.7) and (2.8) together, we obtain the following discrete cell functional

$$W_K(\mathbf{F}) \approx \mathbf{F}_K^T \mathbb{A}_K \mathbf{F}_K - 2\mathbf{F}_K^T (u_K \mathbf{I}_K - \mathbf{U}_K), \qquad (2.9)$$

where  $\mathbf{U}_K = (u_{K,\sigma}, \sigma \in \mathcal{E}_K)^T$ .

Step 3. Establishment of the relation among the unknowns.

Under the condition that  $\mathbb{A}_K$  is symmetric positive definite, the discrete cell functional (2.9) achieves its minimum if and only if

$$\mathbb{A}_K \mathbf{F}_K = u_K \mathbf{I}_K - \mathbf{U}_K \tag{2.10}$$

or equivalently,

$$\mathbf{F}_K = \mathbb{A}_K^{-1}(u_K \mathbf{I}_K - \mathbf{U}_K). \tag{2.11}$$

**Remark 2.2.** Relation (2.10) is obtained by the idea of cell functional minimization. In the rest of paper, (2.10) is the starting point for both the construction of finite volume schemes and theoretical analysis through a certain discrete functional approach whether  $\mathbb{A}_K$  is symmetric or not. A criterion (3.3) for the construction of  $\mathbb{A}_K$  will be derived and allows it to be non-symmetric as shown in Section 3.4. Besides, the analysis in this paper may contribute to any algorithm which has the same local relation as (2.10).

### 2.2. Finite volume scheme based on the cell functional minimization

As can be seen from the discussion in the previous subsection, there are  $2n_K+1$  unknowns that are introduced simultaneously on a single cell, and by (2.10) or (2.11), only  $n_K+1$  of them are independent. Thus, in selecting the independent unknowns for a cell, we have mainly two choices, *i.e.*,

- case (i):  $(u_K, F_{K,\sigma}, \sigma \in \mathcal{E}_K)$ ;
- case (ii):  $(u_K, u_{K,\sigma}, \sigma \in \mathcal{E}_K)$ .

L. YIN ET AL.

Obviously, case (i) leads to a finite volume scheme of mixed type while case (ii) yields a hybrid one. Since case (i) usually involves the solution of a linear system of saddle-point type, we prefer to choose case (ii) in our practice.

The equation corresponding to the cell centered unknown  $u_K$  can be obtained from (2.3) by the standard finite volume spatial discretization,

$$\mathbf{I}_{K}^{T}\mathbf{F}_{K} = \int_{K} f(\mathbf{x}) \,\mathrm{d}\mathbf{x}.$$
(2.12)

By substituting (2.11) into (2.12), we have

$$\mathbf{I}_{K}^{T} \mathbb{A}_{K}^{-1}(u_{K} \mathbf{I}_{K} - \mathbb{D}_{K} \mathbf{U}_{K}) = \mathbf{I}_{K}^{T} \mathbb{A}_{K}^{-1}(\mathbb{I}_{K} - \mathbb{D}_{K})\mathbf{U}_{K} + \int_{K} f(\mathbf{x}) \,\mathrm{d}\mathbf{x}, \quad \forall \ K \in \mathcal{M},$$
(2.13)

where the term relevant to Dirichlet boundary data is moved to the right-hand side. The equation corresponding to an interior edge  $\sigma \in \mathcal{E}_K \cap \mathcal{E}_L$  is obtained by using the continuity of the flux, *i.e.*,  $F_{K,\sigma} + F_{L,\sigma} = 0$ . Specifically, we have

$$-\mathbf{I}_{K,\sigma}^{T} \mathbb{A}_{K}^{-1}(u_{K} \mathbf{I}_{K} - \mathbf{U}_{K}) - \mathbf{I}_{L,\sigma}^{T} \mathbb{A}_{L}^{-1}(u_{L} \mathbf{I}_{L} - \mathbf{U}_{L}) = 0, \quad \sigma \in \mathcal{E}_{K} \cap \mathcal{E}_{L},$$
(2.14)

here a minus sign is added to assure the symmetry of the resulting linear system, which will be made clear in later discussion. Still by moving Dirichlet boundary data to the right-hand side, we can rewrite this edge equation as

$$-\mathbf{I}_{K,\sigma}^{T}\mathbb{A}_{K}^{-1}(u_{K}\mathbf{I}_{K}-\mathbb{D}_{K}\mathbf{U}_{K})-\mathbf{I}_{L,\sigma}^{T}\mathbb{A}_{L}^{-1}(u_{L}\mathbf{I}_{L}-\mathbb{D}_{L}\mathbf{U}_{L})$$
  
$$=-\mathbf{I}_{K,\sigma}^{T}\mathbb{A}_{K}^{-1}(\mathbb{I}_{K}-\mathbb{D}_{K})\mathbf{U}_{K}-\mathbf{I}_{L,\sigma}^{T}\mathbb{A}_{L}^{-1}(\mathbb{I}_{L}-\mathbb{D}_{L})\mathbf{U}_{L}, \quad \sigma \in \mathcal{E}_{K} \cap \mathcal{E}_{L}. \quad (2.15)$$

The edge equation corresponding to the boundary edge on  $\Gamma_N$  can be obtained analogously:

$$-\mathbf{I}_{K,\sigma}^{T} \mathbb{A}_{K}^{-1}(u_{K} \mathbf{I}_{K} - \mathbb{D}_{K} \mathbf{U}_{K}) = -\mathbf{I}_{K,\sigma}^{T} \mathbb{A}_{K}^{-1}(\mathbb{I}_{K} - \mathbb{D}_{K}) \mathbf{U}_{K} - \int_{\sigma} f_{N} \mathrm{d}s, \quad \forall \ \sigma \in \mathcal{E}_{K} \cap \Gamma_{N}.$$
(2.16)

**Lemma 2.3.** Let  $\mathbb{M}$  be the coefficient matrix of the linear system that arises from (2.13), (2.15) and (2.16), and  $\mathbb{U}$  be the solution vector, containing all the cell unknowns and the edge intensity unknowns. Then,

$$\mathbf{U}^{T}\mathbb{M}\mathbf{U} = \sum_{K\in\mathcal{M}} (u_{K}\mathbf{I}_{K} - \mathbb{D}_{K}\mathbf{U}_{K})^{T}\mathbb{A}_{K}^{-1}(u_{K}\mathbf{I}_{K} - \mathbb{D}_{K}\mathbf{U}_{K}).$$
(2.17)

*Proof.* Firstly, multiplying the left-hand side of (2.13) with  $u_K$  and summing over all the cells, we have

$$\sum_{K \in \mathcal{M}} u_K \mathbf{I}_K^T \mathbb{A}_K^{-1} (u_K \mathbf{I}_K - \mathbb{D}_K \mathbf{U}_K).$$
(2.18)

Secondly, by multiplying (2.15) and (2.16) with  $u_{\sigma}$ , summing over all the edges  $\sigma \notin \Gamma_D$  and noting that  $u_{\sigma} = u_{K,\sigma} = u_{L,\sigma}$  when  $\sigma \in \mathcal{E}_K \cap \mathcal{E}_L$  (resp.  $u_{\sigma} = u_{K,\sigma}$  when  $\sigma \in \mathcal{E}_K \cap \Gamma_N$ ), we obtain

$$\sum_{\sigma \in \mathcal{E}_{K} \cap \mathcal{E}_{L} \subset \mathcal{E}^{int}} \left[ -u_{K,\sigma} \mathbf{I}_{K,\sigma}^{T} \mathbb{A}_{K}^{-1} (u_{K} \mathbf{I}_{K} - \mathbb{D}_{K} \mathbf{U}_{K}) - u_{L,\sigma} \mathbf{I}_{L,\sigma}^{T} \mathbb{A}_{L}^{-1} (u_{L} \mathbf{I}_{L} - \mathbb{D}_{L} \mathbf{U}_{L}) \right] + \sum_{\sigma \in \mathcal{E}_{K} \cap \Gamma_{N} \subset \mathcal{E}^{ext}} \left[ -u_{K,\sigma} \mathbf{I}_{K,\sigma}^{T} \mathbb{A}_{K}^{-1} (u_{K} \mathbf{I}_{K} - \mathbb{D}_{K} \mathbf{U}_{K}) \right] = - \sum_{K \in \mathcal{M}} \left[ \left( \sum_{\sigma \in \mathcal{E}_{K} \cap \mathcal{E}^{int}} + \sum_{\sigma \in \mathcal{E}_{K} \cap \Gamma_{N}} \right) u_{K,\sigma} \mathbf{I}_{K,\sigma}^{T} \right] \mathbb{A}_{K}^{-1} (u_{K} \mathbf{I}_{K} - \mathbb{D}_{K} \mathbf{U}_{K}) = - \sum_{K \in \mathcal{M}} (\mathbb{D}_{K} \mathbf{U}_{K})^{T} \mathbb{A}_{K}^{-1} (u_{K} \mathbf{I}_{K} - \mathbb{D}_{K} \mathbf{U}_{K}),$$
(2.19)

where we have used the relation

$$\left(\sum_{\sigma\in\mathcal{E}_K\cap\mathcal{E}^{int}}+\sum_{\sigma\in\mathcal{E}_K\cap\Gamma_N}\right)u_{K,\sigma}\mathbf{I}_{K,\sigma}=\mathbb{D}_K\mathbf{U}_K.$$
(2.20)

Finally, combing (2.18) with (2.19), we reach (2.17) and complete the proof.

**Theorem 2.4.** Let  $\mathbb{M}$  be the coefficient matrix of the linear system that arises from the finite volume equations (2.13), (2.15) and (2.16). If all the cell matrices  $\mathbb{A}_K$  are symmetric positive definite and  $\Gamma_D$  has a nonzero one dimensional measure, then  $\mathbb{M}$  is symmetric positive definite.

*Proof.* Since  $\mathbb{A}_K$  ( $K \in \mathcal{M}$ ) is symmetric positive definite, so is  $\mathbb{A}_K^{-1}$ . Following almost the same procedure in the derivation of (2.17), we have, for any vectors **U** and **V** whose dimensions are equal to the number of intensity unknowns,

$$\mathbf{V}^T \mathbb{M} \mathbf{U} = \mathbf{U}^T \mathbb{M} \mathbf{V},\tag{2.21}$$

which indicates that  $\mathbb{M}$  is symmetric. Besides, (2.17) indicates that  $\mathbb{M}$  is also positive semidefinite. What remains is to prove the nonsingularity of  $\mathbb{M}$ . Suppose that there exists some U such that

$$\mathbb{M}\mathbf{U} = 0$$

It follows from (2.17) and the positive definiteness of  $\mathbb{A}_{K}^{-1}$  that

$$u_K \mathbf{I}_K - \mathbb{D}_K \mathbf{U}_K = 0, \quad \forall \ K \in \mathcal{M}.$$

For  $K \in \mathcal{M}$  and  $\mathcal{E}_K \cap \Gamma_D = \phi$ , it holds that  $\mathbb{D}_K = \mathbb{I}_K$  and consequently,

$$u_K = u_\sigma, \quad \forall \ \sigma \in \mathcal{E}_K.$$

For  $K \in \mathcal{M}$  and  $\mathcal{E}_K \cap \Gamma_D \neq \phi$ , we have

$$u_K = u_\sigma = 0, \ \forall \ \sigma \in \mathcal{E}_K \text{ and } \sigma \notin \Gamma_D$$

Since  $\Gamma_D$  has a non-zero one-dimensional measure, there exists at least one cell K satisfying the above relation. Note also that  $\Omega$  is connected. By summarizing these results, we deduce that  $\mathbf{U} = 0$ , which implies the nonsingularity of  $\mathbb{M}$  and completes the proof.

**Remark 2.5.** When the mesh is a structured quadrilateral one, and the cell matrix is given by formula (3.11) in [14], the cell functional minimization scheme is identical to the hybrid finite volume scheme in [10]. In this case, the symmetry and positive definiteness of  $\mathbb{M}$  can be proved by a technique suggested in Appendix B of [10], where the splitting of  $\mathbb{M}$  into a sum of matrices having Cholesky decompositions was employed and the resulting proof occupied over five pages. By contrast, our method relays on the splitting of the corresponding quadratic form  $\mathbf{U}^T \mathbb{M} \mathbf{U}$ . The present approach not only shortens the proof a great deal, but also provides the possibility to reveal the mechanism how the spectral properties of the cell matrices influence on the stability result and error estimate of the cell functional minimization scheme.

199

## 3. Construction and analysis of the cell matrix $\mathbb{A}_{\mathsf{K}}$

## 3.1. Construction of $\mathbb{A}_K$

For the definition of the cell functional minimization scheme to be complete, it now remains to specify the cell matrix  $\mathbb{A}_K$  in (2.10). Here we follow the practice in [13] and employ the so-called linearity-preserving criterion. More explicitly,  $\mathbb{A}_K$  is selected in a way such that (2.10) is exact for the linear case where

$$u = \alpha_K x + \beta_K y + \gamma_K, \quad \Lambda(\mathbf{x}) = \Lambda_K, \quad \forall \ K \in \mathcal{M},$$
(3.1)

here  $\alpha_K, \beta_K, \gamma_K$  are constants and  $\Lambda_K$  is a 2 × 2 constant symmetric tensor.

Introduce the  $n_K \times 2$  matrices  $\mathscr{F}_K = (f_{k,j})$  and  $\mathscr{U}_K = (u_{\sigma,j})$  whose entries are given by

$$f_{\sigma,j} = -|\sigma|\mathbf{e}_j^T \Lambda_K \mathbf{n}_{K,\sigma} \quad \text{and} \quad u_{\sigma,j} = \mathbf{e}_j^T (\mathbf{x}_K - \mathbf{x}_{K,\sigma}), \tag{3.2}$$

where  $\mathbf{e}_1 = (1,0)^T$ ,  $\mathbf{e}_2 = (0,1)^T$ . By requiring (2.10) to be exact for the linear case (3.1) and through some straightforward calculations, we get

$$\mathbb{A}_K \mathscr{F}_K = \mathscr{U}_K. \tag{3.3}$$

For the structured quadrilateral mesh, a special technique is suggested in [14] to get a unique cell matrix that satisfies (3.3). However, this technique no longer works for the general polygonal mesh. Here, motivated by [3, 4], we choose

$$\mathbb{A}_{K} = \frac{1}{|K|} \mathscr{U}_{K} \Lambda_{K}^{-1} \mathscr{U}_{K}^{T} + \mathscr{C}_{K} \mathscr{D}_{K} \mathscr{C}_{K}^{T}, \qquad (3.4)$$

where  $\mathscr{D}_K$  is an arbitrary  $(n_K - 2) \times (n_K - 2)$  symmetric positive definite matrix, and  $\mathscr{C}_K$  is an  $n_K \times (n_K - 2)$  matrix whose column vectors span the null space of the matrix  $\mathscr{F}_K^T$  so that  $\mathscr{F}_K^T \mathscr{C}_K = 0$ .

**Theorem 3.1.** The cell matrix  $\mathbb{A}_K$  defined by (3.4) satisfies (3.3) and is symmetric positive definite.

*Proof.* Firstly, it can be checked by the divergence theorem that, for any constant vector  $\mathbf{v}$  and any point  $\mathbf{x}_*$ , there holds the identity (see, *e.g.*, Thm. 3.2 in [14]),

$$|K|\mathbf{v} = -\sum_{\sigma \in \mathcal{E}_K} |\sigma| \left(\mathbf{v} \cdot \mathbf{n}_{K,\sigma}\right) (\mathbf{x}_* - \mathbf{x}_{K,\sigma}).$$
(3.5)

By replacing **v** and  $\mathbf{x}_*$  with  $\Lambda_K \mathbf{e}_j$  and  $\mathbf{x}_K$ , respectively, we get

$$|K|\Lambda_K \mathbf{e}_j = \sum_{\sigma \in \mathcal{E}_K} f_{\sigma,j}(\mathbf{x}_K - \mathbf{x}_{K,\sigma}), \ j = 1, 2.$$

Multiplying the above identity with  $\mathbf{e}_i (i = 1, 2)$ , we arrive at

$$\mathscr{U}_K^T \mathscr{F}_K = |K| \Lambda_K. \tag{3.6}$$

This identity indicates that  $\mathscr{C}_K$  is well defined and moreover, the cell matrix  $\mathbb{A}_K$  defined by (3.4) satisfies (3.3).

Secondly,  $\mathbb{A}_K$  is obviously symmetric and positive semidefinite. What remains is to prove the nonsingularity of  $\mathbb{A}_K$ , which can be done by using the same procedure in the proof of Theorem 3.1 in [3] and is omitted here for simplicity.

**Remark 3.2.** With different choices of cell center  $\mathbf{x}_K$  and edge point  $\mathbf{x}_{K,\sigma}$ , we can construct three types of finite volume schemes.



FIGURE 2. Left: a polygonal cell that satisfies (M2); right: an example that (M2) is violated while (H1) still holds.

- Type A.  $\mathbf{x}_K$  is chosen to be the mass center and  $\mathbf{x}_{K,\sigma}$  is the edge midpoint;
- Type B.  $\mathbf{x}_K$  is chosen to be any point in the cell and  $\mathbf{x}_{K,\sigma}$  is the edge midpoint.

Although these schemes are derived through a different approach, they are actually coincide with some existing schemes. As will be seen clearly, schemes in Type A are algebraically equivalent to the mimetic finite difference schemes in [3] while those in Type B are identical to the generalized mimetic finite difference schemes in [5]. Moreover, by introducing a new cell matrix (3.20), we are able to obtain a new type of schemes, *i.e.*,

• Type C.  $\mathbf{x}_K$  is chosen to be any point in the cell and  $\mathbf{x}_{K,\sigma}$  can be any interior point on edge  $\sigma$ .

Compared with those in Type A and B, the cell matrices in Type C are generally non-symmetric. Some new results for Type A and B are given in the next two subsections while Type C will be exploited in Section 3.4.

#### **3.2.** Spectral analysis for $\mathbb{A}_{\mathbf{K}}$ in (3.4)

The spectral results of the cell matrix  $\mathbb{A}_K$  play an important role in the stability analysis and error estimate for the cell functional minimization scheme. We note that the authors in [3] obtained some spectral results for a certain counterpart of  $\mathbb{A}_K$  under a number of geometry assumptions, among which the most important one in the two-dimensional case is

• (M2) There exists a positive number  $\tau_*$ , such that every cell K is star-shaped with respect to every point of a disk with radius  $\tau_* h_K$ .

We recall that the  $\mathbf{x}_K$ -star-shaped condition employed in the related hybrid finite volume method [2, 7] can be viewed as a special case of (**M2**) with  $\tau_* = 0$ . Geometry assumptions of this type are a little stronger and exclude many meshes on which the cell functional minimization scheme performs fairly well. The new ingredient of this section is to study the spectral properties of  $\mathbb{A}_K$  under the following geometry assumption:

• (H1) There exists a positive constant  $\underline{\alpha}$ , such that the measure of K satisfies  $|K| \geq \underline{\alpha} h_K^2$ .

Obviously, (M2) implies (H1) with  $\underline{\alpha} \geq \pi \tau_*^2$  but not vice versa, see Figure 2. Geometry assumption (H1) is not only weaker than (M2) but also natural for the present setting, which will be seen clearly in the subsequential discussion. In the following, we shall adopt, for any real matrix  $\mathbb{A}$ , the matrix norm  $\|\cdot\|$  that is induced by the Euclidean vector norm and equals to the square root of the maximum eigenvalue of  $\mathbb{A}^T \mathbb{A}$ . In this context,  $\|\mathbb{A}\| = \|\mathbb{A}^T\|$  holds since  $\mathbb{A}^T \mathbb{A}$  and  $\mathbb{A}\mathbb{A}^T$  have the same non-zero eigenvalues. First, we give the following result without any geometry assumption on the mesh.

**Lemma 3.3.** For the matrices  $\mathscr{F}_K$  and  $\mathscr{U}_K$  defined by (3.2), we have

$$\|\mathscr{U}_K\| \le \sqrt{n_K} h_K, \quad \|\mathscr{F}_K\| \le \sqrt{n_K} \overline{\kappa} h_K,$$
(3.7)

where  $h_K$  denotes the diameter of K.

*Proof.* Although the proof is similar to that of Lemma 3.2 in [3], we still present it for the sake of completeness. For any vector  $\mathbf{v} = (v_1, v_2)^T \in \mathbb{R}^2$ , we have

$$\|\mathscr{U}_{K}\mathbf{v}\|^{2} = \sum_{\sigma\in\mathscr{E}_{K}} \left(\sum_{j=1}^{2} v_{j} \mathbf{e}_{j} \cdot (\mathbf{x}_{K} - \mathbf{x}_{K,\sigma})\right)^{2} \le \|\mathbf{v}\|^{2} \sum_{\sigma\in\mathscr{E}_{K}} \|\mathbf{x}_{K} - \mathbf{x}_{K,\sigma}\|^{2} \le n_{K}h_{K}^{2} \|\mathbf{v}\|^{2},$$

which leads to the first estimate in (3.7). Now, let  $\widetilde{\mathscr{F}}_{K} = (\widetilde{f}_{\sigma,j})$  be an  $n_{K} \times 2$  matrix, given by

$$\widetilde{f}_{\sigma,j} = -|\sigma|\mathbf{e}_j \cdot \mathbf{n}_{K,\sigma}.$$

From (3.2), we have

$$\mathscr{F}_K = \widetilde{\mathscr{F}}_K \Lambda_K^T = \widetilde{\mathscr{F}}_K \Lambda_K.$$
(3.8)

From this identity and (1.2), we deduce that

$$\|\mathscr{F}_{K}\mathbf{v}\|^{2} \leq \|\Lambda_{K}\mathbf{v}\|^{2} \sum_{\sigma\in\mathcal{E}_{K}}\sum_{j=1}^{2} \left(|\sigma|\mathbf{e}_{j}\cdot\mathbf{n}_{K,\sigma}\right)^{2} = \|\Lambda_{K}\mathbf{v}\|^{2} \sum_{\sigma\in\mathcal{E}_{K}}|\sigma|^{2} \leq n_{K}\overline{\kappa}^{2}h_{K}^{2}\|\mathbf{v}\|^{2}.$$

The second estimate in (3.7) follows immediately and the proof is complete.

Theorem 3.4. Assume that

$$\underline{\lambda} \| \mathscr{C}_K \mathbf{w} \|^2 \le \left\| \mathscr{D}_K^{1/2} \mathscr{C}_K^T \mathscr{C}_K \mathbf{w} \right\|^2 \le \overline{\lambda} \| \mathscr{C}_K \mathbf{w} \|^2, \quad \forall \ \mathbf{w} \in \mathbb{R}^{n_K - 2}.$$
(3.9)

Then, under the geometry assumption (H1), we have

$$\underline{\sigma}_{K} \|\mathbf{v}\|^{2} \leq \left\|\mathbb{A}_{K}^{1/2} \mathbf{v}\right\|^{2} \leq \overline{\sigma}_{K} \|\mathbf{v}\|^{2}, \quad \forall \ \mathbf{v} \in \mathbb{R}^{n_{K}},$$
(3.10)

where

$$\underline{\sigma}_{K} = \min\left\{\frac{1}{2\underline{\lambda}}, \ \frac{\underline{\alpha}^{2}\underline{\lambda}\,\underline{\kappa}^{2}}{n_{K}\overline{\kappa}^{2}(2n_{K} + \underline{\alpha}\,\underline{\lambda}\,\underline{\kappa})}\right\}, \quad \overline{\sigma}_{K} = \overline{\lambda} + \frac{n_{K}}{\underline{\alpha}\,\underline{\kappa}}.$$
(3.11)

*Proof.* Compared with Theorem 3.3 in [3], the upper and lower bounds in (3.10) are updated due to the change of geometry assumption and the replacing of  $\mathbb{M}_E$  in [3] with  $\mathbb{A}_K$  (note that  $\mathbb{M}_E \neq \mathbb{A}_K$ ), however, the proof can be conducted analogously. Let  $\mathbf{v}$  be any vector in  $\mathbb{R}^{n_K}$ . The definition of  $\mathscr{C}_K$  implies that there exist  $\mathbf{v}_1 \in \mathbb{R}^2$  and  $\mathbf{v}_2 \in \mathbb{R}^{n_K-2}$ , such that

$$\mathbf{v} = \mathscr{F}_K \mathbf{v}_1 + \mathscr{C}_K \mathbf{v}_2$$
 and  $\|\mathbf{v}\|^2 = \|\mathscr{F}_K \mathbf{v}_1\|^2 + \|\mathscr{C}_K \mathbf{v}_2\|^2$ .

Then, by (3.9),

$$\left\|\mathscr{D}_{K}^{1/2}\mathscr{C}_{K}^{T}\mathbf{v}\right\|^{2} = \left\|\mathscr{D}_{K}^{1/2}\mathscr{C}_{K}^{T}\mathscr{C}_{K}\mathbf{v}_{2}\right\|^{2} \leq \overline{\lambda} \|\mathscr{C}_{K}\mathbf{v}_{2}\|^{2} \leq \overline{\lambda} \|\mathbf{v}\|^{2}.$$

It follows from (3.4) that

$$\left\|\mathbb{A}_{K}^{1/2}\mathbf{v}\right\|^{2} = \frac{1}{|K|} \left\|\boldsymbol{\Lambda}_{K}^{-1/2}\mathscr{U}_{K}^{T}\mathbf{v}\right\|^{2} + \left\|\mathscr{D}_{K}^{1/2}\mathscr{C}_{K}^{T}\mathbf{v}\right\|^{2} \le \frac{1}{|K|} \left\|\boldsymbol{\Lambda}_{K}^{-1/2}\mathscr{U}_{K}^{T}\mathbf{v}\right\|^{2} + \overline{\lambda} \|\mathbf{v}\|^{2}.$$
(3.12)

From (1.2), (3.7) and the assumption (H1), we see that

$$\frac{1}{|K|} \left\| \Lambda_K^{-1/2} \mathscr{U}_K^T \mathbf{v} \right\|^2 \le \frac{n_K}{\underline{\alpha} \, \underline{\kappa}} \| \mathbf{v} \|^2.$$

Combing this estimate with (3.12) gives the upper bound in (3.10). Now we begin to obtain the lower bound. Starting from the identity in (3.12) and by using (3.6), (3.9) and the definition of  $\mathscr{C}_K$ , we find that

$$\begin{split} \left\|\mathbb{A}_{K}^{1/2}\mathbf{v}\right\|^{2} &= \frac{1}{|K|} \left\||K|\Lambda_{K}^{1/2}\mathbf{v}_{1} + \Lambda_{K}^{-1/2}\mathscr{U}_{K}^{T}\mathscr{C}_{K}\mathbf{v}_{2}\right\|^{2} + \left\|\mathscr{D}_{K}^{1/2}\mathscr{C}_{K}^{T}\mathscr{C}_{K}\mathbf{v}_{2}\right\|^{2} \\ &\geq |K|(1-\varepsilon)\left\|\Lambda_{K}^{1/2}\mathbf{v}_{1}\right\|^{2} + \frac{1}{|K|}\left(1-\frac{1}{\varepsilon}\right)\left\|\Lambda_{K}^{-1/2}\mathscr{U}_{K}^{T}\mathscr{C}_{K}\mathbf{v}_{2}\right\|^{2} + \underline{\lambda}\left\|\mathscr{C}_{K}\mathbf{v}_{2}\right\|^{2}, \end{split}$$

where we have used the inequality

$$\|\mathbf{w}_1 + \mathbf{w}_2\|^2 \ge (1 - \varepsilon) \|\mathbf{w}_1\|^2 + \left(1 - \frac{1}{\varepsilon}\right) \|\mathbf{w}_2\|^2, \quad \varepsilon > 0.$$

Since

$$|K| \left\| \Lambda_K^{1/2} \mathbf{v}_1 \right\|^2 \ge \underline{\kappa} \, \underline{\alpha} h_K^2 \| \mathbf{v}_1 \|^2 \ge \frac{\underline{\alpha} \, \underline{\kappa}}{n_K \overline{\kappa}^2} \, \| \mathscr{F}_K \mathbf{v}_1 \|^2$$

and

$$\frac{1}{|K|} \left\| \Lambda_K^{-1/2} \mathscr{U}_K^T \mathscr{C}_K \mathbf{v}_2 \right\|^2 \le \frac{1}{|K|\underline{\kappa}|} \left\| \mathscr{U}_K^T \mathscr{C}_K \mathbf{v}_2 \right\|^2 \le \frac{n_K h_K^2}{|K|\underline{\kappa}|} \| \mathscr{C}_K \mathbf{v}_2 \|^2 \le \frac{n_K}{\underline{\alpha} |\underline{\kappa}|} \| \mathscr{C}_K \mathbf{v}_2 \|^2,$$

under the condition that  $0 < \varepsilon < 1$ , we can proceed with

$$\left\|\mathbb{A}_{K}^{1/2}\mathbf{v}\right\|^{2} \geq (1-\varepsilon)\frac{\underline{\alpha}\,\underline{\kappa}}{n_{K}\overline{\kappa}^{2}}\left\|\mathscr{F}_{K}\mathbf{v}_{1}\right\|^{2} + \left(1-\frac{1}{\varepsilon}\right)\frac{n_{K}}{\underline{\alpha}\,\underline{\kappa}}\left\|\mathscr{C}_{K}\mathbf{v}_{2}\right\|^{2} + \underline{\lambda}\left\|\mathscr{C}_{K}\mathbf{v}_{2}\right\|^{2}.$$

Finally, by choosing  $\varepsilon = 2n_K/(2n_K + \underline{\alpha} \underline{\lambda} \underline{\kappa})$ , we obtain the desired lower bound in (3.10), which completes the proof.

## 3.3. Direct computation of the inverse of $\mathbb{A}_{\mathsf{K}}$ in (3.4)

As mentioned before, what we prefer in practice is the hybrid case, *i.e.*, case (ii), where  $\mathbb{A}_{K}^{-1}$  must be explicitly known. One straightforward way is to invert the cell matrix  $\mathbb{A}_{K}$  given in (3.4). For the issue of computational cost, the direct computation of  $\mathbb{A}_{K}^{-1}$  is of great interest. We recall that the cell matrix  $\mathbb{A}_{K}$  is a symmetric positive definite matrix which satisfies the linearity preserving condition (3.3) where  $\mathscr{F}_{K}$  and  $\mathscr{U}_{K}$  are given by (3.2) and subjected to (3.6). Recalling that  $\Lambda_{K}$  is symmetric, we have the splitting

$$\mathbb{A}_{K}^{-1} = \frac{1}{|K|} \mathscr{F}_{K} \Lambda_{K}^{-1} \mathscr{F}_{K}^{T} + \mathbb{X}_{K}, \tag{3.13}$$

where  $X_K$  is an  $n_K \times n_K$  symmetric matrix, satisfying

$$\mathbb{X}_K \mathscr{U}_K = 0. \tag{3.14}$$

Now the construction of  $\mathbb{A}_{K}^{-1}$  reduces to the problem of finding a symmetric matrix  $\mathbb{X}_{K}$  such that (3.14) holds and at the same time, the positive definiteness of  $\mathbb{A}_{K}^{-1}$  defined by (3.13) is maintained.

L. YIN ET AL.

Let  $\widetilde{\mathscr{C}}_K$  be an  $n_K \times (n_K - 2)$  matrix whose column vectors span the null space of the matrix  $\mathscr{U}_K^T$  so that  $\mathscr{U}_K^T \widetilde{\mathscr{C}}_K = 0$ . By (3.6),  $\widetilde{\mathscr{C}}_K$  is well defined since  $\mathscr{U}_K^T$  has a full rank 2. We give the following two construction algorithms,

$$\mathbb{X}_{K}^{(1)} = \widetilde{\mathscr{C}}_{K} \widetilde{\mathscr{D}}_{K} \ \widetilde{\mathscr{C}}_{K}^{T} \tag{3.15}$$

and

$$\mathbb{X}_{K}^{(2)} = \left(\mathbb{I}_{K} - \frac{1}{|K|}\mathscr{F}_{K}\Lambda_{K}^{-1}\mathscr{U}_{K}^{T}\right)\widehat{\mathscr{D}}_{K}\left(\mathbb{I}_{K} - \frac{1}{|K|}\mathscr{F}_{K}\Lambda_{K}^{-1}\mathscr{U}_{K}^{T}\right)^{T},$$
(3.16)

where  $\widehat{\mathscr{D}}_K$  (resp.  $\widehat{\mathscr{D}}_K$ ) denotes an arbitrary  $(n_K - 2) \times (n_K - 2)$  (resp.  $n_K \times n_K$ ) symmetric positive definite matrix. It is trivial to check that the matrix  $\mathbb{X}_K$  defined in either way satisfies (3.14) and moreover, assures the symmetric positive definiteness of  $\mathbb{A}_K^{-1}$  defined in (3.13). We remark that the first construction algorithm for  $\mathbb{X}_K$  is inspired by [3] while the second one can be obtained by following the hybrid finite volume approach [5,7], where  $\widehat{\mathscr{D}}_K$  is a certain diagonal matrix in [7] and a general symmetric positive definite matrix in [5]. It appears that formula (3.15) is simpler than (3.16), however, their computational costs do not have so much difference, since the latter only involves  $\mathscr{F}_K$ ,  $\mathscr{U}_K$  and  $\Lambda_K^{-1}$  and the computation of the null space of  $\mathscr{U}_K^T$  is avoided. Moreover, it has been shown in [5] that  $\mathbb{X}_K^{(1)}$  and  $\mathbb{X}_K^{(2)}$  are actually identical if  $\widetilde{\mathscr{D}}_K$  and  $\widehat{\mathscr{D}}_K$  are properly chosen. The key ingredient of the argumentation is the fact that  $\widetilde{\mathscr{C}}_K^T$  and  $(\mathbb{I}_K - \frac{1}{|K|} \mathscr{F}_K \Lambda_K^{-1} \mathscr{U}_K^T)^T$  have the same kernel, which can be easily proved in the present setting.

Another explicit expression for  $\mathbb{A}_{K}^{-1}$  is given by

$$\mathbb{X}_{K}^{(3)} = \left(\mathbb{I}_{K} - \frac{1}{|K|}\mathscr{F}_{K}\Lambda_{K}^{-1}\mathscr{U}_{K}^{T}\right)\mathscr{C}_{K}\widetilde{\mathscr{D}}_{K}\mathscr{C}_{K}^{T}\left(\mathbb{I}_{K} - \frac{1}{|K|}\mathscr{F}_{K}\Lambda_{K}^{-1}\mathscr{U}_{K}^{T}\right)^{T},$$
(3.17)

where  $\mathscr{C}_K$  is defined in (3.4). Although this new expression is a little complicated, it is still subjected to all the constraints on  $\mathbb{X}_K$ . However, we can prove that  $\mathbb{X}_K^{(3)} = \mathbb{X}_K^{(2)}$  since  $(\mathbb{I}_K - \frac{1}{|K|} \mathscr{F}_K \Lambda_K^{-1} \mathscr{U}_K^T)^T$  and  $\mathscr{C}_K^T (\mathbb{I}_K - \frac{1}{|K|} \mathscr{F}_K \Lambda_K^{-1} \mathscr{U}_K^T)^T$  have the same kernel. In fact, suppose that there exists a vector  $\mathbf{v} \in \mathbb{R}^{n_K}$  such that

$$\mathscr{C}_{K}^{T}\left(\mathbb{I}_{K}-\frac{1}{|K|}\mathscr{F}_{K}\Lambda_{K}^{-1}\mathscr{U}_{K}^{T}\right)^{T}\mathbf{v}=0$$

or equivalently,

$$\mathscr{C}_{K}^{T}\left(\mathbf{v}-\frac{1}{|K|}\mathscr{U}_{K}\Lambda_{K}^{-1}\mathscr{F}_{K}^{T}\mathbf{v}\right)=0.$$

By the definition of  $\mathscr{C}_K$ , there exists a vector  $\mathbf{v}_2 \in \mathbb{R}^2$  such that

$$\mathbf{v} - \frac{1}{|K|} \mathscr{U}_K \Lambda_K^{-1} \mathscr{F}_K^T \mathbf{v} = \mathscr{F}_K \mathbf{v}_2.$$
(3.18)

Multiplying both sides with  $\mathscr{F}_K^T$  and using (3.6), we get

$$\mathscr{F}_K^T \mathscr{F}_K \mathbf{v}_2 = 0$$

which implies that  $\mathscr{F}_K \mathbf{v}_2 = 0$ . We then deduce from (3.18) that  $\mathbf{v}$  belongs to the kernel of  $(\mathbb{I}_K - \frac{1}{|K|} \mathscr{F}_K \Lambda_K^{-1} \mathscr{U}_K^T)^T$ .

Here we must point out that, although  $\mathbb{X}_{K}^{(1)}, \mathbb{X}_{K}^{(2)}$  and  $\mathbb{X}_{K}^{(3)}$  are identical if  $\widehat{\mathscr{D}}_{K}$  and  $\widehat{\mathscr{D}}_{K}$  are properly chosen, the numerical performances are not identical in practice, as will be seen in Section 6.

### 3.4. A non-symmetric version of $\mathbb{A}_{\mathsf{K}}$

We have seen that the cell center  $\mathbf{x}_K$  can be any point in K while the edge unknown is defined at the edge midpoint  $\mathbf{x}_{K,\sigma}$ . Actually, under the present framework, we can choose  $\mathbf{x}_{K,\sigma}$  to be any point on the edge  $\sigma$  to obtain a new type of schemes. In this case, it is difficult to get a symmetric matrix  $\mathbb{A}_K$  but a non-symmetric one is possible.

Throughout this subsection, we assume that  $\mathbf{x}_{K,\sigma}$  is an arbitrary point on the edge  $\sigma$  and introduce a  $2 \times 2$  matrix below

$$\mathbb{G}_K = \mathscr{F}_K^T \mathscr{U}_K,\tag{3.19}$$

where  $\mathscr{F}_K$  and  $\mathscr{U}_K$  are given by (3.2). When  $\mathbf{x}_{K,\sigma}$  coincides with the edge midpoint we find from (3.6) that  $\mathbb{G}_K = |K| \Lambda_K^T$ . However, in general case this will not hold and we have to introduce the following assumption

• (A1) For any  $K \in \mathcal{M}$ ,  $\mathbb{G}_K$  is invertible.

Under the above assumption, we have the following construction algorithm

$$\mathbb{A}_{K}^{-1} = \mathscr{F}_{K}\mathbb{G}_{K}^{-1}\mathscr{F}_{K}^{T} + \left(\mathbb{I}_{K} - \mathscr{F}_{K}\mathbb{G}_{K}^{-T}\mathscr{U}_{K}^{T}\right)\bar{\mathscr{D}}_{K}\left(\mathbb{I}_{K} - \mathscr{F}_{K}\mathbb{G}_{K}^{-T}\mathscr{U}_{K}^{T}\right)^{T},$$
(3.20)

where  $\bar{\mathscr{D}}_K$  denotes a generic  $n_K \times n_K$  symmetric positive definite matrix. It is easy to see that the new algorithm satisfies (3.3).

The cell matrix given by (3.20) can be viewed as a further extension of (3.13) with (3.16) in the hybrid finite volume approach [5, 7]. Moreover, the new algorithm is based on assumption (A1) whose verification is not a trivial work for arbitrary polygonal meshes. At the present, we have the result below.

**Theorem 3.5.** For an arbitrary polygonal cell K, if  $\mathbf{x}_{K,\sigma}(\sigma \in \mathcal{E}_K)$  are collinear, then  $\mathbb{G}_K$  defined by (3.19) is singular; For a triangular cell K,  $\mathbb{G}_K$  in (3.19) is singular if and only if  $\mathbf{x}_{K,\sigma}(\sigma \in \mathcal{E}_K)$  are collinear.

*Proof.* From (3.19) and (3.2) and through some direct calculations, we have

$$\mathbb{G}_{K} = -\Lambda_{K}^{T} \sum_{\sigma \in \mathcal{E}_{K}} |\sigma| \mathbf{n}_{K,\sigma} \left( \mathbf{x}_{K} - \mathbf{x}_{K,\sigma} \right)^{T}.$$
(3.21)

Assume that all  $\mathbf{x}_{K,\sigma}(\sigma \in \mathcal{E}_K)$  associated with K are collinear. Then there exists a vector  $\mathbf{v}$ , such that

$$\left(\mathbf{x}_{K,\sigma} - \mathbf{x}_{K,\sigma'}\right)^T \mathbf{v} = 0, \quad \forall \ \sigma, \sigma' \in \mathcal{E}_K.$$
(3.22)

By using (3.21) and the identity  $\sum_{\sigma \in \mathcal{E}_K} |\sigma| \mathbf{n}_{K,\sigma} = \mathbf{0} (\mathbf{0} \in \mathbb{R}^2$  denotes the zero vector), we have

$$\mathbb{G}_{K}\mathbf{v} = -\Lambda_{K}^{T}\sum_{\sigma\in\mathcal{E}_{K}}|\sigma|\left(\mathbf{x}_{K}^{T}\mathbf{v} - \mathbf{x}_{K,\sigma}^{T}\mathbf{v}\right)\mathbf{n}_{K,\sigma} = \Lambda_{K}^{T}\sum_{\sigma\in\mathcal{E}_{K}}|\sigma|\left(\mathbf{x}_{K,\sigma}^{T}\mathbf{v} - \mathbf{x}_{K,\sigma'}^{T}\mathbf{v}\right)\mathbf{n}_{K,\sigma} = \mathbf{0},$$
(3.23)

which implies that  $\mathbb{G}_K$  is singular. Now assume that K is a triangle and  $\mathbb{G}_K$  is singular. Then, there exists a vector **v** such that  $\mathbb{G}_K$ **v** = **0**. From (3.23) and by recalling that  $\Lambda_K$  is symmetric positive definite, we have

$$\sum_{\sigma \in \mathcal{E}_K, \sigma \neq \sigma'} |\sigma| \left( \mathbf{x}_{K,\sigma}^T \mathbf{v} - \mathbf{x}_{K,\sigma'}^T \mathbf{v} \right) \mathbf{n}_{K,\sigma} = \mathbf{0}, \ \forall \sigma' \in \mathcal{E}_K.$$
(3.24)

Since the two vectors  $\mathbf{n}_{K,\sigma}(\sigma \in \mathcal{E}_K, \sigma \neq \sigma')$  are linearly independent, we finally reach (3.22), which implies that  $\mathbf{x}_{K,\sigma}(\sigma \in \mathcal{E}_K)$  are collinear and completes the proof.

The spectral analysis for this new type of  $\mathbb{A}_K$  is not as easy as that of its symmetric counterpart, which is very important for the stability analysis and error estimate. Here, we present a preliminary result.

**Theorem 3.6.** Assume that the symmetric part of  $\mathbb{G}_{K}^{-1}$  is positive definite. Then, for an arbitrary polygonal cell K and the cell matrix  $\mathbb{A}_{K}$  defined in (3.20), we have

$$\mathbf{v}^T \left( \mathbb{A}_K^{-1} + \mathbb{A}_K^{-T} \right) \mathbf{v} > 0, \tag{3.25}$$

where  $\mathbf{v} \in \mathbb{R}^{n_K}$  is an arbitrary non-trivial vector.

*Proof.* Since the symmetric part of  $\mathbb{G}_{K}^{-1}$  is positive definite and  $\overline{\mathscr{D}}_{K}$  is symmetric positive definite, we have from (3.20) that

$$\mathbf{v}^{T}\left(\mathbb{A}_{K}^{-1}+\mathbb{A}_{K}^{-T}\right)\mathbf{v}=\mathbf{v}^{T}\mathscr{F}_{K}\left(\mathbb{G}_{K}^{-1}+\mathbb{G}_{K}^{-T}\right)\mathscr{F}_{K}^{T}\mathbf{v}+2\mathbf{v}^{T}\left(\mathbb{I}_{K}-\mathscr{F}_{K}\mathbb{G}_{K}^{-T}\mathscr{U}_{K}^{T}\right)\bar{\mathscr{D}}_{K}\left(\mathbb{I}_{K}-\mathscr{F}_{K}\mathbb{G}_{K}^{-T}\mathscr{U}_{K}^{T}\right)^{T}\mathbf{v}\geq0.$$

Assume that  $\mathbf{v}^T \left( \mathbb{A}_K^{-1} + \mathbb{A}_K^{-T} \right) \mathbf{v} = 0$  for some  $\mathbf{v}$ , then we have

$$\mathscr{F}_K^T \mathbf{v} = 0$$
 and  $\left(\mathbb{I}_K - \mathscr{F}_K \mathbb{G}_K^{-T} \mathscr{U}_K^T\right)^T \mathbf{v} = 0.$ 

Substituting the first identity into the second one yields  $\mathbf{v} = 0$ . The proof is complete.

In the following two sections, we only provide the numerical analysis results for the case of  $\mathbf{x}_{K,\sigma}$  being the edge midpoint. For the general case, by Theorem 3.6, we need a certain assumption on  $\mathbb{G}_K$  to get the spectral result of cell matrix  $\mathbb{A}_K$ , and the rest proofs of the stability and error estimates are similar.

### 4. The stability result

In this section, we obtain the stability results through a functional approach whose key ingredients are the discrete Poincaré inequality, the discrete trace inequality and the splitting of the quadratic form in (2.17). In this section, we need some more notations and definitions:

- $L(\mathbf{d}, \mathbf{x})$  the semi-line defined by the origin  $\mathbf{x}$  and the fixed direction  $\mathbf{d}$  (a unit vector);
- $\mathbf{y}(\mathbf{d}, \mathbf{x})$  the first point that the semi-line  $L(\mathbf{d}, \mathbf{x})$  meets the boundary  $\partial \Omega$ , and  $\mathbf{y}(\mathbf{d}, \mathbf{x}) \neq \mathbf{x}$  if  $\mathbf{x} \in \partial \Omega$ ;
- $\chi_{\mathbf{d},\sigma}(\mathbf{x})$  a function related to edge  $\sigma$  and the fixed direction  $\mathbf{d}$ , defined on domain  $\overline{\Omega}$  and given by

$$\chi_{\mathbf{d},\sigma}(\mathbf{x}) = \begin{cases} 1, \text{ if } \ \bar{\sigma} \cap [\mathbf{x}, \mathbf{y}(\mathbf{d}, \mathbf{x})] \neq \phi, \\ 0, \text{ otherwise,} \end{cases}$$

where  $[\mathbf{x}, \mathbf{y}(\mathbf{d}, \mathbf{x})]$  stands for the closed line segment connecting  $\mathbf{x}$  and  $\mathbf{y}(\mathbf{d}, \mathbf{x})$ ;

- $d_{K,\sigma}$  the distance between the cell center of K and the cell edge  $\sigma$ ;
- $\mathbb{Q}_K$  an  $n_K \times n_K$  diagonal matrix associated with cell K whose nonzero entry in the  $\sigma$ th row is  $|\sigma|/d_{K,\sigma}$ ;
- $\mathcal{T} = \mathcal{M} \cup \mathcal{E}^{int} \cup (\mathcal{E}^{ext} \cap \Gamma_N)$  the set that contains the cells and edges (not on  $\Gamma_D$ ) of the mesh;
- $X(\mathcal{T})$  the set of discrete functions that are constants corresponding to each element of  $\mathcal{T}$ .

**Definition 4.1** (discrete norms). For  $u_{\mathcal{T}} = \{u_K, u_\sigma\} \in X(\mathcal{T})$ , define the following discrete norms:

• Discrete  $L_2$  norm

$$||u_{\mathcal{T}}||_{0,\mathcal{T}} = \left(\sum_{K \in \mathcal{M}} |K| u_K^2\right)^{1/2}$$

• Discrete trace norm

$$\|u_{\mathcal{T}}\|_{0,\Gamma_N} = \left(\sum_{\sigma\in\Gamma_N} |\sigma| u_{\sigma}^2\right)^{1/2}$$

• Discrete  $H_1$  norm

$$\|u_{\mathcal{T}}\|_{1,\mathcal{T}} = \left(\sum_{K\in\mathcal{M}} \left\|\mathbb{Q}_K^{1/2}\left(u_K\mathbf{I}_K - \mathbb{D}_K\mathbf{U}_K\right)\right\|^2\right)^{1/2},$$

where  $\|\cdot\|$  denotes the Euclidean vector norm.

• Discrete  $H_1$  seminorm

$$|u_{\mathcal{T}}|_{1,\mathcal{T}} = \left(\sum_{K \in \mathcal{M}} \left\| \mathbb{Q}_K^{1/2} (u_K \mathbf{I}_K - \mathbf{U}_K) \right\|^2 \right)^{1/2}.$$

In addition, we introduce the following assumptions:

• (H2) There exist a fixed unit vector **d** and a positive constant  $\overline{L}$ , independent of mesh size h, such that

$$\mathbf{y}(\mathbf{d}, \mathbf{x}) \in \Gamma_D, \quad \forall \ \mathbf{x} \in \Omega \cup \Gamma_N$$

and

$$\sum_{\sigma \in \mathcal{E}^{int}} \chi_{\mathbf{d},\sigma}(\mathbf{x}) (d_{K,\sigma} + d_{L,\sigma}) + \sum_{\sigma \in \mathcal{E}^{ext}} \chi_{\mathbf{d},\sigma}(\mathbf{x}) d_{K,\sigma} \leq \overline{L} \operatorname{diam}(\Omega), \text{ for a.e. } \mathbf{x} \in \Omega \cup \Gamma_N,$$
(4.1)

where diam( $\Omega$ ) is the diameter of  $\Omega$ ,  $\sigma \in \mathcal{E}_K \cap \mathcal{E}_L$  or simply  $\sigma \in \mathcal{E}_K$  when  $\sigma \in \mathcal{E}^{ext}$ .

• (H3) There exists a fixed unit vector **d** and a positive constant  $\overline{\eta}$ , independent of  $\sigma$ , such that

$$\int_{\Gamma_N} \chi_{\mathbf{d},\sigma}(\mathbf{x}) \mathrm{d}s \le \overline{\eta} \ |\sigma|. \tag{4.2}$$

• (H4) There exists positive constants q and  $\overline{q}$ , independent of  $K, \sigma$  and h, such that

$$\underline{q} \leq \frac{|\sigma|}{d_{K,\sigma}} \leq \overline{q} , \quad \forall \sigma \in \mathcal{E}_K \ \forall K \in \mathcal{M}.$$

• (H5) There exists a function  $g \in H^1(\Omega)$ , such that the trace of g on  $\Gamma_D$  is equal to the Dirichlet data  $u_D$ .

The newly introduced (**H2**) can be understood by virtue of Figure 3. The left-hand side of (4.1) is actually the length of the broken line starting from the cell center where **x** is located, connecting alternately the cell centers and edge midpoints in the cells where line segment  $[\mathbf{x}, \mathbf{y}(\mathbf{d}, \mathbf{x})]$  passes through, and ending at the midpoint of the edge on which  $\mathbf{y}(\mathbf{d}, \mathbf{x})$  lies. Obviously, (**H2**) is a weak geometry assumption and it is not an easy work to construct a working mesh to violate it. The left-hand side of (4.2) equals to the length of the part of  $\Gamma_N$  that falls into the support of  $\chi_{\mathbf{d},\sigma}(\mathbf{x})$ , see the left figure in Figure 3. Thus, assumption (**H3**) is actually about the geometry of  $\Omega$ . Compared with geometry assumptions (**H1**)–(**H3**), (**H4**) is the strongest one whose counterparts can be fund in a number of references such as [2,6].

**Lemma 4.2** (discrete Poincaré inequality). Under the assumption (H2), for any  $u_{\mathcal{T}} = \{u_K, u_{\sigma}\} \in X(\mathcal{T})$ , we have

$$\|u_{\mathcal{T}}\|_{0,\mathcal{T}} \le \sqrt{\overline{L}} \operatorname{diam}(\Omega) \|u_{\mathcal{T}}\|_{1,\mathcal{T}}.$$
(4.3)

*Proof.* We recall that in [6] a similar discrete Poincaré inequality was obtained for the so-called admissible finite volume mesh where the line segment between the cell centers of two neighboring cells is orthogonal to their common edge. Here, the argumentation in [6] has to be modified not only to accommodate the non-admissible



FIGURE 3. Left: the shadow part is the support of  $\chi_{\mathbf{d},\sigma}(\mathbf{x})$ ; right: geometric explanation of the left-hand side of (4.1).

mesh but also to include the edge intensity unknowns. The main idea can also be traced back to [16] where the non-admissible meshes were considered but cell edge unknowns were not involved.

By assumption (H2), we have, for a.e.  $\mathbf{x} \in L$ ,

$$|u_L| \leq \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_L \subset \mathcal{E}^{int}} \chi_{\mathbf{d},\sigma}(\mathbf{x}) \left( |u_K - u_{K,\sigma}| + |u_L - u_{L,\sigma}| \right) + \sum_{\sigma \in \mathcal{E}_K \cap \Gamma_D} \chi_{\mathbf{d},\sigma}(\mathbf{x}) |u_K|.$$

Now, by Cauchy–Schwarz inequality and (4.1), we have

$$|u_L|^2 \leq \overline{L} \operatorname{diam}(\Omega) \left[ \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_L \subset \mathcal{E}^{int}} \chi_{\mathbf{d},\sigma}(\mathbf{x}) \left( \frac{|u_K - u_{K,\sigma}|^2}{d_{K,\sigma}} + \frac{|u_L - u_{L,\sigma}|^2}{d_{L,\sigma}} \right) + \sum_{\sigma \in \mathcal{E}_K \cap \Gamma_D} \chi_{\mathbf{d},\sigma}(\mathbf{x}) \frac{|u_K|^2}{d_{K,\sigma}} \right] \cdot$$

Integrating the above inequality over  $\Omega$  gives

$$\begin{split} \sum_{L \in \mathcal{M}} |L| u_L^2 &\leq \overline{L} \operatorname{diam}(\Omega) \left[ \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_L \subset \mathcal{E}^{int}} \left( \frac{|u_K - u_{K,\sigma}|^2}{d_{K,\sigma}} + \frac{|u_L - u_{L,\sigma}|^2}{d_{L,\sigma}} \right) \int_{\Omega} \chi_{\mathbf{d},\sigma}(\mathbf{x}) \mathrm{d}\mathbf{x} \\ &+ \sum_{\sigma \in \mathcal{E}_K \cap \Gamma_D} \frac{|u_K|^2}{d_{K,\sigma}} \int_{\Omega} \chi_{\mathbf{d},\sigma}(\mathbf{x}) \mathrm{d}\Omega \right]. \end{split}$$

Since

$$\int_{\Omega} \chi_{\mathbf{d},\sigma}(\mathbf{x}) \mathrm{d}\Omega \le |\sigma| \mathrm{diam}(\Omega),$$

we conclude that

$$\begin{split} \sum_{L \in \mathcal{M}} |L| u_L^2 &\leq \overline{L} \left[ \operatorname{diam}(\Omega) \right]^2 \left[ \sum_{\sigma \in \mathcal{E}_K \cap \Gamma_D} |\sigma| \frac{|u_K|^2}{d_{K,\sigma}} + \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_L \subset \mathcal{E}^{int}} |\sigma| \left( \frac{|u_K - u_{K,\sigma}|^2}{d_{K,\sigma}} + \frac{|u_L - u_{L,\sigma}|^2}{d_{L,\sigma}} \right) \right] \\ &\leq \overline{L} \left[ \operatorname{diam}(\Omega) \right]^2 \sum_{K \in \mathcal{M}} \left\| \mathbb{Q}_K^{1/2} \left( u_K \mathbf{I}_K - \mathbb{D}_K \mathbf{U}_K \right) \right\|^2, \end{split}$$

which completes the proof.

**Lemma 4.3** (discrete trace inequality). Under assumptions (H2) and (H3), for any  $u_{\mathcal{T}} = \{u_K, u_{\sigma}\} \in X(\mathcal{T})$ , we have

$$\|u_{\mathcal{T}}\|_{0,\Gamma_{N}} \leq \sqrt{\overline{L} \,\overline{\eta} \operatorname{diam}(\Omega)} \,\|u_{\mathcal{T}}\|_{1,\mathcal{T}}.$$
(4.4)

*Proof.* By (H2), for a.e.  $\mathbf{x} \in \sigma' \in \Gamma_N$ , we have,

$$|u_{\sigma'}| \leq \sum_{\sigma \in \mathcal{E}_K \cap \Gamma_N} \chi_{\mathbf{d},\sigma}(\mathbf{x}) |u_{\sigma} - u_K| + \sum_{\sigma \in \mathcal{E}_k \cap \Gamma_D} \chi_{\mathbf{d},\sigma}(\mathbf{x}) |u_K| + \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_L \subset \mathcal{E}^{int}} \chi_{\mathbf{d},\sigma}(\mathbf{x}) \left( |u_K - u_{K,\sigma}| + |u_L - u_{L,\sigma}| \right),$$

where  $\sigma \in \mathcal{E}_K \cap \mathcal{E}_L$  or simply  $\sigma \in \mathcal{E}_K$  when  $\sigma \in \Gamma_D \cup \Gamma_N$ .

Using Cauchy–Schwarz inequality and (4.1), we obtain

$$u_{\sigma'}^{2} \leq \overline{L} \operatorname{diam}(\Omega) \left[ \sum_{\sigma \in \mathcal{E}_{K} \cap \Gamma_{N}} \chi_{\mathbf{d},\sigma}(\mathbf{x}) \frac{|u_{\sigma} - u_{K}|^{2}}{d_{K,\sigma}} + \sum_{\sigma \in \mathcal{E}_{K} \cap \Gamma_{D}} \chi_{\mathbf{d},\sigma}(\mathbf{x}) \frac{|u_{K}|^{2}}{d_{K,\sigma}} \right. \\ \left. + \sum_{\sigma \in \mathcal{E}_{K} \cap \mathcal{E}_{L} \subset \mathcal{E}^{int}} \chi_{\mathbf{d},\sigma}(\mathbf{x}) \left( \frac{|u_{K} - u_{K,\sigma}|^{2}}{d_{K,\sigma}} + \frac{|u_{L} - u_{L,\sigma}|^{2}}{d_{L,\sigma}} \right) \right] \cdot$$

Integrating the above inequality over  $\Gamma_N$  we reach

$$\sum_{\sigma' \in \Gamma_N} |\sigma'| u_{\sigma'}^2 \leq \overline{L} \operatorname{diam}(\Omega) \left[ \sum_{\sigma \in \mathcal{E}_K \cap \Gamma_N} \frac{|u_{\sigma} - u_K|^2}{d_{K,\sigma}} \int_{\Gamma_N} \chi_{\mathbf{d},\sigma}(\mathbf{x}) \mathrm{d}s \right. \\ \left. + \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_L \subset \mathcal{E}^{int}} \left( \frac{|u_K - u_{K,\sigma}|^2}{d_{K,\sigma}} + \frac{|u_L - u_{L,\sigma}|^2}{d_{L,\sigma}} \right) \int_{\Gamma_N} \chi_{\mathbf{d},\sigma}(\mathbf{x}) \mathrm{d}s \right. \\ \left. + \sum_{\sigma \in \mathcal{E}_K \cap \Gamma_D} \frac{|u_K|^2}{d_{K,\sigma}} \int_{\Gamma_N} \chi_{\mathbf{d},\sigma}(\mathbf{x}) \mathrm{d}s \right]$$

and by (**H3**),

$$\sum_{\sigma' \in \Gamma_N} |\sigma'| u_{\sigma}^2 \leq \overline{L} \,\overline{\eta} \operatorname{diam}(\Omega) \left[ \sum_{\sigma \in \mathcal{E}_K \cap \Gamma_N} |\sigma| \frac{|u_{\sigma} - u_K|^2}{d_{K,\sigma}} + \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_L \subset \mathcal{E}^{int}} |\sigma| \left( \frac{|u_K - u_{K,\sigma}|^2}{d_{K,\sigma}} + \frac{|u_L - u_{L,\sigma}|^2}{d_{L,\sigma}} \right) + \sum_{\sigma \in \mathcal{E}_K \cap \Gamma_D} |\sigma| \frac{|u_K|^2}{d_{K,\sigma}} \right]$$
$$\leq \overline{L} \,\overline{\eta} \operatorname{diam}(\Omega) \sum_{K \in \mathcal{M}} \left\| \mathbb{Q}_K^{1/2} \left( u_K \mathbf{I}_K - \mathbb{D}_K \mathbf{U}_K \right) \right\|^2,$$

which leads to (4.4) and completes the proof.

**Theorem 4.4** (stability for the case where  $u_D = 0$ ). Let  $u_T = \{u_K, u_\sigma\} \in X(T)$  with  $u_\sigma = 0$  ( $\sigma \in \Gamma_D$ ) be the solution to system of (2.13), (2.15) and (2.16). Then, under assumptions (H1)–(H4) and  $u_D = 0$ , we have

$$\|u_{\mathcal{T}}\|_{1,\mathcal{T}} \leq \overline{q} \sqrt{\overline{L} \operatorname{diam}(\Omega)} \max_{K \in \mathcal{M}} \overline{\sigma}_{K} \left( \sqrt{\operatorname{diam}(\Omega)} \|f\|_{0,\Omega} + \sqrt{\overline{\eta}} \|f_{N}\|_{0,\Gamma_{N}} \right).$$

$$(4.5)$$

209

*Proof.* Let  $\mathbb{M}$  be the coefficient matrix of the linear system that arises from (2.13), (2.15) and (2.16), and  $\mathbb{U}$  be the solution vector. Multiplying both sides of (2.13), (2.15) and (2.16) with  $u_K$  and  $u_{\sigma}$ , respectively, and summing over all cells and edges  $\sigma \notin \Gamma_D$ , we get

$$\mathbf{U}^{T} \mathbb{M} \mathbf{U} = \sum_{K \in \mathcal{M}} u_{K} \left[ \mathbf{I}_{K}^{T} \mathbb{A}_{K}^{-1} (\mathbb{I}_{K} - \mathbb{D}_{K}) \mathbf{U}_{K} + \int_{K} f d\mathbf{x} \right]$$

$$+ \sum_{\sigma \in \mathcal{E}_{K} \cap \mathcal{E}_{L} \subset \mathcal{E}^{int}} \left[ -u_{K,\sigma} \mathbf{I}_{K,\sigma}^{T} \mathbb{A}_{K}^{-1} (\mathbb{I}_{K} - \mathbb{D}_{K}) \mathbf{U}_{K} - u_{L,\sigma} \mathbf{I}_{L,\sigma}^{T} \mathbb{A}_{L}^{-1} (\mathbb{I}_{L} - \mathbb{D}_{L}) \mathbf{U}_{L} \right]$$

$$+ \sum_{\sigma \in \mathcal{E}_{K} \cap \Gamma_{N}} \left[ -u_{K,\sigma} \mathbf{I}_{K,\sigma}^{T} \mathbb{A}_{K}^{-1} (\mathbb{I}_{K} - \mathbb{D}_{K}) \mathbf{U}_{K} - u_{\sigma} \int_{\sigma} f_{N} ds \right]$$

$$= \sum_{K \in \mathcal{M}, \mathcal{E}_{K} \cap \Gamma_{D} \neq \phi} (u_{K} \mathbf{I}_{K} - \mathbb{D}_{K} \mathbf{U}_{K})^{T} \mathbb{A}_{K}^{-1} (\mathbb{I}_{K} - \mathbb{D}_{K}) \mathbf{U}_{K}$$

$$+ \sum_{K \in \mathcal{M}} u_{K} \int_{K} f d\mathbf{x} - \sum_{\sigma \in \Gamma_{N}} u_{\sigma} \int_{\sigma} f_{N} ds,$$

$$(4.6)$$

where we have used  $u_{\sigma} = u_{K,\sigma} = u_{L,\sigma}$  and (2.20). By (2.17), (3.10), assumptions (H1)–(H4), Cauchy–Schwartz inequality, Lemma 4.2 and Lemma 4.3, we proceed with

$$\begin{split} &\frac{1}{q} \min_{K \in \mathcal{M}} \frac{1}{\overline{\sigma_{K}}} \|u_{\mathcal{T}}\|_{1,\mathcal{T}}^{2} \leq \sum_{K \in \mathcal{M}} \frac{1}{\overline{\sigma_{K}}} \|(u_{K}\mathbf{I}_{K} - \mathbb{D}_{K}\mathbf{U}_{K})\|^{2} \leq \mathbf{U}^{T} \mathbb{M} \mathbf{U} \\ &\leq \max_{K \in \mathcal{M}} \frac{1}{\underline{\sigma_{K}}} \left( \sum_{K \in \mathcal{M}, \mathcal{E}_{K} \cap \Gamma_{D} \neq \phi} \|u_{K}\mathbf{I}_{K} - \mathbb{D}_{K}\mathbf{U}_{K}\|^{2} \right)^{1/2} \left( \sum_{K \in \mathcal{M}, \mathcal{E}_{K} \cap \Gamma_{D} \neq \phi} \|(\mathbb{I}_{K} - \mathbb{D}_{K})\mathbf{U}_{K}\|^{2} \right)^{1/2} \\ &+ \left( \sum_{K \in \mathcal{M}} |K|u_{K}^{2} \right)^{1/2} \left( \sum_{K \in \mathcal{M}} \frac{1}{|K|} \left( \int_{K} f \mathrm{d} \mathbf{x} \right)^{2} \right)^{1/2} + \left( \sum_{\sigma \in \Gamma_{N}} |\sigma|u_{\sigma}^{2} \right)^{1/2} \left( \sum_{\sigma \in \Gamma_{N}} \frac{1}{|\sigma|} \left( \int_{\sigma} f_{N} \mathrm{d} s \right)^{2} \right)^{1/2} \\ &\leq \frac{1}{\sqrt{\underline{q}}} \max_{K \in \mathcal{M}} \frac{1}{\underline{\sigma_{K}}} \|u_{\mathcal{T}}\|_{1,\mathcal{T}} \left( \sum_{K \in \mathcal{M}, \mathcal{E}_{K} \cap \Gamma_{D} \neq \phi} \|(\mathbb{I}_{K} - \mathbb{D}_{K})\mathbf{U}_{K}\|^{2} \right)^{1/2} \\ &+ \sqrt{\underline{L}} \operatorname{diam}(\Omega) \|u_{\mathcal{T}}\|_{1,\mathcal{T}} \left( \sum_{K \in \mathcal{M}} \int_{K} f^{2} \mathrm{d} \mathbf{x} \right)^{1/2} + \sqrt{\underline{L}} \,\overline{\eta} \operatorname{diam}(\Omega) \|u_{\mathcal{T}}\|_{1,\mathcal{T}} \left( \sum_{\sigma \in \Gamma_{N}} \int_{\sigma} f_{N}^{2} \mathrm{d} s \right)^{1/2}. \end{split}$$

It follows that

$$\|u_{\mathcal{T}}\|_{1,\mathcal{T}} \leq \frac{\overline{q}}{\sqrt{\underline{q}}} \max_{K \in \mathcal{M}} \frac{\overline{\sigma}_{K}}{\underline{\sigma}_{K}} \left( \sum_{K \in \mathcal{M}, \mathcal{E}_{K} \cap \Gamma_{D} \neq \phi} \|(\mathbb{I}_{K} - \mathbb{D}_{K})\mathbf{U}_{K}\|^{2} \right)^{1/2} + \overline{q} \max_{K \in \mathcal{M}} \overline{\sigma}_{K} \left( \sqrt{\overline{L}} \operatorname{diam}(\Omega) \|f\|_{0,\Omega} + \sqrt{\overline{L} \,\overline{\eta} \operatorname{diam}(\Omega)} \|f_{N}\|_{0,\Gamma_{N}} \right).$$

$$(4.7)$$

The assumption  $u_D = 0$  implies that

$$\sum_{K \in \mathcal{M}, \mathcal{E}_K \cap \Gamma_D \neq \phi} \| (\mathbb{I}_K - \mathbb{D}_K) \mathbf{U}_K \|^2 = 0.$$

Then, by substituting this result into (4.7) we obtain the desired stability result.

We now consider the case where  $u_D \neq 0$ . Let  $g_{\mathcal{T}} = \{g_K, g_\sigma\} \in X(\mathcal{T})$  be defined by

$$g_{K} = \frac{1}{|K|} \int_{K} g \mathrm{d}\mathbf{x}, \quad \forall K \in \mathcal{M}$$
$$g_{\sigma} = \frac{1}{|\sigma|} \int_{\sigma} g \mathrm{d}s, \quad \forall \sigma \in \mathcal{E},$$

where g is the function defined in (H5). Then we have the following stability result.

**Theorem 4.5** (stability for the case where  $u_D \neq 0$ ). Let  $u_T = \{u_K, u_\sigma\} \in X(T)$  with  $u_\sigma = 1/|\sigma| \int_{\sigma} u_D ds$  $(\sigma \in \Gamma_D)$  be the solution to (2.13), (2.15) and (2.16). Then, under assumptions (H1)-(H5), we have

$$\|u_{\mathcal{T}}\|_{1,\mathcal{T}} \leq \overline{q}\sqrt{\overline{L}\operatorname{diam}(\Omega)} \max_{K \in \mathcal{M}} \overline{\sigma}_{K} \left(\sqrt{\operatorname{diam}(\Omega)} \|f\|_{0,\Omega} + \sqrt{\overline{\eta}} \|f_{N}\|_{0,\Gamma_{N}}\right) + \|g_{\mathcal{T}}\|_{1,\mathcal{T}} + \frac{\overline{q}}{\underline{q}} \max_{K \in \mathcal{M}} \frac{\overline{\sigma}_{K}}{\underline{\sigma}_{K}} |g_{\mathcal{T}}|_{1,\mathcal{T}}.$$
(4.8)

*Proof.* Define  $\tilde{u}_{\mathcal{T}} = u_{\mathcal{T}} - g_{\mathcal{T}} = \{\tilde{u}_K, \tilde{u}_\sigma\} \in X(\mathcal{T})$ . From (2.13), (2.15) and (2.16), we have

$$\mathbf{I}_{K}^{T} \mathbb{A}_{K}^{-1} \left( \tilde{u}_{K} \mathbf{I}_{K} - \mathbb{D}_{K} \tilde{\mathbf{U}}_{K} \right) = \int_{K} f \, \mathrm{d}\mathbf{x} - \mathbf{I}_{K}^{T} \mathbb{A}_{K}^{-1} \left( g_{K} \mathbf{I}_{K} - \mathbf{G}_{K} \right), \tag{4.9}$$

$$-\mathbf{I}_{K,\sigma}^{T}\mathbb{A}_{K}^{-1}\left(\tilde{u}_{K}\mathbf{I}_{K}-\mathbb{D}_{K}\tilde{\mathbf{U}}_{K}\right)-\mathbf{I}_{L,\sigma}^{T}\mathbb{A}_{L}^{-1}\left(\tilde{u}_{L}\mathbf{I}_{L}-\mathbb{D}_{L}\tilde{\mathbf{U}}_{L}\right)=\mathbf{I}_{K,\sigma}^{T}\mathbb{A}_{K}^{-1}\left(g_{K}\mathbf{I}_{K}-\mathbf{G}_{K}\right)+\mathbf{I}_{L,\sigma}^{T}\mathbb{A}_{L}^{-1}\left(g_{L}\mathbf{I}_{L}-\mathbf{G}_{L}\right),$$

$$(4.10)$$

$$-\mathbf{I}_{K,\sigma}^{T}\mathbb{A}_{K}^{-1}\left(\tilde{u}_{K}\mathbf{I}_{K}-\mathbb{D}_{K}\tilde{\mathbf{U}}_{K}\right)=-\int_{\sigma}f_{N}\mathrm{d}s+\mathbf{I}_{K,\sigma}^{T}\mathbb{A}_{K}^{-1}\left(g_{K}\mathbf{I}_{K}-\mathbf{G}_{K}\right),\qquad(4.11)$$

where we have used

$$(\mathbb{I}_K - \mathbb{D}_K) \widetilde{\mathbf{U}}_K = 0, \quad \forall \ K \in \mathcal{M}$$

 $\tilde{\mathbf{U}}_K$  and  $\mathbf{G}_K$  are defined in the same way as  $\mathbf{U}_K$ . Following the derivation of (4.6), we get

$$\tilde{\mathbf{U}}^T \mathbb{M} \tilde{\mathbf{U}} = \sum_{K \in \mathcal{M}} \tilde{u}_K \int_K f \mathrm{d} \mathbf{x} - \sum_{\sigma \in \Gamma_N} \tilde{u}_\sigma \int_\sigma f_N \mathrm{d} s - \sum_{K \in \mathcal{M}} \left( \tilde{u}_K \mathbf{I}_K - \mathbb{D}_K \tilde{\mathbf{U}}_K \right)^T \mathbb{A}_K^{-1} (g_K \mathbf{I}_K - \mathbf{G}_K).$$

Now, performing almost the same derivation in (4.7), we reach

$$\|\tilde{u}_{\mathcal{T}}\|_{1,\mathcal{T}} \leq \overline{q}\sqrt{\overline{L}}\operatorname{diam}(\Omega) \max_{K \in \mathcal{M}} \overline{\sigma}_{K} \left(\sqrt{\operatorname{diam}(\Omega)} \|f\|_{0,\Omega} + \sqrt{\overline{\eta}} \|f_{N}\|_{0,\Gamma_{N}}\right) + \frac{\overline{q}}{\underline{q}} \max_{K \in \mathcal{M}} \frac{\overline{\sigma}_{K}}{\underline{\sigma}_{K}} |g_{\mathcal{T}}|_{1,\mathcal{T}}.$$

Finally, (4.8) is obtained by using the triangle inequality.

**Remark 4.6.** A similar stability result can be directly obtained from (4.7). Actually, by the definition of g and  $g_{\mathcal{T}}$ , we have

$$(\mathbb{I}_K - \mathbb{D}_K)\mathbf{G}_K = \begin{cases} (\mathbb{I}_K - \mathbb{D}_K)\mathbf{U}_K, \text{ if } \mathcal{E}_K \cap \Gamma_D \neq \phi, \\ 0, & \text{otherwise.} \end{cases}$$

Hence,

$$\sum_{K \in \mathcal{M}, \mathcal{E}_K \cap \Gamma_D \neq \phi} \| (\mathbb{I}_K - \mathbb{D}_K) \mathbf{U}_K \|^2 = \sum_{K \in \mathcal{M}} \| (\mathbb{I}_K - \mathbb{D}_K) \mathbf{G}_K \|^2$$
$$\leq \frac{2}{q} \sum_{K \in \mathcal{M}} \left( \left\| \mathbb{Q}_K^{1/2} (g_K \mathbf{I}_K - \mathbb{D}_K \mathbf{G}_K) \right\|^2 + \left\| \mathbb{Q}_K^{1/2} (g_K \mathbf{I}_K - \mathbf{G}_K) \right\|^2 \right)$$
$$= \frac{2}{q} \left( \| g_\mathcal{T} \|_{1,\mathcal{T}}^2 + |g_\mathcal{T}|_{1,\mathcal{T}}^2 \right).$$

L. YIN ET AL.

Substituting this estimate into (4.7) gives

$$\|u_{\mathcal{T}}\|_{1,\mathcal{T}} \leq \frac{\sqrt{2} \,\overline{q}}{\underline{q}} \max_{K \in \mathcal{M}} \frac{\overline{\sigma}_{K}}{\underline{\sigma}_{K}} \left( \|g_{\mathcal{T}}\|_{1,\mathcal{T}} + |g_{\mathcal{T}}|_{1,\mathcal{T}} \right) + \overline{q} \sqrt{\overline{L} \operatorname{diam}(\Omega)} \max_{K \in \mathcal{M}} \overline{\sigma}_{K} \left( \sqrt{\operatorname{diam}(\Omega)} \|f\|_{0,\Omega} + \sqrt{\overline{\eta}} \|f_{N}\|_{0,\Gamma_{N}} \right).$$

$$(4.12)$$

Obviously, this stability result is not so good as (4.8).

### 5. The discrete $H_1$ error estimate

In the following discussion, we derive the discrete  $H_1$  error estimate of the cell functional minimization scheme.

**Definition 5.1.** Let  $u \in H^1(\Omega) \cap C^0(\overline{\Omega})$  be the solution of (1.1), and  $\Pi_{\mathcal{T}} u = (u(\mathbf{x}_K), u(\mathbf{x}_{\sigma})) \in X(\mathcal{T})$  be the interpolation function of u, where  $\mathbf{x}_K$  and  $\mathbf{x}_{\sigma} (= \mathbf{x}_{K,\sigma}$  if  $\sigma \in \mathcal{E}_K)$  denote the cell center and edge midpoint, respectively. Denote by  $\mathbf{R}_K = (r_{K,\sigma}, \sigma \in \mathcal{E}_K)^T$  the consistent error vector of the flux variable and  $r_{K,\sigma}$  is specified by

$$r_{K,\sigma} = \mathbf{I}_{K,\sigma}^T \mathbb{A}_K^{-1}(u(\mathbf{x}_K)\mathbf{I}_K - \mathbf{U}_K^{ext}) + \int_{\sigma} (\Lambda \nabla u) \cdot \mathbf{n}_{K,\sigma} \,\mathrm{d}s,$$
(5.1)

where  $\mathbf{U}_{K}^{ext} = (u(\mathbf{x}_{K,\sigma}), \sigma \in \mathcal{E}_{K})^{T}$ .

**Theorem 5.2.** Let  $u \in H^1(\Omega) \cap C^0(\overline{\Omega})$  be the solution of (1.1). Assume that  $u_{\mathcal{T}} = \{u_K, u_{\sigma}\} \in X(\mathcal{T})$  is the solution to system of equations (2.13), (2.15) and (2.16), where  $u_{\sigma} = u(\mathbf{x}_{\sigma})(\sigma \in \Gamma_D)$  and  $\mathbf{x}_{\sigma}$  denotes the midpoint of edge  $\sigma$ . Then, under assumptions (H1) and (H4),

$$\|\Pi_{\mathcal{T}}u - u_{\mathcal{T}}\|_{1,\mathcal{T}} \le \sqrt{\overline{q}} \max_{K \in \mathcal{M}} \overline{\sigma}_K \left( \sum_{K \in \mathcal{M}} \|\mathbf{R}_K\|^2 \right)^{1/2},$$
(5.2)

where  $\Pi_{\mathcal{T}} u$  and  $\mathbf{R}_K$  are defined in Definition 5.1.

*Proof.* Denote by  $\varepsilon_{\mathcal{T}} = \{\varepsilon_K, \varepsilon_\sigma\} := \Pi_{\mathcal{T}} u - u_{\mathcal{T}} \in X(\mathcal{T})$  the discrete error function and define

$$\mathbf{E}_K = (\varepsilon_{K,\sigma}, \sigma \in \mathcal{E}_K)^T, \ \varepsilon_{K,\sigma} = u(\mathbf{x}_{K,\sigma}) - u_{K,\sigma}.$$

Obviously,

$$(\mathbb{I}_K - \mathbb{D}_K)\mathbf{E}_K = 0, \quad \forall \ K \in \mathcal{M}.$$
(5.3)

Now, rewrite (2.3) for exact solution u,

$$-\sum_{\sigma\in\mathcal{E}_K}\int_{\sigma}(A\nabla u)\cdot\mathbf{n}_{K,\sigma}\,\mathrm{d}s = \int_K f\,\mathrm{d}\mathbf{x}.$$
(5.4)

By using notations  $\mathbf{U}_{K}^{ext}$  and  $\mathbf{R}_{K}$ , we can further rewrite (5.4) as

$$\mathbf{I}_{K}^{T} \mathbb{A}_{K}^{-1}(u(\mathbf{x}_{K})\mathbf{I}_{K} - \mathbb{D}_{K}\mathbf{U}_{K}^{ext}) = \mathbf{I}_{K}^{T} \mathbb{A}_{K}^{-1}(\mathbb{I}_{K} - \mathbb{D}_{K})\mathbf{U}_{K}^{ext} + \int_{K} f \,\mathrm{d}\mathbf{x} + \mathbf{I}_{K}^{T} \mathbf{R}_{K}.$$
(5.5)

Subtracting (2.13) from (5.5) and noting  $\mathbf{E}_K = \mathbf{U}_K^{ext} - \mathbf{U}_K$  and (5.3), we get

$$\mathbf{I}_{K}^{T} \mathbb{A}_{K}^{-1}(\varepsilon_{K} \mathbf{I}_{K} - \mathbb{D}_{K} \mathbf{E}_{K}) = \mathbf{I}_{K}^{T} \mathbf{R}_{K}.$$
(5.6)

Similarly, by using (2.15) and (2.16), respectively, we obtain

$$-\mathbf{I}_{K,\sigma}^{T}\mathbb{A}_{K}^{-1}(\varepsilon_{K}\mathbf{I}_{K}-\mathbb{D}_{K}\mathbf{E}_{K})-\mathbf{I}_{L,\sigma}^{T}\mathbb{A}_{L}^{-1}(\varepsilon_{L}\mathbf{I}_{L}-\mathbb{D}_{L}\mathbf{E}_{L})=-\mathbf{I}_{K,\sigma}^{T}\mathbf{R}_{K}-\mathbf{I}_{L,\sigma}^{T}\mathbf{R}_{L}$$
(5.7)

and

$$-\mathbf{I}_{K,\sigma}^{T}\mathbb{A}_{K}^{-1}(\varepsilon_{K}\mathbf{I}_{K}-\mathbb{D}_{K}\mathbf{E}_{K})=-\mathbf{I}_{K,\sigma}^{T}\mathbf{R}_{K}.$$
(5.8)

Multiplying (5.6), (5.7) and (5.8) with  $\varepsilon_K$  and  $\varepsilon_{\sigma} (= \varepsilon_{K,\sigma} = \varepsilon_{L,\sigma})$ , respectively, and firstly summing up the terms in the left-hand sides,

$$\sum_{K \in \mathcal{M}} \varepsilon_{K} \mathbf{I}_{K}^{T} \mathbb{A}_{K}^{-1} (\varepsilon_{K} \mathbf{I}_{K} - \mathbb{D}_{K} \mathbf{E}_{K}) + \sum_{\sigma \in \Gamma_{N}} \left[ -\varepsilon_{K,\sigma} \mathbf{I}_{K,\sigma}^{T} \mathbb{A}_{K}^{-1} (\varepsilon_{K} \mathbf{I}_{K} - \mathbb{D}_{K} \mathbf{E}_{K}) \right] + \sum_{\sigma \in \mathcal{E}^{int}} \left[ -\varepsilon_{K,\sigma} \mathbf{I}_{K,\sigma}^{T} \mathbb{A}_{K}^{-1} (\varepsilon_{K} \mathbf{I}_{K} - \mathbb{D}_{K} \mathbf{E}_{K}) - \varepsilon_{L,\sigma} \mathbf{I}_{L,\sigma}^{T} \mathbb{A}_{L}^{-1} (\varepsilon_{L} \mathbf{I}_{L} - \mathbb{D}_{K} \mathbf{E}_{L}) \right] = \sum_{K \in \mathcal{M}} (\varepsilon_{K} \mathbf{I}_{K} - \mathbb{D}_{K} \mathbf{E}_{K})^{T} \mathbb{A}_{K}^{-1} (\varepsilon_{K} \mathbf{I}_{K} - \mathbb{D}_{K} \mathbf{E}_{K}), \quad (5.9)$$

where we have used a relation similar to (2.20) on  $\mathbf{E}_K$ . Secondly summing up the right-hand side terms,

$$\sum_{K \in \mathcal{M}} \varepsilon_K \mathbf{I}_K^T \mathbf{R}_K + \sum_{\sigma \in \Gamma_N} \varepsilon_{K,\sigma} \left( -\mathbf{I}_{K,\sigma}^T \mathbf{R}_K \right) + \sum_{\sigma \in \mathcal{E}^{int}} \left( -\varepsilon_{K,\sigma} \mathbf{I}_{K,\sigma}^T \mathbf{R}_K - \varepsilon_{L,\sigma} \mathbf{I}_{L,\sigma}^T \mathbf{R}_L \right) = \sum_{K \in \mathcal{M}} (\varepsilon_K \mathbf{I}_K - \mathbb{D}_K \mathbf{E}_K)^T \mathbf{R}_K.$$
(5.10)

Combining (5.9) with (5.10), and using Theorem 3.4 and Cauchy–Schwarz inequality, we have

$$\sum_{K \in \mathcal{M}} \frac{1}{\overline{\sigma}_{K}} \| \varepsilon_{K} \mathbf{I}_{K} - \mathbb{D}_{K} \mathbf{E}_{K} \|^{2} \leq \sum_{K \in \mathcal{M}} (\varepsilon_{K} \mathbf{I}_{K} - \mathbb{D}_{K} \mathbf{E}_{K})^{T} \mathbb{A}_{K}^{-1} (\varepsilon_{K} \mathbf{I}_{K} - \mathbb{D}_{K} \mathbf{E}_{K}) = \sum_{K \in \mathcal{M}} (\varepsilon_{K} \mathbf{I}_{K} - \mathbb{D}_{K} \mathbf{E}_{K})^{T} \mathbf{R}_{K}$$
$$\leq \left( \sum_{K \in \mathcal{M}} \| \varepsilon_{K} \mathbf{I}_{K} - \mathbb{D}_{K} \mathbf{E}_{K} \|^{2} \right)^{1/2} \left( \sum_{K \in \mathcal{M}} \| \mathbf{R}_{K} \|^{2} \right)^{1/2}.$$
(5.11)

It follows from assumption (H4) and Definition 4.1 that

$$\|\varepsilon_{\mathcal{T}}\|_{1,\mathcal{T}} \leq \sqrt{\overline{q}} \left( \sum_{K \in \mathcal{M}} \|\varepsilon_{K} \mathbf{I}_{K} - \mathbb{D}_{K} \mathbf{E}_{K}\|^{2} \right)^{1/2} \leq \sqrt{\overline{q}} \max_{K \in \mathcal{M}} \overline{\sigma}_{K} \left( \sum_{K \in \mathcal{M}} \|\mathbf{R}_{K}\|^{2} \right)^{1/2},$$
  
es the proof.

which completes the proof.

In practical computation, we usually have to approximate the source term in (2.13) and the flux boundary data in (2.16), respectively, *i.e.*,

$$\mathbf{I}_{K}^{T} \mathbb{A}_{K}^{-1}(u_{K} \mathbf{I}_{K} - \mathbb{D}_{K} \mathbf{U}_{K}) = \mathbf{I}_{K}^{T} \mathbb{A}_{K}^{-1}(\mathbb{I}_{K} - \mathbb{D}_{K})\mathbf{U}_{K} + |K|\Pi_{s}f, \quad \forall K \in \mathcal{M},$$
(5.12)

$$-\mathbf{I}_{K,\sigma}^{T}\mathbb{A}_{K}^{-1}(u_{K}\mathbf{I}_{K}-\mathbb{D}_{K}\mathbf{U}_{K})=-\mathbf{I}_{K,\sigma}^{T}\mathbb{A}_{K}^{-1}(\mathbb{I}_{K}-\mathbb{D}_{K})\mathbf{U}_{K}-|\sigma|\Pi_{f}f_{N}, \quad \forall \ \sigma \in \mathcal{E}^{ext}\cap \Gamma_{N}.$$
(5.13)

where  $\Pi_s$  and  $\Pi_f$  denote certain interpolation operators, for example,

$$\Pi_s f = \Pi_{\mathcal{T}} f = f(\mathbf{x}_K), \ \forall \ K \in \mathcal{M}; \quad \Pi_f f_N = f_N(\mathbf{x}_\sigma), \ \forall \ \sigma \in \mathcal{E}^{ext} \cap \Gamma_N.$$

In this case, we have the result below. The proof can be conducted analogously, except that the discrete Poincaré inequality and discrete trace inequality have to be used.

**Theorem 5.3.** Let  $u \in H^1(\Omega) \cap C^0(\overline{\Omega})$  be the solution of (1.1). Assume that  $u_{\mathcal{T}} = \{u_K, u_{\sigma}\} \in X(\mathcal{T})$  is the solution to system of equations (5.12), (2.15) and (5.13), where  $u_{\sigma} = u(\mathbf{x}_{\sigma})(\sigma \in \Gamma_D)$  and  $\mathbf{x}_{\sigma}$  denotes the midpoint of edge  $\sigma$ . Then, under assumptions (H1)–(H4),

$$\|\Pi_{\mathcal{T}} u - u_{\mathcal{T}}\|_{1,\mathcal{T}} \leq \sqrt{\overline{q}} \max_{K \in \mathcal{M}} \overline{\sigma}_K \left( \sum_{K \in \mathcal{M}} \|\mathbf{R}_K\|^2 \right)^{1/2} + \overline{q} \sqrt{\overline{L} \operatorname{diam}(\Omega)} \max_{K \in \mathcal{M}} \overline{\sigma}_K \left( \sqrt{\operatorname{diam}(\Omega)} \|f - \Pi_s f\|_{0,\Omega} + \sqrt{\overline{\eta}} \|f_N - \Pi_f f_N\|_{0,\Gamma_N} \right).$$
(5.14)

**Theorem 5.4** (error estimate). Let  $u \in C^2(\overline{\Omega})$  be the solution of (1.1). Assume that  $u_{\mathcal{T}} = \{u_K, u_{\sigma}\} \in X(\mathcal{T})$  is the solution to equations (2.13), (2.15) and (2.16), where  $u_{\sigma} = u(\mathbf{x}_{\sigma})(\sigma \in \Gamma_D)$  and  $\mathbf{x}_{\sigma}$  denotes the midpoint of edge  $\sigma$ . Then, under assumptions (H1) and (H4),

$$\|\Pi_{\mathcal{T}} u - u_{\mathcal{T}}\|_{1,\mathcal{T}} \le C |u|_{2,\infty} h, \tag{5.15}$$

where  $h = \max_{K \in \mathcal{M}} h_K$ , and C is a constant independent of h and u. Proof. Since  $u \in C^2(\overline{\Omega})$ , by Taylor expansion,

$$u(\mathbf{x}_K)\mathbf{I}_K - \mathbf{U}_K^{ext} = \mathscr{U}_K \nabla u(\mathbf{x}_K) + \mathbf{R}_K(u), \qquad (5.16)$$

where  $\mathscr{U}_K$  is defined in (3.2) and

$$\|\mathbf{R}_K(u)\| \le 2\sqrt{n_K} \ |u|_{2,\infty} h_K^2$$

By (5.1) and (3.2),

$$\begin{aligned} r_{K,\sigma} &= \mathbf{I}_{K,\sigma}^{T} \mathbb{A}_{K}^{-1} \mathbf{R}_{K}(u) + \mathbf{I}_{K,\sigma}^{T} \mathscr{F}_{K} \nabla u(\mathbf{x}_{K}) + \int_{\sigma} (\Lambda \nabla u) \cdot \mathbf{n}_{K,\sigma} \, \mathrm{d}s \\ &= \mathbf{I}_{K,\sigma}^{T} \mathbb{A}_{K}^{-1} \mathbf{R}_{K}(u) - |\sigma| (\Lambda_{K} \nabla u(\mathbf{x}_{K})) \cdot \mathbf{n}_{K,\sigma} + \int_{\sigma} (\Lambda \nabla u) \cdot \mathbf{n}_{K,\sigma} \, \mathrm{d}s \\ &= \mathbf{I}_{K,\sigma}^{T} \mathbb{A}_{K}^{-1} \mathbf{R}_{K}(u) + \int_{\sigma} [(\Lambda - \Lambda_{K}) \nabla u(\mathbf{x}_{K})] \cdot \mathbf{n}_{K,\sigma} \, \mathrm{d}s + \int_{\sigma} [\Lambda (\nabla u - \nabla u(\mathbf{x}_{K})] \cdot \mathbf{n}_{K,\sigma} \, \mathrm{d}s, \end{aligned}$$

which leads to

$$|r_{K,\sigma}| \leq Ch_K^2$$

The proof is complete.

#### 6. Numerical experiments

In this section, we shall present several examples to validate the theoretical results for the cell functional minimization algorithm (CFM for short), and the notations of the algorithms used in this section are shown in Table 1.

For the scheme CFM-V, edge unknown is defined at the arbitrary point  $\mathbf{x}_{K,\sigma}$  on the edge  $\sigma$ , here we choose  $\mathbf{x}_{K,\sigma}$  to be a randomly perturbation of the midpoint with a magnitude up to 10% of the whole edge.

The solution errors and edge normal flux errors are investigated in the discrete  $L_2$  norms, which are defined by [1, 13]

$$E_u(h) := \left(\sum_{K \in \mathcal{M}} |K| |u(\mathbf{x}_K) - u_K|^2\right)^{1/2},$$
  
$$E_q(h) := \left(\sum_{K \in \mathcal{M}} |K| \sum_{\sigma \in \mathcal{E}_K} \frac{1}{|\sigma|^2} \left| \int_{\sigma} \mathbf{F} \cdot \mathbf{n}_{K,\sigma} \, \mathrm{d}s - F_{K,\sigma} \right|^2 / \sum_{K \in \mathcal{M}} n_K |K| \right)^{1/2}$$

214

TABLE 1. The notations for the various schemes used in the numerical computation.

Notation	Algorithm description
CFM-I	use $\mathbb{A}_K$ given by (3.4), identical to the scheme in [3] if mass center is used
CFM-II	use $\mathbb{A}_{K}^{-1}$ given by (3.13) and (3.15), identical to the scheme in [3] if mass center is used
CFM-III	use $\mathbb{A}_{K}^{-1}$ given by (3.13) and (3.16), identical to the generalized MFD in [5]
CFM-IV	A new symmetric CFM scheme with $\mathbb{A}_{K}^{-1}$ given by (3.13) and (3.17)
CFM-V	A new non-symmetric CFM scheme with $\mathbb{A}_K^{-1}$ given by (3.20)

The rate of convergence can be obtained by a least squares fit on the ones computed on each two successive meshes by the following formula

$$R_{\alpha} = \frac{\log[E_{\alpha}(h_2)/E_{\alpha}(h_1)]}{\log(h_2/h_1)}, \quad \alpha = u, q,$$

where  $h_1, h_2$  denote the mesh sizes of the two successive meshes, and  $E_{\alpha}(h_1), E_{\alpha}(h_2)$  the corresponding  $L_2$  errors of the solution or edge normal flux.

For the discrete  $H_1$  error estimate (5.15), the solution errors in the  $H_1$ -norm and its convergence rates are denoted by  $E_u^{H_1}(h)$  and  $R_u^{H_1}$ , respectively. We also use the following notation

$$\texttt{ratioH1} = \frac{\|\Pi_T u - u_T\|_{1,T}}{\sqrt{\overline{q}} \max_{K \in \mathcal{M}} \overline{\sigma}_K |u|_{2,\infty} h}$$

We use GMRES method [12] to solve the linear systems in all experiments. Throughout, we shall choose case (ii) for unknowns, employ the geometric center (whose coordinates are the simple average of those of the cell vertices) as the cell center, and  $\mathscr{D}_K$  in (3.4),  $\widetilde{\mathscr{D}}_K$  in (3.15) and (3.17),  $\hat{\mathscr{D}}_K$  in (3.16) and  $\overline{\mathscr{D}}_K$  in (3.20) are always chosen to be identity matrices if not specified.

#### 6.1. Test 1: mild anisotropy

We consider the linear diffusion equation with full Dirichlet boundary condition and  $\Omega = [0, 1]^2$ . A homogeneous anisotropic tensor and the exact solution are given below:

$$\Lambda = \begin{pmatrix} 1.5 & 0.5 \\ 0.5 & 1.5 \end{pmatrix}, \quad u(x, y) = 16x(1-x)y(1-y),$$

where the exact solution is located in the interval [0, 1]. This numerical test is classical, and can be found in [8] as a benchmark with a slight modification for the exact solution.

We use a sequence of the uniform square mesh (Mesh1), the strictly acute triangular mesh (Mesh2), two distorted quadrilateral meshes (Mesh3 and Mesh4) which can be found in [8], randomly perturbed quadrilateral mesh (Mesh5) and a jigsaw puzzle mesh (Mesh6) in this numerical test (see Fig. 4), and each mesh was used with 5 successive mesh levels. Numbers of unknowns on each mesh level for six mesh types are given in Table 2, and Table 3 shows the mesh size of each mesh level.

Figure 4f shows a jigsaw puzzle mesh, which is obtained from a uniform square mesh by modifying the horizontal interior cell edges. More explicitly, we first divide a horizontal interior cell edge into three equal parts and then cast down the middle part to reach the square center. Since most of the cells are not star shape ones, the assumption (M2) no longer holds and as a result, the corresponding theoretical results based on this geometry assumption are spoiled. By contrast, the geometry assumption (H1) holds very well with  $\alpha = 36/97$  and the theoretical results in this paper can still be guaranteed, which is confirmed by the numerical results presented in Figure 5. On the jigsaw puzzle mesh (Mesh6), the convergence rates for the solution and edge



FIGURE 4. Samples of the six mesh types used for simulations: each mesh was used with 5 successive mesh levels.

Mesh	$h_1$	$h_2$	$h_3$	$h_4$	$h_5$
Mesh1	208	800	3136	12416	49408
Mesh2	148	576	2272	9024	35968
Mesh3	901	3536	7905	14008	21845
Mesh4	3333	13200	29601	52536	82005
Mesh5	208	800	3136	12416	49408
Mesh6	432	1760	7104	28544	114432

TABLE 2. Numbers of unknowns on five mesh levels for Mesh1–Mesh6.

TABLE 3. Mesh size on five mesh levels for Mesh1–Mesh6.

Mesh	$h_1$	$h_2$	$h_3$	$h_4$	$h_5$
Mesh1	$1.77 \times 10^{-1}$	$8.84 \times 10^{-2}$	$4.42 \times 10^{-2}$	$2.21 \times 10^{-2}$	$1.10 \times 10^{-2}$
Mesh2	$2.50 \times 10^{-1}$	$1.25 \times 10^{-1}$	$6.25 \times 10^{-2}$	$3.13 \times 10^{-2}$	$1.56 \times 10^{-2}$
Mesh3	$3.29 \times 10^{-1}$	$1.67 \times 10^{-1}$	$1.12 \times 10^{-1}$	$8.39 \times 10^{-2}$	$6.72 \times 10^{-2}$
Mesh4	$1.70 \times 10^{-1}$	$8.52 \times 10^{-2}$	$5.69 \times 10^{-2}$	$4.27 \times 10^{-2}$	$3.42 \times 10^{-2}$
Mesh5	$2.85 \times 10^{-1}$	$1.39 \times 10^{-1}$	$7.39 \times 10^{-2}$	$3.82 \times 10^{-2}$	$1.91 \times 10^{-2}$
Mesh6	$2.05\times10^{-1}$	$1.03 \times 10^{-1}$	$5.13 \times 10^{-2}$	$2.26\times 10^{-2}$	$1.28\times 10^{-2}$

normal flux errors are graphically depicted in Figure 5 as log-log plots of the discrete  $L_2$  norm errors versus the characteristic mesh size h. The actual convergence order is reflected by the slopes of the experimental error curves.

In Table 4, we give the comparison of five schemes CFM-I, CFM-II, CFM-III, CFM-IV and CFM-V on the six mesh types Mesh1–Mesh6, and it should be noted that  $\mathscr{D}_K$ ,  $\widetilde{\mathscr{D}}_K$ ,  $\widehat{\mathscr{D}}_K$  and  $\overline{\mathscr{D}}_K$  are chosen to be identity matrices except that for the scheme CFM-IV on Mesh6, we can not obtain the results in the case that  $\widetilde{\mathscr{D}}_K$  is an identity matrix, so we set the diagonal elements of  $\widetilde{\mathscr{D}}_K$  to be 2 and the others to be 1. Table 4 shows the following:

- The discrete  $H_1$  errors defined in Definition 4.1 and the ratio ratioH1 for  $H_1$  estimate on the finest mesh with mesh levels  $h = h_5$  are given in the third and fourth columns, respectively. One can see that the  $H_1$  error bound is less than or equal to 1 numerically, and then theoretical  $H_1$ -error estimate (5.15) is confirmed in this test.
- The convergence rates of discrete  $H_1$  estimates are given in the fifth columns, and all schemes have first order convergence rate on six meshes Mesh1–Mesh6.
- The convergence rates of the five schemes are about  $h^{1.7-2.0}$  with respect to the  $L_2$  norm of the solution on the six types of meshes.
- All schemes have first order convergence rate with respect to the  $L_2$  norm of the edge normal flux.

#### 6.2. Test 2: strong anisotropy

In this test, diffusion equation (1.1a) with the full Dirichlet boundary condition (1.1b) is defined on a unit square domain  $\Omega = [0, 1]^2$  with the anisotropic and heterogeneous diffusion tensor field. The tensor coefficient is a rotating anisotropic tensor:

$$\Lambda = \begin{pmatrix} \alpha(x-x_0)^2 + (y-y_0)^2 & (\alpha-1)(x-x_0)(y-y_0) \\ (\alpha-1)(x-x_0)(y-y_0) & (x-x_0)^2 + \alpha(y-y_0)^2 \end{pmatrix}$$

and we consider the following exact solution in this test

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

#### L. YIN ET AL.

Mesh	Scheme	$E_u^{H_1}(h_5)$	ratioH1	$R_u^{H_1}$	$R_u$	$R_q$
Mesh1	CFM-I	$1.87 \times 10^{-1}$	0.999912	0.994	1.987	1.875
	CFM-II	$3.55 \times 10^{-2}$	1.000007	0.998	1.995	1.973
	CFM-III	$8.59 \times 10^{-2}$	0.999974	0.996	1.992	1.935
	CFM-IV	$3.55 \times 10^{-2}$	1.000007	0.998	1.995	1.973
	CFM-V	$8.60\times10^{-2}$	$6.92 \times 10^{-1}$	0.997	2.019	1.353
	CFM-I	$4.74 \times 10^{-1}$	$7.68 \times 10^{-1}$	1.000	2.001	0.939
Moch9	CFM-II	$4.52\times 10^{-2}$	$7.10 \times 10^{-1}$	1.001	2.014	0.939
Wiesh2	CFM-III	$1.50 \times 10^{-1}$	$7.68 \times 10^{-1}$	1.000	2.004	0.939
	CFM-IV	$4.52 \times 10^{-2}$	$7.10 \times 10^{-1}$	1.002	2.014	0.939
	CFM-V	$1.51 \times 10^{-1}$	$7.80 \times 10^{-2}$	1.001	2.010	1.012
	CFM-I	$4.68 \times 10^{-1}$	$6.65\times10^{-2}$	0.904	1.805	1.654
Mach 9	CFM-II	$1.20 \times 10^{-1}$	$1.17 \times 10^{-2}$	0.984	1.981	1.856
Mesno	CFM-III	$2.29 \times 10^{-1}$	$3.13 \times 10^{-2}$	0.928	1.860	1.741
	CFM-IV	$1.20 \times 10^{-1}$	$1.16 \times 10^{-2}$	0.980	1.977	1.845
	CFM-V	$2.31 \times 10^{-1}$	$9.97 \times 10^{-1}$	0.927	1.859	1.651
	CFM-I	$2.51 \times 10^{-1}$	$1.11 \times 10^{-1}$	0.952	1.905	1.843
Moch4	CFM-II	$6.36 \times 10^{-2}$	$1.87 \times 10^{-2}$	0.990	1.992	1.942
WIE8114	CFM-III	$1.22 \times 10^{-1}$	$5.16 \times 10^{-2}$	0.967	1.937	1.893
	CFM-IV	$6.36 \times 10^{-2}$	$1.86 \times 10^{-2}$	0.988	1.990	1.944
	CFM-V	$9.51 \times 10^{-2}$	$9.02 \times 10^{-1}$	0.972	1.985	1.671
	CFM-I	$2.11 \times 10^{-1}$	$6.34 \times 10^{-7}$	1.008	2.038	1.049
Moch 5	CFM-II	$4.51 \times 10^{-2}$	$7.78 \times 10^{-2}$	1.007	2.025	1.058
Mesho	CFM-III	$9.75 \times 10^{-2}$	$1.54 \times 10^{-1}$	1.014	2.042	1.104
	CFM-IV	$4.37 \times 10^{-2}$	$6.96 \times 10^{-5}$	1.014	1.695	0.942
	CFM-V	$9.77 \times 10^{-2}$	$1.76 \times 10^{-1}$	1.015	2.067	1.082
	CFM-I	$8.29 \times 10^{-2}$	$6.32 \times 10^{-2}$	0.978	1.986	0.984
Mesh6	CFM-II	$6.61 \times 10^{-2}$	$8.96 \times 10^{-2}$	0.998	2.037	0.966
	CFM-III	$6.83 \times 10^{-2}$	$2.36 \times 10^{-1}$	0.996	2.015	1.113
	CFM-IV	$5.94 \times 10^{-2}$	$3.76 \times 10^{-2}$	0.997	2.033	0.949
	CFM-V	$6.84 \times 10^{-2}$	$1.85 \times 10^{-1}$	0.996	2.025	1.114
-2		CFM-I CFM-II CFM-III CFM-IV CFM-V CFM-V d order				CFM-I CFM-II CFM-III CFM-IV CFM-V

TABLE 4. Test 1:  $H_1$  solution error  $E_u^{H_1}(h_5)$  and ratio ratioH1 on the finest mesh with mesh levels  $h = h_5$ , and the convergence rates for the solution and edge normal flux errors on various meshes.



FIGURE 5. Test 1:  $L_2$  errors versus mesh size h on the jigsaw puzzle mesh (Mesh6).

TABLE 5. Test 2: solution behaviors on the uniform square meshes (Mesh1) and jigsaw puzzle mesh (Mesh6).

	Mesh1					Mesh6					
$\alpha$	Scheme	umin	umax	$R_u^{H_1}$	$R_u$	$R_q$	umin	umax	$R_u^{H_1}$	$R_u$	$R_q$
	CFM-I	0.000	1.080	0.995	1.990	1.055	$-2.10 \times 10^{-3}$	1.057	0.980	2.041	0.999
	CFM-II	0.000	0.980	1.000	1.999	1.008	0.000	1.052	1.011	2.082	0.958
1	CFM-III	0.000	1.010	0.997	1.994	1.018	0.000	1.053	1.008	2.071	1.035
	CFM-IV	0.000	0.980	1.000	1.999	1.008	0.000	0.964	1.061	1.804	0.943
	CFM-V	0.000	1.017	1.000	2.034	1.022	0.000	1.062	1.009	2.076	1.032
	CFM-I	0.000	0.997	0.999	1.996	1.010	$-2.49 \times 10^{-2}$	0.993	1.009	1.732	0.983
	CFM-II	0.000	0.977	0.999	1.995	1.009	0.000	0.972	1.051	1.752	0.954
$10^{-2}$	CFM-III	0.000	0.981	0.999	1.996	1.009	0.000	0.976	1.042	1.745	0.993
	CFM-IV	0.000	0.977	0.999	1.995	1.009	0.000	0.850	1.157	1.782	0.918
	CFM-V	0.000	1.015	0.915	1.777	1.007	0.000	0.982	1.035	1.719	0.989
$10^{-4}$	CFM-I	0.000	0.998	1.000	1.994	1.009	$-2.62 \times 10^{-2}$	0.991	1.004	1.664	0.984
	CFM-II	0.000	0.977	0.999	1.995	1.009	0.000	0.969	1.047	1.709	0.955
	CFM-III	0.000	0.982	0.999	1.995	1.009	0.000	0.974	1.036	1.692	0.993
	CFM-IV	0.000	0.977	0.999	1.995	1.009	0.000	0.847	1.158	1.775	0.917
	CFM-V	0.000	0.985	0.793	1.807	1.009	0.000	0.980	1.030	1.668	0.989

The eigenvalues of  $\Lambda$  are  $\lambda_1(x, y) = \alpha[(x - x_0)^2 + (y - y_0)^2]$  and  $\lambda_2(x, y) = (x - x_0)^2 + (y - y_0)^2$ . The anisotropy ratio is therefore  $\frac{1}{\alpha}$  in the whole domain. In this test we choose  $(x_0, y_0) = (-0.1, -0.1)$ , we have  $\inf_{(x,y)\in\Omega}\lambda_1 = 0.02\alpha$  and  $\sup_{(x,y)\in\Omega}\lambda_1 = 2.42\alpha$  (similar relations hold for  $\lambda_2$ ) which corresponds to a highly heterogeneous case from both the point of view of the magnitude of the ratio of the eigenvalues and that of the directions of anisotropy.

Table 5 reports the minimum and maximum solutions, the convergence rates of  $H_1$  estimates, solution errors and edge normal flux errors for  $\alpha = 1, 10^{-2}, 10^{-4}$  on the uniform square meshes (Mesh1) and jigsaw puzzle mesh (Mesh6). It should be noted that  $\mathscr{D}_K$ ,  $\widehat{\mathscr{D}}_K$ ,  $\widehat{\mathscr{D}}_K$  and  $\overline{\mathscr{D}}_K$  are chosen to be identity matrices, and for the scheme CFM-IV on Mesh6, we can not obtain the results in this case, here we choose the diagonal elements of  $\widehat{\mathscr{D}}_K$  to be 4 and the others to be 1. The results in this table reveal the following:

- The values of the numerical solution should be within the range [0, 1]. The solutions on Mesh1 satisfy discrete extremum principle for various  $\alpha$  except that CFM-I, CFM-III and CFM-V for  $\alpha = 1$  and CFM-V for  $\alpha = 10^{-2}$ . On Mesh6 for  $\alpha = 1$ , only scheme CFM-IV satisfies discrete extremum principle. Solutions for schemes CFM-II, CFM-III, CFM-IV and CFM-V on Mesh6 are located in [0, 1] for  $\alpha = 10^{-2}$ ,  $10^{-4}$ .
- The expected convergence rates of  $H_1$  estimates, solution errors and edge normal flux errors for  $\alpha = 1, 10^{-2}, 10^{-4}$  on the uniform square meshes (Mesh1) and jigsaw puzzle mesh (Mesh6) are obtained in this test.

**Remark 6.1.** If we choose  $(x_0, y_0) = (0, 0)$  in the definition of diffusion tensor, similar results can be obtained. However, in this case one has  $\inf_{(x,y)\in\Omega} \lambda_1 = 0$ , no theoretical estimate may be obtained in this case.

## 7. Conclusion

In this paper, we have constructed a new finite volume scheme through a cell functional minimization approach for the anisotropic diffusion problem on general polygonal meshes. This scheme has a local stencil, allow arbitrary diffusion tensors, yields a symmetric positive definite diffusion matrix in case that edge unknowns are defined at the midpoints of edges, and is linearity-preserving, *i.e.*, preserves linear solutions.

Moreover, we have proved the stability of the new scheme through a discrete functional approach. The discrete  $H_1$  error estimate for the new scheme is also given. One of the key elements of our methodology is the very weak

geometry assumption (H1). All the theoretical results have been validated in numerical examples on various meshes, and all the numerical results show the good performance of the proposed scheme (quadratic convergence rate for the approximate solution and higher than first order accuracy for the discrete flux).

Acknowledgements. The authors want to thank the referees whose remarks have improved this paper.

#### References

- I. Aavatsmark, G.T. Eigestad, B.T. Mallison and J.M. Nordbotten, A compact multipoint flux approximation method with improved robustness. *Numer. Methods Partial Differ. Eq.* 24 (2008) 1329–1360.
- [2] O. Angelint, C. Chavant, E. Chénier and R. Eymard, A finite volume scheme for diffusion problem on general meshes applying monitry constraints. SIAM J. Numer. Anal. 47 (2010) 4193–4213.
- [3] F. Brezzi, K. Lipnikov and V. Simoncini, A family of mimetic finite difference methods on polygonal and polyhedral meshes. Math. Model. Methods. Appl. Sci. 15 (2005) 1533–1551.
- [4] F. Brezzi, K. Lipnikov, M. Shashkov and V. Simoncini, A new discretization methodology for diffusion problems on generalized polyhedral meshes. *Comput. Methods Appl. Mech. Engrg.* 196 (2007) 3682–3692.
- [5] J. Droniou, R. Eymard, T. Gallouët and R. Herbin, A unified approach to mimetic finite difference, hybrid finite volume and mixed finite volume methods. Math. Model. Methods. Appl. Sci. 20 (2010) 265–295.
- [6] R. Eymard, T. Gallouët and R. Herbin, Finite volume methods. Handb. Numer. Anal. Elsevier Sciences (2000).
- [7] R. Eymard, T. Gallouët and R. Herbin, A new finite volume scheme for anisotropic diffusion problems on general grids: Convergence analysis. C. R. Math. Acad. Sci. Paris 344 (2007) 403–406.
- [8] R. Herbin and F. Hubert, Benchmark on discretization schemes for anisotropic diffusion problems on general grids. In *Finite volumes for complex applications*. V. ISTE, London (2008) 659–692.
- [9] D. Li and G. Chen, Introduction to difference methods for parabolic equation. Beijing, Science Press (1995) (Chinese).
- [10] J.E. Morel, R.M. Roberts and M.J. Shashkov, A local support-operators diffusion discretization scheme for quadrilateral r-z meshes. J. Comput. Phys. 144 (1998) 17–51.
- [11] J.E. Roberts and J.-M.Thomas, Mixed and hybrid finite element methods. Handb. Numer. Anal. Elsevier Sciences (1987).
- [12] Y. Saad, Iterative method for sparse linear systems. PWS publishing, New York (1996).
- [13] J. Wu, Z. Dai, Z. Gao and G. Yuan, Linearity preserving nine-point schemes for diffusion equation on distorted quadrilateral meshes, J. Comput. Phys. 229 (2010) 3382–3401.
- [14] L. Yin, J. Wu and Y. Yao, A cell functional minimization scheme for parabolic problem. J. Comput. Phys. 229 (2010) 8935–8951.
- [15] L. Yin, J. Wu and Y. Yao, A cell functional minimization scheme for domain decomposition method on non-orthogonal and non-matching meshes. *Numer. Math.* **128** (2014) 773–804.
- [16] Y. Coudiére, J.-P. Vila and P. Villedieu, Convergence rate of a finite volume scheme for a two-dimensional diffusion convection problem. Math. Model. Numer. Anal. 33 (1999) 493–516.