Mathematics

http://www.aimspress.com/journal/Math

AIMS Mathematics, 8(10): 22507–22537.
DOI: 10.3934/math.20231147
Received: 27 May 2023
Revised: 02 July 2023
Accepted: 05 July 2023
Published: 14 July 2023

Research article

An analysis of the isoparametric bilinear finite volume element method by applying the Simpson rule to quadrilateral meshes

Shengying Mu[1] **and Yanhui Zhou**[2,*]

[1] School of Mathematics and Statistics, Northeast Normal University, Changchun 130024, China
[2] School of Mathematics and Systems Science, Guangdong Polytechnic Normal University, Guangzhou 510665, China

* **Correspondence:** Email: zhouyh9@mail2.sysu.edu.cn.

Abstract: In this work, we construct and study a special isoparametric bilinear finite volume element scheme for solving anisotropic diffusion problems on general convex quadrilateral meshes. The new scheme is obtained by employing the Simpson rule to approximate the line integrals in the classical isoparametric bilinear finite volume element method. By using the cell analysis approach, we suggest a sufficient condition to ensure the coercivity of the new scheme. The sufficient condition has an analytic expression, which only involves the anisotropic diffusion tensor and the geometry of quadrilateral mesh. This yields that for any diffusion tensor and quadrilateral mesh, we can directly judge whether this sufficient condition is satisfied. Specifically, this condition covers the traditional $h^{1+\gamma}$-parallelogram and some trapezoidal meshes with any full anisotropic diffusion tensor. An optimal H^1 error estimate of the proposed scheme is also obtained for a quasi-parallelogram mesh. The theoretical results are verified by some numerical experiments.

Keywords: isoparametric bilinear FVEM; Simpson rule; coercivity result; optimal H^1 error estimate; anisotropic diffusion problem
Mathematics Subject Classification: 65N08, 65N12

1. Introduction

Diffusion equations have a wide range of applications, such as radiation hydrodynamics and reservoir simulation. It is a challenging task to find the true solution of a real problem, while it is easy and effective to find the numerical solution. In the past decades, there have been many numerical methods for solving this kind of problem, such as the finite difference method, finite element method (FEM) and finite volume method (FVM). The FVM is easy to implement, and it can also deal with domains with complex geometries. More interestingly, it is locally conserved for physical quantities;

thus, the FVM has now become one of the most widely used numerical methods for solving partial differential equations.

This paper focuses on a special type of FVM, i.e., the finite volume element method (FVEM), which is also called the generalized difference method [1], box method [2] or covolume method [3]. The FVEM has been studied by many researchers [4–7]; also, see the book [8] and the review papers [9,10] for example references. The linear FVEM (P_1-FVEM) is closed to the linear FEM on triangular mesh. Since its cell stiffness matrix coincides with the linear FEM for Poisson equations, the coercivity result then follows [2,4,11] and also leads to an optimal H^1 error estimate. However, the optimal L^2 error analysis depends on the regularity of source term f additionally [12,13], which is different from the FEM. That is, we cannot obtain the optimal convergence rate of two if we only assume that the exact solution $u \in H^2(\Omega)$. Besides, the adaptive linear FVEMs are studied in [14].

Compared with the P_1-FVEM, the theoretical analysis of the classical isoparametric bilinear FVEM (Q_1-FVEM) on quadrilateral mesh is not easy, and most existing works need the quasi-parallelogram mesh assumption. The main reason is that the nodal basis function of Q_1-FVEM is not polynomial on general convex quadrilateral cells, and it has a complicated expression. Moreover, the trial function space and the test function space are different, which leads to an asymmetric bilinear form that is also difficult to analyze. In particular, unlike the P_1-FVEM, the cell stiffness matrix of the Q_1-FVEM is not a trivial perturbation of the corresponding Q_1-FEM. Thus, in order to guarantee the existence and uniqueness of the Q_1-FVEM on quadrilateral mesh, various sufficient conditions have been proposed, and the following geometric assumptions are seen in the literature.

(G1) h^2-parallelogram assumption: $m_K \leq Ch^2$, where C is a constant independent of h and h is small enough;

(G2) $h^{1+\gamma}$-parallelogram assumption: $m_K \leq Ch^{1+\gamma}$, where C is a constant independent of h and h is sufficiently small;

(G3) $m_K \leq Ch$, where C is a constant small enough but independent of h;

here, m_K denotes the distance between the midpoints of the two diagonals of the quadrilateral mesh cell K. Based on **(G1)**, for the condition that the diffusion tensor is an identity matrix, [15] presented the coercivity result and optimal H^1 error estimate. Regarding **(G2)**, [16] analyzed an arbitrary order FVEM on quadrilateral mesh and presented a unified proof of the inf-sup condition with a scalar diffusion coefficient. Regarding **(G3)**, [17] studied the coercivity of the Q_1-FVEM for the full diffusion tensor, and the existence of C was proved. However, for a specific diffusion tensor and a specific mesh cell, it is not easy to judge whether the assumption **(G3)** is satisfied. Recently, the authors of [18] proposed another sufficient condition which covers the traditional quasi-parallelogram mesh assumption. However, the stiffness matrix of the classical Q_1-FVEM cannot be computed exactly by our computers since the nodal basis function is not polynomial. This yields that the sufficient condition presented in [18] is not so accurate in practice.

Given the coercivity result and H^1 error estimate, the L^2, L^∞ and superconvergence are studied in [19–22], which can be referenced for a non-exhaustive list of references. The relevant studies of the FVEM can be found in [23–26] (triangle), [27,28] (quadrilateral), [29] (polygon), [30,31] (three dimensions) and so on. We mention that by postprocessing a high order FEM solution, the authors of [32] obtained a new FVEM solution and proved the stability and optimal convergence results for arbitrary triangular and quadrilateral meshes.

From another aspect, in order to analyze the Q_1-FVEM more easily, and by combining the characteristics in practical calculation, some researchers have employed the numerical integration methods to approximate the line integrals of the stiffness matrix in the classical Q_1-FVEM. For example, by approximating the line integrals at the geometric center of the quadrilateral, the authors of [33] constructed a symmetric scheme, and the error analysis is obtained on uniform rectangular meshes. Recently, by using the trapezoidal (resp. midpoint) rule to approximate the line integrals, the authors of [34] (resp. [35]) constructed a modified scheme. The authors proposed a sufficient condition to guarantee the coercivity result, and this condition covers the traditional quasi-parallelogram mesh. We mention that for the computation of line integrals, [33] is only for constant functions, while [34, 35] are for linear functions. This yields that the three numerical integration methods may not satisfy the practical computation. Therefore, it is necessary to employ another high precision numerical integration method to approximate the line integrals, and at the same time study the coercivity and optimal error estimate of the new scheme.

In this work, we employ the Simpson rule (which is explicitly for cubic functions) to approximate the line integrals in the classical Q_1-FVEM to solve anisotropic diffusion problems on general convex quadrilateral meshes, and the new scheme is called as Q_1-FVEM-SR for short. Different from the previous analysis techniques in [33–35], here for the proposed scheme, we transform the 4×4 cell singular matrix $\mathbb{A}_K$ of the bilinear form into a 3×3 symmetric matrix $\mathbb{B}_K^s$. Then, a necessary and sufficient condition (3.34) is obtained to ensure the positive definiteness of $\mathbb{B}_K^s$. Based on this result, in Theorem 3.1 a sufficient condition is suggested to guarantee the coercivity. More interestingly, this sufficient condition has an analytic expression, which only involves the anisotropic diffusion tensor and the geometry of the mesh. This implies that for an arbitrary full diffusion tensor and quadrilateral mesh, we can directly judge whether this sufficient condition is satisfied. In particular, this condition covers the traditional $h^{1+\gamma}$-parallelogram and some trapezoidal meshes with any full anisotropic diffusion tensor. Finally, by analyzing the difference between the bilinear forms of the Q_1-FVEM-SR scheme and classical Q_1-FVEM, we prove an optimal H^1 error estimate on $h^{1+\gamma}$-parallelogram mesh with $\gamma \geq 1$.

The rest of this paper is organized as follows. In Section 2, we briefly introduce some necessary notations and assumptions which will be used throughout the paper, and we present the construction of the Q_1-FVEM-SR scheme. In Section 3, we propose a sufficient condition to guarantee the coercivity result of the new scheme. Moreover, in Section 4 we discuss the sufficient condition on some special meshes with any full diffusion tensor. An optimal H^1 error estimate of the constructed scheme is given in Section 5. Several numerical examples are presented in Section 6 to validate the theoretical findings, and some concluding remarks are given in the last section.

2. Preliminary

2.1. Problems, meshes and notations

We consider the following anisotropic diffusion problem

$$-\nabla \cdot (\Lambda \nabla u) = f, \;\text{ in } \Omega, \tag{2.1}$$

$$u = 0, \;\text{ on } \partial\Omega, \tag{2.2}$$

where $\Omega \subset \mathbb{R}^2$ is an open bounded connected polygonal domain, $f \in L^2(\Omega)$ is the source term and $\Lambda = \Lambda(\boldsymbol{x})$ is a 2×2 symmetric diffusion tensor that is uniformly bounded above and below, i.e., there exist two positive constants $\underline{\lambda}$ and $\overline{\lambda}$ such that

$$\underline{\lambda}\|\boldsymbol{v}\|^2 \le \boldsymbol{v}^T\Lambda\boldsymbol{v} \le \overline{\lambda}\|\boldsymbol{v}\|^2, \quad \forall\boldsymbol{v}\in\mathbb{R}^2, \quad \forall\boldsymbol{x}\in\Omega, \tag{2.3}$$

where $\|\boldsymbol{v}\|$ is the Euclidean norm of the vector $\boldsymbol{v}$. For simplicity, here we only consider the homogeneous Dirichlet boundary condition.

Suppose that Ω is partitioned into a finite number of non-overlapped and strictly convex quadrilateral cells that form the so-called *primary mesh*. Each primary cell is further partitioned into four quadrilateral subcells by connecting the cell center with the four edge midpoints. All subcells sharing a common vertex of the primary mesh form a polygonal cell of the dual mesh; see Figure 1. For simplicity of exposition, we introduce the following notations, some of which are depicted in Figure 1.

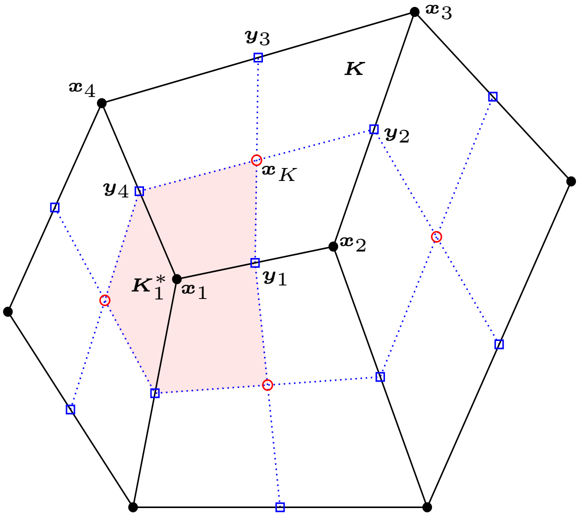


Figure 1. The primary mesh $\mathcal{T}_h$ (solid lines) and its associated dual mesh $\mathcal{T}_h^*$ (dotted lines).

- K a generic primary cell whose cell center, measure and diameter are respectively denoted as $\boldsymbol{x}_K$, $|K|$ and h_K;
- $\boldsymbol{x}_i$ $(1 \le i \le 4)$ the four vertices of K that are ordered anticlockwise. $\boldsymbol{y}_i$ is the midpoint of edge $\boldsymbol{x}_i\boldsymbol{x}_{i+1}$; here and hereafter i denotes, without special mention, a periodic index with period 4;
- $S_{i-1,i,i+1}$ the area of $\triangle\boldsymbol{x}_{i-1}\boldsymbol{x}_i\boldsymbol{x}_{i+1}$;
- $\mathcal{T}_h$ the set of primary cells in $\overline{\Omega}$ and $h = \max_{K\in\mathcal{T}_h} h_K$ is the mesh size;
- $\mathcal{T}_h^*$ the set of dual cells in $\overline{\Omega}$;
- $\overset{\circ}{\mathcal{V}}_h$ the set of all interior vertices;
- K_i^* the dual cell associated with $\boldsymbol{x}_i$;
- $\boldsymbol{n}_i^*$ the unit outward normal vector along the boundary of K_i^*.

In this paper, we assume that Λ is piecewise constant with respect to the primary mesh $\mathcal{T}_h$, and Λ_K is the constant restriction of Λ on K, namely $\Lambda_K = \Lambda$ in each K. Moreover, suppose that $\boldsymbol{x}_K$ is the geometric center of K, i.e.,

$$\boldsymbol{x}_K = \frac{1}{4}(\boldsymbol{x}_1 + \boldsymbol{x}_2 + \boldsymbol{x}_3 + \boldsymbol{x}_4), \tag{2.4}$$

and assume that $\mathcal{T}_h$ is regular, i.e., there exists a positive constant C_r independent of h such that

$$\frac{h_K}{\rho_K} \leq C_r, \quad \forall K \in \mathcal{T}_h, \tag{2.5}$$

where $\rho_K = \min_{1\leq i\leq 4}\{\text{diameter of the circle inscribed in } \Delta \boldsymbol{x}_{i-1}\boldsymbol{x}_i\boldsymbol{x}_{i+1}\}$. Sometimes, we use the quasi-regular assumption of the primary mesh, i.e., there exists a positive constant C_{qr} independent of h such that

$$|K| \geq C_{qr}h_K^2, \quad \forall K \in \mathcal{T}_h. \tag{2.6}$$

One can show that (2.5) implies (2.6), but not vice versa (see Theorem 2.1 in [36]).

2.2. *The classical Q_1-FVEM*

Assume that $\widehat{K} = \widehat{\boldsymbol{x}}_1\widehat{\boldsymbol{x}}_2\widehat{\boldsymbol{x}}_3\widehat{\boldsymbol{x}}_4 = [-1,1]^2$ is the reference rectangular element on the (ξ,η) plane, where the coordinates of four vertices are given by

$$\widehat{\boldsymbol{x}}_1 = (-1,-1)^T, \quad \widehat{\boldsymbol{x}}_2 = (1,-1)^T, \quad \widehat{\boldsymbol{x}}_3 = (1,1)^T, \quad \widehat{\boldsymbol{x}}_4 = (-1,1)^T.$$

Moreover, on $\widehat{K}$, we define the four bilinear nodal basis functions as

$$\widehat{\phi}_1 = \frac{(1-\xi)(1-\eta)}{4}, \quad \widehat{\phi}_2 = \frac{(1+\xi)(1-\eta)}{4}, \quad \widehat{\phi}_3 = \frac{(1+\xi)(1+\eta)}{4}, \quad \widehat{\phi}_4 = \frac{(1-\xi)(1+\eta)}{4}. \tag{2.7}$$

Obviously, we have that $\widehat{\phi}_i(\widehat{\boldsymbol{x}}_j) = \delta_{ij}$, where δ_{ij} denotes the Kronecker delta. For each strictly convex quadrilateral K, there exists a unique invertible bilinear mapping $\mathcal{J}_K$ which maps $\widehat{K}$ onto K that $\mathcal{J}_K(\widehat{\boldsymbol{x}}_i) = \boldsymbol{x}_i$, $i = 1,2,3,4$; see Figure 2. Precisely, this mapping can be written as

$$\mathcal{J}_K(\xi,\eta) = \boldsymbol{x}_K + \frac{1}{2}(\boldsymbol{m}_1\xi + \boldsymbol{m}_2\eta + \boldsymbol{m}_K\xi\eta),$$

where

$$\boldsymbol{m}_1 = \frac{\boldsymbol{x}_2 + \boldsymbol{x}_3 - \boldsymbol{x}_1 - \boldsymbol{x}_4}{2}, \quad \boldsymbol{m}_2 = \frac{\boldsymbol{x}_3 + \boldsymbol{x}_4 - \boldsymbol{x}_1 - \boldsymbol{x}_2}{2}, \quad \boldsymbol{m}_K = \frac{\boldsymbol{x}_1 + \boldsymbol{x}_3 - \boldsymbol{x}_2 - \boldsymbol{x}_4}{2}. \tag{2.8}$$

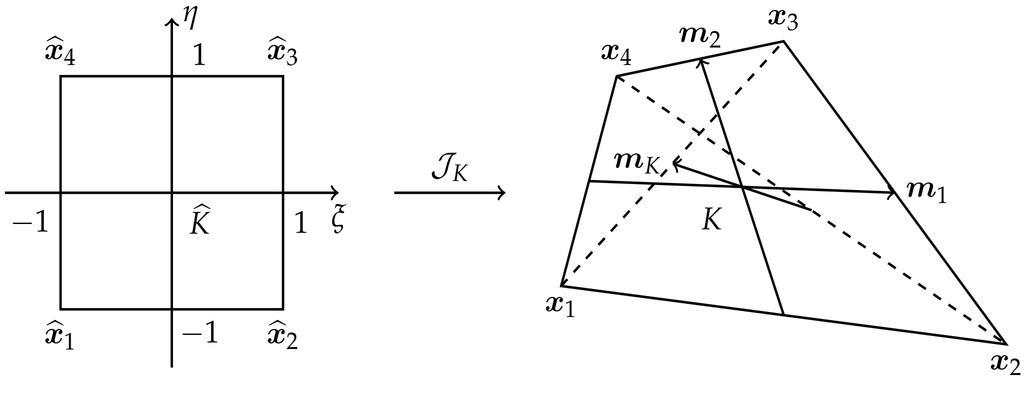


Figure 2. The bilinear mapping $\mathcal{J}_K$.

Then, the Jacobian matrix of the mapping $\mathcal{J}_K$ is given by

$$\mathbb{J}_K(\xi,\eta)=\frac{1}{2}\left(\boldsymbol{m}_1+\boldsymbol{m}_K\eta,\ \boldsymbol{m}_2+\boldsymbol{m}_K\xi\right)^T,$$

and by a direct calculation, we obtain the determinant of the Jacobian matrix

$$\det\mathbb{J}_K(\xi,\eta)=\frac{1}{4}(\boldsymbol{m}_1+\boldsymbol{m}_K\eta)\cdot(\mathcal{R}\boldsymbol{m}_2+\mathcal{R}\boldsymbol{m}_K\xi)=\frac{1}{4}|K|\left(1+\overline{\beta}_K\xi+\overline{\gamma}_K\eta\right),$$

where

$$\mathcal{R}=\begin{pmatrix}0&1\\-1&0\end{pmatrix},\quad \overline{\beta}_K=\frac{\boldsymbol{m}_1\cdot(\mathcal{R}\boldsymbol{m}_K)}{|K|}=\frac{S_{123}-S_{124}}{|K|},\quad \overline{\gamma}_K=\frac{\boldsymbol{m}_K\cdot(\mathcal{R}\boldsymbol{m}_2)}{|K|}=\frac{S_{134}-S_{124}}{|K|},\tag{2.9}$$

and we have used the fact that $\boldsymbol{m}_1\cdot(\mathcal{R}\boldsymbol{m}_2)=|K|$. This leads to

$$\mathbb{J}_K^{-1}(\xi,\eta)=\frac{2}{|K|\left(1+\overline{\beta}_K\xi+\overline{\gamma}_K\eta\right)}\mathcal{R}(\boldsymbol{m}_2+\xi\boldsymbol{m}_K,\ -\boldsymbol{m}_1-\eta\boldsymbol{m}_K).\tag{2.10}$$

Thanks to the mapping $\mathcal{J}_K$, on the primary mesh $\mathcal{T}_h$, we define the trial function space U_h as

$$U_h=\left\{u_h\in C^0(\overline{\Omega}):u_h|_K=\widehat{u}_h\circ\mathcal{J}_K^{-1},\widehat{u}_h|_{\widehat{K}}\text{ is bilinear function},\forall K\in\mathcal{T}_h,u_h|_{\partial\Omega}=0\right\},$$

and on the dual mesh $\mathcal{T}_h^*$, the test function space V_h is defined as

$$V_h=\left\{v_h\in L^2(\Omega):v_h|_{K_i^*}=\text{constant},\forall K_i^*\in\mathcal{T}_h^*,v_h|_{\partial\Omega}=0\right\}.$$

For each $v_h\in V_h$, we have

$$v_h=\sum_{\boldsymbol{x}_i\in\overset{\circ}{\mathcal{V}}_h}v_i\chi_i,$$

where $v_i=v_h(\boldsymbol{x}_i)$ and χ_i is the characteristic function on K_i^*, satisfying $\chi_i(\boldsymbol{x})=1$ if $\boldsymbol{x}\in K_i^*$ and $\chi_i(\boldsymbol{x})=0$ if $\boldsymbol{x}\in\Omega\backslash K_i^*$.

The classical Q_1-FVEM for solving (2.1) and (2.2) is as follows: find $u_h\in U_h$ such that

$$a_h(u_h,v_h)=(f,v_h),\quad\forall v_h\in V_h,$$

where

$$a_h(u_h,v_h)=\sum_{K_i^*\in\mathcal{T}_h^*}v_i\int_{\partial K_i^*}(-\Lambda\nabla u_h)\cdot\boldsymbol{n}_i^*\,\mathrm{d}s,\quad(f,v_h)=\sum_{K_i^*\in\mathcal{T}_h^*}v_i\int_{K_i^*}f\,\mathrm{d}x\mathrm{d}y.$$

By rearranging the summation of a_h, we have

$$a_h(u_h,v_h)=\sum_{K\in\mathcal{T}_h}a_{K,h}(u_h,v_h),$$

where

$$a_{K,h}(u_h,v_h)=\sum_{i=1}^{4}v_i\left(\int_{\boldsymbol{x}_K\boldsymbol{y}_{i-1}}(-\Lambda_K\nabla u_h)\cdot\boldsymbol{n}_{K,i-1}^*\,\mathrm{d}s-\int_{\boldsymbol{x}_K\boldsymbol{y}_i}(-\Lambda_K\nabla u_h)\cdot\boldsymbol{n}_{K,i}^*\,\mathrm{d}s\right),\tag{2.11}$$

and

$$\boldsymbol{n}_{K,i}^*=\frac{1}{\|\boldsymbol{y}_i-\boldsymbol{x}_K\|}\mathcal{R}(\boldsymbol{y}_i-\boldsymbol{x}_K).$$

2.3. The Q_1-FVEM-SR scheme

By employing the Simpson rule to approximate the line integrals in (2.11), we get the so-called Q_1-FVEM-SR scheme, given by

$$\widetilde{a}_h(u_h, v_h) = (f, v_h), \quad \forall v_h \in V_h, \tag{2.12}$$

where

$$\widetilde{a}_h(u_h, v_h) = \sum_{K \in \mathcal{T}_h} \widetilde{a}_{K,h}(u_h, v_h), \tag{2.13}$$

and

$$\begin{aligned} \widetilde{a}_{K,h}(u_h, v_h) = & \frac{1}{6}\sum_{i=1}^{4} v_i \Big[(\boldsymbol{x}_K - \boldsymbol{y}_{i-1})^T \mathcal{R}^T \Lambda_K \left(\nabla u_h(\boldsymbol{x}_K) + 4\nabla u_h(\boldsymbol{z}_{i-1}) + \nabla u_h(\boldsymbol{y}_{i-1}) \right) \\ & - (\boldsymbol{x}_K - \boldsymbol{y}_i)^T \mathcal{R}^T \Lambda_K \left(\nabla u_h(\boldsymbol{x}_K) + 4\nabla u_h(\boldsymbol{z}_i) + \nabla u_h(\boldsymbol{y}_i) \right) \Big], \end{aligned} \tag{2.14}$$

with $\boldsymbol{z}_i = (\boldsymbol{x}_K + \boldsymbol{y}_i)/2$.

In Section 3, we will present a proof of the coercivity result, which is based on the study of cell bilinear form defined by (2.14).

3. The coercivity result of the Q_1-FVEM-SR scheme

For convenience of exposition, in each K, we define the following notations

$$m_{ij} = \frac{1}{4|K|} (\mathcal{R}\boldsymbol{m}_i)^T \Lambda_K \left(\mathcal{R}\boldsymbol{m}_j\right), \quad i, j = 1, 2. \tag{3.1}$$

$$\upsilon_1 = \upsilon_2 + \upsilon_3, \quad \upsilon_2 = \frac{4m_{11}}{4 - \overline{\beta}_K^2}, \quad \upsilon_3 = \frac{4m_{22}}{4 - \overline{\gamma}_K^2}. \tag{3.2}$$

$$\mu_1 = \mu_2 + \mu_3, \quad \mu_2 = \frac{2\left(4 - 3\overline{\beta}_K^2\right)}{\left(1 - \overline{\beta}_K^2\right)\left(4 - \overline{\beta}_K^2\right)} m_{11}, \quad \mu_3 = \frac{2\left(4 - 3\overline{\gamma}_K^2\right)}{\left(1 - \overline{\gamma}_K^2\right)\left(4 - \overline{\gamma}_K^2\right)} m_{22}. \tag{3.3}$$

$$\begin{aligned} \zeta_1 &= m_{11} - \frac{1}{4\mu_1}\left(\frac{\mu_3}{2} + \frac{\upsilon_2}{3}\right)^2 \overline{\beta}_K^2, \quad \zeta_2 = m_{22} - \frac{1}{4\mu_1}\left(\frac{\mu_2}{2} + \frac{\upsilon_3}{3}\right)^2 \overline{\gamma}_K^2, \\ \zeta_3 &= m_{12} + \frac{1}{4\mu_1}\left(\frac{\mu_3}{2} + \frac{\upsilon_2}{3}\right)\left(\frac{\mu_2}{2} + \frac{\upsilon_3}{3}\right)\overline{\beta}_K \overline{\gamma}_K. \end{aligned} \tag{3.4}$$

In addition, we introduce the following assumption.

(A1) There exists a positive constant ϱ, independent of K and h, such that

$$\zeta_1\zeta_2 - \zeta_3^2 \geq \varrho. \tag{3.5}$$

The main result of this section is given in the following Theorem 3.1.

Theorem 3.1. *Let $\Pi_h^* : U_h \to V_h$ be the interpolation operator from U_h to V_h, satisfying $\Pi_h^* u_h(\boldsymbol{x}_i) = u_h(\boldsymbol{x}_i)$. Then, under assumptions* (2.3), (2.4), (2.5) *and **(A1)**, we have*

$$\widetilde{a}_h\left(u_h, \Pi_h^* u_h\right) \geq \kappa |u_h|_1^2, \quad \forall u_h \in U_h, \tag{3.6}$$

where κ is a positive constant, independent of h, and $|\cdot|_1$ denotes the standard H^1 semi-norm.

For the proof of the above Theorem 3.1, we need some preliminary results.

Lemma 3.1. *Assume that K is a strictly convex quadrilateral; then, we have*

$$\left|\overline{\beta}_K\right| + \left|\overline{\gamma}_K\right| < 1, \tag{3.7}$$

and

$$\boldsymbol{m}_K = \overline{\gamma}_K \boldsymbol{m}_1 + \overline{\beta}_K \boldsymbol{m}_2. \tag{3.8}$$

Proof. It follows from (2.9) that

$$\begin{aligned}
1 + \overline{\beta}_K + \overline{\gamma}_K &= 1 + \frac{1}{|K|}(S_{123} + S_{134} - 2S_{124}) = \frac{2S_{234}}{|K|} > 0,\\
1 - \overline{\beta}_K + \overline{\gamma}_K &= 1 + \frac{1}{|K|}(-S_{123} + S_{134}) = \frac{2S_{134}}{|K|} > 0,\\
1 + \overline{\beta}_K - \overline{\gamma}_K &= 1 + \frac{1}{|K|}(S_{123} - S_{134}) = \frac{2S_{123}}{|K|} > 0,\\
1 - \overline{\beta}_K - \overline{\gamma}_K &= 1 + \frac{1}{|K|}(2S_{124} - S_{134} - S_{123}) = \frac{2S_{124}}{|K|} > 0,
\end{aligned}$$

which leads to (3.7). Suppose that $\boldsymbol{m}_K = c_1 \boldsymbol{m}_1 + c_2 \boldsymbol{m}_2$, where c_1 and c_2 are two coefficients to be determined. Then, we have

$$\boldsymbol{m}_K \cdot (\mathcal{R}\boldsymbol{m}_2) = c_1 \boldsymbol{m}_1 \cdot (\mathcal{R}\boldsymbol{m}_2), \quad \boldsymbol{m}_K \cdot (\mathcal{R}\boldsymbol{m}_1) = c_2 \boldsymbol{m}_2 \cdot (\mathcal{R}\boldsymbol{m}_1).$$

Noticing (2.9), we obtain that $c_1 = \overline{\gamma}_K$ and $c_2 = \overline{\beta}_K$. The proof is complete. □

Lemma 3.2. *For the m_{ij} defined by* (3.1)*, we have*

$$m_{12} = m_{21}, \quad m_{11}m_{22} - m_{12}^2 = \frac{1}{16}\det(\Lambda_K). \tag{3.9}$$

Moreover, under the assumptions (2.3) *and* (2.6)*,*

$$|m_{12}| < \frac{\overline{\lambda}}{4C_{qr}}, \quad \frac{C_{qr}\underline{\lambda}}{4} < m_{ii} < \frac{\overline{\lambda}}{4C_{qr}}, \quad i = 1, 2. \tag{3.10}$$

Proof. Recalling that Λ_K is a symmetric and positive definite matrix, we get that $m_{12} = m_{21}$, and by (3.1), we find that

$$\begin{aligned}
m_{11}m_{22} - m_{12}^2 &= \frac{1}{16|K|^2}\det\begin{pmatrix} (\mathcal{R}\boldsymbol{m}_1)^T \Lambda_K (\mathcal{R}\boldsymbol{m}_1) & (\mathcal{R}\boldsymbol{m}_1)^T \Lambda_K (\mathcal{R}\boldsymbol{m}_2) \\ (\mathcal{R}\boldsymbol{m}_2)^T \Lambda_K (\mathcal{R}\boldsymbol{m}_1) & (\mathcal{R}\boldsymbol{m}_2)^T \Lambda_K (\mathcal{R}\boldsymbol{m}_2) \end{pmatrix}\\
&= \frac{1}{16|K|^2}\det\left(\begin{pmatrix} (\mathcal{R}\boldsymbol{m}_1)^T \\ (\mathcal{R}\boldsymbol{m}_2)^T \end{pmatrix} \Lambda_K (\mathcal{R}\boldsymbol{m}_1, \mathcal{R}\boldsymbol{m}_2)\right)\\
&= \frac{\det(\Lambda_K)}{16|K|^2}\left[\det(\boldsymbol{m}_1, \boldsymbol{m}_2)\right]^2 = \frac{\det(\Lambda_K)}{16},
\end{aligned}$$

which implies (3.9). From (2.3) and (2.6), we have

$$\begin{aligned} m_{11} &\geq \frac{\underline{\lambda}\|\boldsymbol{m}_1\|^2}{4|K|} \geq \frac{\underline{\lambda}\|\boldsymbol{m}_1\|^2}{4\|\boldsymbol{m}_1\|\|\boldsymbol{m}_2\|} = \frac{\underline{\lambda}\|\boldsymbol{m}_1\|\|\boldsymbol{m}_2\|}{4\|\boldsymbol{m}_2\|^2} > \frac{\underline{\lambda}|K|}{4h_K^2} \geq \frac{C_{qr}\underline{\lambda}}{4}, \\ m_{11} &\leq \frac{\overline{\lambda}\|\boldsymbol{m}_1\|^2}{4|K|} \leq \frac{\overline{\lambda}\|\boldsymbol{m}_1\|^2}{4C_{qr}h_K^2} < \frac{\overline{\lambda}}{4C_{qr}}. \end{aligned}$$

By the same arguments, we obtain the estimate of m_{22}. Finally, it follows from (3.9) that

$$|m_{12}| = \frac{1}{4}\sqrt{16m_{11}m_{22} - \det(\Lambda_K)} \leq \max\{m_{11}, m_{22}\} < \frac{\overline{\lambda}}{4C_{qr}}.$$

□

Lemma 3.3. *For the* ζ_1 *and* ζ_2 *defined in* (3.4)*, we have*

$$\zeta_1 + \zeta_2 > 0. \tag{3.11}$$

Proof. By (3.2) and (3.3), we find that

$$0 < \upsilon_i < \frac{2}{3}\mu_i, \quad i = 2, 3, \tag{3.12}$$

and then

$$\frac{\mu_3}{2} + \frac{\upsilon_2}{3} < \frac{\mu_3}{2} + \frac{2\mu_2}{9} < \frac{\mu_1}{2}. \tag{3.13}$$

It follows that

$$\zeta_1 > m_{11} - \frac{1}{8}\left(\frac{\mu_3}{2} + \frac{\upsilon_2}{3}\right)\overline{\beta}_K^2.$$

Note that

$$\upsilon_2\overline{\beta}_K^2 = \frac{4\overline{\beta}_K^2}{4 - \overline{\beta}_K^2}m_{11} < \frac{4}{3}m_{11}$$

and

$$\mu_3\overline{\beta}_K^2 = \frac{2\overline{\beta}_K^2\left(4 - 3\overline{\gamma}_K^2\right)}{\left(1 - \overline{\gamma}_K^2\right)\left(4 - \overline{\gamma}_K^2\right)}m_{22} < \frac{2\left(4 - 3\overline{\gamma}_K^2\right)}{4 - \overline{\gamma}_K^2}m_{22} < 2m_{22},$$

which implies that

$$\zeta_1 > m_{11} - \frac{1}{8}(m_{11} + m_{22}).$$

By the same arguments

$$\zeta_2 > m_{22} - \frac{1}{8}(m_{11} + m_{22}).$$

That is

$$\zeta_1 + \zeta_2 > m_{11} + m_{22} - \frac{1}{4}(m_{11} + m_{22}) = \frac{3}{4}(m_{11} + m_{22}) > 0.$$

The proof is complete. □

Moreover, we introduce the following new basis functions:

$$\widetilde{\phi}_1 = \frac{1}{2}, \quad \widetilde{\phi}_2 = \frac{\xi\eta}{2}, \quad \widetilde{\phi}_3 = -\frac{\eta}{2}, \quad \widetilde{\phi}_4 = -\frac{\xi}{2}. \tag{3.14}$$

By (2.7), we deduce that

$$\left(\widehat{\phi}_1, \widehat{\phi}_2, \widehat{\phi}_3, \widehat{\phi}_4\right) = \left(\widetilde{\phi}_1, \widetilde{\phi}_2, \widetilde{\phi}_3, \widetilde{\phi}_4\right) \mathbb{P}, \tag{3.15}$$

where

$$\mathbb{P} = \frac{1}{2}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}.$$

Obviously, $\mathbb{P}$ is a symmetric and orthogonal matrix, i.e.,

$$\mathbb{P} = \mathbb{P}^T = \mathbb{P}^{-1}. \tag{3.16}$$

Lemma 3.4. *For any $u_h \in U_h$, in each K, we denote*

$$\boldsymbol{u}_K = (u_1, u_2, u_3, u_4)^T, \quad \widetilde{\boldsymbol{u}}_K = (\widetilde{u}_1, \widetilde{u}_2, \widetilde{u}_3, \widetilde{u}_4)^T = \mathbb{P}\boldsymbol{u}_K, \tag{3.17}$$

where $u_i = u_h(\boldsymbol{x}_i)$. Then, under the assumption (2.5)*, there exists a positive constant $\widetilde{C}$ independent of h such that*

$$|u_h|_{1,K} \leq \widetilde{C}\|\widetilde{\boldsymbol{w}}_K\|, \quad \forall u_h \in U_h, \quad \forall K \in \mathcal{T}_h, \tag{3.18}$$

where $\widetilde{\boldsymbol{w}}_K = (\widetilde{u}_2, \widetilde{u}_3, \widetilde{u}_4)^T$.

Proof. A proof of (3.18) can be found in Lemma 6 of [18]. □

Remark 3.1. *By Proposition* 1 *of [15], there exist two positive constants $\underline{C}$ and $\overline{C}$ such that*

$$\underline{C}|u_h|_{1,K,h} \leq |u_h|_{1,K} \leq \overline{C}|u_h|_{1,K,h}, \quad \forall u_h \in U_h, \quad \forall K \in \mathcal{T}_h, \tag{3.19}$$

where

$$|u_h|^2_{1,K,h} = \sum_{i=1}^{4} \left[u_h(\boldsymbol{x}_{i+1}) - u_h(\boldsymbol{x}_i)\right]^2. \tag{3.20}$$

Moreover, it is easy to verify that

$$\|\widetilde{\boldsymbol{w}}_K\| \leq |u_h|_{1,K,h}, \quad \forall u_h \in U_h, \quad \forall K \in \mathcal{T}_h. \tag{3.21}$$

Lemma 3.5. *In each K, assume that $\boldsymbol{\vartheta}_K = (\vartheta_1, \vartheta_2, \vartheta_3, \vartheta_4)^T$ with the entry*

$$\vartheta_i = \frac{1}{6}(\boldsymbol{x}_K - \boldsymbol{y}_{i-1})^T \mathcal{R}^T \Lambda_K \left[\boldsymbol{\nu}(\widehat{\boldsymbol{x}}_K) + 4\boldsymbol{\nu}(\widehat{\boldsymbol{z}}_{i-1}) + \boldsymbol{\nu}(\widehat{\boldsymbol{y}}_{i-1})\right] - \frac{1}{6}(\boldsymbol{x}_K - \boldsymbol{y}_i)^T \mathcal{R}^T \Lambda_K \left[\boldsymbol{\nu}(\widehat{\boldsymbol{x}}_K) + 4\boldsymbol{\nu}(\widehat{\boldsymbol{z}}_i) + \boldsymbol{\nu}(\widehat{\boldsymbol{y}}_i)\right],$$

where $\widehat{\boldsymbol{x}} = \mathcal{J}_K^{-1}(\boldsymbol{x})$ and

$$\boldsymbol{\nu}(\xi, \eta) = \frac{\mathcal{R}(\eta\boldsymbol{m}_2 - \xi\boldsymbol{m}_1)}{4\det\mathbb{J}_K(\xi, \eta)} = \frac{\mathcal{R}(\eta\boldsymbol{m}_2 - \xi\boldsymbol{m}_1)}{|K|\left(1 + \overline{\beta}_K\xi + \overline{\gamma}_K\eta\right)}. \tag{3.22}$$

Then, under the assumption (2.4)*, we have*

$$\boldsymbol{\vartheta}_K = \mathbb{P}\mathrm{diag}\left(0, \mu_1, \mu_2 - \frac{2}{3}\upsilon_2, \mu_3 - \frac{2}{3}\upsilon_3\right)\boldsymbol{S}_K, \tag{3.23}$$

where $\boldsymbol{S}_K = (0, 1, \overline{\beta}_K, \overline{\gamma}_K)^T$.

Proof. It follows from (2.4) and (2.8) that

$$\boldsymbol{x}_K - \boldsymbol{y}_1 = \boldsymbol{y}_3 - \boldsymbol{x}_K = \frac{1}{2}\boldsymbol{m}_2, \quad \boldsymbol{x}_K - \boldsymbol{y}_4 = \boldsymbol{y}_2 - \boldsymbol{x}_K = \frac{1}{2}\boldsymbol{m}_1,$$

and by (3.22), we obtain

$$\begin{aligned}\vartheta_1 &= \frac{1}{12}(\mathcal{R}\boldsymbol{m}_1)^T \Lambda_K \left[\boldsymbol{\nu}(0,0) + 4\boldsymbol{\nu}\left(-1/2, 0\right) + \boldsymbol{\nu}(-1,0)\right] - \frac{1}{12}(\mathcal{R}\boldsymbol{m}_2)^T \Lambda_K \left[\boldsymbol{\nu}(0,0) + 4\boldsymbol{\nu}\left(0, -1/2\right) + \boldsymbol{\nu}(0,-1)\right] \\ &= \frac{1}{3}\left(\frac{4}{2-\overline{\beta}_K} + \frac{1}{1-\overline{\beta}_K}\right)m_{11} + \frac{1}{3}\left(\frac{4}{2-\overline{\gamma}_K} + \frac{1}{1-\overline{\gamma}_K}\right)m_{22}.\end{aligned}$$

Similarly

$$\begin{aligned}\vartheta_2 &= -\frac{1}{3}\left(\frac{4}{2+\overline{\beta}_K} + \frac{1}{1+\overline{\beta}_K}\right)m_{11} - \frac{1}{3}\left(\frac{4}{2-\overline{\gamma}_K} + \frac{1}{1-\overline{\gamma}_K}\right)m_{22}, \\ \vartheta_3 &= \frac{1}{3}\left(\frac{4}{2+\overline{\beta}_K} + \frac{1}{1+\overline{\beta}_K}\right)m_{11} + \frac{1}{3}\left(\frac{4}{2+\overline{\gamma}_K} + \frac{1}{1+\overline{\gamma}_K}\right)m_{22}, \\ \vartheta_4 &= -\frac{1}{3}\left(\frac{4}{2-\overline{\beta}_K} + \frac{1}{1-\overline{\beta}_K}\right)m_{11} - \frac{1}{3}\left(\frac{4}{2+\overline{\gamma}_K} + \frac{1}{1+\overline{\gamma}_K}\right)m_{22}.\end{aligned}$$

It follows that

$$(\vartheta_1, \vartheta_2, \vartheta_3, \vartheta_4)\mathbb{P} = \left(0, \mu_1, \left(\mu_2 - \frac{2}{3}\upsilon_2\right)\overline{\beta}_K, \left(\mu_3 - \frac{2}{3}\upsilon_3\right)\overline{\gamma}_K\right) = \boldsymbol{S}_K^T \mathrm{diag}\left(0, \mu_1, \mu_2 - \frac{2}{3}\upsilon_2, \mu_3 - \frac{2}{3}\upsilon_3\right),$$

which implies (3.23) by noticing (3.16). □

Lemma 3.6. *Under the assumption* (2.4)*, we have*

$$\widetilde{a}_{K,h}\left(u_h, \Pi_h^* u_h\right) = \widetilde{\boldsymbol{u}}_K^T \mathbb{A}_K \widetilde{\boldsymbol{u}}_K, \tag{3.24}$$

where

$$\mathbb{A}_K = \frac{1}{|K|}\mathbb{R}_K^T \Lambda_K \mathbb{R}_K + \mathrm{diag}\left(0, \mu_1, \mu_2 - \frac{2}{3}\upsilon_2, \mu_3 - \frac{2}{3}\upsilon_3\right)\boldsymbol{S}_K \boldsymbol{S}_K^T, \tag{3.25}$$

$\mathbb{R}_K = \mathcal{R}(\boldsymbol{0}, \boldsymbol{0}, \boldsymbol{m}_1, -\boldsymbol{m}_2)$ *and* $\boldsymbol{0}$ *is a zero vector.*

Proof. In each K, we have

$$\nabla\widetilde{\phi}_i = \mathbb{J}_K^{-1}(\xi, \eta)\widehat{\nabla}\widetilde{\phi}_i \quad \text{with} \quad \widehat{\nabla} = \left(\frac{\partial}{\partial\xi}, \frac{\partial}{\partial\eta}\right)^T, \tag{3.26}$$

and by (3.14),

$$\widehat{\nabla}\widetilde{\phi}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \widehat{\nabla}\widetilde{\phi}_2 = \frac{1}{2}\begin{pmatrix} \eta \\ \xi \end{pmatrix}, \quad \widehat{\nabla}\widetilde{\phi}_3 = \frac{1}{2}\begin{pmatrix} 0 \\ -1 \end{pmatrix}, \quad \widehat{\nabla}\widetilde{\phi}_4 = \frac{1}{2}\begin{pmatrix} -1 \\ 0 \end{pmatrix}. \tag{3.27}$$

It follows from (3.8) that

$$\boldsymbol{m}_2 + \xi \boldsymbol{m}_K = \left(1 + \overline{\beta}_K \xi\right) \boldsymbol{m}_2 + \xi \overline{\gamma}_K \boldsymbol{m}_1 = \left(1 + \overline{\beta}_K \xi + \overline{\gamma}_K \eta\right) \boldsymbol{m}_2 - \overline{\gamma}_K (\eta \boldsymbol{m}_2 - \xi \boldsymbol{m}_1)$$

and

$$-\boldsymbol{m}_1 - \eta \boldsymbol{m}_K = -(1 + \overline{\gamma}_K \eta)\, \boldsymbol{m}_1 - \eta \overline{\beta}_K \boldsymbol{m}_2 = -\left(1 + \overline{\beta}_K \xi + \overline{\gamma}_K \eta\right) \boldsymbol{m}_1 - \overline{\beta}_K (\eta \boldsymbol{m}_2 - \xi \boldsymbol{m}_1).$$

As a consequence, and by (2.10), it holds that

$$\mathbb{J}_K^{-1}(\xi, \eta) = \frac{2}{|K|} \mathcal{R}(\boldsymbol{m}_2, -\boldsymbol{m}_1) - 2\boldsymbol{\nu}(\xi, \eta) \left(\overline{\gamma}_K, \overline{\beta}_K\right).$$

Then, from (3.26) and (3.27),

$$\nabla \widetilde{\phi}_1 = \boldsymbol{0}, \qquad \nabla \widetilde{\phi}_2 = \frac{\mathcal{R}(\eta \boldsymbol{m}_2 - \xi \boldsymbol{m}_1)}{|K|} - \boldsymbol{\nu}(\xi, \eta) \left(\overline{\beta}_K \xi + \overline{\gamma}_K \eta\right) = \boldsymbol{\nu}(\xi, \eta),$$

$$\nabla \widetilde{\phi}_3 = \frac{\mathcal{R} \boldsymbol{m}_1}{|K|} + \boldsymbol{\nu}(\xi, \eta) \overline{\beta}_K, \qquad \nabla \widetilde{\phi}_4 = -\frac{\mathcal{R} \boldsymbol{m}_2}{|K|} + \boldsymbol{\nu}(\xi, \eta) \overline{\gamma}_K.$$

In other words, we obtain

$$\left(\nabla \widetilde{\phi}_1, \nabla \widetilde{\phi}_2, \nabla \widetilde{\phi}_3, \nabla \widetilde{\phi}_4\right) = \frac{1}{|K|} \mathbb{R}_K + \boldsymbol{\nu}(\xi, \eta) \boldsymbol{S}_K^T. \tag{3.28}$$

By (3.15) and (3.17), it holds that

$$u_h = \left(\widehat{\phi}_1, \widehat{\phi}_2, \widehat{\phi}_3, \widehat{\phi}_4\right) \boldsymbol{u}_K = \left(\widetilde{\phi}_1, \widetilde{\phi}_2, \widetilde{\phi}_3, \widetilde{\phi}_4\right) \mathbb{P}\mathbb{P}^{-1} \widetilde{\boldsymbol{u}}_K = \left(\widetilde{\phi}_1, \widetilde{\phi}_2, \widetilde{\phi}_3, \widetilde{\phi}_4\right) \widetilde{\boldsymbol{u}}_K,$$

and it follows from (3.28) that

$$\nabla u_h = \left(\nabla \widetilde{\phi}_1, \nabla \widetilde{\phi}_2, \nabla \widetilde{\phi}_3, \nabla \widetilde{\phi}_4\right) \widetilde{\boldsymbol{u}}_K = \frac{1}{|K|} \mathbb{R}_K \widetilde{\boldsymbol{u}}_K + \boldsymbol{\nu}(\xi, \eta) \boldsymbol{S}_K^T \widetilde{\boldsymbol{u}}_K. \tag{3.29}$$

Substituting the above equality into (2.14), we have

$$\widetilde{a}_{K,h}\left(u_h, \Pi_h^* u_h\right) = I_1 + I_2,$$

where the first part is given by

$$I_1 = \frac{1}{|K|} \sum_{i=1}^{4} u_i (\boldsymbol{y}_i - \boldsymbol{y}_{i-1})^T \mathcal{R}^T \Lambda_K \mathbb{R}_K \widetilde{\boldsymbol{u}}_K,$$

and the second part is defined in (3.30). For I_1, since

$$\mathcal{R}(\boldsymbol{y}_i - \boldsymbol{y}_{i-1}) = \mathbb{R}_K \mathbb{P}_i,$$

where $\mathbb{P}_i$ denotes the i-th column of matrix $\mathbb{P}$, we obtain

$$I_1 = \frac{1}{|K|}\boldsymbol{u}_K^T\mathbb{P}^T\mathbb{R}_K^T\Lambda_K\mathbb{R}_K\widetilde{\boldsymbol{u}}_K = \frac{1}{|K|}\widetilde{\boldsymbol{u}}_K^T\mathbb{R}_K^T\Lambda_K\mathbb{R}_K\widetilde{\boldsymbol{u}}_K.$$

For I_2, from Lemma 3.5, we deduce that

$$I_2 = \left(\sum_{i=1}^{4} u_i\vartheta_i\right)\boldsymbol{S}_K^T\widetilde{\boldsymbol{u}}_K = \boldsymbol{u}_K^T\boldsymbol{\vartheta}_K\boldsymbol{S}_K^T\widetilde{\boldsymbol{u}}_K = \widetilde{\boldsymbol{u}}_K^T\text{diag}\left(0, \mu_1, \mu_2 - \frac{2}{3}\upsilon_2, \mu_3 - \frac{2}{3}\upsilon_3\right)\boldsymbol{S}_K\boldsymbol{S}_K^T\widetilde{\boldsymbol{u}}_K. \tag{3.30}$$

Combining the above results, we get the desired equality (3.24). □

By a direct calculation, it follows from (3.25) that

$$\mathbb{A}_K = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \mu_1 & \mu_1\overline{\beta}_K & \mu_1\overline{\gamma}_K \\ 0 & \left(\mu_2 - \frac{2}{3}\upsilon_2\right)\overline{\beta}_K & 4m_{11} + \left(\mu_2 - \frac{2}{3}\upsilon_2\right)\overline{\beta}_K^2 & -4m_{12} + \left(\mu_2 - \frac{2}{3}\upsilon_2\right)\overline{\beta}_K\overline{\gamma}_K \\ 0 & \left(\mu_3 - \frac{2}{3}\upsilon_3\right)\overline{\gamma}_K & -4m_{12} + \left(\mu_3 - \frac{2}{3}\upsilon_3\right)\overline{\beta}_K\overline{\gamma}_K & 4m_{22} + \left(\mu_3 - \frac{2}{3}\upsilon_3\right)\overline{\gamma}_K^2 \end{pmatrix}.$$

Let

$$\mathbb{A}_K^s = \frac{1}{2}\left(\mathbb{A}_K + \mathbb{A}_K^T\right) = \begin{pmatrix} 0 & \boldsymbol{0}^T \\ \boldsymbol{0} & \mathbb{B}_K^s \end{pmatrix}$$

be the symmetric part of $\mathbb{A}_K$, where

$$\mathbb{B}_K^s = \begin{pmatrix} \mu_1 & \left(\frac{\mu_1+\mu_2}{2} - \frac{\upsilon_2}{3}\right)\overline{\beta}_K & \left(\frac{\mu_1+\mu_3}{2} - \frac{\upsilon_3}{3}\right)\overline{\gamma}_K \\ \left(\frac{\mu_1+\mu_2}{2} - \frac{\upsilon_2}{3}\right)\overline{\beta}_K & 4m_{11} + \left(\mu_2 - \frac{2}{3}\upsilon_2\right)\overline{\beta}_K^2 & -4m_{12} + \left(\frac{\mu_1}{2} - \frac{\upsilon_1}{3}\right)\overline{\beta}_K\overline{\gamma}_K \\ \left(\frac{\mu_1+\mu_3}{2} - \frac{\upsilon_3}{3}\right)\overline{\gamma}_K & -4m_{12} + \left(\frac{\mu_1}{2} - \frac{\upsilon_1}{3}\right)\overline{\beta}_K\overline{\gamma}_K & 4m_{22} + \left(\mu_3 - \frac{2}{3}\upsilon_3\right)\overline{\gamma}_K^2 \end{pmatrix}. \tag{3.31}$$

It follows that

$$\widetilde{a}_{K,h}\left(u_h, \Pi_h^* u_h\right) = \widetilde{\boldsymbol{u}}_K^T\mathbb{A}_K\widetilde{\boldsymbol{u}}_K = \widetilde{\boldsymbol{u}}_K^T\mathbb{A}_K^s\widetilde{\boldsymbol{u}}_K = \widetilde{\boldsymbol{w}}_K^T\mathbb{B}_K^s\widetilde{\boldsymbol{w}}_K, \tag{3.32}$$

where $\widetilde{\boldsymbol{w}}_K$ is defined in Lemma 3.4, i.e.,

$$\widetilde{\boldsymbol{u}}_K = \begin{pmatrix} \widetilde{u}_1 \\ \widetilde{\boldsymbol{w}}_K \end{pmatrix}.$$

Lemma 3.7. *Assume that* $\mathbb{T} = \mathbb{T}_1\mathbb{T}_2$*, where*

$$\mathbb{T}_1 = \begin{pmatrix} 1 & -\frac{1}{\mu_1}\left(\frac{\mu_1+\mu_2}{2} - \frac{\upsilon_2}{3}\right)\overline{\beta}_K & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbb{T}_2 = \begin{pmatrix} 1 & 0 & -\frac{1}{\mu_1}\left(\frac{\mu_1+\mu_3}{2} - \frac{\upsilon_3}{3}\right)\overline{\gamma}_K \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Then, for the matrix $\mathbb{B}_K^s$ *defined by* (3.31)*, it holds that*

$$\mathbb{T}^T\mathbb{B}_K^s\mathbb{T} = \begin{pmatrix} \mu_1 & 0 & 0 \\ 0 & 4\zeta_1 & -4\zeta_3 \\ 0 & -4\zeta_3 & 4\zeta_2 \end{pmatrix}. \tag{3.33}$$

Consequently, $\mathbb{B}_K^s$ *is a positive definite matrix if and only if*

$$\zeta_1\zeta_2 - \zeta_3^2 > 0. \tag{3.34}$$

Proof. By a direct calculation, it holds that

$$\mathbb{T}_1^T\mathbb{B}_K^s\mathbb{T}_1 = \begin{pmatrix} \mu_1 & 0 & \left(\frac{\mu_1+\mu_3}{2} - \frac{\upsilon_3}{3}\right)\overline{\gamma}_K \\ 0 & 4\zeta_1 & -4\zeta_3 \\ \left(\frac{\mu_1+\mu_3}{2} - \frac{\upsilon_3}{3}\right)\overline{\gamma}_K & -4\zeta_3 & 4m_{22} + \left(\mu_3 - \frac{2}{3}\upsilon_3\right)\overline{\gamma}_K^2 \end{pmatrix},$$

and still through some straightforward calculations with $\mathbb{T}_2$, we obtain (3.33). Since $\mu_1 > 0$, $\mathbb{T}_1$ and $\mathbb{T}_2$ are invertible matrices; we find that $\mathbb{B}_K^s$ is a positive definite matrix if and only if the roots of the characteristic equation

$$\lambda^2 - (\zeta_1 + \zeta_2)\lambda + \zeta_1\zeta_2 - \zeta_3^2 = 0 \tag{3.35}$$

are all positive, and we get the desired equivalent condition (3.34) by noticing (3.11). □

Lemma 3.8. *For the matrix* $\mathbb{T}$ *defined in Lemma 3.7, we have*

$$||\mathbb{T}|| < 3, \tag{3.36}$$

where $||\mathbb{T}||$ *denotes the spectral norm of* $\mathbb{T}$.

Proof. Let

$$c_1 = -\frac{1}{\mu_1}\left(\frac{\mu_1+\mu_2}{2} - \frac{\upsilon_2}{3}\right)\overline{\beta}_K, \quad c_2 = -\frac{1}{\mu_1}\left(\frac{\mu_1+\mu_3}{2} - \frac{\upsilon_3}{3}\right)\overline{\gamma}_K;$$

then we deduce that

$$\mathbb{T}_1^T\mathbb{T}_1 = \begin{pmatrix} 1 & c_1 & 0 \\ c_1 & c_1^2+1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbb{T}_2^T\mathbb{T}_2 = \begin{pmatrix} 1 & 0 & c_2 \\ 0 & 1 & 0 \\ c_2 & 0 & c_2^2+1 \end{pmatrix}.$$

It follows that

$$||\mathbb{T}_i|| = \lambda_{\max}\left(\mathbb{T}_i^T\mathbb{T}_i\right) = \left(\frac{c_i^2 + 2 + \sqrt{\left(c_i^2+2\right)^2 - 4}}{2}\right)^{1/2} < \sqrt{c_i^2+2}, \quad i = 1, 2,$$

where $\lambda_{\max}(\mathbb{T}_i^T\mathbb{T}_i)$ is the maximum eigenvalue of $\mathbb{T}_i^T\mathbb{T}_i$. Moreover, by (3.7), (3.12) and (3.3), we have

$$|c_1| < \frac{1}{\mu_1}\left(\frac{\mu_1+\mu_2}{2} - \frac{\upsilon_2}{3}\right) < \frac{\mu_1+\mu_2}{2\mu_1} < 1,$$

and we also have $|c_2| < 1$. As a result,

$$\|\mathbb{T}_i\| < \sqrt{3}, \quad i = 1, 2,$$

which leads to (3.36) by using the fact that $\|\mathbb{T}\| \leq \|\mathbb{T}_1\|\|\mathbb{T}_2\|$. □

Lemma 3.9. *Under the assumption (A1), $\mathbb{B}_K^s$ is uniformly positive definite, that is*

$$\boldsymbol{v}^T \mathbb{B}_K^s \boldsymbol{v} \geq \frac{8C_{qr}\varrho}{9\overline{\lambda}} \|\boldsymbol{v}\|^2, \quad \forall \boldsymbol{v} \in \mathbb{R}^3. \tag{3.37}$$

Proof. It follows from (3.33) that

$$\boldsymbol{v}^T \mathbb{B}_K^s \boldsymbol{v} = \left(\mathbb{T}^{-1}\boldsymbol{v}\right)^T \left(\mathbb{T}^T \mathbb{B}_K^s \mathbb{T}\right)\left(\mathbb{T}^{-1}\boldsymbol{v}\right) \geq \lambda_K \left\|\mathbb{T}^{-1}\boldsymbol{v}\right\|^2,$$

where $\lambda_K = \min\{\mu_1, 4\lambda_K'\}$ and λ_K' is the minimum root of characteristic equation (3.35), given by

$$\lambda_K' = \frac{\zeta_1 + \zeta_2 - \sqrt{(\zeta_1 + \zeta_2)^2 - 4\left(\zeta_1\zeta_2 - \zeta_3^2\right)}}{2}.$$

From (3.7) and (3.3), we have

$$\mu_2 \geq 2m_{11}, \quad \mu_3 \geq 2m_{22},$$

and by using (3.4),

$$\lambda_K' \leq \frac{\zeta_1 + \zeta_2}{2} \leq \frac{m_{11} + m_{22}}{2} \leq \frac{\mu_1}{4},$$

which implies that $\lambda_K = 4\lambda_K'$. Moreover, it holds that

$$\lambda_K' = \frac{2\left(\zeta_1\zeta_2 - \zeta_3^2\right)}{\zeta_1 + \zeta_2 + \sqrt{(\zeta_1 + \zeta_2)^2 - 4\left(\zeta_1\zeta_2 - \zeta_3^2\right)}} \geq \frac{\zeta_1\zeta_2 - \zeta_3^2}{\zeta_1 + \zeta_2} \geq \frac{\zeta_1\zeta_2 - \zeta_3^2}{m_{11} + m_{22}} > \frac{2C_{qr}\varrho}{\overline{\lambda}},$$

where we have used the facts of (3.5) and (3.10) in the last inequality. By (3.36), we find that

$$\|\mathbb{T}^{-1}\boldsymbol{v}\| \geq \frac{1}{\|\mathbb{T}\|}\|\boldsymbol{v}\| \geq \frac{1}{3}\|\boldsymbol{v}\|.$$

Combining the above facts, we get the desired result (3.37). □

The proof of Theorem 3.1. It follows from (2.13), (3.32), (3.37) and (3.18) that

$$\widetilde{a}_h\left(u_h, \Pi_h^* u_h\right) = \sum_{K\in\mathcal{T}_h} \widetilde{a}_{K,h}\left(u_h, \Pi_h^* u_h\right) = \sum_{K\in\mathcal{T}_h} \widetilde{\boldsymbol{w}}_K^T \mathbb{B}_K^s \widetilde{\boldsymbol{w}}_K \geq \frac{8C_{qr}\varrho}{9\overline{\lambda}} \sum_{K\in\mathcal{T}_h} \|\widetilde{\boldsymbol{w}}_K\|^2 \geq \frac{8C_{qr}\varrho}{9\overline{\lambda}\widetilde{C}^2}|u_h|_1^2;$$

we obtain (3.6) with $\kappa = (8C_{qr}\varrho)/(9\overline{\lambda}\widetilde{C}^2)$ and complete the proof of Theorem 3.1. □

4. Discussions on some special meshes

By Theorem 3.1, one can see that the assumption **(A1)** plays an important role in our coercivity result of the Q_1-FVEM-SR scheme. However, the meaning of (3.5) is not so straightforward, since it involves the anisotropic diffusion tensor Λ_K and the geometry of the general convex quadrilateral cell K. In this section, we employ some special meshes to explore the meaning of **(A1)**, including the parallelogram, $h^{1+\gamma}$-parallelogram and trapezoidal meshes.

4.1. Parallelogram mesh

Theorem 4.1. *Suppose that $\mathcal{T}_h$ consists of parallelograms; then, under the assumption* (2.3), ***(A1)*** *holds with*

$$\varrho = \min_{K\in\mathcal{T}_h}\left[\frac{1}{16}\det(\Lambda_K)\right] \geq \frac{1}{16}\underline{\lambda}^2. \tag{4.1}$$

Proof. If $K \in \mathcal{T}_h$ is a parallelogram, then by (2.8) and (2.9), we obtain

$$\boldsymbol{m}_K = \boldsymbol{0}, \quad \overline{\beta}_K = \overline{\gamma}_K = 0. \tag{4.2}$$

It follows from (3.4) and (3.9) that

$$\zeta_1\zeta_2 - \zeta_3^2 = m_{11}m_{22} - m_{12}^2 = \frac{1}{16}\det(\Lambda_K).$$

Thus, by recalling (2.3), we obtain (4.1) and complete the proof. □

4.2. $h^{1+\gamma}$-parallelogram mesh

Theorem 4.2. *Suppose that $\mathcal{T}_h$ consists of $h^{1+\gamma}$-parallelograms, namely there exists a positive constant C_1 such that*

$$\|\boldsymbol{m}_K\| \leq C_1 h_K^{1+\gamma}, \quad \forall K \in \mathcal{T}_h, \tag{4.3}$$

where $\gamma > 0$ is a constant. Moreover, we assume that (2.3) *and* (2.6) *hold. Consequently, when h is sufficiently small, we have*

$$\left|\left(\zeta_1\zeta_2 - \zeta_3^2\right) - \frac{1}{16}\det\left(\Lambda_K\right)\right| \leq C_2 h_K^{2\gamma}, \quad \forall K \in \mathcal{T}_h,$$

where C_2 is a positive constant independent of K and h.

Proof. Let

$$a_1 = -\frac{1}{4\mu_1}\left(\frac{\mu_3}{2} + \frac{\upsilon_2}{3}\right)^2 \overline{\beta}_K^2, \quad a_2 = -\frac{1}{4\mu_1}\left(\frac{\mu_2}{2} + \frac{\upsilon_3}{3}\right)^2 \overline{\gamma}_K^2, \quad a_3 = \frac{1}{4\mu_1}\left(\frac{\mu_3}{2} + \frac{\upsilon_2}{3}\right)\left(\frac{\mu_2}{2} + \frac{\upsilon_3}{3}\right)\overline{\beta}_K\overline{\gamma}_K.$$

Then, by (3.4) and (3.9), we find that

$$\zeta_1\zeta_2 - \zeta_3^2 = (m_{11} + a_1)(m_{22} + a_2) - (m_{12} + a_3)^2 = \frac{1}{16}\det(\Lambda_K) + Res,$$

where

$$Res = a_2m_{11} + a_1m_{22} + a_1a_2 - 2a_3m_{12} - a_3^2.$$

It follows from (2.9), (4.3) and (2.6) that

$$\left|\overline{\beta}_K\right|, \left|\overline{\gamma}_K\right| \leq \frac{C_1}{C_{qr}} h_K^{\gamma},$$

and when h is sufficiently small, we deduce from (3.2) and (3.3) that

$$\upsilon_2 < \frac{4}{3} m_{11}, \quad \mu_3 < \frac{2}{1-\overline{\gamma}_K^2} m_{22} < 4m_{22}.$$

As a result, by (3.13) and (3.10),

$$|a_1| < \frac{C_1^2}{8C_{qr}^2}\left(\frac{\mu_3}{2} + \frac{\upsilon_2}{3}\right) h_K^{2\gamma} < \frac{C_1^2}{4C_{qr}^2}\left(\frac{2}{9} m_{11} + m_{22}\right) h_K^{2\gamma} < \frac{11\overline{\lambda} C_1^2}{144 C_{qr}^3} h_K^{2\gamma}.$$

Similarly, $|a_2|$ and $|a_3|$ can be bounded by $h_K^{2\gamma}$. By using (3.10) again, there exists a constant $C_2 > 0$ such that $|Res| \leq C_2 h_K^{2\gamma}$, and this completes the proof. □

Remark 4.1. *By Theorem 4.2, we find that for the $h^{1+\gamma}$-parallelogram mesh, **(A1)** holds with $\varrho = C_0 \underline{\lambda}^2$, where $0 < C_0 < 1/16$ is a constant.*

4.3. Trapezoidal mesh

Theorem 4.3. *Suppose that $\mathcal{T}_h$ consists of trapezoids, and that, for each $K \in \mathcal{T}_h$, the lengths of the two bottoms are denoted as L_b and L_t; see Figure 3. Moreover, we define the ratio $\tau = L_b/L_t$ or $\tau = L_t/L_b$ and assume that* (2.3) *and* (2.6) *hold. Then, for any trapezoidal cell K, if*

$$\left|\frac{1-\tau}{1+\tau}\right| < \frac{6\sqrt{26} C_{qr} \underline{\lambda}}{13\overline{\lambda}}, \tag{4.4}$$

$\mathbb{B}_K^s$ is a positive definite matrix.

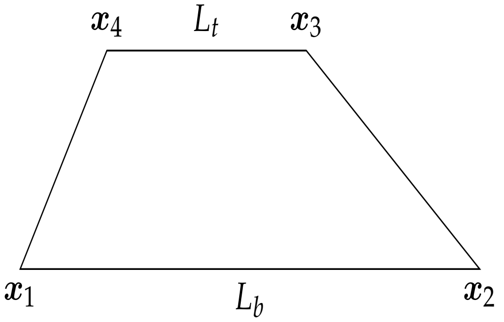


Figure 3. A general trapezoidal cell used in Theorem 4.3.

Proof. Without loss of generality, we assume that $x_1x_2 // x_4x_3$. Then, by (2.9) and (3.3), we obtain

$$\overline{\beta}_K = 0, \quad \overline{\gamma}_K = \frac{L_t - L_b}{L_t + L_b}, \quad \mu_2 = 2m_{11}.$$

It follows from (3.4), (3.9), (2.3), (3.12), (3.7) and (3.10) that

$$\begin{aligned}\zeta_1\zeta_2-\zeta_3^2 =& m_{11}\left(m_{22}-\frac{1}{4\mu_1}\left(\frac{\mu_2}{2}+\frac{\upsilon_3}{3}\right)^2\overline{\gamma}_K^2\right)-m_{12}^2=\frac{1}{16}\det\left(\Lambda_K\right)-\frac{m_{11}}{4\mu_1}\left(\frac{\mu_2}{2}+\frac{\upsilon_3}{3}\right)^2\overline{\gamma}_K^2\\ \geq&\frac{1}{16}\underline{\lambda}^2-\frac{m_{11}}{8}\left(\frac{\mu_2}{2}+\frac{\upsilon_3}{3}\right)\overline{\gamma}_K^2=\frac{1}{16}\underline{\lambda}^2-\frac{m_{11}}{8}\left(m_{11}+\frac{4}{3\left(4-\overline{\gamma}_K^2\right)}m_{22}\right)\overline{\gamma}_K^2\\ \geq&\frac{1}{16}\underline{\lambda}^2-\frac{m_{11}}{8}\left(m_{11}+\frac{4}{9}m_{22}\right)\overline{\gamma}_K^2\geq\frac{1}{16}\left(\underline{\lambda}^2-\frac{13\overline{\lambda}^2}{72C_{qr}^2}\overline{\gamma}_K^2\right).\end{aligned}$$

Therefore, by Lemma 3.7, we deduce that $\mathbb{B}_K^s$ is a positive definite matrix provided that (4.4) holds. The proof is complete. □

Remark 4.2. *We mention that in Theorem 4.3,* (4.4) *is just a sufficient condition to ensure the positive definiteness of* $\mathbb{B}_K^s$*. As a special case, if K is a parallelogram, then $\tau = 1$, implies that* (4.4) *holds. In other words, the result of Theorem 4.3 covers parallelogram mesh.*

5. H^1 error estimate

Under the assumption (2.5), there exist two positive constants C_3 and C_4 such that

$$|u-\Pi_h u|_1\leq C_3h|u|_2,\quad \forall u\in H_0^1(\Omega)\cap H^2(\Omega),\tag{5.1}$$

and

$$|a_h(\Pi_hu-u,\Pi_h^*w_h)|\leq C_4h|u|_2|w_h|_1,\quad \forall u\in H_0^1(\Omega)\cap H^2(\Omega),\ w_h\in H_0^1(\Omega),\tag{5.2}$$

where $\Pi_h u\in U_h$ is the isoparametric bilinear interpolation of u, satisfying $\Pi_h u(\boldsymbol{x}_i)=u(\boldsymbol{x}_i)$. A proof of (5.1) can be found in [37], while that for (5.2) is given in [15]. Moreover, in order to present the optimal H^1 error estimate, we need the following assumption.

(A2) There exists a positive constant C_5, independent of h, such that

$$\left|\widetilde{a}_h(u_h,\Pi_h^*w_h)-a_h(u_h,\Pi_h^*w_h)\right|\leq C_5h|u_h|_1|w_h|_1,\quad \forall u_h,w_h\in U_h.\tag{5.3}$$

Theorem 5.1. *Assume that $u\in H_0^1(\Omega)\cap H^2(\Omega)$ is the exact solution of* (2.1) *and* (2.2)*, $u_h\in U_h$ is the Q_1-FVEM-SR solution of* (2.12) *and $h\leq 1$. Then, under the assumptions* (2.3), (2.5)*, **(A1)** and **(A2)**, we have*

$$|u-u_h|_1\leq C_6h\|u\|_2,$$

where

$$C_6=C_3+\frac{1}{\kappa}(C_4+C_5+C_3C_5).$$

Proof. It follows from (3.6) that

$$|\Pi_hu-u_h|_1^2\leq\frac{1}{\kappa}\widetilde{a}_h(\Pi_hu-u_h,\Pi_h^*(\Pi_hu-u_h))$$

and then

$$|\Pi_hu-u_h|_1\leq\frac{1}{\kappa}\sup_{w_h\in U_h}\frac{\widetilde{a}_h(\Pi_hu-u_h,\Pi_h^*w_h)}{|w_h|_1}.\tag{5.4}$$

By (5.2), (5.3) and (5.1), we obtain

$$\begin{aligned}\left|\widetilde{a}_h(\Pi_h u - u_h, \Pi_h^* w_h)\right| &\leq |a_h(\Pi_h u - u, \Pi_h^* w_h)| + |\widetilde{a}_h(\Pi_h u - u_h, \Pi_h^* w_h) - a_h(\Pi_h u - u, \Pi_h^* w_h)| \\ &= |a_h(\Pi_h u - u, \Pi_h^* w_h)| + |\widetilde{a}_h(\Pi_h u, \Pi_h^* w_h) - a_h(\Pi_h u, \Pi_h^* w_h)| \\ &\leq h|w_h|_1 \left(C_4|u|_2 + C_5|\Pi_h u|_1\right) \leq h|w_h|_1 \left[C_4|u|_2 + C_5(|\Pi_h u - u|_1 + |u|_1)\right] \\ &\leq h|w_h|_1 \left[C_4|u|_2 + C_5(C_3 h|u|_2 + |u|_1)\right] \\ &\leq (C_4 + C_5 + C_3C_5)h\|u\|_2|w_h|_1,\end{aligned}$$

where the fact that

$$\widetilde{a}_h(u_h, \Pi_h^* w_h) = (f, \Pi_h^* w_h) = a_h(u, \Pi_h^* w_h)$$

is used in the second equality. Then, we deduce from (5.4) that

$$|\Pi_h u - u_h|_1 \leq \frac{1}{\kappa}(C_4 + C_5 + C_3C_5)h\|u\|_2,$$

which implies that

$$|u - u_h|_1 \leq |u - \Pi_h u|_1 + |\Pi_h u - u_h|_1 \leq C_6 h\|u\|_2.$$

The proof is complete. □

By Theorem 5.1, we observe that the assumption **(A2)** plays an important role in the optimal H^1 error estimate of the Q_1-FVEM-SR scheme. In the rest of this section, we explore the meaning of **(A2)** for some special meshes, including the parallelogram and $h^{1+\gamma}$-parallelogram; see Theorem 5.2 and 5.3, respectively.

Theorem 5.2. *Suppose that $\mathcal{T}_h$ consists of parallelograms; then,* ***(A2)*** *holds with*

$$\widetilde{a}_h(u_h, \Pi_h^* w_h) = a_h(u_h, \Pi_h^* w_h), \quad \forall u_h, w_h \in U_h. \tag{5.5}$$

Proof. If $K \in \mathcal{T}_h$ is a parallelogram, then it follows from (4.2) and (2.10) that

$$\mathbb{J}_K^{-1}(\xi, \eta) = \frac{2}{|K|}\mathcal{R}(\boldsymbol{m}_2, -\boldsymbol{m}_1).$$

Note that

$$\nabla u_h = \mathbb{J}_K^{-1}(\xi, \eta)\widehat{\nabla u_h}$$

and Λ_K is a constant matrix, which leads to $(\Lambda_K \nabla u_h) \cdot \boldsymbol{n}_i^*$ is a linear function on each edge of K_i^*. Since the Simpson rule is exact for polynomials of degree not greater than 3, then we obtain

$$\widetilde{a}_{K,h}(u_h, \Pi_h^* w_h) = a_{K,h}(u_h, \Pi_h^* w_h),$$

which implies (5.5) and completes the proof. □

Theorem 5.3. *Suppose that $\mathcal{T}_h$ consists of $h^{1+\gamma}$-parallelograms, namely* (4.3) *is satisfied. Moreover, we assume that* (2.3) *and* (2.6) *hold. Consequently, when h is sufficiently small, we have*

$$\left|\widetilde{a}_h(u_h, \Pi_h^* w_h) - a_h(u_h, \Pi_h^* w_h)\right| \leq \left(3 + 2\sqrt{3}\right)\frac{C_1\overline{\lambda}}{C_{qr}^2\underline{C}^2}h^\gamma|u_h|_1|w_h|_1, \quad \forall u_h, w_h \in U_h. \tag{5.6}$$

Proof. For the $h^{1+\gamma}$-parallelogram K, assume that its two vectors $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ form a parallelogram K', and that $\boldsymbol{x}'_i$ $(i = 1,2,3,4)$ denotes the four vertices of K'; see Figure 4. For simplicity, we denote u_h^K as the restriction of u_h on K. Moreover, let $u_h^{K'}$ be the isoparametric bilinear function on K, satisfying $u_h^{K'}(\boldsymbol{x}'_i) = u_h^K(\boldsymbol{x}_i)$, $i = 1,2,3,4$, and $\Lambda_{K'} = \Lambda_K$. Then, by using the facts that

$$|K'| = \boldsymbol{m}_1 \cdot (\mathcal{R}\boldsymbol{m}_2) = |K|, \quad \overline{\beta}_{K'} = \overline{\gamma}_{K'} = 0,$$

we deduce from (3.29) that

$$\nabla u_h^{K'} = \frac{1}{|K|}\left(\mathbb{R}_K \widetilde{\boldsymbol{u}}_K + \mathcal{R}(\eta \boldsymbol{m}_2 - \xi \boldsymbol{m}_1)\boldsymbol{S}_{K'}^T \widetilde{\boldsymbol{u}}_K\right),$$

which leads to

$$\nabla u_h^K - \nabla u_h^{K'} = \frac{\mathcal{R}(\eta \boldsymbol{m}_2 - \xi \boldsymbol{m}_1)}{|K|}\left(\frac{\boldsymbol{S}_K^T \widetilde{\boldsymbol{u}}_K}{1 + \overline{\beta}_K \xi + \overline{\gamma}_K \eta} - \boldsymbol{S}_{K'}^T \widetilde{\boldsymbol{u}}_K\right).$$

A direct calculation yields that

$$J_1 := \left|a_{K,h}(u_h, \Pi_h^* w_h) - a_{K',h}(u_h, \Pi_h^* w_h)\right| = \left|\sum_{i=1}^4 E_i\right|,$$

where

$$E_i = \frac{w_h(\boldsymbol{x}_{i+1}) - w_h(\boldsymbol{x}_i)}{\|\boldsymbol{x}_K - \boldsymbol{y}_i\|}\int_{\boldsymbol{x}_K\boldsymbol{y}_i} (\boldsymbol{x}_K - \boldsymbol{y}_i)^T \mathcal{R}^T \Lambda_K \left(\nabla u_h^K - \nabla u_h^{K'}\right) \mathrm{d}s.$$

When the mesh size h is sufficiently small, it follows from (3.10), (3.19) and (3.21) that

$$\begin{aligned} |E_1| &= \left| m_{22}(w_h(\boldsymbol{x}_2) - w_h(\boldsymbol{x}_1)) \int_{-1}^0 \eta \left(\frac{\boldsymbol{S}_K^T \widetilde{\boldsymbol{u}}_K}{1 + \overline{\gamma}_K \eta} - \boldsymbol{S}_{K'}^T \widetilde{\boldsymbol{u}}_K\right) \mathrm{d}\eta \right| \\ &\le \frac{\overline{\lambda}}{2C_{qr}} |w_h|_{1,K,h} \int_{-1}^0 \left|\boldsymbol{S}_K^T \widetilde{\boldsymbol{u}}_K - (1 + \overline{\gamma}_K \eta)\, \boldsymbol{S}_{K'}^T \widetilde{\boldsymbol{u}}_K\right| \mathrm{d}\eta \\ &\le \frac{\sqrt{3} C_1 \overline{\lambda}}{2 C_{qr}^2 \underline{C}^2} h_K^\gamma |u_h|_{1,K} |w_h|_{1,K}, \end{aligned} \tag{5.7}$$

where we have used the fact that

$$\left|\boldsymbol{S}_K^T \widetilde{\boldsymbol{u}}_K - (1 + \overline{\gamma}_K \eta) \boldsymbol{S}_{K'}^T \widetilde{\boldsymbol{u}}_K\right| = \left|(-\overline{\gamma}_K \eta,\, \overline{\beta}_K,\, \overline{\gamma}_K) \widetilde{\boldsymbol{w}}_K\right| \le \frac{\sqrt{3} C_1}{C_{qr}} h_K^\gamma \|\widetilde{\boldsymbol{w}}_K\|, \quad \forall \eta \in [-1, 0].$$

Similarly, the above inequality (5.7) holds for any E_i. This yields that

$$J_1 \le \sum_{i=1}^4 |E_i| \le \frac{2\sqrt{3} C_1 \overline{\lambda}}{C_{qr}^2 \underline{C}^2} h_K^\gamma |u_h|_{1,K} |w_h|_{1,K}.$$

On the other hand, we have

$$J_2 := \left|\widetilde{a}_{K,h}(u_h, \Pi_h^* w_h) - a_{K',h}(u_h, \Pi_h^* w_h)\right| = \left|\sum_{i=1}^4 (w_h(\boldsymbol{x}_{i+1}) - w_h(\boldsymbol{x}_i))(F_i - G_i)\right|,$$

where

$$F_i = \frac{1}{6}(\boldsymbol{x}_K - \boldsymbol{y}_i)^T\mathcal{R}^T\Lambda_K\left(\nabla u_h^K(\boldsymbol{x}_K) + 4\nabla u_h^K(\boldsymbol{z}_i) + \nabla u_h^K(\boldsymbol{y}_i)\right)$$

and

$$G_i = \frac{1}{\|\boldsymbol{x}_K - \boldsymbol{y}_i\|}\int_{\boldsymbol{x}_K\boldsymbol{y}_i}(\boldsymbol{x}_K - \boldsymbol{y}_i)^T\mathcal{R}^T\Lambda_K\nabla u_h^{K'}\,\mathrm{d}s.$$

By a direct calculation, we find that

$$F_1 - G_1 = m_{22}\left(\widetilde{u}_2 - a_4\boldsymbol{S}_K^T\widetilde{\boldsymbol{u}}_K\right),$$

where

$$a_4 = \frac{1}{3}\left(\frac{4}{2-\overline{\gamma}_K} + \frac{1}{1-\overline{\gamma}_K}\right).$$

Note that $0 < a_4 < 2$ and

$$|1 - a_4| = \left|\frac{4 - 3\overline{\gamma}_K}{3\,(1-\overline{\gamma}_K)\,(2-\overline{\gamma}_K)}\overline{\gamma}_K\right| < \left|\overline{\gamma}_K\right| \leq \frac{C_1}{C_{qr}}h_K^\gamma.$$

As a result

$$|F_1 - G_1| \leq \frac{\overline{\lambda}}{4C_{qr}}\left|\widetilde{u}_2 - a_4\boldsymbol{S}_K^T\widetilde{\boldsymbol{u}}_K\right| \leq \frac{3C_1\overline{\lambda}}{4C_{qr}^2\underline{C}}h_K^\gamma|u_h|_{1,K}, \tag{5.8}$$

where we have used the fact that

$$\left|\widetilde{u}_2 - a_4\boldsymbol{S}_K^T\widetilde{\boldsymbol{u}}_K\right| = \left|(1-a_4)\widetilde{u}_2 - a_4\overline{\beta}_K\widetilde{u}_3 - a_4\overline{\gamma}_K\widetilde{u}_4\right| \leq \frac{3C_1}{C_{qr}}h_K^\gamma\|\widetilde{\boldsymbol{w}}_K\|.$$

By the same arguments, the estimate (5.8) holds for any $F_i - G_i$. This yields that

$$J_2 \leq |w_h|_{1,K,h}\sum_{i=1}^4|F_i - G_i| \leq \frac{3C_1\overline{\lambda}}{C_{qr}^2\underline{C}^2}h_K^\gamma|u_h|_{1,K}|w_h|_{1,K}.$$

Combining the above results, we obtain

$$\left|\widetilde{a}_{K,h}(u_h,\Pi_h^*w_h) - a_{K,h}(u_h,\Pi_h^*w_h)\right| \leq J_1 + J_2 \leq \left(3 + 2\sqrt{3}\right)\frac{C_1\overline{\lambda}}{C_{qr}^2\underline{C}^2}h_K^\gamma|u_h|_{1,K}|w_h|_{1,K}.$$

Finally, by using the fact that

$$\left|\widetilde{a}_h(u_h,\Pi_h^*w_h) - a_h(u_h,\Pi_h^*w_h)\right| \leq \sum_{K\in\mathcal{T}_h}\left|\widetilde{a}_{K,h}(u_h,\Pi_h^*w_h) - a_{K,h}(u_h,\Pi_h^*w_h)\right|$$

and the Cauchy-Schwarz inequality, we obtain (5.6) and complete the proof. □

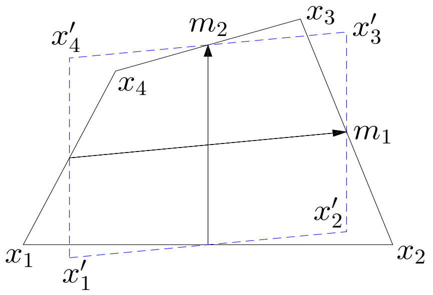


Figure 4. The $h^{1+\gamma}$-parallelogram $K = \square \boldsymbol{x}_1\boldsymbol{x}_2\boldsymbol{x}_3\boldsymbol{x}_4$ (solid lines) and associated parallelogram $K' = \square \boldsymbol{x}_1'\boldsymbol{x}_2'\boldsymbol{x}_3'\boldsymbol{x}_4'$ (dotted lines), which is used in Theorem 5.3.

Remark 5.1. *In Theorem 5.3, if $\gamma \geq 1$, then* (5.6) *implies* (5.3). *That is, the assumption* **(A2)** *holds on $h^{1+\gamma}$-parallelogram mesh with $\gamma \geq 1$.*

Remark 5.2. *We mention that the coercivity results in [34, 35] do not cover arbitrary trapezoidal meshes, and, based on the coercivity results, [34, 35] proved the optimal H^1 error estimate. Thus, the error analysis in [34, 35] does not hold for arbitrary convex quadrilateral meshes; it also needs some mesh assumptions.*

6. Numerical examples

We present several examples to verify the theoretical findings of an isoparametric bilinear FVEM based on the Simpson formula, including the H^1 error and coercivity result. Examples 6.1, 6.2, 6.3 and 6.4 have been designed for scalar, discontinuous, anisotropic diffusion and variable coefficients, respectively. However, Example 6.5 has been constructed to show that the assumption **(A1)** is just a sufficient condition to guarantee the coercivity result. For simplicity, we denote $e_i = u(\boldsymbol{x}_i) - u_h(\boldsymbol{x}_i)$ as the error of the solution at vertex $\boldsymbol{x}_i$. Then, the discrete H^1 error and convergence rate are respectively defined by

$$E_u = \left(\sum_{K \in \mathcal{T}_h} \sum_{i=1}^{4} (e_{i+1} - e_i)^2 \right)^{1/2}, \quad R_u = \frac{\log[E_u(h_2)/E_u(h_1)]}{\log(h_2/h_1)},$$

where h_1, h_2 denote the mesh sizes of two successive meshes and $E_u(h_1)$, $E_u(h_2)$ are the corresponding errors. Moreover, in order to investigate the coercivity of the scheme numerically, we define

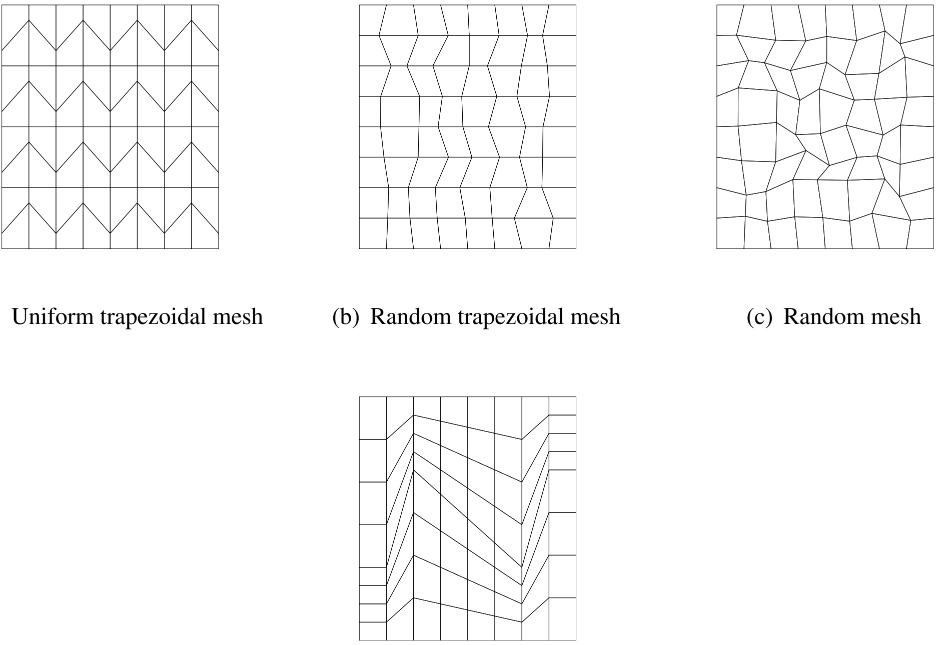
$$\varrho = \min_{K \in \mathcal{T}_h} \left\{ \zeta_1\zeta_2 - \zeta_3^2 \right\}, \quad Coer = \frac{\widetilde{a}_h(u_h, \Pi_h^* u_h)}{|u_h|_{1,h}^2}, \quad |u_h|_{1,h}^2 = \sum_{K \in \mathcal{T}_h} |u_h|_{1,K,h}^2,$$

where $|u_h|_{1,K,h}$ is defined by (3.20). Then E_u and $|u_h|_{1,h}$ are equivalent to $|\Pi_h u - u_h|_1$ and $|u_h|_1$ respectively.

Four types of meshes were used in our experiments; see Figure 5. The first type is a uniform trapezoidal mesh (Figure 5(a)), which is obtained by moving some interior vertices of the corresponding uniform square meshes along the longitudinal direction. The random mesh (Figure 5(c)) was constructed from the uniform square mesh by applying a random distortion of the interior vertices as follows

$$x := x + \omega r_x h, \quad y := y + \omega r_y h,$$

where $\omega \in (0, 0.5)$ is the degree of distortion, r_x and r_y are two random numbers that belong to $[-1, 1]$ and h is the mesh size of the uniform square mesh. The random trapezoidal mesh (Figure 5(b)) is distorted only in the x direction. The last is the Kershaw mesh (Figure 5(d)); its description can be found in [38], and it is a quasi-parallelogram mesh. It can be checked that the meshes in Figure 5(a), (b) and (c) are not $h^{1+\gamma}$-parallelogram. Here we generally choose $\omega = 0.3$ and $\Omega = [0, 1]^2$.



(a) Uniform trapezoidal mesh

(b) Random trapezoidal mesh

(c) Random mesh

(d) Kershaw mesh

Figure 5. Four mesh types used in the numerical tests.

Example 6.1. *Solve* (2.1)*, where* Λ *is the identity matrix. The exact solution is given by*

$$u(x, y) = \sin(2\pi x)\sin(\pi y)e^{x^2+y},$$

and the source term f *is determined by* Λ *and* u*.*

The values of ϱ *are presented in Table 1; one can see that they are all greater than* 0 *and do not tend to* 0 *with the refinement of grids. That is,* ***(A1)*** *is satisfied for the four mesh types. From Table 1, we also find that the values of Coer have a positive lower bound that is independent of h, i.e., the theoretical finding in Theorem 3.1 is verified. The numerical results of* H^1 *error are given in Table 2, where a first order convergence can be explicitly observed, which validates the theoretical result of Theorem 5.1. Note that for the Kershaw mesh, the* H^1 *error order is 2, and there is a superconvergence phenomenon. However, for the uniform trapezoidal mesh, the superconvergence cannot always be expected; see the following examples. The reason is that Kershaw mesh is a quasi-parallelogram mesh, but the uniform trapezoidal mesh is not. We remark that the scheme constructed in this work is identical to the classical* Q_1*−FVEM for uniform rectangular mesh, and the corresponding superconvergence has been proved by some researchers (e.g., [19]).*

Table 1. Numerical coercivity results for Example 6.1.

Mesh		8×8	16×16	32×32	64×64	128×128
Uniform trapezoidal mesh	*Coer*	0.443	0.435	0.433	0.433	0.433
	ϱ	6.083e-002	6.083e-002	6.083e-002	6.083e-002	6.083e-002
Random trapezoidal mesh	*Coer*	0.443	0.465	0.457	0.459	0.458
	ϱ	6.115e-002	6.093e-002	6.070e-002	6.054e-002	6.051e-002
Random mesh	*Coer*	0.435	0.457	0.447	0.449	0.448
	ϱ	5.930e-002	5.968e-002	5.743e-002	5.599e-002	5.717e-002
Kershaw mesh	*Coer*	0.486	0.467	0.453	0.448	0.446
	ϱ	6.056e-002	6.178e-002	6.227e-002	6.243e-002	6.248e-002

Table 2. H^1 errors and convergence rates for Example 6.1.

Mesh		8×8	16×16	32×32	64×64	128×128
Uniform trapezoidal mesh	E_u	0.590	0.152	3.946e-002	1.057e-002	3.179e-003
	R_u	-	1.957	1.946	1.900	1.733
Random trapezoidal mesh	E_u	0.549	0.165	7.801e-002	3.967e-002	2.003e-002
	R_u	-	1.709	1.150	0.975	0.984
Random mesh	E_u	0.677	0.218	0.105	5.599e-002	2.792e-002
	R_u	-	1.589	1.137	0.946	1.001
Kershaw mesh	E_u	2.188	1.095	0.401	0.120	3.194e-002
	R_u	-	1.155	1.548	1.802	1.933

Example 6.2. *To verify the validity and efficiency of the numerical scheme, many researchers adopted the discontinuous coefficient for some general quadrilateral meshes [29, 34, 39]. Here we also solve the problem* (2.1) *with the following discontinuous coefficient and exact solution*

$$\Lambda(x,y) = \begin{cases} 1, & x \le 0.5, \\ 4, & x > 0.5. \end{cases}, \quad u(x,y) = \begin{cases} y^4 - 2y^2 + 4xy + 2y + 6x + 1, & x \le 0.5, \\ y^4 - 2y^2 + xy + 3.5y + 1.5x + 3.25, & x > 0.5. \end{cases}$$

Note that Λ *is discontinuous across the line* $x = 0.5$*. Thus, in this example, for the random trapezoidal and random quadrilateral meshes, all of the vertices on the line* $x = 0.5$ *are only allowed to be distorted in the y direction. The coercivity results,* H^1 *errors and the corresponding convergence rates are presented in Tables 3 and 4. One can see that although the diffusion coefficient is discontinuous, the numerical performance of the* Q_1*-FVEM-SR scheme is similar to that of the previous Example 6.1. Moreover, the superconvergence result can be observed for the Kershaw mesh. The reason is that the discontinuity of the diffusion coefficient is fitted with the boundary of quadrilaterals.*

Table 3. Numerical coercivity results for Example 6.2.

Mesh		8×8	16×16	32×32	64×64	128×128
Uniform trapezoidal mesh	$Coer$	0.688	0.680	0.676	0.674	0.673
	ϱ	6.083e-002	6.083e-002	6.083e-002	6.083e-002	6.083e-002
Random trapezoidal mesh	$Coer$	0.692	0.700	0.694	0.696	0.696
	ϱ	6.123e-002	6.093e-002	6.096e-002	6.054e-002	6.056e-002
Random mesh	$Coer$	0.676	0.688	0.682	0.684	0.683
	ϱ	6.036e-002	5.968e-002	5.912e-002	5.790e-002	5.717e-002
Kershaw mesh	$Coer$	0.627	0.633	0.635	0.635	0.635
	ϱ	6.056e-002	6.178e-002	6.227e-002	6.243e-002	6.248e-002

Table 4. H^1 errors and convergence rates for Example 6.2.

Mesh		8×8	16×16	32×32	64×64	128×128
Uniform trapezoidal mesh	E_u	1.968e-002	6.017e-003	2.180e-003	9.503e-004	4.556e-004
	R_u	-	1.709	1.467	1.198	1.061
Random trapezoidal mesh	E_u	1.956e-002	9.294e-003	5.434e-003	2.835e-003	1.415e-003
	R_u	-	1.058	0.824	0.938	1.001
Random mesh	E_u	2.979e-002	1.729e-002	8.919e-003	4.720e-003	2.383e-003
	R_u	-	0.764	1.003	0.982	0.983
Kershaw mesh	E_u	0.151	8.139e-002	3.368e-002	1.065e-002	2.916e-003
	R_u	-	1.036	1.360	1.714	1.897

Example 6.3. *We still solve the problem* (2.1)*, choosing the anisotropic diffusion tensor and analytic solution as follows*

$$\Lambda(x,y) = \begin{pmatrix} 1.5 & 0.5 \\ 0.5 & 1.5 \end{pmatrix}, \quad u(x,y) = \frac{1}{2}\left[\frac{\sin((1-x)(1-y))}{\sin 1} + (1-x)^3(1-y)^2\right].$$

*The numerical results are presented in Tables 5 and 6, showing the first order convergence for H^1 errors and the satisfaction of **(A1)**. One can see that the numerical performance is similar to the previous two examples.*

Table 5. Numerical coercivity results for Example 6.3.

Mesh		8×8	16×16	32×32	64×64	128×128
Uniform trapezoidal mesh	$Coer$	0.672	0.696	0.709	0.715	0.718
	ϱ	0.121	0.121	0.121	0.121	0.121
Random trapezoidal mesh	$Coer$	0.924	0.908	0.900	0.903	0.902
	ϱ	0.122	0.121	0.121	0.121	0.121
Random mesh	$Coer$	0.903	0.872	0.867	0.866	0.862
	ϱ	0.119	0.117	0.115	0.110	0.112
Kershaw mesh	$Coer$	0.543	0.545	0.545	0.544	0.544
	ϱ	0.121	0.124	0.125	0.125	0.125

Table 6. H^1 errors and convergence rates for Example 6.3.

Mesh		8×8	16×16	32×32	64×64	128×128
Uniform trapezoidal mesh	E_u	9.073e-003	4.362e-003	2.163e-003	1.082e-003	5.421e-004
	R_u	-	1.057	1.012	0.999	0.997
Random trapezoidal mesh	E_u	8.356e-003	5.037e-003	2.797e-003	1.393e-003	7.348e-004
	R_u	-	0.720	0.903	1.004	0.921
Random mesh	E_u	1.888e-002	6.826e-003	3.350e-003	1.785e-003	9.430e-004
	R_u	-	0.777	1.106	0.947	0.918
Kershaw mesh	E_u	7.837e-002	3.140e-002	9.812e-003	2.692e-003	7.014e-004
	R_u	-	1.527	1.792	1.925	1.970

Example 6.4. *Consider the problem* (2.1) *and we choose the following variable coefficient*

$$\Lambda(x,y)=\begin{pmatrix} 1+x & \frac{1}{4}(x+y) \\ \frac{1}{4}(x+y) & 1+y \end{pmatrix}.$$

The analytic solution and corresponding right-hand side function are respectively given by

$$u(x,y)=e^{x+y},\quad f(x,y)=-\frac{3}{2}(3+x+y)e^{x+y}.$$

Since in this example, Λ is a variable coefficient, in our numerical experiments, we let $\Lambda_K=\Lambda(\boldsymbol{x}_K)$. The numerical results are presented in Tables 7 and 8, where we can observe that the numerical performance is similar to that of the previous examples. For comparison, in Tables 9 and 10 we present the numerical results by employing the trapezoidal rule, where the definitions of Coer and ϱ are the same as in [34]. We find that the numerical performance of the Simpson rule is similar to that of the trapezoidal rule.

Table 7. Numerical coercivity results for Example 6.4 by employing the Simpson rule.

Mesh		8×8	16×16	32×32	64×64	128×128
Uniform trapezoidal mesh	*Coer*	0.753	0.738	0.730	0.726	0.724
	ϱ	6.861e-002	6.468e-002	6.275e-002	6.179e-002	6.131e-002
Random trapezoidal mesh	*Coer*	0.930	0.949	0.936	0.939	0.937
	ϱ	7.059e-002	6.640e-002	6.446e-002	6.332e-002	6.285e-002
Random mesh	*Coer*	0.864	0.917	0.895	0.892	0.889
	ϱ	7.071e-002	6.650e-002	6.429e-002	6.310e-002	6.285e-002
Kershaw mesh	*Coer*	0.570	0.576	0.577	0.577	0.577
	ϱ	6.886e-002	6.565e-002	6.407e-002	6.328e-002	6.289e-002

Table 8. H^1 errors and convergence rates for Example 6.4 by employing the Simpson rule.

Mesh		8×8	16×16	32×32	64×64	128×128
Uniform trapezoidal mesh	E_u	3.206e-002	1.313e-002	6.109e-003	2.998e-003	1.494e-003
	R_u	-	1.288	1.104	1.027	1.005
Random trapezoidal mesh	E_u	2.662e-002	1.548e-002	8.679e-003	4.295e-003	2.205e-003
	R_u	-	0.771	0.888	1.014	0.960
Random mesh	E_u	3.284e-002	2.203e-002	1.174e-002	6.227e-003	3.182e-003
	R_u	-	0.560	0.979	0.953	0.966
Kershaw mesh	E_u	0.442	0.189	6.069e-002	1.689e-002	4.440e-003
	R_u	-	1.421	1.747	1.904	1.958

Table 9. Numerical coercivity results for Example 6.4 by employing the trapezoidal rule.

Mesh		8×8	16×16	32×32	64×64	128×128
Uniform trapezoidal mesh	$Coer$	0.753	0.738	0.730	0.726	0.724
	ϱ	6.829e-002	6.437e-002	6.245e-002	6.150e-002	6.102e-002
Random trapezoidal mesh	$Coer$	0.930	0.949	0.936	0.939	0.937
	ϱ	7.059e-002	6.640e-002	6.444e-002	6.331e-002	6.283e-002
Random mesh	$Coer$	0.864	0.917	0.895	0.892	0.889
	ϱ	7.071e-002	6.649e-002	6.425e-002	6.295e-002	6.285e-002
Kershaw mesh	$Coer$	0.570	0.576	0.577	0.577	0.577
	ϱ	6.886e-002	6.565e-002	6.407e-002	6.328e-002	6.289e-002

Table 10. H^1 errors and convergence rates for Example 6.4 by employing the trapezoidal rule.

Mesh		8×8	16×16	32×32	64×64	128×128
Uniform trapezoidal mesh	E_u	3.529e-002	1.564e-002	7.552e-003	3.750e-003	1.874e-003
	R_u	-	1.174	1.050	1.010	1.001
Random trapezoidal mesh	E_u	2.662e-002	1.548e-002	8.679e-003	4.295e-003	2.205e-003
	R_u	-	0.771	0.888	1.014	0.960
Random mesh	E_u	2.985e-002	1.954e-002	1.041e-002	5.557e-003	2.836e-003
	R_u	-	0.595	0.978	0.944	0.968
Kershaw mesh	E_u	0.460	0.198	6.418e-002	1.791e-002	4.712e-003
	R_u	-	1.402	1.739	1.900	1.956

Example 6.5. *From Lemma 3.7, one can see that* (3.34) *is a necessary and sufficient condition to ensure the positive definiteness of cell matrix* $\mathbb{B}_K^s$. *However, we mention that* ***(A1)*** *is just a sufficient condition for the coercivity result, since in this work we use the cell analysis approach to prove* (3.6). *Thus, in the last example, we choose the diffusion tensor and exact solution as below*

$$\Lambda(x,y) = \begin{pmatrix} 10 & 3 \\ 3 & 1 \end{pmatrix}, \quad u(x,y) = 4.3 - 0.6x + 3.2y + 1.6xy - 2y^2.$$

From Table 11, we observe that ***(A1)*** *is invalid on the uniform trapezoidal, random and Kershaw meshes, but* $Coer > 0$ *indicates that the scheme is still coercive. Moreover, in Table 12 one can find that the numerical solution still converges to the exact solution with the optimal convergence rate under the* H^1 *norm. Therefore,* ***(A1)*** *is only a sufficient but unnecessary condition for the coercivity result.*

Table 11. Numerical coercivity results for Example 6.5.

Mesh		8 × 8	16 × 16	32 × 32	64 × 64	128 × 128
Uniform trapezoidal mesh	$Coer$	0.376	0.368	0.366	0.366	0.366
	ϱ	-0.129	-0.129	-0.129	-0.129	-0.129
Random trapezoidal mesh	$Coer$	0.629	0.626	0.625	0.625	0.624
	ϱ	5.988e-002	5.913e-002	5.897e-002	5.861e-002	5.861e-002
Random mesh	$Coer$	0.584	0.559	0.567	0.564	0.562
	ϱ	-0.210	-0.262	-0.256	-0.595	-0.446
Kershaw mesh	$Coer$	0.362	0.348	0.340	0.337	0.336
	ϱ	-9.826e-002	4.689e-003	4.407e-002	5.720e-002	6.107e-002

Table 12. H^1 errors and convergence rates for Example 6.5.

Mesh		8 × 8	16 × 16	32 × 32	64 × 64	128 × 128
Uniform trapezoidal mesh	E_u	0.124	6.481e-002	3.306e-002	1.669e-002	8.383e-003
	R_u	-	0.933	0.971	0.986	0.993
Random trapezoidal mesh	E_u	3.391e-002	1.669e-002	8.937e-003	4.487e-003	2.341e-003
	R_u	-	1.009	0.959	0.993	0.937
Random mesh	E_u	7.659e-002	4.741e-002	2.521e-002	1.198e-002	6.301e-003
	R_u	-	0.673	0.982	1.119	0.924
Kershaw mesh	E_u	0.417	0.248	0.106	3.647e-002	1.113e-002
	R_u	-	0.863	1.306	1.594	1.738

7. Conclusions

We have analyzed the coercivity and H^1 error estimate of the Q_1-FVEM-SR scheme that is obtained by using the Simpson rule to approximate the line integrals in the classical Q_1-FVEM. Based on assumption **(A1)**, we have obtained the coercivity result for the constructed scheme. More interestingly, we find that **(A1)** covers the traditional $h^{1+\gamma}$-parallelogram and some trapezoidal meshes with any full anisotropic diffusion tensor. As a result, under assumption **(A2)**, we proved that the numerical solution converges to the exact solution with the optimal convergence rate under the H^1 norm. In particular, **(A2)** covers arbitrary parallelogram and $h^{1+\gamma}$-parallelogram meshes with any anisotropic diffusion tensor, where $\gamma \geq 1$.

A counterexample is given in Example 6.5 which implies that, even if the cell matrix $\mathbb{B}_K^s$ is not positive definite, the proposed scheme can still be coercive. That is, there exists one unique numerical solution even though the assumption **(A1)** is violated. Furthermore, in Section 6 the numerical results also indicate that the Q_1-FVEM-SR solution preserves the optimal convergence rate under the H^1 error norm even though the meshes consist of trapezoids or general convex quadrilaterals (i.e., **(A2)** is not satisfied). In summary, the relaxation of mesh requirements in assumptions **(A1)** and **(A2)** should be explored in future works.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors thank the reviewers for their careful readings and valuable suggestions. This work was partially supported by the Guangdong Basic and Applied Basic Research Foundation (Nos. 2022A1515012106, 2023A1515010885), the Project of Guangdong Polytechnic Normal University (No. 2022SDKYA023) and the Scientific Research Capacity Improvement Project of the Doctoral Program Construction Unit of Guangdong Polytechnic Normal University in 2022 (No. 22GPNUZDJS31).

Conflict of interest

The authors declared that they have no conflicts of interest regarding the publication of this work.

References

1. P. Zhu, R. Li, Generalized difference methods for second order elliptic partial differential equations. II. Quadrilateral subdivision, *Numer. Math. J. Chin. Univ.*, **4** (1982), 360–375.
2. R. E. Bank, D. J. Rose, Some error estimates for the box method, *SIAM J. Numer. Anal.*, **24** (1987), 777–787. https://doi.org/10.1137/0724050
3. S. Chou, Q. Li, Error estimates in L^2, H^1 and L^∞ in covolume methods for elliptic and parabolic problems: A unified approach, *Math. Comput.*, **69** (2000), 103–120. https://doi.org/10.1090/S0025-5718-99-01192-8
4. Z. Cai, On the finite volume element method, *Numer. Math.*, **58** (1990), 713–735. https://doi.org/10.1007/BF01385651
5. I. Mishev, Finite volume element methods for non-definite problems, *Numer. Math.*, **83** (1999), 161–175. https://doi.org/10.1007/s002110050443
6. P. Chatzipantelidis, R. Lazarov, Error estimates for a finite volume element method for elliptic PDEs in nonconvex polygonal domains, *SIAM J. Numer. Anal.*, **42** (2005), 1932–1958. https://doi.org/10.1137/S0036142903427639
7. S. Chou, X. Ye, Unified analysis of finite volume methods for second order elliptic problems, *SIAM J. Numer. Anal.*, **45** (2007), 1639–1653. https://doi.org/10.1137/050643994
8. R. Li, Z. Chen, W. Wu, Generalized difference methods for differential equations: Numerical analysis of finite volume methods, New York: Marcel Dekker, 2000.
9. Y. Lin, J. Liu, M. Yang, Finite volume element methods: An overview on recent developments, *Int. J. Numer. Anal. Mod. B*, **4** (2013), 14–34.

10. Z. Zhang, Q. Zou, Some recent advances on vertex centered finite volume element methods for elliptic equations, *Sci. China Math.*, **56** (2013), 2507–2522. https://doi.org/10.1007/s11425-013-4740-8

11. J. Xu, Q. Zou, Analysis of linear and quadratic simplicial finite volume methods for elliptic equations, *Numer. Math.*, **111** (2009), 469–492. https://doi.org/10.1007/s00211-008-0189-z

12. Z. Chen, R. Li, A. Zhou, A note on the optimal L^2-estimate of the finite volume element method, *Adv. Comput. Math.*, **16** (2002), 291–303. https://doi.org/10.1023/A:1014577215948

13. R. E. Ewing, T. Lin, Y. Lin, On the accuracy of the finite volume element method based on piecewise linear polynomials, *SIAM J. Numer. Anal.*, **39** (2002), 1865–1888. https://doi.org/10.1137/S0036142900368873

14. C. Erath, D. Praetorius, Adaptive vertex-centered finite volume methods for general second-order linear elliptic partial differential equations, *IMA J. Numer. Anal.*, **39** (2019), 983–1008. https://doi.org/10.1093/imanum/dry006

15. Y. Li, R. Li, Generalized difference methods on arbitrary quadrilateral networks, *J. Comput. Math.*, **17** (1999), 653–672.

16. Z. Zhang, Q. Zou, Vertex-centered finite volume schemes of any order over quadrilateral meshes for elliptic boundary value problems, *Numer. Math.*, **130** (2015), 363–393. https://doi.org/10.1007/s00211-014-0664-7

17. T. Schmidt, Box schemes on quadrilateral meshes, *Computing*, **51** (1993), 271–292. https://doi.org/10.1007/BF02238536

18. Q. Hong, J. Wu, A Q_1-finite volume element scheme for anisotropic diffusion problems on general convex quadrilateral mesh, *J. Comput. Appl. Math.*, **372** (2020), 112732. https://doi.org/10.1016/j.cam.2020.112732

19. J. Lv, Y. Li, L^2 error estimates and superconvergence of the finite volume element methods on quadrilateral meshes, *Adv. Comput. Math.*, **37** (2012), 393–416. https://doi.org/10.1007/s10444-011-9215-2

20. Y. Lin, M. Yang, Q. Zou, L^2 error estimates for a class of any order finite volume schemes over quadrilateral meshes, *SIAM J. Numer. Anal.*, **53** (2015), 2030–2050. https://doi.org/10.1137/140963121

21. C. Nie, S. Shu, H. Yu, W. Xia, Superconvergence and asymptotic expansions for bilinear finite volume element approximation on non-uniform grids, *J. Comput. Appl. Math.*, **321** (2017), 323–335. https://doi.org/10.1016/j.cam.2016.12.024

22. W. He, Z. Zhang, Q. Zou, Maximum-norms error estimates for high-order finite volume schemes over quadrilateral meshes, *Numer. Math.*, **138** (2018), 473–500. https://doi.org/10.1007/s00211-017-0912-8

23. Z. Chen, J. Wu, Y. Xu, Higher-order finite volume methods for elliptic boundary value problems, *Adv. Comput. Math.*, **37** (2012), 191–253. https://doi.org/10.1007/s10444-011-9201-8

24. X. Wang, Y. Li, L^2 error estimates for high order finite volume methods on triangular meshes, *SIAM J. Numer. Anal.*, **54** (2016), 2729–2749. https://doi.org/10.1137/140988486

25. Y. Zhou, J. Wu, A unified analysis of a class of quadratic finite volume element schemes on triangular meshes, *Adv. Comput. Math.*, **46** (2020), 71. https://doi.org/10.1007/s10444-020-09809-8

26. X. Wen, Y. Zhou, A coercivity result of quadratic finite volume element schemes over triangular meshes, *Adv. Appl. Math. Mech.*, **15** (2023), 901–931. https://doi.org/10.4208/aamm.OA-2021-0311

27. M. Yang, A second-order finite volume element method on quadrilateral meshes for elliptic equations, *ESAIM: M2AN*, **40** (2006), 1053–1067. https://doi.org/10.1051/m2an:2007002

28. J. Lv, Y. Li, Optimal biquadratic finite volume element methods on quadrilateral meshes, *SIAM J. Numer. Anal.*, **50** (2012), 2379–2399. https://doi.org/10.1137/100805881

29. Y. Zhou, Y. Zhang, J. Wu, A polygonal finite volume element method for anisotropic diffusion problems, *Comput. Math. Appl.*, **140** (2023), 225–236. https://doi.org/10.1016/j.camwa.2023.04.025

30. Y. Zhang, X. Wang, Unified construction and L^2 analysis for the finite volume element method over tensorial meshes, *Adv. Comput. Math.*, **49** (2023), 2. https://doi.org/10.1007/s10444-022-10004-0

31. Y. Zhou, Y. Jiang, Q. Zou, Three dimensional high order finite volume element schemes for elliptic equations, *Numer. Methods Partial Differ. Eq.*, **39** (2023), 1672–1705. https://doi.org/10.1002/num.22950

32. Y. Zhou, J. Wu, A new high order finite volume element solution on arbitrary triangular and quadrilateral meshes, *Appl. Math. Lett.*, **134** (2022), 108354. https://doi.org/10.1016/j.aml.2022.108354

33. S. Shu, H. Yu, Y. Huang, C. Nie, A symmetric finite volume element scheme on quadrilateral grids and superconvergence, *Int. J. Numer. Anal. Mod.*, **3** (2006), 348–360.

34. Q. Hong, J. Wu, Coercivity results of a modified Q_1-finite volume element scheme for anisotropic diffusion problems, *Adv. Comput. Math.*, **44** (2018), 897–922. https://doi.org/10.1007/s10444-017-9567-3

35. F. Fang, Q. Hong, J. Wu, Analysis of a special Q_1-finite volume element scheme for anisotropic diffusion problems, *Numer. Math. Theor. Meth. Appl.*, **12** (2019), 1141–1167. https://doi.org/10.4208/nmtma.OA-2018-0080

36. S. Chou, S. He, On the regularity and uniformness conditions on quadrilateral grids, *Comput. Methods Appl. Mech. Eng.*, **191** (2002), 5149–5158. https://doi.org/10.1016/S0045-7825(02)00357-2

37. P. Ciarlet, The finite element method for elliptic problems, Amsterdam: North-Holland, 1978.

38. D. Kershaw, Differencing of the diffusion equation in Lagrangian hydrodynamic codes, *J. Comput. Phys.*, **39** (1981), 375–395. https://doi.org/10.1016/0021-9991(81)90158-3

39. G. Yuan, Z. Sheng, Monotone finite volume schemes for diffusion equations on polygonal meshes, *J. Comput. Phys.*, **227** (2008), 6288–6312. https://doi.org/10.1016/j.jcp.2008.03.007

AIMS Press