IMA Journal of Numerical Analysis (2011) **31**, 1038−1061
doi:10.1093/imanum/drp054
Advance Access publication on April 28, 2010

A quadratic finite volume element method for parabolic problems on quadrilateral meshes

MIN YANG*
Department of Mathematics, Yantai University, Yantai, Shandong 264005, China
*Corresponding author: yang@ytu.edu.cn

AND

JIANGGUO LIU
Department of Mathematics, Colorado State University, Fort Collins, CO 80523-1874, USA
liu@math.colostate.edu

[Received on 26 April 2009; revised on 17 September 2009]

In this paper we utilize affine biquadratic elements and a two-step temporal discretization to develop a finite volume element method for parabolic problems on quadrilateral meshes. The method is proved to have an optimal order convergence rate in $L^2(0, T; H^1(\Omega))$ under the 'asymptotically parallelogram' mesh assumption. Numerical experiments that corroborate the theoretical analysis are also presented.

Keywords: error estimation; finite volume element methods; parabolic equations; quadrilateral meshes; second-order accuracy.

1. Introduction

In this paper we develop a second-order finite-volume method for the following model parabolic initial boundary-value problem:

$$\begin{cases} u_t - \nabla \cdot (a(\mathbf{x})\nabla u) = f(\mathbf{x}, t), & (\mathbf{x}, t) \in \Omega \times (0, T], \\ u = 0, & (\mathbf{x}, t) \in \partial\Omega \times (0, T], \\ u(\mathbf{x}, 0) = u_0(\mathbf{x}), & \mathbf{x} \in \Omega, \end{cases} \tag{1.1}$$

where Ω is a bounded polygonal domain in $\mathbb{R}^2$ with boundary $\partial\Omega$ and $\mathbf{x} = (x, y)$. It is assumed that $f(\mathbf{x}, t) \in L^2(\Omega)$ for $t \in [0, T]$ and $a(\mathbf{x})$ is Lipschitz continuous and bounded almost everywhere with positive lower and upper bounds a_* and a^*. For the purpose of error analysis, we further assume that the initial data $u_0 \in H^3(\Omega)$ and the solution of (1.1) satisfies

$$u_t \in L^1(0, T; H^3(\Omega)), \quad u_{tt} \in L^\infty(0, T; L^2(\Omega)), \quad u_{ttt} \in L^1(0, T; L^2(\Omega)). \tag{1.2}$$

Finite-volume methods have been widely used in sciences and engineering, for example, computational fluid mechanics and petroleum reservoir simulations. The methods can be formulated in the finite-difference framework, known as cell-centred methods, or in the Petrov–Galerkin framework, categorized as finite volume element methods. We refer to the monographs Eymard *et al.* (2000) and Li *et al.* (2000) for general presentations of these methods and to the papers Bank & Rose (1987),

Hackbusch (1989), Morton (1998), Ewing *et al.* (2002), Carstensen *et al.* (2005), Nicaise & Djadel (2005), Ye (2006) and Sinha & Geiser (2007) (also the references therein) for more details. Compared to finite-difference and finite-element methods, finite-volume methods are usually easier to implement and offer flexibility in handling complicated domain geometries. More importantly, the methods ensure local mass conservation, a highly desirable property in many applications.

Piecewise linear finite-volume methods have been extensively studied for various parabolic problems. The main idea is to treat the linear finite-volume methods as 'small perturbations' of the linear finite-element methods. Consequently, techniques (such as *a posteriori* error estimates and superconvergence) for the finite elements can also be used for the finite-volume methods. In Chou & Li (2000) a unified approach was presented to derive error estimates in the L^2-, H^1- and L^∞-norms. Later, error estimates and superconvergence results in the L^p-norm ($2 \leqslant p < \infty$) were obtained by Chou *et al.* (2003). Finite volume element methods for parabolic problems in convex polygonal domains were studied by Chatzipantelidis *et al.* (2004) and error estimates in the H^1-, L^2- and L^∞-norms under limited regularities of exact solutions were established. In order to solve the discrete equations more efficiently, Rui (2002) and Ma *et al.* (2003) constructed several symmetric finite-volume schemes. Ohlberger & Rohde (2002) obtained *a posteriori* error estimates in the L^1-norm for finite-volume approximations of weakly coupled convection-dominated parabolic systems.

Compared to linear finite-volume methods, quadratic and higher-order elements differ considerably from the corresponding finite-element methods. The development of higher-order finite-volume methods is a challenging task and has been attracting many researchers' attention in recent years. Plexousakis & Zouraris (2004) performed an analysis of higher-order finite-volume methods for one-dimensional elliptic problems. Cai *et al.* (2003) presented several higher-order finite-volume schemes from mixed finite-element methods over rectangular meshes for elliptic problems. Based on different dual partitions, two types of quadratic simplicial finite-volume schemes were discussed in Liebau (1996) and Li *et al.* (2000) for two-dimensional elliptic problems, and both schemes were shown to be second-order accurate in the H^1-norm. Xu & Zou (2009) combined the above two schemes together and gave an analysis under some weak conditions on grids. Error estimation of quadratic quadrilateral finite-volume methods for elliptic problems was studied in Yang (2006). The analysis was then extended to three-dimensional right quadrangular prism meshes in Yang *et al.* (2009).

The aforementioned papers on higher-order finite-volume schemes focused on stationary problems. To the best of the authors' knowledge, little progress has been made on time-dependent problems up to now. The main difficulty is due to the fact that the inner product $(\cdot, I_h^*\cdot)$ and the weak form $a_h(\cdot, I_h^*\cdot)$ (see Section 2) of higher-order finite-volume methods are not symmetric even for constant coefficient problems. Therefore it is difficult to handle temporal terms in theoretical analysis. How to control nonsymmetry of the related bilinear forms is generally still not clear.

It is well known that quadrilateral meshes have simpler mesh data structures than triangular meshes but are more flexible than rectangular meshes in handling complicated domain geometries. Shown in Fig. 1 are two examples of domains for which quadrilateral meshes are preferred: a compression channel and a triangular obstacle in a flow (see Schroll & Svensson, 2006).

The purpose of this paper is to develop a second-order finite volume element method on quadrilateral meshes for parabolic problems. It is a continuation of our efforts in Yang (2006) and Yang *et al.* (2009) for two-dimensional and three-dimensional elliptic problems. However, the dual partition used in Yang (2006) for elliptic problems no longer works for parabolic problems (Remark 3.1). Here we introduce a new dual partition for two-dimensional parabolic problems. By exploring properties of affine mappings, we manage to control nonsymmetry of the bilinear form $(\cdot, I_h^*\cdot)$ (Lemma 3.5) and ensure coercivity of the bilinear form $a_h(\cdot, I_h^*\cdot)$ (Lemma 3.8). Our new method also features a well-balanced combination

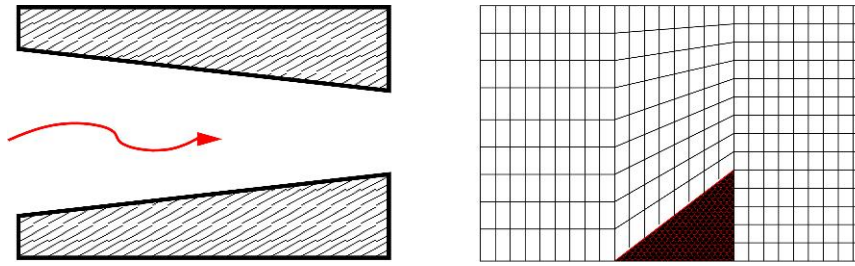

FIG. 1. Two examples that favour quadrilateral meshes.

of a quadratic spatial approximation and a two-step temporal discretization. As an implicit method, it is unconditionally stable (Theorem 4.1), allows relatively large time steps ($\Delta t = \mathcal{O}(h)$) and can generate numerical solutions that have a second-order accuracy in the energy norm (Theorem 4.2).

The rest of this paper is organized as follows. In Section 2 we introduce some notation and establish a finite volume element scheme for the parabolic initial boundary-value problem (1.1). In Section 3 we prove some auxiliary results about quadrilateral meshes and the bilinear forms. A second-order error estimate in the energy norm is proved in Section 4 under certain regularity conditions. Section 5 presents numerical experiments to illustrate the theoretical analysis. The paper is concluded with some remarks in Section 6.

Throughout this paper we use C (with or without subscripts) to denote a generic positive constant that is independent of discretization parameters.

2. A fully discrete finite-volume scheme on quadrilateral meshes

2.1 *Notation*

We shall use the standard notation for the Sobolev spaces $W^{m,p}(\Omega)$ with the norm $\|\cdot\|_{m,p,\Omega}$ and the seminorm $|\cdot|_{m,p,\Omega}$. We also denote $W^{m,2}(\Omega)$ by $H^m(\Omega)$ and omit the index $p = 2$ and the domain Ω when there is no ambiguity, i.e., $\|u\|_{m,p} = \|u\|_{m,p,\Omega}$ and $\|u\|_m = \|u\|_{m,2,\Omega}$. The same convention is adopted for the seminorms.

Let $\Omega_h = \{Q\}$ be a quadrilateral partition of Ω, where any two closed quadrilaterals share a common edge, vertex or nothing. Let $\widehat{Q} = (0,1)^2$ be the reference element in the $\hat{x}\hat{y}$-plane. For each element $Q \in \Omega_h$ there exists a bijective bilinear mapping $F_Q\colon \widehat{Q} \to Q$ satisfying (see Fig. 2)

$$F_Q(\hat{P}_i) = P_i, \quad 1 \leqslant i \leqslant 4.$$

Let $\mathscr{J}_{F_Q}$ be the Jacobian matrix of F_Q at $\hat{\mathbf{x}}$ and $J_{F_Q} = \det \mathscr{J}_{F_Q}$. Accordingly, $\mathscr{J}_{F_Q^{-1}}$ is the Jacobian matrix of F_Q^{-1} at $\mathbf{x}$ and $J_{F_Q^{-1}} = \det \mathscr{J}_{F_Q^{-1}}$. Based on the partition Ω_h, we define S_h as the standard conforming finite-element space of piecewise affine biquadratic functions

$$S_h = \{v \in H_0^1(\Omega)\colon v|_Q = \hat{v} \circ F_Q^{-1}, \hat{v}|_{\widehat{Q}} \text{ is biquadratic } \forall\, Q \in \Omega_h; v|_{\partial\Omega} = 0\}. \tag{2.1}$$

Before developing the finite volume element method, we make some assumptions on quadrilateral meshes. For any $Q \in \Omega_h$ let h_Q be its diameter, let h'_Q be the smallest length of the edges and let θ_Q be any interior angle. We set $h = \max_{Q\in\Omega_h} h_Q$.

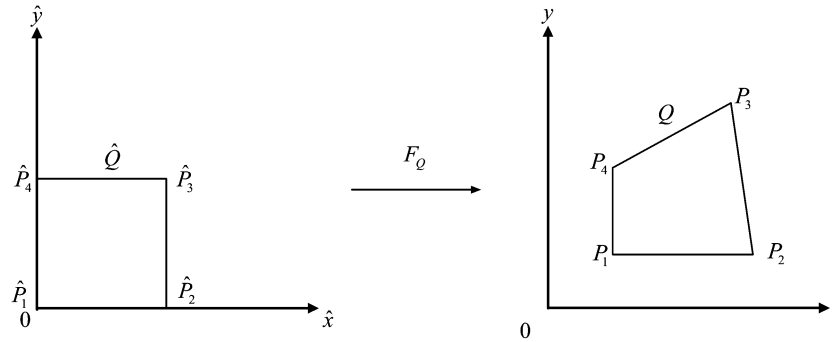

FIG. 2. An invertible bilinear mapping $F_Q\colon \widehat{Q} \to Q$.

- *Mesh assumption A*. The partition $\Omega_h = \{Q\}$ is *regular*, that is, there exist two positive constants σ and γ such that

$$h_Q/h'_Q \leqslant \sigma, \quad |\cos\theta_Q| \leqslant \gamma < 1 \quad \forall\, Q \in \Omega_h. \tag{2.2}$$

- *Mesh assumption B*. The quadrilateral meshes are 'asymptotically parallelogram', that is, for $Q \in \Omega_h$ one has

$$\vartheta_Q = \mathcal{O}(h_Q), \tag{2.3}$$

where $\vartheta_Q = \max(|\pi - \theta_1|, |\pi - \theta_2|)$, θ_1 is the angle between the outward normals of two opposite sides of Q and θ_2 is the angle between the outward normals of the other two sides.

REMARK 2.1 The asymptotically parallelogram assumption has been adopted by many authors, although it takes several different forms in the literature. In Arnold *et al.* (2002) it was called 'asymptotically parallelogram' and defined through the maximal angle difference of the outward normals. It is referred to as 'h^2-parallelogram' and defined through the edge deviation in Ewing *et al.* (1999). A definition based on the distance of the two diagonal midpoints was adopted by Süli (1992). A detailed analysis on the equivalence of these different forms can be found in Chou & He (2002).

REMARK 2.2 It should be pointed out that 'h^2-uniform quadrilateral meshes', a closely related but stronger assumption, was also used in Ewing *et al.* (1999) to prove some superconvergence results. The 'h^2-uniformness' requires that any two adjacent quadrilaterals form an h^2-parallelogram. This stronger assumption was needed to obtain a good cancellation property and then the superconvergence results (see Ewing *et al.*, 1999, p. 782).

REMARK 2.3 We may adopt a more general Assumption B that $\vartheta_Q = \mathcal{O}(h_Q^\tau)$ for some $\tau > 0$. Then the main difference will appear in Lemma 3.5, i.e., $|(u_h, I_h^* v_h) - (v_h, I_h^* u_h)| \leqslant Ch^\tau \|u_h\|_0 \|v_h\|_0$, which will be used in Theorems 4.1 and 4.2. Consequently, we shall assume that $\Delta t = \mathcal{O}(h^\tau)$ to obtain an optimal order convergence rate. Since the expected error is $\mathcal{O}(h^2 + \Delta t^2)$, a linear rate in h_Q is a practical assumption that avoids unnecessary complexities.

In order to describe the finite volume element scheme, we introduce a dual partition Ω_h^*, whose elements are called control volumes. As shown in Fig. 3, each edge of $Q \in \Omega_h$ is partitioned into three segments so that the ratio of these segments is 1:4:1. We connect these partition points with line

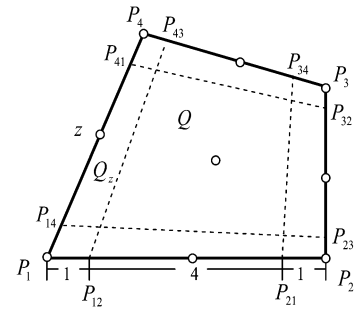

FIG. 3. A generic quadrilateral Q is partitioned into nine subregions.

segments to the corresponding points on the opposite edge. This way, each quadrilateral of Ω_h is divided into nine sub-quadrilaterals Q_z, with $z \in \mathscr{Z}_h(Q)$, where $\mathscr{Z}_h(Q)$ is the set of the vertices, the midpoints of the edges and the centre of Q. For each node $z \in \mathscr{Z}_h = \bigcup_{Q\in\Omega_h} \mathscr{Z}_h(Q)$ we associate a control volume V_z, which is the union of the subregions Q_z containing the node z. Therefore we obtain a collection of control volumes covering the domain Ω. This is the dual partition Ω_h^* of the primal partition Ω_h. We denote the set of interior nodes of $\mathscr{Z}_h$ by $\mathscr{Z}_h^0$.

REMARK 2.4 The dual partition introduced here is different to the one used in Yang (2006), where the sub-quadrilaterals were constructed based on the partition ratio 1:2:1. This new dual partition allows us to control nonsymmetry of the bilinear form $(\cdot, I_h^*\cdot)$ (see Lemma 3.5) and ensure coercivity of the bilinear form $a_h(\cdot, I_h^*\cdot)$ (see Lemma 3.8). Both forms will be defined later.

2.2 *Finite volume element scheme*

Now we formulate the finite volume element method for the model problem (1.1). Given an interpolation node $z \in \mathscr{Z}_h^0$, integrating the first equation in (1.1) over the associated control volume V_z and applying Green's formula, we obtain

$$\int_{V_z} u_t \,\mathrm{d}\mathbf{x} - \int_{\partial V_z} a\nabla u \cdot \mathbf{n}\,\mathrm{d}s = \int_{V_z} f(\mathbf{x}, t)\mathrm{d}\mathbf{x}, \tag{2.4}$$

where $\mathbf{n}$ denotes the unit outer normal vector on ∂V_z. The above formulation also states that we have an integral conservation form on the control volume.

The integral form (2.4) can be further written in a variational form similar to the finite-element method, with the help of a transfer operator $I_h^*\colon S_h \to S_h^*$ from the trial space to the test space defined in Chou & Li (2000) by

$$I_h^* v = \sum_{z\in\mathscr{Z}_h^0} v(z)\Psi_z, \tag{2.5}$$

where

$$S_h^* = \left\{ v \in L^2(\Omega)\colon v|_{V_z} \text{ is constant } \forall\, z \in \mathscr{Z}_h^0;\, v|_{V_z} = 0 \quad \forall\, z \in \partial\Omega \right\} \tag{2.6}$$

and Ψ_z is the characteristic function of the control volume V_z.

We multiply (2.4) by $v_h(z)$ and sum over all $z \in \mathscr{Z}_h^0$ to obtain

$$(u_t, I_h^* v_h) + a_h(u, I_h^* v_h) = (f, I_h^* v_h) \quad \forall v_h \in S_h, \tag{2.7}$$

where the bilinear form $a_h(\cdot, I_h^* \cdot)$ is defined as follows. For any $u \in H_0^1(\Omega)$ and $v_h \in S_h$ we have

$$a_h(u, I_h^* v_h) = -\sum_{z \in \mathscr{Z}_h^0} v_h(z) \int_{\partial V_z} a \nabla u \cdot \mathbf{n} \, \mathrm{d}s. \tag{2.8}$$

Let N be a positive integer. For simplicity of presentation, we consider a uniform time step $\Delta t = T/N$ and set $t_n = n\Delta t$ $(0 \leqslant n \leqslant N)$. Let

$$u^n = u(\cdot, t_n), \quad \bar{\partial} u^n = \frac{u^n - u^{n-1}}{\Delta t}, \quad \bar{D} u^n = \left(\frac{3}{2}u^n - 2u^{n-1} + \frac{1}{2}u^{n-2}\right) \Big/ \Delta t.$$

A fully discrete finite-volume scheme for (1.1) is formulated as follows. Find $u_h^n \in S_h$ $(2 \leqslant n \leqslant N)$ such that, for any $v_h \in S_h$, we have

$$(\bar{D} u_h^n, I_h^* v_h) + a_h(u_h^n, I_h^* v_h) = (f(t_n), I_h^* v_h), \tag{2.9}$$

with the initial approximations u_h^0 and u_h^1 given by

$$u_h^0 = R_h u_0, \tag{2.10}$$

$$(\bar{\partial} u_h^1, I_h^* v_h) + a_h(u_h^1, I_h^* v_h) = (f(t_1), I_h^* v_h) \quad \forall v_h \in S_h, \tag{2.11}$$

where $R_h \colon H_0^1(\Omega) \cap H^3(\Omega) \to S_h$ is the elliptic (Ritz) projection defined by

$$a_h(R_h u, I_h^* v_h) = a_h(u, I_h^* v_h) \quad \forall v_h \in S_h. \tag{2.12}$$

REMARK 2.5 The scheme (2.9) can be written as

$$\frac{3}{2}(u_h^n, I_h^* v_h) + \Delta t a_h(u_h^n, I_h^* v_h) = 2(u_h^{n-1}, I_h^* v_h) - \frac{1}{2}(u_h^{n-2}, I_h^* v_h) + \Delta t (f(t_n), I_h^* v_h), \quad n \geqslant 2.$$

Set $A(\cdot, \cdot) = \frac{3}{2}(\cdot, I_h^* \cdot) + \Delta t a_h(\cdot, I_h^* \cdot)$. From Lemmas 3.4, 3.6 and 3.8, we know that the bilinear form $A(\cdot, \cdot)$ is bounded and coercive on S_h for sufficiently small h. The right-hand side of the above equation defines a continuous linear functional on S_h, for given u_h^{n-1}, u_h^{n-2} and $f(t_n)$. Therefore there exists a unique solution u_h^n at each time step by the Lax–Milgram theorem (see Brenner & Scott, 2002).

3. Properties of quadrilateral meshes and the bilinear forms

In this section we prove some lemmas regarding properties of quadrilateral meshes and the bilinear forms defined in Section 2. Let P_1 and P_2 be two points. We use P_1P_2 to denote the line segment, $|P_1P_2|$ to denote its length and $\overrightarrow{P_1P_2}$ to denote the vector from point P_1 to point P_2.

According to Lemmas 3.1–3.4 in Yang (2006), we have the following results.

LEMMA 3.1 Let $Q = P_1P_2P_3P_4 \in \Omega_h$. Mesh assumption B is equivalent to the following condition:

$$|\overrightarrow{P_1P_2} + \overrightarrow{P_3P_4}| = \mathcal{O}(h_Q^2). \tag{3.1}$$

LEMMA 3.2 Suppose that Ω_h satisfies Mesh assumption B, $Q = P_1P_2P_3P_4 \in \Omega_h$ and

$$\frac{|P_2P_5|}{|P_2P_1|} = \frac{|P_3P_6|}{|P_3P_4|} = d,$$

where $0 \leqslant d \leqslant 1$ is an arbitrary constant. Then we have

$$|\overrightarrow{P_6P_5} + \overrightarrow{P_1P_4}| = \mathcal{O}(h_Q^2). \tag{3.2}$$

LEMMA 3.3 Under the same assumptions as in Lemma 3.2, we have

$$|\angle P_5P_6P_4 - \angle P_4P_1P_5| = \mathcal{O}(h_Q), \tag{3.3}$$

$$\angle P_5P_6P_4 + \angle P_6P_4P_1 = \pi + \mathcal{O}(h_Q). \tag{3.4}$$

It is assumed that there exist a pair of integers n_x and n_y such that the cardinality of Ω_h is equal to n_xn_y, and we can assign each $Q \in \Omega_h$ a pair of integers (i, j), where $0 \leqslant i \leqslant n_x-1$ and $0 \leqslant j \leqslant n_y-1$. Thus we label Q with the subscript (i, j) and denote its vertices by $\mathbf{x}_{i,j}$, $\mathbf{x}_{i+1,j}$, $\mathbf{x}_{i+1,j+1}$ and $\mathbf{x}_{i,j+1}$, corresponding to P_1, P_2, P_3 and P_4 in Fig. 2. Let $\nu_i, \nu_j = 0$ or $\frac{1}{2}$. Then the midpoints of the edges of Q are denoted by $\mathbf{x}_{i+\nu_i,j+\nu_j}$, where $\nu_i + \nu_j = \frac{1}{2}$,and the centre of Q is denoted by $\mathbf{x}_{i+\frac{1}{2},j+\frac{1}{2}}$.

We now define some discrete norms on S_h. For any $u_h \in S_h$ we have

$$|||v_h|||_0^2 = (v_h, I_h^*v_h), \quad \|v_h\|_{0,h}^2 = (I_h^*v_h, I_h^*v_h), \tag{3.5}$$

$$\begin{aligned} |u_h|_{1,h}^2 &= \sum_{Q\in\Omega_h} |u_h|_{1,h,Q}^2 \\ &= \sum_{Q\in\Omega_h} \left[\sum_{\nu_i=\frac{1}{2},1} \sum_{\nu_j=0,\frac{1}{2},1} \left(\delta_x u_h\left(\mathbf{x}_{i+\nu_i,j+\nu_j}\right)\right)^2 + \sum_{\nu_i=0,\frac{1}{2},1} \sum_{\nu_j=\frac{1}{2},1} \left(\delta_y u_h\left(\mathbf{x}_{i+\nu_i,j+\nu_j}\right)\right)^2 \right], \end{aligned} \tag{3.6}$$

where

$$\delta_x u_h\left(\mathbf{x}_{i+\nu_i,j+\nu_j}\right) = u_h\left(\mathbf{x}_{i+\nu_i,j+\nu_j}\right) - u_h\left(\mathbf{x}_{i+\nu_i-\frac{1}{2},j+\nu_j}\right),$$

$$\delta_y u_h\left(\mathbf{x}_{i+\nu_i,j+\nu_j}\right) = u_h\left(\mathbf{x}_{i+\nu_i,j+\nu_j}\right) - u_h\left(\mathbf{x}_{i+\nu_i,j+\nu_j-\frac{1}{2}}\right).$$

The following lemma indicates that the discrete norms are equivalent to the corresponding L^2-norm or H^1-seminorm.

LEMMA 3.4 Assume that Ω_h satisfies Mesh assumption A. Then there exist positive constants C_0 and C_1, independent of h, such that, for any $u_h \in S_h$, we have

$$C_0\|u_h\|_0 \leqslant |||u_h|||_0 \leqslant C_1\|u_h\|_0, \tag{3.7}$$

$$C_0\|u_h\|_0 \leqslant \|u_h\|_{0,h} \leqslant C_1\|u_h\|_0, \tag{3.8}$$

$$C_0|u_h|_1 \leqslant |u_h|_{1,h} \leqslant C_1|u_h|_1. \tag{3.9}$$

Proof. Since the partition is regular, we have

$$\|\hat{u}_h\|_{L^2(\widehat{Q})} \leqslant \left\| J_{F_Q^{-1}} \right\|_{L^\infty(Q)}^{1/2} \|u_h\|_{L^2(Q)},$$

$$\|u_h\|_{L^2(Q)} \leqslant \left\| J_{F_Q} \right\|_{L^\infty(\widehat{Q})}^{1/2} \|\hat{u}_h\|_{L^2(\widehat{Q})},$$

where

$$\left\| J_{F_Q^{-1}} \right\|_{L^\infty(Q)} \leqslant Ch_Q^{-2}, \quad \left\| J_{F_Q} \right\|_{L^\infty(\widehat{Q})} \leqslant Ch_Q^2.$$

Therefore

$$C^{-1} h_Q \|\hat{u}_h\|_{L^2(\widehat{Q})} \leqslant \|u_h\|_{L^2(Q)} \leqslant C h_Q \|\hat{u}_h\|_{L^2(\widehat{Q})}. \tag{3.10}$$

For $\int_Q u_h I_h^* u_h \, \mathrm{d}\mathbf{x}$ we have a similar estimate as follows:

$$C^{-1} h_Q^2 \int_{\widehat{Q}} \hat{u}_h{}^2 \widehat{I_h^* u_h} \, \mathrm{d}\hat{\mathbf{x}} \leqslant \int_Q u_h I_h^* u_h \, \mathrm{d}\mathbf{x} \leqslant C h_Q^2 \int_{\widehat{Q}} \hat{u}_h{}^2 \widehat{I_h^* u_h} \, \mathrm{d}\hat{\mathbf{x}}. \tag{3.11}$$

Let

$$\varphi_1(x) = (x-1)(2x-1), \quad \varphi_2(x) = 4x(1-x), \quad \varphi_3(x) = x(2x-1)$$

be the local quadratic basis in one dimension. Let $\psi_i(x)$, where $1 \leqslant i \leqslant 3$, be the characteristic functions associated with the partitions $[0, 1/6]$, $[1/6, 5/6]$ and $[5/6, 1]$, respectively. Then, using the standard tensor-product basis and the resulting interpolation form of $\hat{u}_h$ on the reference element, we immediately obtain

$$\|\hat{u}_h\|_{L^2(\widehat{Q})}^2 = \mathbf{u}_Q (\mathscr{G}_1 \otimes \mathscr{G}_1) \mathbf{u}_Q^{\mathrm{T}}, \quad \int_{\widehat{Q}} \hat{u}_h \widehat{I_h^* u_h} \, \mathrm{d}\hat{\mathbf{x}} = \mathbf{u}_Q (\mathscr{G}_2 \otimes \mathscr{G}_2) \mathbf{u}_Q^{\mathrm{T}},$$

where $\mathbf{u}_Q \in \mathbb{R}^9$ is a vector consisting of the nodal values of u_h on Q and

$$\mathscr{G}_1 = \left[\int_0^1 \varphi_i \varphi_j \, \mathrm{d}x \right]_{i,j=1}^3 = \frac{1}{30} \begin{bmatrix} 4 & 2 & -1 \\ 2 & 16 & 2 \\ -1 & 2 & 4 \end{bmatrix},$$

$$\mathscr{G}_2 = \left[\int_0^1 \varphi_i \psi_j \, \mathrm{d}x \right]_{i,j=1}^3 = \frac{1}{648} \begin{bmatrix} 83 & 32 & -7 \\ 32 & 368 & 32 \\ -7 & 32 & 83 \end{bmatrix}.$$

Since the matrices $\mathscr{G}_1$ and $\mathscr{G}_2$ are symmetric and positive definite, it is not difficult to see that $\|\hat{u}_h\|_{L^2(\widehat{Q})}^2$ and $\int_{\widehat{Q}} \hat{u}_h \widehat{I_h^* u_h} \, \mathrm{d}\hat{\mathbf{x}}$ are equivalent. Applying (3.10) and (3.11) and summing the result over Ω_h yields estimate (3.7). Estimate (3.8) can be obtained similarly. Estimate (3.9) was proved in Yang (2006). □

The bilinear form $(\cdot, I_h^* \cdot)$ is generally nonsymmetric. The next lemma measures how far it is from being symmetric.

LEMMA 3.5 Assume that Ω_h satisfies Mesh assumptions A and B. For any $u_h, v_h \in S_h$ we have

$$|(u_h, I_h^* v_h) - (v_h, I_h^* u_h)| \leqslant Ch\|u_h\|_0\|v_h\|_0. \tag{3.12}$$

Proof. For any $Q \in \Omega_h$ a change of variables in multiple integrals gives us

$$\int_Q u_h I_h^* v_h \,\mathrm{d}\mathbf{x} = \int_{\hat{Q}} \hat{u_h}\widehat{I_h^* v_h} J_{F_Q} \,\mathrm{d}\hat{\mathbf{x}}.$$

A similar calculation to that in the proof of Lemma 3.4 yields

$$m(Q)\int_{\hat{Q}} \hat{u_h}\widehat{I_h^* v_h} \,\mathrm{d}\hat{\mathbf{x}} = m(Q)\mathbf{v}_Q(\mathscr{G}_2 \otimes \mathscr{G}_2)\mathbf{u}_Q^{\mathrm{T}},$$

where $m(Q)$ is the measure of Q. Since the matrix $\mathscr{G}_2$ is symmetric, we have

$$m(Q)\int_{\hat{Q}} \hat{u_h}\widehat{I_h^* v_h} \,\mathrm{d}\hat{\mathbf{x}} - m(Q)\int_{\hat{Q}} \hat{v_h}\widehat{I_h^* u_h} \,\mathrm{d}\hat{\mathbf{x}} = 0. \tag{3.13}$$

Direct calculations yield (see Fig. 2)

$$J_{F_Q} = 2A_{124} + 2(A_{123} - A_{124})\hat{x} + 2(A_{134} - A_{124})\hat{y}, \quad 0 \leqslant \hat{x}, \hat{y} \leqslant 1,$$

where A_{ijk} denotes the area of the triangle with vertices P_i, P_j and P_k. It is obvious that $m(Q) = J_{F_Q}\left(\frac{1}{2}, \frac{1}{2}\right)$. Therefore, from Lemmas 3.1 and 3.3, we have

$$\begin{aligned}
\left|J_{F_Q} - m(Q)\right| &\leqslant |A_{123} - A_{124}| + |A_{134} - A_{124}| \\
&\leqslant \frac{1}{2}|P_1P_2|\,||P_2P_3|\sin(\angle P_1P_2P_3) - |P_4P_1|\sin(\angle P_4P_1P_2)| \\
&\quad + \frac{1}{2}|P_4P_1|\,||P_1P_3|\sin(\angle P_4P_1P_3) - |P_3P_4|\sin(\angle P_3P_4P_1)| \\
&\leqslant Ch_Q^3.
\end{aligned}$$

Combining the above estimates, we obtain

$$\begin{aligned}
\left|\int_Q u_h I_h^* v_h \,\mathrm{d}\mathbf{x} - \int_Q v_h I_h^* u_h \,\mathrm{d}\mathbf{x}\right| &\leqslant \left|\int_Q u_h I_h^* v_h \,\mathrm{d}\mathbf{x} - m(Q)\int_{\hat{Q}} \hat{u_h}\widehat{I_h^* v_h} \,\mathrm{d}\hat{\mathbf{x}}\right| \\
&\quad + \left|\int_Q v_h I_h^* u_h \,\mathrm{d}\mathbf{x} - m(Q)\int_{\hat{Q}} \hat{v_h}\widehat{I_h^* u_h} \,\mathrm{d}\hat{\mathbf{x}}\right| \\
&\leqslant Ch_Q^3\left(\int_{\hat{Q}} |\hat{u_h}\widehat{I_h^* v_h}|\mathrm{d}\hat{\mathbf{x}} + \int_{\hat{Q}} |\hat{v_h}\widehat{I_h^* u_h}|\mathrm{d}\hat{\mathbf{x}}\right).
\end{aligned}$$

By (3.10), we have

$$\int_{\hat{Q}} |\hat{u_h}\widehat{I_h^* v_h}|\mathrm{d}\hat{\mathbf{x}} \leqslant C\|\hat{u_h}\|_{L^2(\hat{Q})}\|\hat{v_h}\|_{L^2(\hat{Q})} \leqslant C/h_Q^2\|u_h\|_{L^2(Q)}\|v_h\|_{L^2(Q)},$$

and hence

$$\left|\int_Q u_h I_h^* v_h \,\mathrm{d}\mathbf{x} - \int_Q v_h I_h^* u_h \,\mathrm{d}\mathbf{x}\right| \leqslant C h_Q \|u_h\|_{L^2(Q)} \|v_h\|_{L^2(Q)}.$$

Therefore

$$\begin{aligned}|(u_h, I_h^* v_h) - (v_h, I_h^* u_h)| &\leqslant \sum_{Q\in\Omega_h} \left|\int_Q u_h I_h^* v_h \,\mathrm{d}\mathbf{x} - \int_Q v_h I_h^* u_h \,\mathrm{d}\mathbf{x}\right| \\ &\leqslant C \sum_{Q\in\Omega_h} h_Q \|u_h\|_{L^2(Q)} \|v_h\|_{L^2(Q)} \\ &\leqslant C h \|u_h\|_0 \|v_h\|_0\end{aligned}$$

by the Cauchy–Schwarz inequality. □

REMARK 3.1 As shown in the proof of Lemma 3.5, the symmetry of the matrix $\mathscr{G}_2$ is needed for (3.13) to hold. The symmetry relies on the dual partition ratio 1:4:1 introduced in this paper. If the ratio of the dual segments is set as 1:2:1 then one can verify that

$$\mathscr{G}_2 = \frac{1}{48}\begin{bmatrix} 8 & 1 & -1 \\ 5 & 22 & 5 \\ -1 & 1 & 8 \end{bmatrix},$$

which is not symmetric. Therefore

$$m(Q)\int_{\hat{Q}} \hat{u_h}\widehat{I_h^* v_h}\,\mathrm{d}\hat{\mathbf{x}} - m(Q)\int_{\hat{Q}} \hat{v_h}\widehat{I_h^* u_h}\,\mathrm{d}\hat{\mathbf{x}} = m(Q)\mathbf{v}_Q(\mathscr{G}_2\otimes\mathscr{G}_2 - \mathscr{G}_2^{\mathrm{T}}\otimes\mathscr{G}_2^{\mathrm{T}})\mathbf{u}_Q^{\mathrm{T}}.$$

If the maximum eigenvalue of the matrix $\mathscr{G}_2\otimes\mathscr{G}_2 - \mathscr{G}_2^{\mathrm{T}}\otimes\mathscr{G}_2^{\mathrm{T}}$ is $\mathcal{O}(h_Q)$ then we would obtain an almost symmetric result as in Lemma 3.5. But a simple calculation gives

$$\mathscr{G}_2\otimes\mathscr{G}_2 - \mathscr{G}_2^{\mathrm{T}}\otimes\mathscr{G}_2^{\mathrm{T}} = \frac{1}{576}\begin{bmatrix} 0 & -8 & 0 & -8 & -6 & 1 & 0 & 1 & 0 \\ 8 & 0 & 8 & 0 & -22 & 0 & -1 & 0 & -1 \\ 0 & -8 & 0 & 1 & -6 & -8 & 0 & 1 & 0 \\ 8 & 0 & -1 & 0 & -22 & 0 & 8 & 0 & -1 \\ 6 & 22 & 6 & 22 & 0 & 22 & 6 & 22 & 6 \\ -1 & 0 & 8 & 0 & -22 & 0 & -1 & 0 & 8 \\ 0 & 1 & 0 & -8 & -6 & 1 & 0 & -8 & 0 \\ -1 & 0 & -1 & 0 & -22 & 0 & 8 & 0 & 8 \\ 0 & 1 & 0 & 1 & -6 & -8 & 0 & -8 & 0 \end{bmatrix},$$

which could not produce a factor $\mathcal{O}(h_Q)$ as we expected! Then it is very difficult to measure how far the bilinear form $(\cdot, I_h^*\cdot)$ is from being symmetric based on the 1:2:1 dual partition ratio.

Let $I_h\colon H_0^1(\Omega)\cap H^3(\Omega)\to S_h$ be the usual nodal interpolation operator satisfying the approximation property (see, e.g., Brenner & Scott, 2002)

$$\|u-I_hu\|_r\leqslant Ch^{3-r}\|u\|_3,\quad 0\leqslant r\leqslant 2. \tag{3.14}$$

Next we consider the continuity of the bilinear form $a_h(\cdot, I_h^*\cdot)$. The proof is very similar to that for Lemma 3.6 in Yang (2006). We provide a short one for completeness.

LEMMA 3.6 There exists a constant $C>0$, independent of h, such that

$$|a_h(u_h, I_h^*v_h)|\leqslant C\|u_h\|_1\|v_h\|_1\quad \forall\, u_h, v_h\in S_h, \tag{3.15}$$

$$|a_h(u-I_hu, I_h^*v_h)|\leqslant Ch^2\|u\|_3\|v_h\|_1\quad \forall\, u\in H^3(\Omega), v_h\in S_h. \tag{3.16}$$

Proof. Let w denote u_h or $u-I_hu$. In view of definition (2.8), we reorder by edges to get

$$|a_h(w, I_h^*v_h)|=\sum_{Q\in\mathscr{T}_h}\sum_{z_1,z_2\in\mathscr{Z}_h^0\cap Q}\left|(v_h(z_1)-v_h(z_2))\int_{\partial V_{z_1}\cap\partial V_{z_2}} a\nabla w\cdot\mathbf{n}\,\mathrm{d}s\right|,$$

where z_1 and z_2 are chosen in Q with no repetition. It is obvious from (3.6) that

$$|v_h(z_1)-v_h(z_2)|\leqslant |v_h|_{1,h,Q}.$$

It follows from the trace inequality (see, e.g., Theorem 1.6.6 in Brenner & Scott, 2002) that

$$\begin{aligned}\left|\int_{\partial V_{z_1}\cap\partial V_{z_2}} a\nabla w\cdot\mathbf{n}\,\mathrm{d}s\right| &\leqslant Ca^*h_Q^{1/2}\|w\|_{H^1(\partial V_{z_1}\cap\partial V_{z_2})}\\ &\leqslant Ch_Q^{1/2}\|w\|_{H^1(Q)}^{1/2}\|w\|_{H^2(Q)}^{1/2}.\end{aligned}$$

If $w=u_h$ then we use the inverse estimate $\|u_h\|_{H^2(Q)}\leqslant Ch_Q^{-1}\|u_h\|_{H^1(Q)}$(see Brenner & Scott, 2002) to get

$$\left|\int_{\partial V_{z_1}\cap\partial V_{z_2}} a\nabla u_h\cdot\mathbf{n}\,\mathrm{d}s\right|\leqslant C\|u_h\|_{H^1(Q)}.$$

If $w=u-I_hu$ then we use the approximation property (3.14) to get

$$\left|\int_{\partial V_{z_1}\cap\partial V_{z_2}} a\nabla(u-I_hu)\cdot\mathbf{n}\,\mathrm{d}s\right|\leqslant Ch_Q^2\|u_h\|_{H^3(Q)}.$$

Combining the above estimates and using Lemma 3.4 gives

$$|a_h(u_h, I_h^*v_h)|\leqslant C\sum_{Q\in\mathscr{T}_h}\|u_h\|_{H^1(Q)}|v_h|_{1,h,Q}\leqslant C\|u_h\|_1|v_h|_1$$

and

$$|a_h(u-I_hu, I_h^*v_h)|\leqslant C\sum_{Q\in\mathscr{T}_h}h_Q^2\|u\|_{H^3(Q)}|v_h|_{1,h,Q}\leqslant Ch^2\|u\|_3|v_h|_1,$$

as desired. □

The following lemma about partitioned matrices was proved in Yang (2006) and will be used in this paper for the proof of the coercivity of $a_h(\cdot, I_h^*\cdot)$.

LEMMA 3.7 Let $\mathscr{A} = \begin{bmatrix} \mathscr{A}_1 & \mathscr{A}_2 \\ \mathscr{A}_2^{\mathrm{T}} & \mathscr{A}_1 \end{bmatrix}$ and $\mathscr{B} = \begin{bmatrix} \mathscr{B}_1 & \mathscr{B}_2 \\ \mathscr{B}_2 & \mathscr{B}_1 \end{bmatrix}$ be two partitioned matrices and let $\kappa \neq 0$ be a constant. Then the matrix $\begin{bmatrix} \mathscr{A} & \kappa\mathscr{B} \\ \kappa\mathscr{B}^{\mathrm{T}} & \kappa^2\mathscr{A} \end{bmatrix}$ is positive definite if and only if the matrices

$$\left[\mathscr{A}_1 + \frac{\mathscr{B}_1 + \mathscr{B}_1^{\mathrm{T}}}{2}\right] \pm \frac{1}{2}[(\mathscr{A}_2 + \mathscr{A}_2^{\mathrm{T}}) + (\mathscr{B}_2 + \mathscr{B}_2^{\mathrm{T}})]$$

and

$$\left[\mathscr{A}_1 - \frac{\mathscr{B}_1 + \mathscr{B}_1^{\mathrm{T}}}{2}\right] \pm \frac{1}{2}[(\mathscr{A}_2 + \mathscr{A}_2^{\mathrm{T}}) - (\mathscr{B}_2 + \mathscr{B}_2^{\mathrm{T}})]$$

are positive definite.

LEMMA 3.8 Assume that Ω_h satisfies Mesh assumptions A and B. There exists a constant $C_0 > 0$, independent of h, such that, for sufficiently small h, we have

$$a_h(u_h, I_h^* u_h) \geqslant C_0 \|u_h\|_1^2 \quad \forall\, u_h \in S_h. \tag{3.17}$$

Proof. We first study some properties of the auxiliary bilinear form

$$\bar{a}_h(u, I_h^* v_h) = -\sum_{z \in \mathscr{Z}_h^0} v_h(z) \int_{\partial V_z} (\bar{a}\nabla u) \cdot \mathbf{n}\, \mathrm{d}s \quad \forall\, u \in H_0^1(\Omega), v_h \in S_h, \tag{3.18}$$

where $\bar{a}|_Q = a\big(\mathbf{x}_{i+\frac{1}{2}, j+\frac{1}{2}}\big)$. We can rewrite (3.18) as

$$\begin{aligned} \bar{a}_h(u, I_h^* v_h) &= \sum_{Q \in \Omega_h} \bar{a}_{Q,h}(u, I_h^* v_h) \\ &= \sum_{Q \in \Omega_h} \left(-\sum_{z \in \mathscr{Z}_h(Q)} v_h(z) \bar{a} \int_{\partial V_z \cap Q} \nabla u \cdot \mathbf{n}\, \mathrm{d}s \right). \end{aligned} \tag{3.19}$$

The Piola transformation maps a vector-valued function on $\hat{Q}$ to one on Q by

$$\mathbf{v} = \frac{1}{J_{F_Q}} \mathscr{J}_{F_Q} \hat{\mathbf{v}} \circ F_Q^{-1}.$$

This transformation has the following well-known property (see Brezzi & Fortin, 1991):

$$\int_e \mathbf{v} \cdot \mathbf{n}\, \mathrm{d}s = \int_{\hat{e}} \hat{\mathbf{v}} \cdot \hat{\mathbf{n}}\, \mathrm{d}\hat{s},$$

where s and $\hat{s}$ denote the arc lengths along the edges e and $\hat{e}$, respectively, with $\mathbf{n}$ and $\hat{\mathbf{n}}$ as the unit normal vectors. For any $u_h \in S_h$ let $\alpha = (\mathbf{u}_{x,Q}, \mathbf{u}_{y,Q})$ be a vector with $\mathbf{u}_{x,Q}, \mathbf{u}_{y,Q} \in \mathbb{R}^6$ defined by

$$\mathbf{u}_{x,Q} = \left(\delta_x u_h\left(\mathbf{x}_{i+\frac{1}{2},j}\right), \delta_x u_h\left(\mathbf{x}_{i+\frac{1}{2},j+\frac{1}{2}}\right), \delta_x u_h\left(\mathbf{x}_{i+\frac{1}{2},j+1}\right), \ldots, \delta_x u_h(\mathbf{x}_{i+1,j+1})\right),$$

$$\mathbf{u}_{y,Q} = \left(\delta_y u_h\left(\mathbf{x}_{i,j+\frac{1}{2}}\right), \delta_y u_h\left(\mathbf{x}_{i+\frac{1}{2},j+\frac{1}{2}}\right), \delta_y u_h\left(\mathbf{x}_{i+1,j+\frac{1}{2}}\right), \ldots, \delta_y u_h(\mathbf{x}_{i+1,j+1})\right).$$

Based on the Piola transformation, we integrate equation (3.19) along the edges of the reference element to obtain

$$\frac{1}{a\left(\mathbf{x}_{i+\frac{1}{2},j+\frac{1}{2}}\right)}\bar{a}_{Q,h}(u_h, I_h^* u_h) = \alpha \mathscr{A} \alpha^{\mathrm{T}} = \frac{1}{2}\alpha(\mathscr{A} + \mathscr{A}^{\mathrm{T}})\alpha^{\mathrm{T}}, \tag{3.20}$$

which involves the partition matrix

$$\mathscr{A} = \begin{bmatrix} \mathscr{A}_1 & \mathscr{A}_2 \\ \underline{\mathscr{A}}_2 & \underline{\mathscr{A}}_1 \end{bmatrix}.$$

Here the matrices $(\mathscr{A}_1)_{6\times 6}$ and $(\underline{\mathscr{A}}_1)_{6\times 6}$ represent the contraction distortion, while the matrices $(\mathscr{A}_2)_{6\times 6}$ and $(\underline{\mathscr{A}}_2)_{6\times 6}$ characterize the rotational distortion. Let

$$\varphi_1(x) = (x-1)(2x-1), \quad \varphi_2(x) = 4x(1-x), \quad \varphi_3(x) = x(2x-1)$$

be the local quadratic basis in one dimension. The entries of these matrices are specified as follows:

$$(\mathscr{A}_1)_{ij} = \begin{cases} \frac{7}{3}\displaystyle\int_{I_i} \frac{\varphi_j(\hat{y})}{J_{F_Q}\left(\frac{1}{6},\hat{y}\right)}|P_{12}P_{43}|^2, & 1 \leqslant i, j \leqslant 3, \\ -\frac{1}{3}\displaystyle\int_{I_i} \frac{\varphi_{j-3}(\hat{y})}{J_{F_Q}\left(\frac{1}{6},\hat{y}\right)}|P_{12}P_{43}|^2, & 1 \leqslant i, j-3 \leqslant 3, \\ -\frac{1}{3}\displaystyle\int_{I_{i-3}} \frac{\varphi_j(\hat{y})}{J_{F_Q}\left(\frac{1}{6},\hat{y}\right)}|P_{21}P_{34}|^2, & 1 \leqslant i-3, j \leqslant 3, \\ \frac{7}{3}\displaystyle\int_{I_{i-3}} \frac{\varphi_{j-3}(\hat{y})}{J_{F_Q}\left(\frac{1}{6},\hat{y}\right)}|P_{21}P_{34}|^2, & 4 \leqslant i, j \leqslant 6, \end{cases}$$

$$(\mathscr{A}_2)_{ij} = \begin{cases} C_j \displaystyle\int_{I_i} \frac{(4\hat{y}-3)(1-\hat{y})}{J_{F_Q}\left(\frac{1}{6},\hat{y}\right)}\overrightarrow{P_1P_2}\cdot\overrightarrow{P_{12}P_{43}} + \frac{(4\hat{y}-3)\hat{y}}{J_{F_Q}\left(\frac{1}{6},\hat{y}\right)}\overrightarrow{P_4P_3}\cdot\overrightarrow{P_{12}P_{43}}, & 1 \leqslant i, j \leqslant 3, \\ -C_{j-3}\displaystyle\int_{I_i} \frac{(4\hat{y}-1)(1-\hat{y})}{J_{F_Q}\left(\frac{1}{6},\hat{y}\right)}\overrightarrow{P_1P_2}\cdot\overrightarrow{P_{12}P_{43}} + \frac{(4\hat{y}-1)\hat{y}}{J_{F_Q}\left(\frac{1}{6},\hat{y}\right)}\overrightarrow{P_4P_3}\cdot\overrightarrow{P_{12}P_{43}}, & 1 \leqslant i, j-3 \leqslant 3, \\ -C_j\displaystyle\int_{I_{i-3}} \frac{(4\hat{y}-1)(1-\hat{y})}{J_{F_Q}\left(\frac{5}{6},\hat{y}\right)}\overrightarrow{P_1P_2}\cdot\overrightarrow{P_{21}P_{34}} + \frac{(4\hat{y}-1)\hat{y}}{J_{F_Q}\left(\frac{5}{6},\hat{y}\right)}\overrightarrow{P_4P_3}\cdot\overrightarrow{P_{21}P_{34}}, & 1 \leqslant i-3, j \leqslant 3, \\ C_{j-3}\displaystyle\int_{I_{i-3}} \frac{(4\hat{y}-3)(1-\hat{y})}{J_{F_Q}\left(\frac{5}{6},\hat{y}\right)}\overrightarrow{P_1P_2}\cdot\overrightarrow{P_{21}P_{34}} + \frac{(4\hat{y}-3)\hat{y}}{J_{F_Q}\left(\frac{5}{6},\hat{y}\right)}\overrightarrow{P_4P_3}\cdot\overrightarrow{P_{21}P_{34}}, & 4 \leqslant i, j \leqslant 6, \end{cases}$$

where $C_1 = C_2 = 5/9$ and $C_3 = -1/9$. The integral intervals are $I_1 = [0, 1/6]$, $I_2 = [1/6, 5/6]$ and $I_3 = [5/6, 1]$. The matrices $\underline{\mathscr{A}}_1$ and $\underline{\mathscr{A}}_2$ take a similar form to those above.

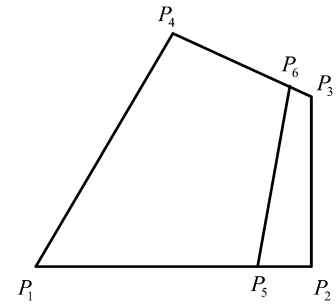

FIG. 4. A figure for Lemmas 3.1–3.3.

We first investigate an auxiliary matrix $\tilde{\mathscr{A}}$ for a parallelogram. Without loss of generality, we choose $\theta_Q = \angle P_4P_1P_2$ (see Fig. 4). Let $\kappa = |P_1P_4|/|P_1P_2|$. Note that the Jacobian determinant J_{F_Q} is a constant for a parallelogram. Then

$$\tilde{\mathscr{A}} = \frac{|P_1P_2|^2}{m(Q)}\tilde{\mathscr{M}}, \quad \tilde{\mathscr{M}} = \begin{bmatrix} \tilde{\mathscr{A}}_1 & \kappa\tilde{\mathscr{A}}_2 \\ \kappa\tilde{\mathscr{A}}_2 & \kappa^2\tilde{\mathscr{A}}_1 \end{bmatrix}$$

with

$$\tilde{\mathscr{A}}_1 = \begin{bmatrix} \tilde{\mathscr{A}}_{11} & \tilde{\mathscr{A}}_{12} \\ \tilde{\mathscr{A}}_{12} & \tilde{\mathscr{A}}_{11} \end{bmatrix}, \quad \tilde{\mathscr{A}}_2 = \begin{bmatrix} \tilde{\mathscr{A}}_{21} & \tilde{\mathscr{A}}_{22} \\ \tilde{\mathscr{A}}_{22} & \tilde{\mathscr{A}}_{21} \end{bmatrix}$$

and

$$\tilde{\mathscr{A}}_{11} = \frac{7}{1944}\begin{bmatrix} 83 & 32 & -7 \\ 32 & 368 & 32 \\ -7 & 32 & 83 \end{bmatrix}, \quad \tilde{\mathscr{A}}_{12} = \frac{-1}{1944}\begin{bmatrix} 83 & 32 & -7 \\ 32 & 368 & 32 \\ -7 & 32 & 83 \end{bmatrix},$$

$$\tilde{\mathscr{A}}_{21} = \frac{\cos\theta_Q}{81}\begin{bmatrix} -20 & -20 & 4 \\ -30 & -30 & 6 \\ 5 & 5 & -1 \end{bmatrix}, \quad \tilde{\mathscr{A}}_{22} = \frac{\cos\theta_Q}{81}\begin{bmatrix} 5 & 5 & -1 \\ -30 & -30 & 6 \\ -20 & -20 & 4 \end{bmatrix}.$$

The minimum principal major of the matrices

$$\left(\tilde{\mathscr{A}}_{11} + \frac{\tilde{\mathscr{A}}_{21} + \tilde{\mathscr{A}}_{21}^{\mathrm{T}}}{2}\right) \pm \left(\tilde{\mathscr{A}}_{12} + \frac{\tilde{\mathscr{A}}_{22} + \tilde{\mathscr{A}}_{22}^{\mathrm{T}}}{2}\right)$$

and

$$\left(\tilde{\mathscr{A}}_{11} - \frac{\tilde{\mathscr{A}}_{21} + \tilde{\mathscr{A}}_{21}^{\mathrm{T}}}{2}\right) \pm \left(\tilde{\mathscr{A}}_{12} - \frac{\tilde{\mathscr{A}}_{22} + \tilde{\mathscr{A}}_{22}^{\mathrm{T}}}{2}\right)$$

is

$$\frac{50}{729} - \frac{50}{729}|\cos(\theta_Q)| - \frac{85}{2916}|\cos^2(\theta_Q)|.$$

Let

$$F(\alpha) = \frac{50}{729} - \frac{50}{729}\alpha - \frac{85}{2916}\alpha^2.$$

It is not difficult to verify that $F(\alpha)$ is monotone decreasing for $\alpha \in [0, 1]$ and has a root $\alpha_0 \in (0, 1)$ (approximately 0.75667). We can choose $\gamma < \alpha_0$ in (2.2) so that $F(|\cos(\theta_Q)|) > 0$. Then the principal minors of the above matrices are all positive, and hence these matrices are positive definite. By Lemma 3.7, we know that the matrix $\tilde{\mathscr{M}}$ itself is positive definite as well. The minimum eigenvalue $\lambda_\gamma > 0$ of the matrix $\tilde{\mathscr{M}}$ depends only on the constant γ. From Mesh assumption A we have

$$\lambda_{\min}\left(\frac{\tilde{\mathscr{A}} + \tilde{\mathscr{A}}^{\mathrm{T}}}{2}\right) \geqslant \lambda_\gamma \frac{|P_1P_2|^2}{m(Q)} \geqslant \frac{2\lambda_\gamma}{\sigma}. \tag{3.21}$$

Now we consider the difference between the matrix $\frac{\tilde{\mathscr{A}}+\tilde{\mathscr{A}}^{\mathrm{T}}}{2}$ on a parallelogram and the matrix $\frac{\mathscr{A}+\mathscr{A}^{\mathrm{T}}}{2}$ on an asymptotically parallelogram. Let $\mathscr{D} = \frac{\mathscr{A}+\mathscr{A}^{\mathrm{T}}}{2} - \frac{\tilde{\mathscr{A}}+\tilde{\mathscr{A}}^{\mathrm{T}}}{2}$. By Lemmas 3.2 and 3.3, we have

$$\begin{aligned} |P_{12}P_{43}| &= |P_{21}P_{34}| = |P_1P_4| + \mathcal{O}(h_Q^2), \\ |P_{14}P_{23}| &= |P_{41}P_{32}| = |P_1P_2| + \mathcal{O}(h_Q^2) \end{aligned} \tag{3.22}$$

and

$$\begin{aligned} \overrightarrow{P_1P_2}\cdot\overrightarrow{P_{12}P_{43}} &= \overrightarrow{P_4P_3}\cdot\overrightarrow{P_{12}P_{43}} = \overrightarrow{P_1P_2}\cdot\overrightarrow{P_{21}P_{34}} = \overrightarrow{P_4P_3}\cdot\overrightarrow{P_{21}P_{34}} \\ &= \overrightarrow{P_1P_4}\cdot\overrightarrow{P_{14}P_{23}} = \overrightarrow{P_2P_3}\cdot\overrightarrow{P_{14}P_{43}} = \overrightarrow{P_1P_4}\cdot\overrightarrow{P_{41}P_{32}} = \overrightarrow{P_2P_3}\cdot\overrightarrow{P_{41}P_{32}} \\ &= |P_1P_2||P_1P_4|\cos\theta_Q\cos h_Q + \mathcal{O}(h_Q^3). \end{aligned} \tag{3.23}$$

By (2.3), (3.22) and (3.23), we can verify that, when h_Q is small enough, we have

$$|(\mathscr{D})_{ij}| \leqslant C\frac{(1-\cos h_Q)h_Q^2 + h_Q^3}{m(Q)}, \quad 1 \leqslant i, j \leqslant 12.$$

Mesh assumption A implies that $h_Q^2/m(Q) \leqslant 2\sigma/\sqrt{1-\gamma^2}$. So we have

$$|(\mathscr{D})_{ij}| \leqslant C(1-\cos h_Q + h_Q), \quad 1 \leqslant i, j \leqslant 12,$$

and hence

$$\lambda_{\max}(\mathscr{D}) \leqslant \|\mathscr{D}\|_\infty \leqslant 12C(1-\cos h_Q + h_Q) \leqslant Ch_Q \leqslant Ch.$$

Combining the above results with (3.20), we obtain

$$\begin{aligned} \bar{a}_{Q,h}(u_h, I_h^* u_h) &= a\left(\mathbf{x}_{i+\frac{1}{2},j+\frac{1}{2}}\right)\alpha\frac{\mathscr{A}+\mathscr{A}^{\mathrm{T}}}{2}\alpha^{\mathrm{T}} \\ &= a\left(\mathbf{x}_{i+\frac{1}{2},j+\frac{1}{2}}\right)\alpha[(\tilde{\mathscr{A}}+\tilde{\mathscr{A}}^{\mathrm{T}})/2 + \mathscr{D}]\alpha^{\mathrm{T}} \end{aligned}$$

$$\begin{aligned}
&\geqslant a_*[\lambda_{\min}((\tilde{\mathscr{A}}+\tilde{\mathscr{A}}^{\mathrm{T}})/2)-\lambda_{\max}(\mathscr{D})]\alpha\alpha^{\mathrm{T}}\\
&=a_*[\lambda_{\min}((\tilde{\mathscr{A}}+\tilde{\mathscr{A}}^{\mathrm{T}})/2)-\lambda_{\max}(\mathscr{D})]|u_h|^2_{1,h,Q}\\
&\geqslant a_*(2\lambda_\gamma/\sigma-Ch)|u_h|^2_{1,h,Q}.
\end{aligned}$$

Since $a(\mathbf{x})$ is Lipschitz continuous, we have (see Yang, 2006)

$$|a(u_h, I_h^* u_h)-\bar{a}(u_h, I_h^* u_h)| \leqslant Ch|u_h|_1^2. \tag{3.24}$$

Applying Poincaré's inequality, we obtain for sufficiently small h that

$$\begin{aligned}
a_h(u_h, I_h^* u_h) &= \sum_{Q\in\Omega_h} \bar{a}_{Q,h}(u_h, I_h^* u_h) + [a(u_h, I_h^* u_h)-\bar{a}(u_h, I_h^* u_h)]\\
&\geqslant a_* \sum_{Q\in\Omega_h} (2\lambda_\gamma/\sigma - Ch)|u_h|^2_{1,h,Q} - Ch|u_h|_1^2\\
&\geqslant C_0\|u_h\|_1^2,
\end{aligned}$$

which completes the proof. □

4. Stability and convergence analysis

In this section we present a stability and convergence analysis for the finite volume element method. First, we derive some elementary inequalities that will be used in the proofs of the two main theorems. For any $u_h, v_h \in S_h$ one has

$$\begin{aligned}
0 \leqslant |||u_h - v_h|||_0^2 &= (u_h - v_h, I_h^*(u_h - v_h))\\
&= |||u_h|||_0^2 + |||v_h|||_0^2 - (u_h, I_h^* v_h) - (v_h, I_h^* u_h).
\end{aligned} \tag{4.1}$$

Together with Lemma 3.5, this implies that

$$\begin{aligned}
(u_h, I_h^* v_h) &\leqslant \frac{1}{2}(|||u_h|||_0^2 + |||v_h|||_0^2 + (u_h, I_h^* v_h) - (v_h, I_h^* u_h))\\
&\leqslant \frac{1}{2}(|||u_h|||_0^2 + |||v_h|||_0^2) + \frac{Ch}{2}\|u_h\|_0\|v_h\|_0.
\end{aligned} \tag{4.2}$$

From (4.1) we have

$$\begin{aligned}
|||u_h + v_h|||_0^2 &= |||u_h|||_0^2 + |||v_h|||_0^2 + (u_h, I_h^* v_h) + (v_h, I_h^* u_h)\\
&\leqslant 2(|||u_h|||_0^2 + |||v_h|||_0^2).
\end{aligned} \tag{4.3}$$

The following theorem indicates that the finite volume element method is stable in the L^2-norm.

THEOREM 4.1 Let u_h^n be the numerical solution of the finite-volume scheme (2.9)–(2.11). There exists a positive constant C, independent of the discretization parameters, such that

$$\|u_h^M\|_0 \leqslant C\,\mathrm{e}^{T(h/\Delta t+1)}\left(\|u_h^0\|_0 + \Delta t\sum_{n=1}^{M}\|f(t_n)\|_0\right), \quad 1 \leqslant M \leqslant N. \tag{4.4}$$

Proof. Introducing two difference operators $\delta_i u_h^n = u_h^n - u_h^{n-i}$ for $i = 1, 2$, we rewrite $\Delta t \bar{D} u_h^n = 2\delta_1 u_h^n - \frac{1}{2}\delta_2 u_h^n$. Note that, for $i = 1, 2$, we have

$$2(\delta_i u_h^n, I_h^* u_h^n) = \delta_i |||u_h^n|||_0^2 + |||\delta_i u_h^n|||_0^2 + [(u_h^n, I_h^* u_h^{n-i}) - (u_h^{n-i}, I_h^* u_h^n)],$$

where $\delta_i |||u_h^n|||_0^2 = |||u_h^n|||_0^2 - |||u_h^{n-i}|||_0^2$. So we have

$$\begin{aligned}\Delta t(\bar{D} u_h^n, I_h^* u_h^n) &= \delta_1 |||u_h^n|||_0^2 - \frac{1}{4}\delta_2 |||u_h^n|||_0^2 + |||\delta_1 u_h^n|||_0^2 - \frac{1}{4}|||\delta_2 u_h^n|||_0^2 \\ &\quad + [(u_h^n, I_h^* u_h^{n-1}) - (u_h^{n-1}, I_h^* u_h^n)] - \frac{1}{4}[(u_h^n, I_h^* u_h^{n-2}) - (u_h^{n-2}, I_h^* u_h^n)]. \end{aligned} \tag{4.5}$$

It is clear that

$$\sum_{n=2}^{M}\left(\delta_1 |||u_h^n|||_0^2 - \frac{1}{4}\delta_2 |||u_h^n|||_0^2\right) = \frac{3}{4}|||u_h^M|||_0^2 - \frac{1}{4}|||u_h^{M-1}|||_0^2 - \frac{3}{4}|||u_h^1|||_0^2 + \frac{1}{4}|||u_h^0|||_0^2. \tag{4.6}$$

Since $\delta_2 u_h^n = \delta_1 u_h^n + \delta_1 u_h^{n-1}$, it follows from (4.3) that

$$\begin{aligned}\sum_{n=2}^{M}\left(|||\delta_1 u_h^n|||_0^2 - \frac{1}{4}|||\delta_2 u_h^n|||_0^2\right) &\geqslant \frac{1}{2}\sum_{n=2}^{M}(|||\delta_1 u_h^n|||_0^2 - |||\delta_1 u_h^{n-1}|||_0^2) \\ &= \frac{1}{2}(|||\delta_1 u_h^M|||_0^2 - |||\delta_1 u_h^1|||_0^2). \end{aligned} \tag{4.7}$$

By Lemma 3.5, we have

$$\begin{aligned}&\sum_{n=2}^{M}\left([(u_h^n, I_h^* u_h^{n-1}) - (u_h^{n-1}, I_h^* u_h^n)] - \frac{1}{4}[(u_h^n, I_h^* u_h^{n-2}) - (u_h^{n-2}, I_h^* u_h^n)]\right) \\ &\quad \leqslant Ch\sum_{n=2}^{M}(\|u_h^{n-1}\|_0 + \|u_h^{n-2}\|_0)\|u_h^n\|_0. \end{aligned} \tag{4.8}$$

Since $|||u_h^{M-1}|||_0^2 \leqslant 2|||u_h^M|||_0^2 + 2|||\delta_1 u_h^M|||_0^2$, we have from (4.5)–(4.8) that

$$\begin{aligned}\Delta t\sum_{n=2}^{M}(\bar{D} u_h^n, I_h^* u_h^n) &\geqslant \frac{3}{4}|||u_h^M|||_0^2 - \frac{1}{4}|||u_h^{M-1}|||_0^2 - \frac{3}{4}|||u_h^1|||_0^2 + \frac{1}{4}|||u_h^0|||_0^2 \\ &\quad + \frac{1}{2}(|||\delta_1 u_h^M|||_0^2 - |||\delta_1 u_h^1|||_0^2) - Ch\sum_{n=2}^{M}(\|u_h^{n-1}\|_0 + \|u_h^{n-2}\|_0)\|u_h^n\|_0 \\ &\geqslant \frac{1}{4}|||u_h^M|||_0^2 - \frac{3}{4}|||u_h^1|||_0^2 + \frac{1}{4}|||u_h^0|||_0^2 - \frac{1}{2}|||\delta_1 u_h^1|||_0^2 \\ &\quad - Ch\sum_{n=2}^{M}(\|u_h^{n-1}\|_0 + \|u_h^{n-2}\|_0)\|u_h^n\|_0. \end{aligned} \tag{4.9}$$

On the other hand,

$$\Delta t(\bar{\partial} u_h^1, I_h^* u_h^1) \geqslant \frac{1}{2}(|||u_h^1|||_0^2 - |||u_h^0|||_0^2) + \frac{1}{2}|||\delta_1 u_h^1|||_0^2 - Ch\|u_h^0\|_0\|u_h^1\|_0. \tag{4.10}$$

Combining (4.9) and (4.10) yields

$$\Delta t(\bar{\partial} u_h^1, I_h^* u_h^1) + \Delta t\sum_{n=2}^{M}(\bar{D} u_h^n, I_h^* u_h^n) \geqslant \frac{1}{4}|||u_h^M|||_0^2 - \frac{1}{4}|||u_h^1|||_0^2 - \frac{1}{4}|||u_h^0|||_0^2$$
$$- Ch\sum_{n=1}^{M}\|u_h^{n-1}\|_0\|u_h^n\|_0 - Ch\sum_{n=2}^{M}\|u_h^{n-2}\|_0\|u_h^n\|_0. \tag{4.11}$$

From (2.9) and (2.11) we have

$$\Delta t(\bar{\partial} u_h^1, I_h^* u_h^1) + \Delta t\sum_{n=2}^{M}(\bar{D} u_h^n, I_h^* u_h^n) + \Delta t\sum_{n=1}^{M} a_h(u_h^n, I_h^* u_h^n)$$
$$\leqslant \Delta t\sum_{n=1}^{M}\|f(t_n)\|_0\|I_h^* u_h\|_0 \leqslant C_1\Delta t\sum_{n=1}^{M}\|f(t_n)\|_0\|u_h^n\|_0.$$

Applying (3.7) and (4.11), we obtain

$$\frac{1}{4}|||u_h^M|||_0^2 + \Delta t\sum_{n=1}^{M} a_h(u_h^n, I_h^* u_h^n) \leqslant \frac{1}{4}|||u_h^1|||_0^2 + \frac{1}{4}|||u_h^0|||_0^2 + Ch\sum_{n=1}^{M}(\|u_h^{n-1}\|_0 + \|u_h^{n-2}\|_0)\|u_h^n\|_0$$
$$+ C_1\Delta t\sum_{n=1}^{M}\|f(t_n)\|_0\|u_h^n\|_0$$
$$\leqslant \frac{1}{4}|||u_h^1|||_0^2 + \frac{1}{4}|||u_h^0|||_0^2 + Ch\sum_{n=1}^{M}|||u_h^{n-1}|||_0|||u_h^n|||_0$$
$$+ Ch\sum_{n=2}^{M}|||u_h^{n-2}|||_0|||u_h^n|||_0 + C\Delta t\sum_{n=1}^{M}\|f(t_n)\|_0|||u_h^n|||_0.$$

Let J be such that $|||u_h^J|||_0 = \max_{1\leqslant n\leqslant M}|||u_h^n|||_0$. Then

$$|||u_h^J|||_0^2 \leqslant \left(|||u_h^1|||_0 + |||u_h^0|||_0 + Ch\sum_{n=1}^{M}|||u_h^{n-1}|||_0 + C\Delta t\sum_{n=1}^{M}\|f(t_n)\|_0\right)|||u_h^J|||_0.$$

Therefore

$$|||u_h^M|||_0 \leqslant |||u_h^J|||_0 \leqslant |||u_h^1|||_0 + |||u_h^0|||_0 + Ch\sum_{n=1}^{M}|||u_h^{n-1}|||_0 + C\Delta t\sum_{n=1}^{M}\|f(t_n)\|_0.$$

Using (4.2), we have

$$|||u_h^1|||_0 \leqslant (1+Ch)|||u_h^0|||_0 + C\,\Delta t\,\|f(t_1)\|_0.$$

Applying the discrete Gronwall inequality yields

$$\begin{aligned}|||u_h^M|||_0 &\leqslant C\,\mathrm{e}^{M(h+\Delta t)}\left(|||u_h^0|||_0 + \Delta t\sum_{n=1}^{M}\|f(t_n)\|_0\right)\\ &\leqslant C\,\mathrm{e}^{\mathrm{T}(h/\Delta t+1)}\left(|||u_h^0|||_0 + \Delta t\sum_{n=1}^{M}\|f(t_n)\|_0\right), \quad 1\leqslant M\leqslant N.\end{aligned}$$

The proof is completed by another use of (3.7). □

THEOREM 4.2 Let u be the solution of (1.1) and let u_h^n be the numerical solution of the finite-volume scheme (2.9)–(2.11). Assume that (1.2) and Mesh assumptions A and B hold. If $\Delta t = \mathcal{O}(h)$ then we have the following error estimate in the energy norm:

$$\|u^M - u_h^M\|_0 + \left(\Delta t\sum_{n=0}^{M}\|u^n - u_h^n\|_1^2\right)^{\frac{1}{2}} \leqslant C(\Delta t^2 + h^2), \quad 0\leqslant M\leqslant N, \tag{4.12}$$

for sufficiently small h and Δt, where the positive constant C is independent of the discretization parameters.

Proof. We decompose the error as $u_h^n - u^n = \xi^n - \eta^n$, where $\xi^n = u_h^n - R_h u^n$ and $\eta^n = u^n - R_h u^n$, with R_h being the elliptic projection defined in (2.12). By (2.7), (2.9) and (2.11), we have the following error equations:

$$\begin{aligned}(\bar{D}\xi^n, I_h^* v_h) + a_h(\xi^n, I_h^* v_h) &= (\omega^n, I_h^* v_h), \quad n\geqslant 2,\\ (\bar{\partial}\xi^1, I_h^* v_h) + a_h(\xi^1, I_h^* v_h) &= (\omega^1, I_h^* v_h),\end{aligned}$$

where

$$\begin{aligned}\omega^n &= \bar{D}\eta^n + u_t^n - \bar{D}u^n, \quad n\geqslant 2,\\ \omega^1 &= \bar{\partial}\eta^1 + u_t^1 - \bar{\partial}u^1.\end{aligned}$$

By Lemma 3.8, definition (2.12) and the Galerkin orthogonality, we have

$$\begin{aligned}C_0\|I_h u - R_h u\|_1^2 &\leqslant a_h(I_h u - R_h u, I_h^*(I_h u - R_h u))\\ &= a_h(I_h u - u, I_h^*(I_h u - R_h u)).\end{aligned}$$

It follows from Lemma 3.6 that

$$\|I_h u - R_h u\|_1 \leqslant Ch^2\|u\|_3. \tag{4.13}$$

Using (3.14) and a triangle inequality, we obtain

$$\|\eta^n\|_1 \leqslant \|u^n - I_h u^n\|_1 + \|I_h u^n - R_h u^n\|_1 \leqslant Ch^2\|u^n\|_3. \tag{4.14}$$

Suppose that J is such that $|||\xi^J|||_0 = \max_{1\leqslant n\leqslant M} |||\xi^n|||_0$, where $1 \leqslant M \leqslant N$. Since $\xi^0 = 0$, a similar argument as that in the proof of Theorem 4.1 leads to

$$\begin{aligned}\frac{1}{4}|||\xi^J|||_0^2 + C_0\Delta t\sum_{n=1}^{M}\|\xi^n\|_1^2 &\leqslant Ch\sum_{n=1}^{M}\|\xi^{n-1}\|_0\|\xi^n\|_0 + C\Delta t\sum_{n=1}^{M}\|\omega^n\|_0|||\xi^J|||_0 + C\Delta t^2\|\omega^1\|_0^2\\ &\leqslant C(h+\Delta t)\sum_{n=1}^{M}|||\xi^n|||_0^2 + 8C\left(\Delta t\sum_{n=1}^{M}\|\omega^n\|_0\right)^2\\ &\quad + \frac{1}{8}|||\xi^J|||_0^2 + C\Delta t^2\|\omega^1\|_0^2. \end{aligned} \tag{4.15}$$

Applying Taylor's expansion with the remainder term in integral form, we get

$$\begin{aligned}\Delta t\sum_{n=2}^{M}\|\omega^n\|_0 &\leqslant \Delta t\sum_{n=2}^{M}(\|\bar{D}\eta^n\|_0 + \|u_t^n - \bar{D}u^n\|_0)\\ &\leqslant C\sum_{n=2}^{M}\left(\int_{t_{n-2}}^{t_n}\|\eta_t\|_0\,\mathrm{d}\tau + \Delta t^2\int_{t_{n-2}}^{t_n}\|u_{ttt}\|_0\,\mathrm{d}\tau\right)\\ &\leqslant C\left(\int_0^T\|\eta_t\|_0\,\mathrm{d}\tau + \Delta t^2\int_0^T\|u_{ttt}\|_0\,\mathrm{d}\tau\right)\\ &\leqslant C\left(h^2\int_0^T\|u_t\|_3\,\mathrm{d}\tau + \Delta t^2\int_0^T\|u_{ttt}\|_0\,\mathrm{d}\tau\right). \end{aligned} \tag{4.16}$$

Similarly, we have

$$\begin{aligned}\Delta t\|\omega^1\|_0 &\leqslant \Delta t\|\bar{\partial}\eta^1\|_0 + \Delta t\|u_t^1 - \bar{\partial}u^1\|_0\\ &\leqslant Ch^2\int_0^{\Delta t}\|u_t\|_3\,\mathrm{d}\tau + \Delta t\int_0^{\Delta t}\|u_{tt}\|_0\,\mathrm{d}\tau. \end{aligned} \tag{4.17}$$

Combining (4.15)–(4.17), we obtain

$$|||\xi^J|||_0^2 + \Delta t\sum_{n=1}^{M}\|\xi^n\|_1^2 \leqslant C(h+\Delta t)\sum_{n=1}^{M}|||\xi^n|||_0^2 + C_u(\Delta t^4 + h^4),$$

where $C_u = C(\|u_t\|_{L^1(0,T;H^3(\Omega))} + \|u_{tt}\|_{L^\infty(0,T;L^2(\Omega))} + \|u_{ttt}\|_{L^1(0,T;L^2(\Omega))})^2$.

To balance spatial and temporal errors it is sufficient to have $h = \mathcal{O}(\Delta t)$, which allows us to use (3.7) and the discrete Gronwall inequality to derive

$$\|\xi^M\|_0^2 + \Delta t\sum_{n=1}^{M}\|\xi^n\|_1^2 \leqslant C_0^{-2}|||\xi^J|||_0^2 + \Delta t\sum_{n=1}^{M}\|\xi^n\|_1^2 \leqslant C(\Delta t^4 + h^4), \quad M \geqslant 1,$$

for sufficiently small h and Δt. By (4.14), we obtain

$$\|\eta^n(t)\|_1 \leqslant Ch^2\|u^n\|_3 \leqslant Ch^2\left(\|u_0\|_3 + \int_0^t \|u_t\|_3\,\mathrm{d}\tau\right).$$

Combining the above two estimates, we get the desired estimate. □

5. Numerical experiments

In this section we present numerical experiments to illustrate the theoretical results presented in the previous sections. The domain is $\Omega = [0, \pi]^2$, the permeability coefficient $a(x, y) = (x + 1)^2 + y^2$, the exact solution $u(x, y, t) = \mathrm{e}^{-0.1t}\sin(x)\sin(y)$ and the final time $T = 0.2$. The right-hand side $f(x, y, t)$ is computed accordingly.

We test the finite-volume method first on a family of rectangular meshes (see Fig. 5) that have $M = 4, 8, 16, 32$ or 64 partitions in both the x- and y-directions and then on a set of quadrilateral meshes generated by

$$x_{i,j} = \frac{\pi}{M}i + \frac{\pi}{4M}\sin(j)\mathrm{rand}(), \quad y_{i,j} = \frac{\pi}{M}j + \frac{\pi}{4M}\sin(i)\mathrm{rand}(), \quad 1 \leqslant i, j \leqslant M - 1,$$

where $\sin(i)$ and $\sin(j)$ use the radian unit for angles and rand() generates a random number in $(0, 1)$. An implementation of the random function rand() depends on the computer platform used but is available in the standard library of most C/C++ compilers.

Given a spatial mesh parameter $h = \pi/M$, the time step size is chosen as $\Delta t = T/(2\pi)h$. We use the first-order backward Euler as a starter scheme (see, e.g., Thomée, 2006). The preconditioned bi-conjugate gradient stabilized method is employed to solve the nonsymmetric discrete linear systems.

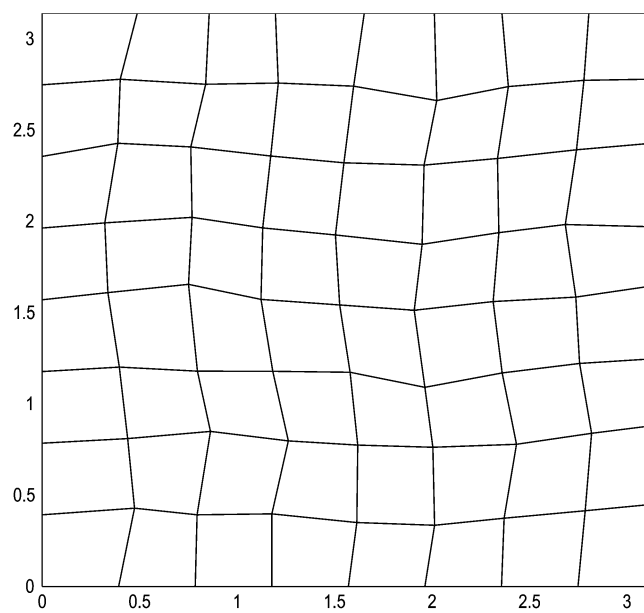

FIG. 5. An example for asymptotically parallelogram quadrilateral meshes (used in our numerical experiments).

TABLE 1 *Errors and convergence rates on the rectangular meshes*

M	$\|I_h u^N - u_h^N\|_{l^\infty}$	Rate	$\|I_h u^N - u_h^N\|_{1,h}$	Rate	$\|u_h^n\|_E$	Rate
4	8.563×10^{-3}	—	2.202×10^{-2}	—	2.032×10^{-2}	—
8	2.074×10^{-3}	2.04	5.546×10^{-3}	1.98	5.226×10^{-3}	1.95
16	5.138×10^{-4}	2.01	1.388×10^{-3}	1.99	1.312×10^{-3}	1.99
32	1.281×10^{-4}	2.00	3.471×10^{-4}	1.99	3.281×10^{-4}	2.00
64	3.200×10^{-5}	2.00	8.676×10^{-5}	2.00	8.241×10^{-5}	1.99

TABLE 2 *Errors and convergence rates on the quadrilateral meshes*

M	$\|I_h u^N - u_h^N\|_{l^\infty}$	Rate	$\|I_h u^N - u_h^N\|_{1,h}$	Rate	$\|u_h^n\|_E$	Rate
4	8.089×10^{-3}	—	1.945×10^{-2}	—	2.126×10^{-2}	—
8	2.274×10^{-3}	1.83	5.883×10^{-3}	1.73	6.397×10^{-3}	1.73
16	5.483×10^{-4}	2.05	1.466×10^{-3}	2.00	1.596×10^{-3}	2.00
32	1.400×10^{-4}	1.96	3.718×10^{-4}	1.97	4.072×10^{-4}	1.97
64	3.426×10^{-5}	2.03	9.316×10^{-5}	1.99	1.024×10^{-4}	1.99

The tolerance for residuals is set as 10^{-11} and simple diagonal preconditioning is used. The numerical order of convergence is then measured by comparing the computed errors on two successive mesh levels.

From the proof of Theorem 4.2 we know that the second-order error estimate in the energy norm is essentially determined by the second-order convergence in Δt and h of the following quantity:

$$\|u_h^N\|_E = |||I_h u^N - u_h^N|||_0 + \left(\Delta t \sum_{n=1}^{N} |I_h u^n - u_h^n|_{1,h}^2\right)^{1/2},$$

due to the approximation property (3.14), the norm equivalences (3.7) and (3.9) and the estimate (4.13). We thus measure $\|u_h^n\|_E$ in the numerical experiments in lieu of the energy norm defined in (4.12). The numerical results listed in Tables 1 and 2 clearly reflect the second-order convergence of $\|u_h^n\|_E$ for both the rectangular and the quadrilateral meshes, which agrees very well with our theoretical findings. As a by-product, we can also obtain second-order accuracy in the max-norm $\|I_h u^N - u_h^N\|_{l^\infty}$ and the discrete H^1-norm $|I_h u^N - u_h^N|_{1,h}$.

6. Concluding remarks

In this paper we presented a quadratic finite volume element method for parabolic problems on quadrilateral meshes. A suitable dual partition was introduced to control the nonsymmetry of the bilinear forms. A second-order convergence in the $L^2(0, T; H^1(\Omega))$-norm was derived and verified in numerical experiments. However, there are technical difficulties in deriving error estimates in $L^\infty(0, T; H^1(\Omega))$ and $L^\infty(0, T; L^2(\Omega))$ similar to those for the finite-element methods. The main reason for this deficiency is that we are not able to derive the optimal L^2-norm error estimate for the Ritz projection for the bilinear form $a_h(\cdot, I_h^* \cdot)$. Such a deficiency exists on triangular and quadrilateral meshes (see Liebau, 1996; Li *et al.*, 2000; Yang, 2006; Xu & Zou, 2009). How to obtain the optimal L^2-norm error estimates for finite volume element methods based on multidimensional higher-order elements is still an open problem.

The coercivity of the bilinear form $a_h(\cdot, I_h^*\cdot)$ plays a critical role in the stability and convergence analysis of the finite volume element scheme. The proof of the coercivity (Lemma 3.8) is demanding. The usual techniques used in the analysis of linear finite-volume methods do not help here. However, the coercivity is related to the positive definiteness of certain matrices at the element level. This challenging issue is then resolved by breaking up the distortion of affine mappings into contraction and rotation and employing partitioned matrices for size reduction. We hope that these techniques could be extended to the analysis of other higher-order finite-volume methods in multiple dimensions.

Quadrilateral meshes have been used in a variety of applications and the asymptotically parallelogram assumption has been widely accepted in the literature (see Süli, 1992; Ewing *et al.*, 1999; Arnold *et al.*, 2002; Flemisch & Wohlmuth, 2007), although it takes several different but equivalent forms. It would be very interesting to see whether the assumption can be significantly weakened.

Apparently, the quadratic finite-volume elements can be combined with the backward Euler temporal discretization. Even though the implicit method is unconditionally stable, the condition $\Delta t = \mathcal{O}(h^2)$ is still needed to balance the spatial and temporal errors. This is prohibitive for fine spatial meshes. With the two-step implicit scheme presented in this paper, the time step size is only $\Delta t = \mathcal{O}(h)$, but a second-order convergence is obtained in numerical solutions with reduced computational costs.

The quadratic finite-volume method can also be applied to convection–diffusion equations, in conjunction with treatments of convection by some upwinding techniques or the characteristic methods. The development of finite-volume elements of order higher than two in multiple dimensions and their analysis are even more challenging. These are currently under investigation and will be reported in our future work.

References

Arnold, D., Boffi, D. & Falk, R. (2002) Approximation by quadrilateral finite elements. *Math. Comput.*, **71**, 909–922.

Bank, R. E. & Rose, D. J. (1987) Some error estimates for the box method. *SIAM J. Numer. Anal.*, **24**, 777–787.

Brenner, S. C. & Scott, L. R. (2002) *The Mathematical Theory of Finite Element Methods*. New York: Springer.

Brezzi, F. & Fortin, M. (1991) *Mixed and Hybrid Finite Element Methods*. New York: Springer.

Cai, Z., Douglas Jr, J. & Park, M. (2003) Development and analysis of higher order finite volume methods over rectangles for elliptic equations. *Adv. Comput. Math.*, **19**, 3–33.

Carstensen, C., Lazarov, R. & Tomov, S. (2005) Explicit and averaging a posteriori error estimates for adaptive finite volume methods. *SIAM J. Numer. Anal.*, **42**, 2496–2521.

Chatzipantelidis, P., Lazarov, R. D. & Thomée, V. (2004) Error estimates for a finite volume element method for parabolic equations in convex polygonal domains. *Numer. Methods Partial Differ. Equ.*, **20**, 650–674.

Chou, S. H. & He, S. (2002) On the regularity and uniformness conditions on quadrilateral grids. *Comput. Methods Appl. Mech. Eng.*, **191**, 5149–5158.

Chou, S. H., Kwak, D. Y. & Li, Q. (2003) L^p error estimates and superconvergence for covolume or finite volume element methods. *Numer. Methods Partial Differ. Equ.*, **19**, 463–486.

Chou, S. H. & Li, Q. (2000) Error estimates in L^2, H^1, and L^∞ in covolume methods for elliptic and parabolic problems: a unified approach. *Math. Comput.*, **69**, 103–120.

Ewing, R. E., Lin, T. & Lin, Y. (2002) On the accuracy of the finite volume element method based on piecewise linear polynomials. *SIAM J. Numer. Anal.*, **39**, 1865–1888.

Ewing, R. E., Liu, M. & Wang, J. (1999) Superconvergence of mixed finite element approximations over quadrilaterals. *SIAM J. Numer. Anal.*, **36**, 772–787.

EYMARD, R., GALLOUËT, T. & HERBIN, R. (2000) *Finite Volume Methods: Handbook of Numerical Analysis*. Amsterdam: North-Holland.

FLEMISCH, B. & WOHLMUTH, B. (2007) Stable Lagrange multipliers for quadrilateral meshes of curved interfaces in 3D. *Comput. Methods Appl. Mech. Eng.*, **196**, 1589–1602.

HACKBUSCH, W. (1989) On first and second order box schemes. *Computing*, **41**, 277–296.

LI, R., CHEN, Z. & WU, W. (2000) *Generalized Difference Methods for Differential Equations: Numerical Analysis of Finite Volume Methods*. New York: Marcel Dekker.

LIEBAU, F. (1996) The finite volume element method with quadratic basis functions. *Computing*, **57**, 281–299.

MA, X., SHU, S. & ZHOU, A. (2003) Symmetric finite volume discretizations for parabolic problems. *Comput. Methods Appl. Mech. Eng.*, **192**, 4467–4485.

MORTON, K. W. (1998) On the analysis of finite volume methods for evolutionary problems. *SIAM J. Numer. Anal.*, **35**, 2195–2222.

NICAISE, S. & DJADEL, K. (2005) Convergence analysis of a finite volume method for the Stokes system using non-conforming arguments. *IMA J. Numer. Anal.*, **25**, 523–548.

OHLBERGER, M. & ROHDE, C. (2002) Adaptive finite volume approximations for weakly coupled convection dominated parabolic systems. *IMA J. Numer. Anal.*, **22**, 253–280.

PLEXOUSAKIS, M. & ZOURARIS, G. E. (2004) On the construction and analysis of high order locally conservative finite volume-type methods for one-dimensional elliptic problems. *SIAM J. Numer. Anal.*, **42**, 1226–1260.

RUI, H. (2002) Symmetric modified finite volume element methods for self-adjoint elliptic and parabolic problems. *J. Comput. Appl. Math.*, **146**, 373–386.

SCHROLL, H. J. & SVENSSON, F. (2006) A bi-hyperbolic finite volume method on quadrilateral meshes. *J. Sci. Comput.*, **26**, 237–260.

SINHA, R. K. & GEISER, J. (2007) Error estimates for finite volume element methods for convection–diffusion–reaction equations. *Appl. Numer. Math.*, **57**, 59–72.

SÜLI, E. (1992) The accuracy of cell vertex finite volume methods on quadrilateral meshes. *Math. Comput.*, **59**, 359–382.

THOMÉE, V. (2006) *Galerkin Finite Element Methods for Parabolic Problems*. Berlin: Springer.

XU, J. C. & ZOU, Q. (2009) Analysis of linear and quadratic simplicial finite volume methods for elliptic equations. *Numer. Math.*, **111**, 469–492.

YANG, M. (2006) A second-order finite volume element method on quadrilateral meshes for elliptic equations. *ESAIM Math. Model. Numer. Anal.*, **40**, 1053–1068.

YANG, M., LIU, J. & CHEN, C. (2009) Error estimation of a quadratic finite volume method on right quadrangular prism grids. *J. Comput. Appl. Math.*, **229**, 274–282.

YE, X. (2006) A discontinuous finite volume method for the Stokes problems. *SIAM J. Numer. Anal.*, **44**, 183–198.