Journal of Computational and Applied Mathematics 229 (2009) 274–282

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Error estimation of a quadratic finite volume method on right quadrangular prism grids

Min Yang [a,*], Jiangguo Liu [b], Chuanjun Chen [a]

[a] Department of Mathematics, Yantai University, Yantai, Shandong 264005, PR China

[b] Department of Mathematics, Colorado State University, Fort Collins, CO 80523-1874, USA

ARTICLE INFO

Article history:
Received 28 March 2007
Received in revised form 6 October 2008

MSC:
65N12
65N30

Keywords:
Error estimates
Elliptic problems
Finite volume element methods
Quadratic bases
Right quadrangular prism meshes
Second order accuracy

ABSTRACT

In this paper, we develop a finite volume element method with affine quadratic bases on right quadrangular prism meshes for three-dimensional elliptic boundary value problems. The optimal H^1-norm error estimate of second order accuracy is proved under certain assumptions about the meshes. Numerical results are presented to illustrate the theoretical analysis.

1. Introduction

We consider the following elliptic boundary value problem:

$$\begin{cases} -\nabla \cdot (a(\mathbf{x})\nabla u) + b(\mathbf{x})u = f(\mathbf{x}), & \mathbf{x} \in \Omega, \\ u(\mathbf{x}) = 0, & \mathbf{x} \in \partial\Omega, \end{cases} \tag{1}$$

where Ω is a bounded polyhedral domain in $\mathbb{R}^3$. It is assumed that $f(\mathbf{x}) \in L^2(\Omega)$, $a(\mathbf{x})$ and $b(\mathbf{x})$ are Lipschitz continuous with a Lipschitz constant L, $0 < a_* \le a(\mathbf{x}) \le a^*$, and $0 \le b(\mathbf{x}) \le b^*$. Here a_*, a^* and b^* are positive constants.

Finite volume methods have been widely used in science and engineering (e.g., computational fluid mechanics and petroleum reservoir simulations). The methods can be formulated in the finite difference framework, known as cell-centered methods, or in the Petrov–Galerkin framework, categorized as finite volume element methods. We refer to the monographs [8,11] for general presentations of these methods, and to the papers [3–5,7,17] (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.

Finite volume methods based on piecewise constant functions or piecewise linear functions are well developed for elliptic problems prototyped as (1), see, e.g., [4,7,10,13]. The development of efficient higher order finite volume methods is important for various applications. In [14], finite volume element methods based on piecewise polynomials of degree higher

* Corresponding author.
E-mail addresses: yang@ytu.edu.cn (M. Yang), liu@math.colostate.edu (J. Liu).

0377-0427/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.cam.2008.10.036

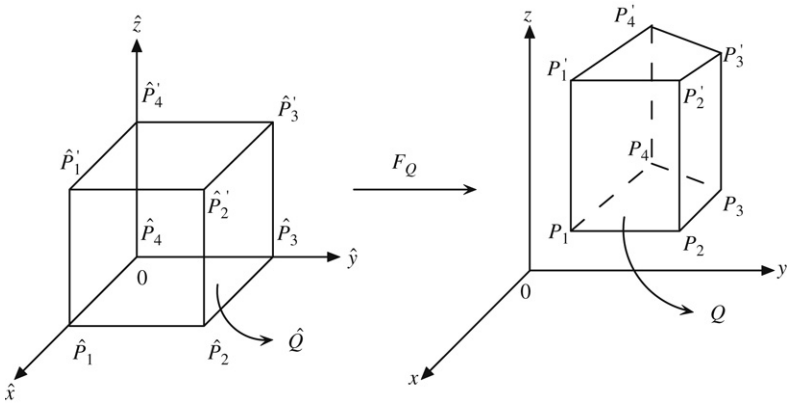

Fig. 1. The invertible affine mapping F_Q from the reference element $\widehat{Q} = [0, 1]^3$ to a generic right quadrangular prism element Q.

than two are investigated for one-dimensional elliptic problems. Optimal error estimates in the L^2-, H^1-, and L^∞-norms are derived there. A systematic way to derive higher order mixed finite volume methods over rectangular meshes for elliptic problems is developed in [2]. In [11,12], based on different dual partitions, two types of finite volume element methods with quadratic basis functions are established for two-dimensional elliptic problems, and both schemes are shown to be second order accurate in the H^1-norm. A similar result is obtained in [19] for elliptic problems on quadrilateral meshes. One may find more general higher order finite volume schemes in [16,18], but the behaviors of those methods are still not well known.

Developing higher order finite volume element methods for three-dimensional problems is nontrivial, due to the complexity of 3-dim geometries and construction of suitable dual meshes for different approximation spaces. Well-posedness and optimal convergence rates of finite volume schemes are more complicated for 3-dim problems. In this paper, we focus on right quadrangular prism meshes. These type of meshes are used in petroleum reservoir simulations to accommodate different geological layers. Similar prism meshes have also been used in [9] to treat convection-diffusion problems by linear finite volumes. In this paper, we develop a second order finite volume scheme with affine quadratic bases on 3-dim right quadrangular prism grids for the elliptic boundary value problem (1). A right quadrangular prism mesh can be viewed as the tensor product of a two-dimensional quadrilateral mesh with a one-dimensional mesh. Certain properties of tensor products will be utilized to simplify the analysis of the finite volume scheme.

The rest of this paper is organized as follows. Section 2 establishes a finite volume element scheme for the elliptic boundary value problem (1) and introduces two assumptions about right quadrangular prism meshes. In Section 3, we first prove continuity and coercivity of the bilinear form in the finite volume scheme. An optimal second order convergence in the H^1-norm is then derived. Section 4 discusses implementation issues and presents numerical results to illustrate the effectiveness of the finite volume method.

Throughout this paper, we use C (with or without subscripts) to denote a generic positive constant that is independent of the spatial mesh size.

2. A finite volume scheme on prism meshes

For ease of presentation, we assume that Ω is a rectangular domain parallel to the coordinate axes. Let $\mathcal{E}_h = \{Q\}$ be a right quadrangular prism partition of Ω, where any two prisms share a face or an edge, or just a node. Let $\widehat{Q} = [0, 1]^3$ be the reference element. For each prism $Q \in \mathcal{E}_h$, there exists a bijective multilinear (bilinear in x, y and linear in z) mapping $F_Q : \widehat{Q} \longrightarrow Q$ satisfying

$$F_Q(\hat{P}_i) = P_i, \qquad F_Q(\hat{P}_i') = P_i', \quad 1 \le i \le 4, \tag{2}$$

as shown in Fig. 1. Let $\mathcal{J}_{F_Q}$ be the Jacobian matrix of F_Q and $J_{F_Q} = \det(\mathcal{J}_{F_Q})$. Accordingly, $\mathcal{J}_{F_Q^{-1}}$ is understood as the Jacobian matrix of F_Q^{-1} and $J_{F_Q^{-1}} = \det(\mathcal{J}_{F_Q^{-1}})$. Based on $\mathcal{E}_h$, we define S_h as the standard conforming finite element space of piecewise affine quadratic functions

$$S_h = \{v \in H_0^1(\Omega) : v|_Q = \hat{v} \circ F_Q^{-1},\ \hat{v}|_{\widehat{Q}} \text{ is quadratic}, \forall Q \in \mathcal{E}_h;\ v|_{\partial\Omega} = 0\}. \tag{3}$$

For each $Q \in \mathcal{E}_h$, let h_Q be its diameter, h_Q' the length of the shortest edge, and θ_Q the minimal acute angle between any two edges. We define $h = \max_{Q\in\mathcal{E}_h} h_Q$.

- *Mesh Assumption A.* The partition $\mathcal{E}_h = \{Q\}$ is *regular*, that is, there exist positive constants σ and γ such that

$$h_Q/h_Q' \le \sigma, \qquad |\cos\theta_Q| \le \gamma < 1, \quad \forall Q \in \mathcal{E}_h. \tag{4}$$

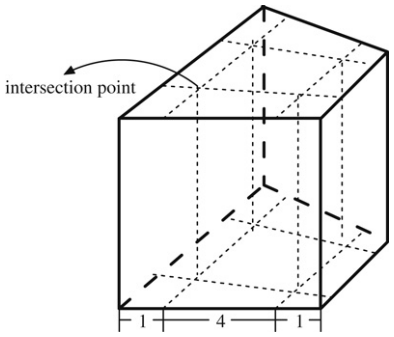

Fig. 2. The dual partitions in the x-, y-directions within a typical element. The dual partition in the z-direction is not shown here to avoid complicating the illustration.

- *Mesh Assumption B.* Each face is *almost a parallelogram*. Namely, there exits a positive constant τ such that

$$|\overrightarrow{P_1P_4} + \overrightarrow{P_3P_2}| = O(h_Q^{1+\tau}), \quad \forall Q \in \mathcal{E}_h. \tag{5}$$

Remark 2.1. Mesh Assumption B is weaker than the "almost parallelogram" condition introduced in [19], where $|\overrightarrow{P_1P_4} + \overrightarrow{P_3P_2}| = O(h_Q^2)$ was used. In fact, we can see from the proof of Lemmas 3.3, 3.4, and 3.9 in [19] that the weaker assumption $O(h_Q^{1+\tau})$ is sufficient for the proof of coercivity of the corresponding bilinear forms.

As discussed in [17], we assume, without loss of generality, that the meshes are *topologically rectangular*. There exists a triplet of integers (n_x, n_y, n_z) such that the cardinality of $\mathcal{E}_h$ is $n_x n_y n_z$. We can associate a 3-index (i, j, k) with each $Q \in \mathcal{E}_h$ for $0 \le i \le n_x - 1$, $0 \le j \le n_y - 1$, $0 \le k \le n_z - 1$. Then we may label Q by the subscripts $\{i, j, k\}$ and write $Q_{i,j,k}$ instead. The vertices of $Q_{i,j,k}$ are denoted by $\mathbf{x}_{i,j,k}$, $\mathbf{x}_{i+1,j,k}$, $\mathbf{x}_{i+1,j+1,k}$, $\mathbf{x}_{i,j+1,k}$, $\mathbf{x}_{i,j,k+1}$, $\mathbf{x}_{i+1,j,k+1}$, $\mathbf{x}_{i+1,j+1,k+1}$, $\mathbf{x}_{i,j+1,k+1}$, corresponding to $P_1, P_2, P_3, P_4, P_1', P_2', P_3', P_4'$, respectively, as shown in Fig. 1. Let ν_i, ν_j, $\nu_k = 0$ or $\frac{1}{2}$. Then the midpoint of each edge of Q will be denoted by $\mathbf{x}_{i+\nu_i,j+\nu_j,k+\nu_k}$, where $\nu_i + \nu_j + \nu_k = \frac{1}{2}$. Similarly, the center of each face will be denoted by $\mathbf{x}_{i+\nu_i,j+\nu_j,k+\nu_k}$, where $\nu_i + \nu_j + \nu_k = 1$. The center of Q will be denoted by $\mathbf{x}_{i+\frac{1}{2},j+\frac{1}{2},k+\frac{1}{2}}$. All these vertices, midpoints, and centers in $\mathcal{E}_h$ together form a set of interpolation points $\mathcal{Z}_h$.

Remark 2.2. For the rest of this paper, we omit the subscripts $\{i, j, k\}$ in $Q_{i,j,k}$ and just write Q, if no ambiguity arises.

Definition 2.1. For an element $Q \in \mathcal{E}_h$, two edges are called *opposite*, if they belong to the same face but do not share a vertex. Two faces of Q are called *opposite*, if they do not share any edge.

To establish a finite volume element scheme, we introduce a dual partition, whose elements are called control (dual) volumes. As shown in Fig. 2, each edge of an element in $\mathcal{E}_h$ is partitioned into three segments so that the ratio of these segments is 1:4:1. We connect these partition points with the corresponding partition points on the opposite edge. Hence there are four intersection points on each face. Next we connect the intersection points on a face with the corresponding ones on the opposite face. This way, each prism of $\mathcal{E}_h$ is divided into twenty seven sub-prisms Q_ω, $\omega \in \mathcal{Z}_h \cap \bar{\Omega}$. For each node $\omega \in \mathcal{Z}_h$, we assign a control volume V_ω, which is the union of the subregions Q_ω containing the node ω. Therefore, we obtain a collection of control volumes that covers the domain Ω. This is the dual partition $\mathcal{E}_h^*$ of the primal partition $\mathcal{E}_h$. We denote the set of interior nodes of $\mathcal{Z}_h$ by $\mathcal{Z}_h^0$.

Remark 2.3. The dual partition introduced here is different than the one used in [19], where the ratio of segments is 1:2:1. The new dual partition ensures the matrices $\mathcal{G}_2$ and $\mathcal{H}$ defined in Lemmas 1 and 2 are symmetric, which plays an important role in the coercivity proof of the bilinear form $a_h(\cdot, I_h^*\cdot)$ (see Lemma 4 in Section 3).

Now we formulate the finite volume element method for the model problem (1). Given an interpolation node $\omega \in Z_h^0$, integrating the first equation in (1) over the control volume V_ω and applying the Green's formula, we obtain

$$-\int_{\partial V_\omega} a\nabla u \cdot \mathbf{n} ds + \int_{V_\omega} bu d\mathbf{x} = \int_{V_\omega} f d\mathbf{x}, \tag{6}$$

where $\mathbf{n}$ denotes the unit outer normal on ∂V_ω. The above formulation implies that we have local mass conservation on the control volume.

The integral form (6) can be further rewritten in a variational form similar to those in finite element methods, with the help of a transfer operator $I_h^* : S_h \to S_h^*$ from the trial space to the test space defined as

$$I_h^* v = \sum_{\omega \in Z_h^0} v(\omega)\Psi_\omega, \tag{7}$$

where

$$S_h^* = \{v \in L^2(\Omega) : v|_{V_\omega} \text{ is constant}, \forall \omega \in Z_h^0; v|_{V_\omega} = 0; \forall \omega \in \partial\Omega\}, \tag{8}$$

and Ψ_ω is the indicator function of the control volume V_ω.

We multiply (6) by $v_h(\omega)$ and sum over all $\omega \in Z_h^0$ to obtain

$$a_h(u, I_h^* v_h) + (bu, I_h^* v_h) = (f, I_h^* v_h), \quad \forall v_h \in S_h, \tag{9}$$

where the bilinear form $a_h(\cdot, I_h^* \cdot)$ is defined as

$$a_h(u, I_h^* v_h) = -\sum_{\omega \in Z_h^0} v_h(\omega) \int_{\partial V_\omega} a\nabla u \cdot \mathbf{n} ds, \tag{10}$$

for any $u \in H_0^1(\Omega)$, $v_h \in S_h$. The finite volume scheme for the model problem (1) is formulated as: Seek $u_h \in S_h$ such that for any $v_h \in S_h$,

$$a_h(u_h, I_h^* v_h) + (bu_h, I_h^* v_h) = (f, I_h^* v_h), \quad \forall v_h \in S_h. \tag{11}$$

3. Convergence analysis

We shall use the standard notations 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 skip the index $p = 2$ and the domain Ω, i.e., $\|u\|_{m,p} = \|u\|_{m,p,\Omega}$, $\|u\|_m = \|u\|_{m,2,\Omega}$, when there is no ambiguity. The same convention is adopted for the seminorms.

In order to analyze the finite volume scheme, we now define three discrete (semi) norms on S_h. For any $u_h \in S_h$,

$$|||u_h|||_0^2 = (u_h, I_h^* u_h), \qquad \|u_h\|_{0,h}^2 = (I_h^* u_h, I_h^* u_h), \tag{12}$$

$$|u_h|_{1,h}^2 = \sum_{Q \in \mathcal{E}_h} |u_h|_{1,h,Q}^2, \tag{13}$$

where

$$\begin{aligned} |u_h|_{1,h,Q}^2 = {} & h_Q \sum_{\nu_i = \frac{1}{2}, 1} \sum_{\nu_j, \nu_k = 0, \frac{1}{2}, 1} \left(\delta_x u_h(\mathbf{x}_{i+\nu_i, j+\nu_j, k+\nu_k})\right)^2 + h_Q \sum_{\nu_j = \frac{1}{2}, 1} \sum_{\nu_i, \nu_k = 0, \frac{1}{2}, 1} \left(\delta_y u_h(\mathbf{x}_{i+\nu_i, j+\nu_j, k+\nu_k})\right)^2 \\ & + h_Q \sum_{\nu_k = \frac{1}{2}, 1} \sum_{\nu_i, \nu_j = 0, \frac{1}{2}, 1} \left(\delta_z u_h(\mathbf{x}_{i+\nu_i, j+\nu_j, k+\nu_k})\right)^2, \end{aligned}$$

and

$$\begin{aligned} \delta_x u_h(\mathbf{x}_{i+\nu_i, j+\nu_j, k+\nu_k}) &= u_h(\mathbf{x}_{i+\nu_i, j+\nu_j, k+\nu_k}) - u_h(\mathbf{x}_{i+\nu_i-\frac{1}{2}, j+\nu_j, k+\nu_k}), \\ \delta_y u_h(\mathbf{x}_{i+\nu_i, j+\nu_j, k+\nu_k}) &= u_h(\mathbf{x}_{i+\nu_i, j+\nu_j, k+\nu_k}) - u_h(\mathbf{x}_{i+\nu_i, j+\nu_j-\frac{1}{2}, k+\nu_k}), \\ \delta_z u_h(\mathbf{x}_{i+\nu_i, j+\nu_j, k+\nu_k}) &= u_h(\mathbf{x}_{i+\nu_i, j+\nu_j, k+\nu_k}) - u_h(\mathbf{x}_{i+\nu_i, j+\nu_j, k+\nu_k-\frac{1}{2}}). \end{aligned}$$

The following two lemmas indicate that the discrete norms defined in (12) and (13) are equivalent to the continuous L^2-norm or H^1-seminorm.

Lemma 1. *Assume that $\mathcal{E}_h$ satisfies Mesh Assumption A. Then there exist positive constants C_0 and C_1 independent of h such that for any $u_h \in S_h$,*

$$C_0\|u_h\|_0 \le |||u_h|||_0 \le C_1\|u_h\|_0, \tag{14}$$

$$C_0\|u_h\|_0 \le \|u_h\|_{0,h} \le C_1\|u_h\|_0. \tag{15}$$

Proof. Let $\hat{u}_h = u_h \circ F_Q$. For any $Q \in \mathcal{E}_h$, by the properties of affine mappings, we have

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

$$\|u_h\|_{L^2(Q)} \le \|J_{F_Q}\|_{L^\infty(\widehat{Q})}^{1/2} \|\hat{u}_h\|_{L^2(\widehat{Q})}.$$

Since the partition is regular, we have [6]

$$\|J_{F_Q^{-1}}\|_{L^\infty(Q)} \le Ch_Q^{-3}, \qquad \|J_{F_Q}\|_{L^\infty(\widehat{Q})} \le Ch_Q^3.$$

Therefore,

$$C^{-1}h_Q^{3/2}\|\hat{u}_h\|_{L^2(\widehat{Q})} \le \|u_h\|_{L^2(Q)} \le Ch_Q^{3/2}\|\hat{u}_h\|_{L^2(\widehat{Q})}. \tag{16}$$

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

$$C^{-1}h_Q^3 \int_{\widehat{Q}} \hat{u}_h \widehat{I_h^* u_h} d\hat{\mathbf{x}} \le \int_Q u_h I_h^* u_h d\mathbf{x} \le Ch_Q^3 \int_{\widehat{Q}} \hat{u}_h \widehat{I_h^* u_h} d\hat{\mathbf{x}}. \tag{17}$$

Using the standard tensor-product basis and the resulting interpolation form of $\hat{u}_h$ on $\widehat{Q}$, we obtain the following matrix forms of $\|\hat{u}_h\|^2_{L^2(\widehat{Q})}$ and $\int_{\widehat{Q}} \hat{u}_h \widehat{I_h^* u_h} d\hat{\mathbf{x}}$:

$$\|\hat{u}_h\|^2_{L^2(\widehat{Q})} = \mathbf{u}_Q(\mathcal{G}_1 \otimes \mathcal{G}_1 \otimes \mathcal{G}_1)\mathbf{u}_Q^T,$$

$$\int_{\widehat{Q}} \hat{u}_h \widehat{I_h^* u_h} d\hat{\mathbf{x}} = \mathbf{u}_Q(\mathcal{G}_2 \otimes \mathcal{G}_2 \otimes \mathcal{G}_2)\mathbf{u}_Q^T,$$

where $\mathbf{u}_Q \in \mathbb{R}^{27}$ is the nodal vector of u_h on Q and

$$\mathcal{G}_1 = \frac{1}{30}\begin{bmatrix} 4 & 2 & -1 \\ 2 & 16 & 2 \\ -1 & 2 & 4 \end{bmatrix}, \qquad \mathcal{G}_2 = \frac{1}{648}\begin{bmatrix} 83 & 32 & -7 \\ 32 & 368 & 32 \\ -7 & 32 & 83 \end{bmatrix}.$$

Since the matrices $\mathcal{G}_1$ and $\mathcal{G}_2$ are symmetric and positive definite, it is not difficult to see that $\|\hat{u}_h\|^2_{L^2(\widehat{Q})}$ and $\int_{\widehat{Q}} \hat{u}_h \widehat{I_h^* u_h} d\hat{\mathbf{x}}$ are equivalent. Applying (16) and (17) and summing the result over $\mathcal{E}_h$ yield estimate (14). Estimate (15) can be derived similarly. □

Lemma 2. *Assume that $\mathcal{E}_h$ satisfies Mesh Assumption A. Then there exist positive constants C_0 and C_1 independent of h such that for any $u_h \in S_h$,*

$$C_0|u_h|_1 \le |u_h|_{1,h} \le C_1|u_h|_1. \tag{18}$$

Proof. By mesh regularity and the properties of affine mappings again, we have

$$|\hat{u}_h|_{H^1(\widehat{Q})} \le C\|J_{F_Q^{-1}}\|^{1/2}_{L^\infty(Q)}|F_Q|_{W^1_\infty(\widehat{Q})}|u_h|_{H^1(Q)},$$

$$|u_h|_{H^1(Q)} \le C\|J_{F_Q}\|^{1/2}_{L^\infty(\widehat{Q})}|F_Q^{-1}|_{W^1_\infty(Q)}|\hat{u}_h|_{H^1(\widehat{Q})},$$

where

$$|F_Q|_{W^1_\infty(\widehat{Q})} \le Ch_Q, \qquad |F_Q^{-1}|_{W^1_\infty(Q)} \le Ch_Q^{-1}.$$

Therefore,

$$C^{-1}h_Q^{1/2}|\hat{u}_h|_{H^1(\widehat{Q})} \le |u_h|_{H^1(Q)} \le Ch_Q^{1/2}|\hat{u}_h|_{H^1(\widehat{Q})}. \tag{19}$$

We see from (13) that $|u_h|^2_{1,h,Q}$ has three parts related to δ_x, δ_y, and δ_z respectively. We now show that the first part

$$D_x(u_h) := h_Q \sum_{v_i=\frac{1}{2},1} \sum_{v_j,v_k=0,\frac{1}{2},1} \left(\delta_x u_h(\mathbf{x}_{i+v_i,j+v_j,k+v_k})\right)^2 \tag{20}$$

is equivalent to $h_Q \int_{\widehat{Q}} \left(\frac{\partial \hat{u}_h}{\partial \hat{x}}\right)^2 d\hat{\mathbf{x}}$. We define a vector $\mathbf{u}_{x,Q} \in \mathbb{R}^{18}$ of the values appearing in (20) as follows

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

Applying the standard tensor-product basis and the resulting interpolation form of $\widehat{u_h}$ on the reference element, we obtain immediately

$$\int_{\widehat{Q}} \left(\frac{\partial \hat{u}_h}{\partial \hat{x}}\right)^2 d\hat{\mathbf{x}} = \mathbf{u}_{x,Q}(\mathcal{H} \otimes \mathcal{G}_1 \otimes \mathcal{G}_1)\mathbf{u}_{x,Q}^T,$$

where $\mathcal{G}_1$ is the matrix defined in the proof of Lemma 1 and

$$\mathcal{H} = \frac{1}{3}\begin{bmatrix} 7 & -1 \\ -1 & 7 \end{bmatrix}.$$

Note that the matrices $\mathcal{G}_1$ and $\mathcal{H}$ are symmetric and positive definite. There exist two positive constants C_2 and C_3 related to the minimal and maximal eigenvalues of $\mathcal{G}_1$ and $\mathcal{H}$ such that

$$C_2 D_x(u_h) = C_2 h_Q \mathbf{u}_{x,Q}\mathbf{u}_{x,Q}^T \le h_Q \int_{\widehat{Q}} \left(\frac{\partial \hat{u}_h}{\partial \hat{x}}\right)^2 \mathrm{d}\hat{\mathbf{x}} \le C_3 h_Q \mathbf{u}_{x,Q}\mathbf{u}_{x,Q}^T = C_3 D_x(u_h).$$

Hence $D_x(u_h)$ is equivalent to $h_Q \int_{\widehat{Q}} (\frac{\partial \hat{u}_h}{\partial \hat{x}})^2 \mathrm{d}\hat{\mathbf{x}}$.

Similarly, we can prove that the other two parts in $|u_h|^2_{1,h,Q}$ are equivalent to $h_Q \int_{\widehat{Q}} (\frac{\partial \hat{u}_h}{\partial \hat{y}})^2 \mathrm{d}\hat{\mathbf{x}}$ and $h_Q \int_{\widehat{Q}} (\frac{\partial \hat{u}_h}{\partial \hat{z}})^2 \mathrm{d}\hat{\mathbf{x}}$, respectively. Therefore,

$$C_2 |u_h|^2_{1,h,Q} \le h_Q |\hat{u}_h|^2_{H^1(\widehat{Q})} \le C_3 |u_h|^2_{1,h,Q}.$$

Combining the above estimate with (19) and summing the result over $\mathcal{E}_h$, we obtain the desired norm equivalence (18). □

Let $I_h : H^3(\Omega) \cap H_0^1(\Omega) \to S_h$ be the usual nodal interpolation operator satisfying the approximation property [1]

$$\|u - I_h u\|_r \le Ch^{3-r}\|u\|_3, \quad 0 \le r \le 2. \tag{21}$$

The following trace theorem [1] will be used in the proof of Lemma 3 about continuity of $a_h(\cdot, I_h^*\cdot)$. For any domain Ω with a Lipschitz boundary, we have

$$\|u\|_{L^2(\partial\Omega)} \le C\|u\|^{1/2}_{L^2(\Omega)}\|u\|^{1/2}_{H^1(\Omega)}, \quad \forall u \in H^1(\Omega). \tag{22}$$

Lemma 3. *If $u \in H^3(\Omega) \cap H_0^1(\Omega)$, then there exists a positive constant C independent of h such that*

$$\left|a_h(u - I_h u, I_h^* v_h)\right| \le Ch^2\|u\|_3\|v_h\|_1, \quad \forall v_h \in S_h. \tag{23}$$

Proof. According to (10), we have

$$\begin{aligned} a_h(u - I_h u, I_h^* v_h) &= \sum_{Q\in\mathcal{E}_h} a_{Q,h}(u - I_h u, I_h^* v_h) \\ &:= \sum_{Q\in\mathcal{E}_h}\left(-\sum_{\omega\in Z_h\cap\bar{Q}} v_h(\omega)\int_{\partial V_\omega\cap Q} a\nabla(u - I_h u)\cdot\mathbf{n}\mathrm{d}s\right). \end{aligned}$$

Reordering by faces yields

$$\left|a_{Q,h}(u - I_h u, I_h^* v_h)\right| = \left|\sum_{\omega_1,\omega_2\in Z_h\cap\bar{Q}} (v_h(\omega_1) - v_h(\omega_2))\int_{\partial V_{\omega_1}\cap\partial V_{\omega_2}} a\nabla(u - I_h u)\cdot\mathbf{n}\mathrm{d}s\right|,$$

where ω_1, ω_2 are chosen in Q with no repetition. It follows from the approximation property (21) and the trace inequality (22) that

$$\begin{aligned} \left|\int_{\partial V_{\omega_1}\cap\partial V_{\omega_2}} a\nabla(u - I_h u)\cdot\mathbf{n}\mathrm{d}s\right| &\le Ca^* h_Q\|u - I_h u\|_{H^1(\partial V_{\omega_1}\cap\partial V_{\omega_2})} \\ &\le Ch_Q\|u - I_h u\|^{1/2}_{H^1(Q)}\|u - I_h u\|^{1/2}_{H^2(Q)} \\ &\le Ch_Q^{5/2}\|u\|_{H^3(Q)}. \end{aligned}$$

It is obvious from (13) that

$$|v_h(\omega_1) - v_h(\omega_2)| \le Ch_Q^{-1/2}|v_h|_{1,h,Q}.$$

Combining the above estimates gives

$$\left|a_{Q,h}(u - I_h u, I_h^* v_h)\right| \le Ch^2\|u\|_{H^3(Q)}|v_h|_{1,h,Q},$$

and therefore,

$$\left|a_h(u - I_h u, I_h^* v_h)\right| \le Ch^2\|u\|_3|v_h|_{1,h} \le Ch^2\|u\|_3\|v_h\|_1,$$

by the Cauchy–Schwarz inequality and Lemma 2. □

The following lemma about coercivity of the bilinear form $a_h(\cdot, I_h^*\cdot)$ plays a critical role in the convergence analysis.

Lemma 4. *Assume that $\mathcal{E}_h$ satisfies Mesh Assumption A & B. There exists a constant $C_0 > 0$ independent of h such that for sufficiently small h,*

$$a_h(u_h, I_h^* u_h) \geq C_0 \|u_h\|_1^2, \quad \forall u_h \in S_h. \tag{24}$$

Proof. Since each element of $\mathcal{E}_h$ is a right prism, any basis function $\alpha(\mathbf{x}) \in S_h$ can be written as $\alpha(\mathbf{x}) = \beta(x, y)\gamma(z)$, where $\beta(x, y)$ is an affine biquadratic basis function in x, y and $\gamma(z)$ a quadratic basis function in z. Tensor products can be used to simplify the proof.

Let

$$\bar{a}_{Q,h}(u_h, I_h^* v_h) = -\sum_{\omega \in Z_h \cap \bar{Q}} v_h(\omega) \int_{\partial V_\omega \cap Q} \bar{a}_Q \nabla u_h \cdot \mathbf{n} \mathrm{d}s,$$

where $\bar{a}_Q$ is the average of $a(x, y, z)$ over element Q. Reordering by faces as in Lemma 3, we obtain

$$\bar{a}_{Q,h}(u_h, I_h^* v_h) = \bar{a}_Q \left[|\mathbf{x}_{i,j,k+1} - \mathbf{x}_{i,j,k}| (\mathbf{v}_{x,Q}, \mathbf{v}_{y,Q})(\mathcal{A} \otimes \mathcal{G}_2)(\mathbf{u}_{x,Q}, \mathbf{u}_{y,Q})^T + \frac{1}{|\mathbf{x}_{i,j,k+1} - \mathbf{x}_{i,j,k}|} \mathbf{v}_{z,Q} (\mathcal{B} \otimes \mathcal{H}) \mathbf{u}_{z,Q}^T \right], \tag{25}$$

where $\mathbf{u}_{x,Q}, \mathbf{u}_{y,Q}, \mathbf{u}_{z,Q}$ are the vectors of the corresponding differences appearing in the definition of $|u_h|_{1,h,Q}$, $\mathcal{G}_2$ and $\mathcal{H}$ the matrices defined in the proofs of Lemmas 1 and 2, $\mathcal{A}$ the matrix of the integral $-\int_{P_1P_2P_3P_4} \Delta u_h I_h^* v_h \mathrm{d}x\mathrm{d}y$ (see Fig. 1), and $\mathcal{B}$ the matrix of $\int_{P_1P_2P_3P_4} u_h I_h^* v_h \mathrm{d}x\mathrm{d}y$, for any $u_h(x, y)$, $v_h(x, y) \in S_h$.

In general, neither $\mathcal{A}$ nor $\mathcal{B}$ is symmetric and positive definite. Making the same argument as the one for Lemma 3.9 in [19], we can prove that under Mesh Assumption A & B, when h is small enough, matrix $(\mathcal{A} + \mathcal{A}^T)$ is positive definite and its minimal eigenvalue $\lambda_1 > 0$ is independent of h. Next we examine the properties of matrix $\mathcal{B}$.

Let $w_h(x, y) \in S_h$ be an arbitrary function independent of z. Let $D = P_1P_2P_3P_4$ be the bottom face of Q. Then we have

$$\int_D w_h I_h^* w_h \mathrm{d}x\mathrm{d}y = \mathbf{w}_D \mathcal{B} \mathbf{w}_D^T = \frac{1}{2} \mathbf{w}_D (\mathcal{B} + \mathcal{B}^T) \mathbf{w}_D^T, \tag{26}$$

where $\mathbf{w}_D \in \mathbb{R}^9$ is a vector consisting of the values of w_h at the interpolation points in domain $\bar{D}$. Let F_D denote the bilinear affine mapping from $\widehat{D} = [0, 1]^2$ to D and $\hat{w}_h = w_h \circ F_D$. Similar to the proof of Lemma 1, we have

$$\int_D w_h I_h^* w_h \mathrm{d}x\mathrm{d}y \geq C^{-1} h_Q^2 \int_{\widehat{D}} \hat{w}_h \widehat{I_h^* w_h} \mathrm{d}\hat{x}\mathrm{d}\hat{y} \tag{27}$$

with

$$\int_{\widehat{D}} \hat{w}_h \widehat{I_h^* w_h} \mathrm{d}\hat{x}\mathrm{d}\hat{y} = \mathbf{w}_D (\mathcal{G}_2 \otimes \mathcal{G}_2) \mathbf{w}_D^T.$$

Let λ_2 be the minimal eigenvalue of $(\mathcal{B} + \mathcal{B}^T)$. Since matrix $\mathcal{G}_2$ is positive definite, we can use (26) and (27) to get

$$\lambda_2 \geq \inf_{0 \neq w_h(x,y) \in S_h} \frac{\mathbf{w}_D (\mathcal{B} + \mathcal{B}^T) \mathbf{w}_D^T}{\mathbf{w}_D \mathbf{w}_D^T} \geq C h_Q^2.$$

It is obvious from (25) that

$$\bar{a}_{Q,h}(v_h, I_h^* u_h) = \bar{a}_Q \left[|\mathbf{x}_{i,j,k+1} - \mathbf{x}_{i,j,k}| (\mathbf{v}_{x,Q}, \mathbf{v}_{y,Q})(\mathcal{A}^T \otimes \mathcal{G}_2^T)(\mathbf{u}_{x,Q}, \mathbf{u}_{y,Q})^T + \frac{1}{|\mathbf{x}_{i,j,k+1} - \mathbf{x}_{i,j,k}|} \mathbf{v}_{z,Q} (\mathcal{B}^T \otimes \mathcal{H}^T) \mathbf{u}_{z,Q}^T \right]. \tag{28}$$

Set $v_h = u_h$ in (25) and (28) and sum the results together. Mesh Assumption A, the symmetry of $\mathcal{G}_2$ and $\mathcal{H}$, and the above estimates combined yield

$$\begin{aligned}
\bar{a}_{Q,h}(u_h, I_h^* u_h) &= \frac{\bar{a}_Q}{2} \left[|\mathbf{x}_{i,j,k+1} - \mathbf{x}_{i,j,k}| (\mathbf{u}_{x,Q}, \mathbf{u}_{y,Q}) \left((\mathcal{A} + \mathcal{A}^T) \otimes \mathcal{G}_2 \right) (\mathbf{u}_{x,Q}, \mathbf{u}_{y,Q})^T \right. \\
&\quad \left. + \frac{1}{|\mathbf{x}_{i,j,k+1} - \mathbf{x}_{i,j,k}|} \mathbf{u}_{z,Q} \left((\mathcal{B} + \mathcal{B}^T) \otimes \mathcal{H} \right) \mathbf{u}_{z,Q}^T \right] \\
&\geq C_1 \lambda_1 \frac{h_Q}{\sigma} (\mathbf{u}_{x,Q}, \mathbf{u}_{y,Q})(\mathbf{u}_{x,Q}, \mathbf{u}_{y,Q})^T + C_2 \lambda_2 \frac{1}{h_Q} \mathbf{u}_{z,Q} \mathbf{u}_{z,Q}^T \\
&\geq C |u_h|_{1,h,Q}^2.
\end{aligned} \tag{29}$$

By Lemma 2, we have

$$\bar{a}_h(u_h, I_h^* u_h) \geq C|u_h|_{1,h}^2 \geq C|u_h|_1^2,$$

where $\bar{a}_h$ is understood as the summation of $\bar{a}_{Q,h}$ over all elements in $\mathcal{E}_h$. The Lipschitz continuity of $a(x, y, z)$ and a similar argument to that in the proof of Lemma 3 lead to

$$\left|a_h(u_h, I_h^* u_h) - \bar{a}_h(u_h, I_h^* u_h)\right| \leq CLh|u_h|_1^2.$$

The desired coercivity (24) follows from the above two estimates and the Poincaré's inequality. □

Remark 3.1. As shown in the proof of Lemma 4, the symmetry of matrices $\mathcal{G}_2$ and $\mathcal{H}$ is needed for (29) to hold. The symmetry relies on the dual partition introduced in this paper. One can verify that $\mathcal{G}_2$ will not be symmetric if the ratio of the dual segments is set as 1:2:1. It is therefore very difficult to prove coercivity of the bilinear form $a_h(\cdot, I_h^* \cdot)$ in this circumstance.

Now we are ready to state and prove the main theorem about the optimal error estimate for the finite volume scheme.

Theorem 1. *Let u be the solution of (1) and u_h the numerical solution of the finite volume scheme (11). Assume that $u \in H^3(\Omega) \cap H_0^1(\Omega)$. If Mesh Assumption A & B are satisfied, then for sufficiently small h,*

$$\|u_h - u\|_1 \leq Ch^2\|u\|_3. \tag{30}$$

Proof. We decompose the error as $u_h - u = \xi - \eta$, where $\xi = u_h - I_h u$ and $\eta = u - I_h u$. By (9) and (11), and Lemma 4 (coercivity), we have

$$C_0\|\xi\|_1^2 + (b\xi, I_h^*\xi) \leq a_h(\xi, I_h^*\xi) + (b\xi, I_h^*\xi) = a_h(\eta, I_h^*\xi) + (b\eta, I_h^*\xi). \tag{31}$$

It follows from Lemma 3 (continuity) and the approximation property (21) that

$$a_h(\eta, I_h^*\xi) + (b\eta, I_h^*\xi) \leq Ch^2\|u\|_3\|\xi\|_1 + Cb^*h^3\|u\|_3\|\xi\|_{0,h}. \tag{32}$$

For any element $Q \in \mathcal{E}_h$, let $\bar{b}_Q$ be the average of $b(x, y, z)$ over the element. We also define $\bar{b}$ as a piecewise constant that takes the value $\bar{b}_Q$ for each $Q \in \mathcal{E}_h$. From the proof of Lemma 1, it is clear that $(\bar{b}_Q\xi, I_h^*\xi) \geq 0$. On the other hand, by the Lipschitz continuity of $b(x, y, z)$, we have

$$|((\bar{b} - b)\xi, I_h^*\xi)| \leq CLh\|\xi\|_0\|\xi\|_{0,h},$$

By (31) and (32), and Lemma 1 (norm equivalence), we obtain

$$\begin{aligned} C_0\|\xi\|_1^2 - CLh\|\xi\|_0^2 &\leq Ch^2\|u\|_3\|\xi\|_1 + Ch^3\|u\|_3\|\xi\|_{0,h} \\ &\leq C(h^2 + h^3)\|u\|_3\|\xi\|_1. \end{aligned}$$

Choosing h small enough yields

$$\|\xi\|_1 \leq Ch^2\|u\|_3. \tag{33}$$

The approximation property (21) and a triangle inequality lead to

$$\|u_h - u\|_1 \leq \|\xi\|_1 + \|\eta\|_1 \leq Ch^2\|u\|_3,$$

which completes the proof. □

4. Numerical experiments

A suite of C++ code has been developed to validate the finite volume scheme. We implement the scheme following the methodology of finite elements, since it is essentially a Petrov–Galerkin method. The third (or higher) order Gaussian quadratures are used to evaluate the surface integral (the first term on the left hand side) and the volume integrals (the second term on the left hand side and the term on the right hand side) in the finite volume scheme (11). Element stiffness and mass matrices are assembled into a large sparse linear system, which is nonsymmetric and can be solved by the preconditioned bi-conjugate gradient stabilized method (BiCGStab) [15].

Now we present numerical results to illustrate the proved error estimates. We test an example on the rectangular domain $\Omega = [0, \pi]^2 \times [0, 1]$ with $a(x, y, z) = ((x + 1)^2 + y^2)e^z$ and $b(x, y, z) \equiv 1$. The exact solution is $u(x, y, z) = \sin(x)\sin(y)z(1 - z)$, and the source term $f(x, y, z)$ is computed accordingly. We test the finite volume method on two sets of meshes. The first set is a family of rectangular meshes that have $M = 4$, 8, 16, 32, or 64 uniform partitions in the x, y, z-directions. The second set consists of right quadrangular prism meshes. In the xy-plane, we have quadrilateral meshes that

Table 1
Errors and convergence rates on the rectangular meshes.

M	$\|I_h u - u_h\|_{l^\infty}$	Rate	$\|I_h u - u_h\|_{0,h}$	Rate	$\|I_h u - u_h\|_{1,h}$	Rate
4	4.405E−4	–	5.071E−5	–	3.947E−3	–
8	1.009E−4	2.12	5.517E−6	3.20	9.440E−4	2.06
16	2.428E−5	2.05	5.203E−7	3.40	2.193E−4	2.10
32	5.974E−6	2.02	4.677E−8	3.47	5.342E−5	2.03
64	1.482E−6	2.01	4.154E−9	3.49	1.329E−5	2.00

Table 2
Errors and convergence rates on the prism meshes.

M	$\|I_h u - u_h\|_{l^\infty}$	Rate	$\|I_h u - u_h\|_{0,h}$	Rate	$\|I_h u - u_h\|_{1,h}$	Rate
4	5.790E−3	–	4.068E−4	–	5.506E−2	–
8	1.458E−3	1.98	4.245E−5	3.26	1.787E−2	1.62
16	3.589E−4	2.02	4.109E−6	3.36	4.414E−3	2.01
32	8.837E−5	2.02	3.740E−7	3.45	1.082E−3	2.02
64	2.199E−5	2.00	3.332E−8	3.48	2.688E−4	2.00

are perturbations of the rectangular meshes. To be specific, the quadrilateral meshes have nodes

$$x_{i,j} = \frac{\pi i}{M}, \qquad y_{i,j} = \frac{\pi j}{M} + \frac{\pi}{2M^2} \sin\left(\frac{2\pi i}{M}\right) \sin\left(\frac{2\pi j}{M}\right), \quad 0 \le i, j \le M.$$

The partitions in the z-direction are the same as those in the rectangular meshes. Obviously, Mesh Assumption A & B are satisfied for these two sets of meshes.

BiCGStab is employed to solve the discrete linear systems with the tolerance for residuals set as 10^{-11}. The simple diagonal preconditioning is used. Based on the approximation property (21) and the norm equivalences (15) and (18), we measure $|I_h u - u_h|_{1,h}$ in lieu of $\|u - u_h\|_1$. The second order convergence in $|I_h u - u_h|_{1,h}$ is clearly exhibited for both rectangular and prism meshes, as proved in Theorem 1. As two by-products, the errors in $\|I_h u - u_h\|_{l^\infty}$ and $\|I_h u - u_h\|_{0,h}$ are also reported. One can observe the second and third order convergence in these two quantities respectively in Tables 1 and 2.

Acknowledgments

The authors would like to express their sincere thanks to the referees for their very valuable comments and suggestions, which greatly improved the quality of this paper.

References

[1] S.C. Brenner, L.R. Scott, The Mathematical Theory of Finite Element Methods, 2nd ed., Springer Verlag, New York, 2002.
[2] Z. Cai, J. Douglas, M. Park, Development and analysis of higher order finite volume methods over rectangles for elliptic equations, Adv. Comput. Math. 19 (2003) 3–33.
[3] P. Chatzipantelidis, Finite volume methods for elliptic PDE's: A new approach, M2AN Math. Model. Numer. Anal. 36 (2002) 307–324.
[4] S.H. Chou, Q. Li, Error estimates in L^2, H^1 and L^∞ in covolume methods for elliptic and parabolic problems: A unified approach, Math. Comp. 69 (2000) 103–120.
[5] S.H. Chou, X. Ye, Unified analysis of finite volume methods for second order elliptic problems, SIAM J. Numer. Anal. 45 (2007) 1639–1653.
[6] P.G. Ciarlet, The Finite Element Method for Elliptic Problems, SIAM, Philadelphia, 2002.
[7] 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.
[8] R. Eymard, T. Galloët, R. Herbin, Finite Volume Methods: Handbook of Numerical Analysis, North-Holland, Amsterdam, 2000.
[9] F. Gao, Y. Yuan, The upwind finite volume element method based on straight triangular prism partition for nonlinear convection-diffusion problems, Appl. Math. Comput. 181 (2006) 1229–1242.
[10] J. Huang, S. Xi, On the finite volume element method for general self-adjoint elliptic problems, SIAM J. Numer. Anal. 35 (1998) 1762–1774.
[11] R. Li, Z. Chen, W. Wu, Generalized difference methods for differential equations, in: Numerical Analysis of Finite Volume Methods, Marcel Dekker, New York, 2000.
[12] F. Liebau, The finite volume element method with quadratic basis functions, Computing 57 (1996) 281–299.
[13] I.D. Mishev, Finite volume element methods for nondefinite problems, Numer. Math. 83 (1999) 161–175.
[14] M. Plexousakis, G.E. Zouraris, On the construction and analysis of high order locally conservative finite volume-type methods for one dimensional elliptic problems, SIAM J. Numer. Anal. 42 (2004) 1226–1260.
[15] Y. Saad, Iterative Methods for Sparse Linear Systems, 2nd ed., SIAM, Philadelphia, 2003.
[16] C.W. Shu, High-order finite difference and finite volume WENO schemes and discontinuous Galerkin methods for CFD, Int. J. Comput. Fluid Dyn. 17 (2003) 107–118.
[17] E. Süli, The accuracy of cell vertex finite volume methods on quadrilateral meshes, Math. Comp. 59 (1992) 359–382.
[18] Z.J. Wang, Spectral (finite) volume methods for conservation laws on unstructured grids: Basic formulation, J. Comput. Phys. 178 (2002) 210–251.
[19] M. Yang, A second-order finite volume element method on quadrilateral meshes for elliptic equations, M2AN Math. Model. Numer. Anal. 40 (2006) 1053–1068.