Numerical integration with Taylor truncations for the quadrilateral and hexahedral finite elements

Shangyou Zhang*

Department of Mathematical Sciences, University of Delaware, Newark, DE 19716, USA

Received 10 January 2005; received in revised form 17 May 2006

Abstract

For general quadrilateral or hexahedral meshes, the finite-element methods require evaluation of integrals of rational functions, instead of traditional polynomials. It remains as a challenge in mathematics to show the traditional Gauss quadratures would ensure the correct order of approximation for the numerical integration in general. However, in the case of nested refinement, the refined quadrilaterals and hexahedra converge to parallelograms and parallelepipeds, respectively. Based on this observation, the rational functions of inverse Jacobians can be approximated by the Taylor expansion with truncation. Then the Gauss quadrature of exact order can be adopted for the resulting integrals of polynomials, retaining the optimal order approximation of the finite-element methods. A theoretic justification and some numerical verification are provided in the paper.

© 2006 Elsevier B.V. All rights reserved.

MSC: 65N30; 65N50; 65D30

Keywords: Quadrilateral finite elements; Hexahedral finite elements; Gauss quadrature; Nested refinement; Multigrid refinement

1. Introduction

Instead of integrals of polynomials in the traditional triangular or tetrahedral finite-element method, we need to evaluate integrals of rational functions in the quadrilateral or hexahedral element method. This is because, for non-parallelogram or non-parallelepiped elements, the reference mappings are bilinear or trilinear functions and the resulting inverse Jacobian determinants and inverse Jacobian matrices would be no longer constants. They are rational functions. Although it is a common practice in engineering that Gauss quadratures are applied to the numerical integration of rational functions arising from quadrilateral or hexahedral finite elements, there is no general mathematical theory. Large errors would be encountered in practice due to the use of usual quadrature rules (for polynomials), cf. [6,8,9]. So far, for some (low order and in 2D, mostly) finite elements, special quadrature rules or even exact rules were proposed and analyzed (cf. [4,13,16,17,19,21,24]). For specific finite elements, one may obtain specific error bounds with explicit dependence on the shape regularity of grids, for a quadrature rule. In general, three methods are commonly used. The method of (product) high-order Gauss quadratures (cf. [22,14]), or Newton–Cotes formulas (cf. [5,25]), applied directly to integrals of rational functions, may not guarantee the required order of accuracy for the numerical integration. For example, some Gauss quadratures are listed in [5] for commonly used low-order quadrilateral elements. The method

* Tel.: +1 302 8310625; fax: +1 302 8314511.
E-mail address: szhang@udel.edu.

0377-0427/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.cam.2006.05.007
of adaptive, Romberg-type integrals (cf. [5,15,23]) is costly, as a low-order quadrature rule would be repeatedly used for a same integral. Furthermore, such extrapolated results lack mathematical justification in order to retain the order of convergence of the finite-element solution. The same difficulty is encountered in the method of Gauss quadratures. Finally, the method of Gauss quadratures for integrals with rational weights [12,17,21] is costly, and not practical currently for general grids, as the quadrature rules differ on each element, except for some special cases [18,20].

In the case of nested refinement, the high-level, refined quadrilaterals and hexahedra converge to parallelograms and parallelepipeds rapidly (cf. [1,26,27]). Based on this observation, it is proposed (cf. [1–3,26]) and shown that the rational functions of inverse Jacobian matrices and Jacobian determinants can be replaced by constants for bilinear or trilinear elements, while retaining the optimal order of convergence. In this paper, we extend the idea in [26] to higher order quadrilateral and hexahedral elements. The rational functions of Jacobians are approximated by their Taylor expansions with truncations of a certain order, depending on the finite-element order. Then the Gauss quadrature of exact order can be adopted for the resulting integrals of polynomials on the reference square or cube, retaining the optimal order approximation of finite elements. For quadrature rules on squares and n-cubes, further, on n-spheres, n-spaces, n-simplex, and even with certain weight functions, we refer readers to [10] and references therein. In the finite-element computation, we usually use the Gauss product formula for multiple integrals, which may not be optimal in the sense that less quadrature points might achieve the same order of accuracy (cf. [11]). The theoretic justification and numerical verification of the proposed Taylor truncation method are provided in this paper. The method of Taylor expansion of rational functions is used in numerical calculation of surface integrals previously [5,15,23].

We should remark that the proposed method here, approximating inverse Jacobians and Jacobian matrices by the first few terms of their Taylor polynomials, may not be necessary mathematically, neither is efficient in computation. It is much easier to evaluate rational functions than their Taylor polynomials. But we do not have a theory to ensure the order of accuracy of the numerical integration for the former method yet.

2. Nested refinement

In this section, we consider the nested refinement of general quadrilaterals and hexahedra (see Fig. 1). Specifically, we use only middle points to subdivide a quadrilateral element into four half-sized elements, or a hexahedral element into eight subelements.

![Fig. 1. The standard nested refinement of a hexahedron or a quadrilateral.](image-url)
We note that a quadrilateral and hexahedral element $K$ is defined by its four or eight vertices, $\{v_i\}$, via the $Q_1$ referencing mapping which maps the reference square or cube (shown in Figs. 2 and 3): $\hat{K} = [-1, 1]^n$, $n = 2, 3$, to $K$ by

$$F(\hat{x}_1, \ldots, \hat{x}_n) = \sum_{i=1}^{2^n} v_i b_i(\hat{x}_1, \ldots, \hat{x}_n), \quad n = 2, 3,$$

where $\{b_i\}$ are nodal basis functions:

$$b_i(\hat{x}_1, \hat{x}_2) = (1 \pm \hat{x}_1)(1 \pm \hat{x}_2)/4, \quad n = 2,$$

$$b_i(\hat{x}_1, \hat{x}_2, \hat{x}_3) = (1 \pm \hat{x}_1)(1 \pm \hat{x}_2)(1 \pm \hat{x}_3)/8, \quad n = 3.$$

In the nested refinement, the shapes of subelements would become more regular, closer to parallelograms or parallelepipeds. To see this, we define a measure $s(K)$ for non-parallelism of quadrilaterals and hexahedra. The $s(K)$ is simply the distance between the middle points of two diagonals of a quadrilateral in 2D (see Fig. 2), or the maximum of all the six face $s(K[i])$.

**Definition 2.1.** The *irregularity* of $K$ is defined as, letting $\| \cdot \|_2$ denote $l_2$ vector norm,

$$s(K) = \begin{cases} \frac{v_1 + v_3}{2} - \frac{v_2 + v_4}{2} \quad &\text{for } 2D \ K \ (\text{see Fig. 2}), \\ \max_{1 \leq i \leq 6} s(K[i]) \quad &\text{for } 3D \ K, \text{ where } K[i] \text{ are faces of } K \ (\text{see Fig. 3}). \end{cases}$$
Proposition 2.1 (Zhang [26]). \( s(K) = 0 \) if and only if \( K \) is a parallelogram in 2D, or a parallelepiped in 3D.

One step of the nested refinement of the reference element \( \hat{K} \) in 2D or 3D is done by subdividing it into four or eight subelements with the coordinate axes or planes, respectively, i.e.,

\[
\hat{K} = (I_1 \times I_1) \cup (I_2 \times I_1) \cup (I_1 \times I_2) \cup (I_1 \times I_2) \quad \text{in 2D},
\]

\[
\hat{K} = (I_1 \times I_1 \times I_1) \cup (I_2 \times I_1 \times I_1) \cup (I_2 \times I_2 \times I_1) \cup (I_1 \times I_2 \times I_1) \cup (I_1 \times I_1 \times I_2) \cup (I_2 \times I_1 \times I_2) \cup (I_2 \times I_2 \times I_2) \cup (I_1 \times I_2 \times I_2) \quad \text{in 3D},
\]

where \( I_1 = [-1, 0] \) and \( I_2 = [0, 1] \). The refinement of a general \( K \) is defined below. We refer to [26] for the correctness of the definition in producing a nested refinement of grids on a given domain.

Definition 2.2. One step of the nested refinement of \( K \) with an associated reference mapping \( F \) is defined by the images of vertices of subelements of \( \hat{K} \) under the mapping \( F \).

We can recursively apply Definition 2.2 to each of the new subelements, to define elements of higher levels. The level \( k \) nested refinement of \( K \) consists \( 2^{n(k-1)} \) subelements of \( K, K_{i,k} \). Here \( K = K_{1,k} \). The subelements converge to parallelograms or parallelepipeds rapidly as stated in the following theorem. It says that after one level of refinement, the size of subelements are about \( \frac{1}{2} \) of the original size, but the irregularity of each subelement is \( \frac{1}{4} \), or less, of that of the original element.

Theorem 2.1 (cf. Zhang [26]). Let \( K_k \) be one of the four or eight subelements nestedly refined from \( K \).

\[
s(K_k) \leq \frac{s(K)}{2^2}. \tag{4}
\]

We next define the finite-element approximation subspaces. Instead of using the grid size \( h \) as the parameter for the finite-element space, we use the space level \( k \) as the index, as in the multigrid method. Let \( \mathcal{T}_1 = \{K\}_{K \in \mathcal{T}_1} \) be a given initial quadrilateral or hexahedral grid as usual, that is,\[
\bigcup K = \Omega, \quad K^0 \cap K^0 = \emptyset \quad \forall K' (\neq K) \in \mathcal{T}_1, \\
K \cap K' = \text{a common face or an edge or a vertex of both } K \text{ and } K', \quad \text{or the empty set.} \tag{5}
\]
The nested refinement defines grids\[
\mathcal{T}_k = \{K_{i,k}\}_{K \in \mathcal{T}_1, i=1,2,...,2^{n(k-1)}}, \quad k = 1, 2, \ldots. \tag{6}
\]
The finite-element spaces in 2D or 3D \((n = 2\) or \(3\)) are continuous (rational) functions defined by piecewise referencing mappings of polynomials of total degrees \( m \) or less:

\[
V_k = \left\{ u \in C(\Omega) \mid u \big|_K = \hat{u} \circ F^{-1}(x_1, x_2, x_n), \hat{u} = \sum_{i_1+i_2+i_n \leq m} u_{i_1i_2i_n} \tilde{x}_1^{i_1} \tilde{x}_2^{i_2} \tilde{x}_n^{i_n} \right\}. \tag{7}
\]

3. Taylor expansion of Jacobians

We will apply Theorem 2.1 to estimate the Taylor expansion of Jacobians of the referencing mappings under the nested refinement. According to the degree of polynomials used in the finite-element method, different truncations will be applied to the Jacobian matrices and Jacobian determinants, so that the correct order of accuracy of a quadrature rule can be achieved to ensure the optimal rate of convergence of the finite element.
We consider the 2D case first and then the 3D case in this section. Given four vertices of a strictly convex quadrilateral \( K \) in 2D, the reference mapping (1) is a \( Q_1 \) function:

\[
F(\hat{x}_1, \hat{x}_2) = \sum_{i=1}^{4} v_i b_i(\hat{x}_1, \hat{x}_2) = a_{00} + a_{10} \hat{x}_1 + a_{01} \hat{x}_2 + a_{11} \hat{x}_1 \hat{x}_2.
\] (8)

The coefficients (vectors) are

\[
\begin{cases}
a_{00} = (v_1 + v_3 + v_2 + v_4)/4, \\
a_{10} = (v_{12} + v_{43})/4, \\
a_{01} = (v_{14} + v_{23})/4, \\
a_{11} = (-v_{12} + v_{43})/4,
\end{cases}
\] (9)

where

\[ v_{ij} = v_j - v_i. \] (10)

We note that (cf. [26]), for \( l_2 \) vector norm,

\[ \|a_{11}\|_2 = \frac{s(K)}{2}. \] (11)

On one element \( K \in \mathcal{T}_k \), a finite-element function in \( V_k \), defined by (7), has the following form:

\[ u|_{K}(x_1, x_2) = \hat{u} \circ F^{-1}(x_1, x_2) = \sum_{j=1}^{(m+1)^2} \hat{u}(v_j^{(m)}) b_j^{(m)}(F^{-1}(x_1, x_2)). \]

Here \( b_j^{(m)} \) are nodal basis functions of \( Q_m \) polynomial on the reference element \( \hat{K} \), and \( \hat{v}_j^{(m)} \) the uniform tensor-product nodes. In particular, \( b_j^{(1)} = b_j \) are listed in (2). For the partial derivatives on an element \( K \), we have (\( n = 2 \) here)

\[
\frac{\partial b_j^{(m)}(F^{-1}(x_1, x_2))}{\partial x_i} = \sum_{j=1}^{n} \frac{\partial \hat{b}_j^{(m)}}{\partial \hat{x}_i} \frac{\partial \hat{x}_i}{\partial x_i}
\]

Since \( F^{-1}(F(\hat{x}_1, \hat{x}_2)) = (\hat{x}_1, \hat{x}_2) \) on \( \hat{K} \), i.e.,

\[
\sum_{i=1}^{n} \frac{\partial \hat{x}_i}{\partial x_i} (F(\hat{x}_1, \hat{x}_2)) \frac{\partial x_i}{\partial \hat{x}_j} (\hat{x}_1, \hat{x}_2) = \delta_{ij}, \quad 1 \leq i, j \leq n,
\]

we have the following relation for the Jacobian matrices:

\[
\left( \frac{\partial \hat{x}_i}{\partial x_j} \right)_{n \times n} = \left( \frac{\partial x_i}{\partial \hat{x}_j} \right)_{n \times n}^{-1} = \left( \frac{\partial F}{\partial \hat{x}} \right)^{-1}_{n \times n}. \] (12)

Here for simplicity \( \hat{x} \) stands for \( (\hat{x}_1, \hat{x}_2) \), or for \( (\hat{x}_1, \hat{x}_2, \hat{x}_3) \) later. We write the Jacobians in terms of the coefficients of (9) and (8) as follows:

\[
S_K = \begin{pmatrix} \frac{\partial x_j}{\partial \hat{x}_i} \end{pmatrix}_{2 \times 2} = (a_{10} + a_{11} \hat{x}_2 \ a_{01} + a_{11} \hat{x}_1),
\] (13)

\[
J_K = \det S_K = \det(s_1 \ s_2) + \det(s_1 \ s_4) + \det(s_3 \ s_2) + \det(s_3 \ s_4)
= d_{K,0} + d_{K,1}, \] (14)

where

\[
s_1 = a_{10}, \quad s_2 = a_{01}, \quad s_3 = a_{11} \hat{x}_2, \quad s_4 = a_{11} \hat{x}_1.
\]
and
\[
d_{K,0} = \det(s_1 \ s_2), \quad d_{K,1} = \det(s_1 \ s_4) + \det(s_3 \ s_2). \tag{15}
\]

We note that \(\det(s_3 \ s_4) = 0\), as its two column vectors are linearly dependent. Therefore, the inverse Jacobian matrix, which is needed in quadrature rules in computing bilinear forms of the finite-element method, is a rational function where both numerators and denominators are \(Q_1\) functions in 2D.

We define the inverse matrix of \(S_K\) by
\[
T_K = S_K^{-1} = \left( \frac{\partial \xi_j}{\partial x_i} \right)_{2 \times 2} = \frac{M_K}{\tilde{J}_K}, \tag{16}
\]
where \(M_K\) is the transpose of the adjoint matrix of \(S_K\),
\[
M_K = \text{adj}(S_K)^T = \text{adj}(s_1 + s_3 \ s_2 + s_4)^T, \tag{17}
\]
\[
M_{K,0} = \text{adj}(s_1 \ s_2)^T, \tag{18}
\]
\[
M_{K,1} = \text{adj}(s_3 \ s_4)^T. \tag{19}
\]

We denote the determinant of the inverse Jacobian matrix by \(\tilde{J}_K\),
\[
\tilde{J}_K = \det(T_k) = \det \left( \frac{\partial \xi_j}{\partial x_i} \right) = \frac{1}{J_K}. \tag{20}
\]

It is shown in [3, 26] that we can drop the non-constant terms in both \(M_K\) and \(\tilde{J}_K\), i.e., in both the numerator and the denominator of (16). In particular,
\[
\tilde{J}_K = \frac{1}{d_{1001} + d_{1011} \hat{x}_1 + d_{1101} \hat{x}_2} \simeq \frac{1}{d_{1001}!}.
\]

But for higher order \(Q_m\) finite elements, we have to use some high-order approximation. This extends our earlier work in [26]. The extension is listed as a definition next.

**Definition 3.1.** Let \(K\) be a general quadrilateral in 2D, \(F\) the reference mapping (defined in (8)), \(T_K\) the inverse Jacobian matrix \((\partial F^{-1}/\partial \xi)\) (defined in (16)), \(J_K\) the Jacobian \(det(\partial F/\partial \xi)\) (defined in (14)), \(M_K = J_K T_K\) (defined in (17)), and \(\tilde{J}_K = J_K^{-1}\) (defined in (20)). The Taylor truncations are defined as follows:

\[
M^{(m)}_K = \begin{cases} 
M_{K,0} & \text{if } m = 1, \\
M_{K,0} + M_{K,1} & \text{if } m \geq 2,
\end{cases} \tag{21}
\]

\[
J^{(m)}_K = \begin{cases} 
d_{K,0} & \text{if } m = 1, \\
d_{K,0}(1 + c_1) & \text{if } m \geq 2,
\end{cases} \tag{22}
\]

\[
\tilde{J}^{(m)}_K = \begin{cases} 
d^{-1}_{K,0} & \text{if } m = 1, \\
d^{-1}_{K,0}(1 - c_1) & \text{if } m = 2, \\
d^{-1}_{K,0} \sum_{i=0}^{k-1} (-1)^i c_1^i & \text{if } m \geq k,
\end{cases} \tag{23}
\]

where \(c_1 = d_{K,1}/d_{K,0}\).
Lemma 3.1. Let \( K \) be a general quadrilateral and \( K_s \) the subquadrilateral of \( K \) having \( v_1 \) as a vertex (see Figs. 2 and 1). The following relations hold:

\[
\frac{1}{4} \min_{1 \leq i \leq 4} J_K(v_i) \lesssim d_{K_s,0} \leq \frac{1}{4} \max_{1 \leq i \leq 4} J_K(v_i),
\]

(24)

\[
d_{K_s,1} \leq \frac{1}{8}(|\det(s_1 s_4)| + |\det(s_3 s_2)|),
\]

(25)

\[
\|M_{K_s,0}\| \leq \frac{1}{2} \max\{\|\text{adj}(v_{12} v_{14})\|^T, \|\text{adj}(v_{12} v_{23})\|^T, \|\text{adj}(v_{43} v_{14})\|^T, \|\text{adj}(v_{43} v_{23})\|^T\},
\]

(26)

\[
M_{K_s,1} = \frac{1}{4} M_{K,1},
\]

(27)

where \( \| \cdot \| \) denotes the \( l^2 \) matrix norm, and variables with and without subscript \( K_s \) are defined for the subquadrilateral \( K_s \) or for \( K \).

Proof. The four vertices of \( K_s \) are (see Figs. 2 and 1)

\[
v_{1,K_s} = v_1, \quad v_{2,K_s} = \frac{v_1 + v_2}{2}, \quad v_{3,K_s} = \frac{v_1 + v_2 + v_3 + v_4}{4}, \quad v_{4,K_s} = \frac{v_3 + v_4}{2}.
\]

(28)

Therefore, we can compare the coefficients of the two reference mappings:

\[
a_{10} = \frac{v_{12} + v_{43}}{4}, \quad a_{01} = \frac{v_{14} + v_{23}}{4}, \quad a_{11} = \frac{v_{43} - v_{12}}{4},
\]

\[
a_{10,k_s} = \frac{3v_{12} + v_{43}}{16}, \quad a_{01,k_s} = \frac{3v_{14} + v_{23}}{16}, \quad a_{11,k_s} = \frac{v_{43} - v_{12}}{16} = \frac{a_{11}}{4}.
\]

Relations (24)–(27) would be verified immediately. \( \square \)

Corollary 3.1. Let \( \mathcal{T}_k = \{K_{i,k}\}_{K \in \mathcal{T}_i, i=1,2,...,2^m(k-1)} \) be the \( k \)th level nested refinement of an initial quadrilateral grid \( \mathcal{T}_1 \). Then

\[
\frac{C_0}{2^n(k-1)} \leq d_{K_{i,k},0} \leq \frac{C_1}{2^n(k-1)},
\]

\[
d_{K_{i,k},j} \leq \frac{C_1}{2^n(k-1)2^j(k-1)}, \quad j = 0, 1,
\]

\[
\|M_{K_{i,k},j}\| \leq \frac{C_1}{2^n(k-1)2^j(k-1)}, \quad j = 0, 1,
\]

where \( C_0 \) and \( C_1 \) are positive constants depending on \( \mathcal{T}_1 \) only.

Next, for 3D elements, given eight vertices \( \{v_i\} \) we define an 8-vertex element by the following 3D \( Q_1 \) mapping, where the \( Q_1 \) basis functions are defined in (3),

\[
F(\hat{x}_1, \hat{x}_2, \hat{x}_3) = \sum_{i=1}^{8} v_i b_i(\hat{x}_1, \hat{x}_2, \hat{x}_3)
\]

\[
= a_{000} + a_{100} \hat{x}_1 + a_{010} \hat{x}_2 + a_{001} \hat{x}_3 + a_{110} \hat{x}_1 \hat{x}_2 + a_{101} \hat{x}_1 \hat{x}_3 + a_{011} \hat{x}_2 \hat{x}_3 + a_{111} \hat{x}_1 \hat{x}_2 \hat{x}_3,
\]

(29)

where \( a_{ijkl} \) are constants defined by the coordinates of the eight vertices of the element. We note that the high-order coefficients in (29) are related to the distances of two midpoints of each pair of diagonals of face quadrilaterals, used in the definition of element irregularity in Definition 2.1:

\[
a_{110} = (v_5 + v_7 - v_6 - v_8)/8 + (v_1 + v_3 - v_2 - v_4)/8 \quad \text{(top and bottom faces)},
\]

\[
a_{101} = (v_4 + v_7 - v_3 - v_8)/8 + (v_1 + v_6 - v_2 - v_5)/8 \quad \text{(left and right faces)},
\]
We also let (similar to the 2D notation)

\[ a_{011} = \frac{(v_2 + v_7 - v_3 - v_6)}{8} + \frac{(v_1 + v_8 - v_4 - v_5)}{8} \]  
(front and back faces),

\[ a_{111} = \frac{(v_5 + v_7 - v_6 - v_8)}{8} - \frac{(v_1 + v_3 - v_2 - v_4)}{8} \]

\[ = \frac{(v_4 + v_7 - v_3 - v_8)}{8} - \frac{(v_1 + v_6 - v_2 - v_5)}{8} \]

\[ = \frac{(v_2 + v_7 - v_3 - v_6)}{8} - \frac{(v_1 + v_8 - v_4 - v_5)}{8}. \]  
(30)

Similar to the 2D case, we have

\[ S_K = \left( \frac{\partial x_i}{\partial x_j} \right)_{3 \times 3} = \begin{pmatrix} (a_{100} + a_{110}\hat{x}_2 + a_{101}\hat{x}_3 + a_{111}\hat{x}_2\hat{x}_3)^T & (a_{101} + a_{110}\hat{x}_1 + a_{101}\hat{x}_3 + a_{111}\hat{x}_1\hat{x}_3)^T \\ \left(a_{001} + a_{010}\hat{x}_1 + a_{011}\hat{x}_2 + a_{111}\hat{x}_1\hat{x}_2\right)^T \end{pmatrix}^T. \]  
(31)

Repeating the 2D definitions (16)–(20), we let

\[ T_K = \left( \frac{\partial \hat{x}_j}{\partial x_i} \right)_{3 \times 3} = \frac{M_K}{J_K}, \]  
(32)

where

\[ J_K = \det S_K = \det \left( \frac{\partial x_i}{\partial x_j} \right), \]  
(33)

\[ M_K = \text{adj} \left( \frac{\partial x_i}{\partial x_j} \right)^T. \]  
(34)

To have shorter notations, we define, for different column vectors of \( S_K \) of different orders,

\[ s_1 = a_{100}, \quad s_4 = a_{110}\hat{x}_2 + a_{101}\hat{x}_3, \quad s_7 = a_{111}\hat{x}_2\hat{x}_3, \]

\[ s_2 = a_{010}, \quad s_5 = a_{110}\hat{x}_1 + a_{011}\hat{x}_3, \quad s_8 = a_{111}\hat{x}_1\hat{x}_3, \]

\[ s_3 = a_{001}, \quad s_6 = a_{101}\hat{x}_1 + a_{011}\hat{x}_2, \quad s_9 = a_{111}\hat{x}_1\hat{x}_2. \]

Please note the order differences of above nine vectors. The determinant of \( S_K \) is supposedly a degree 6 polynomial, but similar to the 2D case above, it is a degree 4 polynomial instead. To be more specific, each column vector of \( S_K \) is a sum of three vectors, and the determinant of \( S_K \) is a sum of \( 3^3 = 27 \) determinants generated by the nine vectors of \( S_K \). Many of the 27 determinants are zero, for example, all degree 5 and degree 6 polynomial terms. We group the non-zero (in general) determinants by their orders,

\[ d_{K,0} = d_{123}, \]

\[ d_{K,1} = d_{126} + d_{153} + d_{423}, \]

\[ d_{K,2} = d_{129} + d_{183} + d_{723} + d_{156} + d_{426} + d_{453}, \]

\[ d_{K,3} = d_{159} + d_{186} + d_{429} + d_{726} + d_{483} + d_{753}, \]

\[ d_{K,4} = d_{459} + d_{486} + d_{756}, \]

\[ d_{K,5} = d_{K,6} = 0, \]

where

\[ d_{ijk} = \det(s_i \ s_j \ s_k), \quad i = 1, 4, 7, \quad j = 2, 5, 8, \quad k = 3, 6, 9. \]  
(35)

We also let (similar to the 2D notation)

\[ c_i = \frac{d_{K,i}}{d_{K,0}}, \quad i = 1, 2, 3, 4. \]
Again, in the above notations, $d_i$ and $c_i$ stand for the terms of different order in the determinant $J_K$, for examples,

$$d_{K,0} = \det(a_{100} \ a_{010} \ a_{001}),$$
$$c_{K,1} = \frac{1}{d_0} \left[ \left( \begin{array}{ccc} a_{100}^T & a_{100}^T \\ a_{110}^T & a_{010}^T \\ a_{001}^T \\ \end{array} \right) \right] \left( \begin{array}{c} \hat{x}_1 \end{array} \right) + \left( \begin{array}{ccc} a_{100}^T & a_{100}^T \\ a_{110}^T & a_{010}^T \\ a_{001}^T \\ \end{array} \right) \left( \begin{array}{c} \hat{x}_2 \end{array} \right) + \left( \begin{array}{ccc} a_{100}^T & a_{100}^T \\ a_{110}^T & a_{010}^T \\ a_{001}^T \\ \end{array} \right) \left( \begin{array}{c} \hat{x}_3 \end{array} \right).$$

Next we let $M_{K,i}$ stand for the “sub”-matrix of $M_K$ formed by the $i$th order terms, i.e., we “separate” $M_K$ such that it is equal to the sum of $M_{K,i}$‘s. We note that the grouping this time is not as obvious as that in separating terms in $S_K$. Because

$$M_K = \text{adj}(S_K)^T = \text{adj}(s_1 + s_4 + s_7 \ s_2 + s_5 + s_8 \ s_3 + s_6 + s_9)^T = \left( ((s_2 + s_5 + s_8) \times (s_3 + s_6 + s_9))^T, \right.$$  
$$\left. ((s_3 + s_6 + s_9) \times (s_1 + s_4 + s_7))^T, \right.$$  
$$\left. ((s_1 + s_4 + s_7) \times (s_2 + s_5 + s_8))^T \right),$$

we let

$$M_{K,0} = m_{123},$$
$$M_{K,1} = m_{126} + m_{153} + m_{423} - m_{123},$$
$$M_{K,2} = m_{456} + m_{129} + m_{183} + m_{723} - m_{123},$$
$$M_{K,3} = m_{459} + d_{486} + m_{756} - m_{456},$$

where

$$m_{ijk} = \text{adj}(s_i \ s_j \ s_k)^T, \quad i = 1, 4, 7, \quad j = 2, 5, 8, \quad k = 3, 6, 9.$$  

We note that in both 2D and 3D, $M_{K,i}$ are the adjoint matrix of the matrix formed by all the $i$th order terms in $M_K$. But there is a fundamental difference. For example, the $M_{K,1}$ in 2D has all entries of the first order, but it has all entries of the second order in 3D. This is because each entry of such a 3D matrix is a determinant of a two-by-two submatrix of the original matrix, but entries in the corresponding 2D matrix are one-by-one determinants. Please note the one order difference in bounding $\|M_{K,i}\|$ in Corollary 3.1 and Lemma 3.2.

**Definition 3.2.** Let $K$ be a general hexahedron in 3D, $F$ the reference mapping (defined in (29)), and $\hat{J}_K = J_K^{-1}$. The Taylor truncations are defined as follows:

$$M_K^{(m)} = \begin{cases} M_{K,0} & \text{if } m = 1, \\
M_{K,0} + M_{K,1} & \text{if } m = 2, \\
M_{K,0} + M_{K,1} + M_{K,2} & \text{if } m = 3, \\
M_{K,0} + M_{K,1} + M_{K,2} + M_{K,3} & \text{if } m \geq 4, \end{cases}$$

(37)

$$J_K^{(m)} = \begin{cases} d_{K,0} & \text{if } m = 1, \\
d_{K,0} + d_{K,1} & \text{if } m = 2, \\
d_{K,0} + d_{K,1} + d_{K,2} & \text{if } m = 3, \\
d_{K,0} + d_{K,1} + d_{K,2} + d_{K,3} & \text{if } m = 4 \\
d_{K,0} + d_{K,1} + d_{K,2} + d_{K,3} + d_{K,4} & \text{if } m \geq 5, \end{cases}$$

(38)
\[ \tilde{J}_K^{(m)} = \begin{cases} 
\frac{d}{2} - 1, & \text{if } m = 1, \\
\frac{d}{2} - 1(1 - c_1), & \text{if } m = 2, \\
\frac{d}{2} - 1(1 - c_1 + (c_1^2 - c_2)), & \text{if } m = 3, \\
\frac{d}{2} - 1(1 - c_1 + c_1^2 - c_2 + 2c_1c_2 - c_1^3 + c_2), & \text{if } m = 4, \\
\sum \text{all Taylor terms } \prod_{j=1}^{4} c_j^p_j \text{ of } \tilde{J}_K & \text{where } \sum j p_j < m. 
\end{cases} \] (39)

Lemma 3.2. Let \( n = 3 \), and \( \mathcal{T}_k = \{ K_{i,k} \}_{K \in \mathcal{T}_1, i = 1, 2, ..., 2^{n(k-1)}} \) be the \( k \)th level nested refinement of an initial hexahedral grid \( \mathcal{T}_1 \). Then

\[ \frac{C_0}{2^{n(k-1)}} \leq d_{K,0} \leq \frac{C_1}{2^{n(k-1)}}, \]

\[ \frac{C_1}{2^{n(k-1)2^j(k-1)}} \leq d_{K,j} \leq \frac{C_1}{2^{n(k-1)2^j(k-1)}}, \quad j = 0, 1, 2, 3, 4, \]

\[ \frac{C_1}{2^{(n-1)(k-1)2^j(k-1)}} \leq \| M_{K,j} \| \leq \frac{C_1}{2^{(n-1)(k-1)2^j(k-1)}}, \quad j = 0, 1, 2, 3, \]

where \( C_0 \) and \( C_1 \) are positive constants depending on \( \mathcal{T}_1 \) only.

Proof. The proof is the same for 2D and for 3D. We only need to compute and estimate the variables under consideration for one subhexahedron at the nested refinement of a general hexahedron. We can compute the coefficients of the reference mapping for the subhexahedron similar to (28), and then the other variables. \( \square \)

4. Truncations of Jacobians

We will apply the nested \( Q_m \) finite element method to solving a model second order elliptic problem. We will show the convergence of the finite-element solutions obtained by replacing rational functions in Jacobians by their Taylor polynomials.

We consider the following second order self-adjoint elliptic problem:

\[ \sum_{i,j} \partial_i (a_{i,j} \partial_j u) = f \quad \text{in } \Omega, \]

\[ u = 0 \quad \text{on } \Omega, \]

where \((a_{i,j})\) is a symmetric matrix over \( \Omega \), and is uniformly positive definite on the domain. We are given an initial grid \( \mathcal{T}_1 = \{ K \} \) on \( \Omega \) and the nested (refined) grids \( \mathcal{T}_k, k = 1, 2, \ldots \). We assume only the first grid \( \mathcal{T}_1 \) is regular, i.e., in addition to (5), for any \( K \in \mathcal{T}_1 \),

\[ \begin{cases} |J_{F,0,\infty,K}| := \max \det(J_K) > C^{-1}, \\
|F^{-1}|_{1,\infty,\Omega} := \max |(T_K)_{ij}| < C, \\
|F^{-1}|_{2,\infty,\Omega} := \max |(\nabla T_K)_{ij}| < C, \end{cases} \] (40)

for some constant \( 0 < C < \infty \), where \((T_K)_{ij}\) is the \((i, j)\) entry of matrix \( T_K \). Then the Galerkin finite-element approximations are defined: find \( u_h \in V_k \) (defined in (7)), such that

\[ a(u_h, w) = (f, w) \quad \forall w \in V_k, \] (41)

where \( a(u_h, w) = \int_\Omega \sum a_{i,j} \partial_j u_h \partial_i w \) and \((\cdot, \cdot)\) is the \( L^2\)-inner product over \( \Omega \).
**Theorem 4.1** (cf. Zhang [26]). Under assumption (40) for the initial grid, the finite-element solutions of second order elliptic problem (41) converge at the optimal order:

\[ \| u - u_h \|_0 + h | u - u_h |_1 \leq Ch^r \| u \|_{r}, \quad r \leq m + 1, \]

for \( u \in H^r(\Omega) \).

Next, we replace the bilinear forms in (41) by the approximations where Jacobians are approximated by their truncated Taylor expansions defined in Definitions 3.1 and 3.2.

\[
\begin{align*}
a(u, w) & = \sum_{K \in T_k} \int_{\hat{K}} \left( \frac{\partial u}{\partial \hat{x}} \right)^T A \left( \frac{\partial u}{\partial \hat{x}} \right) \, dx = \sum_{K \in T_k} \int_{\hat{K}} \left( \frac{\partial \hat{w}}{\partial \hat{x}} \right)^T M_K^T A M_K \left( \frac{\partial \hat{u}}{\partial \hat{x}} \right) \hat{J}_K \hat{d} \hat{x}, \\
a_k(u, w) & = \sum_{K \in T_k} \int_{\hat{K}} \left( \frac{\partial \hat{w}}{\partial \hat{x}} \right)^T M^{(m)}_K A M^{(m)}_K \left( \frac{\partial \hat{u}}{\partial \hat{x}} \right) \hat{J}^{(m)}_K \, d \hat{x}, \quad (42) \\
(f, w) & = \sum_{K \in T_k} \int_{\hat{K}} f(F_K(\hat{x}))w(F_K(\hat{x}))J_K \, d \hat{x}, \\
(f, w)_k & = \sum_{K \in T_k} \int_{\hat{K}} f(F_K(\hat{x}))w(F_K(\hat{x}))J^{(m)}_K \, d \hat{x}. \quad (43)
\end{align*}
\]

To analyze the perturbations, our method is to relate the level \( k \) referencing mapping \( F_{i,k} \) to the level one \( F_{1,1} : K \to \hat{K} \) by the following lemma.

**Lemma 4.1** (cf. Zhang [26]). Under the nested refinement, the \( Q_1 \) mappings for \( K_{i,k} \) (cf. (6)) on level \( k \), \( F_{i,k} \), are uniform scalings of the restrictions of the original \( Q_1 \) mapping from \( \hat{K} \) to \( K = K_{1,1} \), \( F = F_{1,1} \), with shifts. That is, \( F^{-1} \circ F_{i,k} : \hat{K} \to \hat{K} \).

\[
\begin{align*}
F^{-1} \circ F_{i,k} : \hat{K} & \to \hat{K} , \\
F^{-1} \circ F_{i,k}(\hat{x}_1, \hat{x}_2, \hat{x}_n) & = \left( \hat{x}_{1,i,k} + 2^{-k+1}(\hat{x}_1 + 1) \right), \\
& \quad \left( \hat{x}_{2,i,k} + 2^{-k+1}(\hat{x}_2 + 1) \right), \\
& \quad \left( \hat{x}_{n,i,k} + 2^{-k+1}(\hat{x}_n + 1) \right) \quad (45)
\end{align*}
\]

where \((\hat{x}_{1,i,k}, \hat{x}_{2,i,k}, \hat{x}_{n,i,k})\) is the first vertex of the \( i \)th subsquare/cube on the \( k \)th level of \( \hat{K} \), for the space dimension \( n = 2 \) or \( 3 \).

**Lemma 4.2.** For any polynomial degree \( m \), on \( k \)th level, it holds that

\[ |(f, w) - (f, w)_k| \leq C 2^{-mk} \| f \|_0 \| w \|_0 \quad \forall f, w \in V_k. \]

**Proof.** The difference is the few dropped high-order polynomial terms in \( J_K \).

\[
\begin{align*}
| (f, w) - (f, w)_k | & = \left| \sum_{K_{i,k} \in T_k} \int_{\hat{K}} f(F_{i,k}(\hat{x}))w(F_{i,k}(\hat{x}))(J_{K_{i,k}} - J^{(m)}_{K_{i,k}}) \, d \hat{x} \right| \\
& \leq \left( \max_{K_{i,k} \in T_k} \left| 1 - \frac{J^{(m)}_{K_{i,k}}}{J_{K_{i,k}}} \right|_{L^\infty(\hat{K})} \right) \| f \|_0 \| w \|_0.
\end{align*}
\]
In 2D, if $m > 1$, then the difference above is zero, i.e., $J_{K_{i,k}}^{(m)} = J_{K_{i,k}}^{(m+1)}$. When $m = 1$ in 2D, we have, letting $K_{i,k}$ be a subquadrilateral of $K_{1,1}$ in $\mathcal{T}_1$, by Corollary 3.1,

$$
1 - \frac{J_{K_{i,k}}^{(m)}}{J_{K_{i,k}}^{(m+1)}} = \left| \frac{dK_{i,k,1} + dK_{i,k,4}}{dK_{i,k,0} + dK_{i,k,1}} \right| \leq \frac{C_1/(2^{n(k-1)2^{k-1}})}{C_0/2^{n(k-1)} - C_1/(2^{n(k-1)2^{k-1}})} = C_2^{-k},
$$

where $C > 0$ depending on $\mathcal{T}_1$ only.

The proof is the same for the 3D case. In 3D, when $m \geq 5$, the difference is zero. For $m < 5$, say $m = 3$, the difference is, by Lemma 3.2,

$$
1 - \frac{J_{K_{i,k}}^{(m)}}{J_{K_{i,k}}^{(m+1)}} = \left| \frac{dK_{i,k,3} + dK_{i,k,4}}{dK_{i,k,0} + dK_{i,k,1} + dK_{i,k,3} + dK_{i,k,4}} \right| \leq \frac{C_1/2^{3(k-1)} + C_1/2^{4(k-1)}}{C_0 - C_1/2^{k-1} - C_1/2^{4(k-1)}} = C_2^{-3k},
$$

where $C > 0$ depending on $\mathcal{T}_1$ only. □

**Lemma 4.3.** For any polynomial degree $m$, on $k$th level, it holds that

$$
|a(u, w) - a_k(u, w)| \leq C2^{-mk} \|u\| \|w\| \quad \forall u, w \in V_k.
$$

**Proof.** For any $u, w \in V_k$,

$$
a(u, w) - a_k(u, w) = \sum_{K \in \mathcal{T}_k} \int_K \left[ \left( \frac{\hat{u}}{\partial \hat{x}} \right)^T T_K^T A T_K \left( \frac{\hat{u}}{\partial \hat{x}} \right) (1 - J_{K_{i,k}}^{(m)}) \right]
+ \left( \frac{\hat{u}}{\partial \hat{x}} \right)^T T_K^T (A - (J_{K_i}^{(m)} M_{K_i}^{(m)})^T A (J_{K_i}^{(m)} M_{K_i}^{(m)})) T_K \left( \frac{\hat{u}}{\partial \hat{x}} \right) J_{K_{i,k}}^{(m)} \right] \hat{J}_K \, d\hat{x}. \tag{46}
$$

The first term is the perturbation of the Jacobian determinant.

$$
\sum_{K \in \mathcal{T}_k} \left| \int_K \left( \frac{\hat{u}}{\partial \hat{x}} \right)^T T_K^T A T_K \left( \frac{\hat{u}}{\partial \hat{x}} \right) (J_{K_{i,k}}^{(m)} - J_{K_{i,k}}^{(m+1)}) \hat{J}_K \, d\hat{x} \right| \leq \max_{K \in \mathcal{T}_k} \|1 - J_{K_{i,k}}^{(m)}\|_{L^\infty(\hat{K})} \|u\| \|v\| \|w\| \|v\|
\leq \max_{K \in \mathcal{T}_k} \|1 - J_{K_{i,k}}^{(m)}\|_{L^\infty(\hat{K})} C \|u\| \|v\| \|w\|. \tag{47}
$$

Here $\|u\| = \sqrt{a(u, u)}$ is the energy norm. We estimate the perturbation above as we did in Lemma 4.2. For example, when $n = 3$ and $m = 2$, we have, by Lemma 3.2,

$$
|1 - J_{K_{i,k}}^{(m)}| = |1 - (1 + c_1 + c_2 + c_3 + c_4)(1 - c_1 + c_2^2 - c_2)|
\leq | - c_1^3 + 2c_1c_2 - c_3 + (fourth \ order \ terms)|
\leq C 2^{-3k}.
$$

In the second term, the part of perturbation in $\hat{J}_K$ is treated as in the first term. For example, when $m = 2$ in 3D again, we have

$$
\max_{K \in \mathcal{T}_k} \left\| J_{K_{i,k}}^{(m)} \right\|_{L^\infty(\hat{K})} = \max_{K \in \mathcal{T}_k} \left\| J_{K_{i,k}}^{(m)} \right\|_{L^\infty(\hat{K})}
\leq |1 + c_1^3 - 2c_1c_2 + c_3 + (fourth \ order \ terms)| \leq C(1 + 2^{-mk}), \tag{48}
$$

where $C$ depends on $\mathcal{T}_1$ only.
The perturbation to the symmetric matrix in the second integral of (46) is estimated next. Let
\[ \| M \|_{L^\infty(\tilde{K})} \] also stand for the maximal value of \( l_\infty \)-matrix norm of matrix \( M \) on \( \tilde{K} \).

\[
\begin{align*}
\| A - (\tilde{J}_K^{(m)} M_K^{(m)} S_K)^T A (\tilde{J}_K^{(m)} M_K^{(m)} S_K) \|_{L^\infty(\tilde{K})} \\
= \| (\tilde{J}_K^{(m)} M_\delta S_K)^T A + A (\tilde{J}_K^{(m)} M_\delta S_K) + (\tilde{J}_K^{(m)} M_\delta S_K)^T A (\tilde{J}_K^{(m)} M_\delta S_K) \|_{L^\infty(\tilde{K})} \\
\leq \| A \|_{L^\infty(\tilde{K})} (2 \| \tilde{J}_K^{(m)} M_\delta S_K \|_{L^\infty(\tilde{K})} + \| \tilde{J}_K^{(m)} M_\delta S_K \|_{L^\infty(\tilde{K})}^2).
\end{align*}
\]

(49)

Here \( M_\delta \) is the matrix made by the dropped higher order terms of \( M_K \). For example, when \( m = 2 \) and \( n = 2 \), \( M_\delta = 0 \). But when \( m = 1 \) and \( n = 2 \), cf. (19) and (21),

\[
M_\delta = M_{K, 1} = \begin{pmatrix}
a_{11, K}^{(2)} \hat{x}_1 & -a_{11, K}^{(1)} \hat{x}_1 \\
-a_{11, K}^{(2)} \hat{x}_2 & a_{11, K}^{(1)} \hat{x}_2
\end{pmatrix} = \frac{1}{4^{k-1}} \begin{pmatrix}
a_{11, K_1}^{(2)} \hat{x}_1 & -a_{11, K_1}^{(1)} \hat{x}_1 \\
-a_{11, K_1}^{(2)} \hat{x}_2 & a_{11, K_1}^{(1)} \hat{x}_2
\end{pmatrix} = \frac{1}{2^{(n-1)(k-1)2m(k-1)}} M_{K, 1, 1},
\]

(50)

where we suppose \( K \) is a \( k \)th level refinement of a quadrilateral \( K_1 \) of \( \mathcal{T}_1 \). Let us check another case, when \( m = 2 \) and \( n = 3 \). In this case, by Lemma 3.2,

\[
\| M_\delta \| = \| M_{K, 2} + M_{K, 3} \| \leq \frac{C_1}{2(n-1)(k-1)2m(k-1)}.
\]

(51)

Next, for the matrix \( S_K \) in (49), if \( n = 2 \), the entries of \( S_K \) are rearranged from those of \( M_K \) with two sign changes. In both 2D and 3D, by (13) and (31), it is easy to check that, as we did in Lemmas 3.1 and 3.2,

\[
\| S_K \| \leq \frac{C_1}{2^{k-1}}.
\]

(52)

Finally, for the Jacobian determinant \( \tilde{J}_K \) in (49), either by checking area (2D) or volume (3D), or by comparing it with \( d_{K, i} \), we get

\[
\| \tilde{J}_K \|_{L^\infty(\tilde{K})} \leq C_1 2^{nk},
\]

(53)

where \( C_1 \) depends only on level one element, from which \( K \) is refined. Combining (48)–(53), we get the needed estimate for the second term in (46),

\[
\begin{align*}
\sum_{K \in \mathcal{T}_k} \int_{\tilde{K}} \left( \frac{\partial \tilde{u}}{\partial \tilde{x}} \right)^T T_K (A - (\tilde{J}_K^{(m)} M_K^{(m)} S_K)^T A (\tilde{J}_K^{(m)} M_K^{(m)} S_K)) T_K \left( \frac{\partial \tilde{u}}{\partial \tilde{x}} \right) \tilde{J}_K^{(m)} d\tilde{x} \\
\leq C \left( 1 + \frac{1}{2mk} \right) (C_1 2^{nk}) \frac{C_1}{2^{(n-1)(k-1)2m(k-1)}} \frac{C_1}{2^{k-1}} \| u \|_1 \| u \|_1 \\
= C 2^{-nk} \| u \|_1 \| u \|_1.
\end{align*}
\]

(54)

As we have shown the bound for both terms in (46), the lemma is proven.

\[ \square \]

**Corollary 4.1.** For grid level \( k \) high enough, \( k > k_0 \), \( a_k(:, :) \) is \( V \)-elliptic, i.e.,

\[
a_k(u, u) \geq \gamma a(u, u) \quad \forall u \in V_k,
\]

(55)

where \( \gamma > 0 \) is independent of \( k \), but \( k_0 \), \( k_0 \) depends on the initial grid \( \mathcal{T}_1 \) only.
Proof. Similar to the proof of Lemma 4.3, we get
\[
\begin{align*}
    a_k(u, u) &\geq C \frac{1}{1 + 2^{-m_k}} \sum_{K \in \mathcal{T}_h} C \| I - \hat{J}_K M_0 S_K \|_{L^\infty} \int_K \left( \frac{\partial \hat{u}}{\partial \hat{x}} \right)^T T_K^T A K \left( \frac{\partial \hat{u}}{\partial \hat{x}} \right) J_K \, d\hat{x} \\
    &\geq C \frac{1 - C 2^{-m_k}}{1 + 2^{-m_k}} a(u, u) = \gamma a(u, u).
\end{align*}
\]

Corollary 4.1 ensures the existence and uniqueness of finite-element solutions for the perturbed problems: find \( u_k \in V_k \) such that
\[
    a_k(u_k, w) = (f, w)_k \quad \forall w \in V_k. \tag{56}
\]
It is standard now to show that the convergence order for \( u_k \) would be optimal, i.e., the same as that for the regular finite-element solutions \( u_h \) defined in (41).

**Theorem 4.2.** The finite-element solutions of (56), i.e., the bilinear forms of (41) are approximated by truncated Taylor expansions defined in Definitions 3.1 and 3.2, converge at the optimal order:
\[
    \| u - u_k \|_{V} \leq C h_k^r \| u \|_{r},
\]
if \( u \in H^r(\omega) \) for \( r \leq m \).

**Proof.** The proof is standard by the first Strang lemma (cf. Theorem 4.1.1 in [7]):
\[
    \| u - u_k \|_{V} \leq C \inf_{z_k \in V_k} \left\{ \| u - z_k \|_{V} + \sup_{w_k \in V_k} \frac{|a(z_k, w_k) - a_k(z_k, w_k)|}{\| w_k \|_{V}} \right\} + C \sup_{w_k \in V_k} \frac{|(f, w_k) - (f, w_k)_k|}{\| w_k \|_{V}}.
\]
By Lemmas 4.2 and 4.3, the theorem is proved. \( \square \)

5. Numerical tests

In this section, we will show two numerical tests, a 2D one and a 3D one. In both cases, we solve a Poisson equation with homogeneous Dirichlet boundary condition:
\[
    -\Delta u = f \quad \text{in } \Omega, \\
    u = 0 \quad \text{on } \partial \Omega. \tag{57}
\]
Here we select a domain first, then pick up an exact solution \( u \) to generate the right-hand side function \( f \).

The 2D domain is a quadrilateral (shown in Figs. 1 and in 4) with vertices
\[
    (0, 0), \quad (0.8, 0), \quad (0.96, 0.6), \quad (0, 0.8).
\]
The exact solution in 2D is
\[
    u(x, y) = -100x \left( y - 0.6 \frac{x - 0.8}{0.96 - 0.8} \right) y \left( x - 0.96 \frac{y - 0.8}{0.6 - 0.8} \right)
\]
which would define the right-hand side function \( f(x, y) \). The solution is plotted in Fig. 4.

The 3D domain of our numerical test is a generalized hexahedron with non-flat top surface (shown in Fig. 1). The eight vertices of the domain is
\[
    (0, 0, 0), \quad (1, 0, 0), \quad (1, 1, 0), \quad (0, 1, 0), \\
    (0, 0, 1.25), \quad (1, 0, 0.75), \quad (1, 1, 1), \quad (0, 1, 1).
\]
The exact solution in the 3D numerical test is

\[ u(x, y, z) = 2^{1/2}x(1 - x)y(1 - y)z(1.25 - z - 0.5x - 0.25y + 0.5xy) \]

which would again define the right-hand side function \( f(x, y) \).

We use different orders of Taylor truncations and list the maximal nodal errors (most cases) for our numerical tests. We also list the order \( r \) of convergence,

\[ \|e_k\|_\infty = O(h^r_k) \quad \text{where } e_k = u - u_k, \]

for each degree \( m \) of finite elements. In the last two columns in each table, we list the data for applying Gauss quadratures directly to rational functions, as it was commonly done by engineers.

In Table 1, we list the data for bilinear elements, where we used various order of Taylor truncations. The convergence orders verify the theory, i.e., it is enough using \( M_{K,0} \) and \( d_{K,0} \) (see Definition 3.1) to achieve the optimal convergence rate for the bilinear elements. In fact, the theory and numerical tests for this case, \( m = 1 \), were already shown in [3,26].

In Table 2, we list the nodal errors and convergence rates for biquadratic finite elements. We first remark that even the meshes are non-uniform, but it seems we still have a superconvergence rate for the biquadratic finite element, i.e., the rates for last three cases are of order 4 instead of order 3. The meshes in this case do become more regular as the higher level quadrilaterals tend to parallelograms. When we intentionally use one lower order Taylor truncations here, we lose an order of convergence. This can be seen in the third and the fifth columns of Table 2. In fact, comparing the results in Table 1, the biquadratic elements would perform even worse in these two cases. However, our theory says that when using the second order truncation, we would get the second order of convergence in the \( H^1 \)-norm. In general, the nodal error would be of one order higher, the third order in this case, than that for the derivatives. But we only achieve
order 2 convergence rate here. Though there is no contradiction to our theory, we could not explain this phenomenon yet. However, when we check the $H^1$-rate, see Table 6 below, we do get the right order of convergence, as predicted by our theory.

Table 3 is similar to Table 2, where we used bicubic finite elements. The data verify closely the theory.

The numerical data are listed in Table 4 for trilinear elements. The convergence rates are as expected.

The nodal errors and the convergence rates of nodal errors of the triquadratic elements are listed in Table 5. Similar to the case of biquadratic elements, all rates are as expected except the rate for the second order Taylor truncations $M_K^{(2)}$ and $J_K^{(2)}$, i.e., we expect to see 3 in the fifth column of Table 5 instead of 2. To make sure that Theorem 4.2 matches numerical test results, we list again the errors and the convergence orders for the triquadratic element in semi-$H^1$ norm in Table 6. The first two columns for $r$ do verify exactly Theorem 4.2. However, the next two rates are one order too high, higher than we expected. There might be a superconvergence here too. Or these might be caused by the numerical quadrature rules in use for evaluating the semi-$H^1$ norms of the differences between the exact solution and the numerical solutions of piecewise rational functions.

In Table 7, we listed the errors and the convergence order in semi-$H^1$ norm, which is the energy norm in this case, for the 3D cubic elements. The order of convergence in each case matches perfectly the theory provided in Theorem 4.2, i.e., we need the order 3 Taylor truncations for the Jacobian matrix and Jacobian.
The difference is that in the first case, the Taylor expansions are done at the barycentric-center of each element, while in the second case, they are done at one corner point of each element. To show this phenomenon, we list in Table 8 the nodal errors for triquadratic elements.

### Table 5
Nodal errors and rates for triquadratic \((n = 3 \text{ and } m = 2)\) elements

<table>
<thead>
<tr>
<th>(k)</th>
<th>(M^{(1)}_K) and (J^{(1)}_K)</th>
<th>(M^{(2)}_K) and (J^{(2)}_K)</th>
<th>(M^{(3)}_K) and (J^{(3)}_K)</th>
<th>(M^{(4)}_K) and (J^{(4)}_K)</th>
<th>(3 \times 3 \times 3) points</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(</td>
<td>e_k</td>
<td>_{\infty})</td>
<td>(r)</td>
<td>(</td>
</tr>
<tr>
<td>2</td>
<td>2.026</td>
<td>–</td>
<td>0.621</td>
<td>0.96</td>
<td>0.30523</td>
</tr>
<tr>
<td>3</td>
<td>0.977</td>
<td>1.05</td>
<td>0.075</td>
<td>3.04</td>
<td>0.02859</td>
</tr>
<tr>
<td>4</td>
<td>0.334</td>
<td>1.55</td>
<td>0.016</td>
<td>2.18</td>
<td>0.00226</td>
</tr>
<tr>
<td>5</td>
<td>0.092</td>
<td>1.85</td>
<td>0.004</td>
<td>1.89</td>
<td>0.00015</td>
</tr>
<tr>
<td>6</td>
<td>0.024</td>
<td>1.93</td>
<td>0.001</td>
<td>1.99</td>
<td>0.00001</td>
</tr>
</tbody>
</table>

### Table 6
Semi-\(H^1\) errors and rates for triquadratic \((n = 3 \text{ and } m = 2)\) elements

<table>
<thead>
<tr>
<th>(k)</th>
<th>(M^{(1)}_K) and (J^{(1)}_K)</th>
<th>(M^{(2)}_K) and (J^{(2)}_K)</th>
<th>(M^{(3)}_K) and (J^{(3)}_K)</th>
<th>(M^{(4)}_K) and (J^{(4)}_K)</th>
<th>(3 \times 3 \times 3) points</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(</td>
<td>e_k</td>
<td>_{H^1})</td>
<td>(r)</td>
<td>(</td>
</tr>
<tr>
<td>2</td>
<td>3.44071</td>
<td>–</td>
<td>1.12062</td>
<td>0.95</td>
<td>0.81538</td>
</tr>
<tr>
<td>3</td>
<td>2.25060</td>
<td>0.61</td>
<td>0.22459</td>
<td>2.32</td>
<td>0.12545</td>
</tr>
<tr>
<td>4</td>
<td>1.21860</td>
<td>0.89</td>
<td>0.04735</td>
<td>2.25</td>
<td>0.01649</td>
</tr>
<tr>
<td>5</td>
<td>0.62214</td>
<td>0.97</td>
<td>0.01113</td>
<td>2.09</td>
<td>0.00209</td>
</tr>
<tr>
<td>7</td>
<td>0.31272</td>
<td>0.99</td>
<td>0.00273</td>
<td>2.02</td>
<td>0.00026</td>
</tr>
</tbody>
</table>

### Table 7
Semi-\(H^1\) errors and rates for tricubic \((n = 3 \text{ and } m = 3)\) elements

<table>
<thead>
<tr>
<th>(k)</th>
<th>(M^{(1)}_K) and (J^{(1)}_K)</th>
<th>(M^{(2)}_K) and (J^{(2)}_K)</th>
<th>(M^{(3)}_K) and (J^{(3)}_K)</th>
<th>(M^{(4)}_K) and (J^{(4)}_K)</th>
<th>(3 \times 3 \times 3) points</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(</td>
<td>e_k</td>
<td>_{H^1})</td>
<td>(r)</td>
<td>(</td>
</tr>
<tr>
<td>2</td>
<td>4.40992</td>
<td>0.49</td>
<td>0.927979</td>
<td>1.29</td>
<td>0.0855865</td>
</tr>
<tr>
<td>3</td>
<td>2.36533</td>
<td>0.90</td>
<td>0.270747</td>
<td>1.78</td>
<td>0.0140203</td>
</tr>
<tr>
<td>4</td>
<td>1.21954</td>
<td>0.96</td>
<td>0.072885</td>
<td>1.89</td>
<td>0.0019192</td>
</tr>
<tr>
<td>5</td>
<td>0.61841</td>
<td>0.98</td>
<td>0.018940</td>
<td>1.94</td>
<td>0.0002471</td>
</tr>
<tr>
<td>6</td>
<td>0.31127</td>
<td>0.99</td>
<td>0.004829</td>
<td>1.97</td>
<td>0.0000312</td>
</tr>
</tbody>
</table>

### Table 8
Nodal errors and rates for bilinear elements with reference element \([0, 1]^2\)

<table>
<thead>
<tr>
<th>(k)</th>
<th>(M^{(1)}_K) and (J^{(1)}_K)</th>
<th>(M^{(2)}_K) and (J^{(2)}_K)</th>
<th>(M^{(3)}_K) and (J^{(3)}_K)</th>
<th>(M^{(4)}_K) and (J^{(4)}_K)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(</td>
<td>e_k</td>
<td>_{\infty})</td>
<td>(r)</td>
</tr>
<tr>
<td>2</td>
<td>1.679575</td>
<td>–</td>
<td>1.393823</td>
<td>–</td>
</tr>
<tr>
<td>3</td>
<td>0.792917</td>
<td>1.08</td>
<td>0.647108</td>
<td>1.11</td>
</tr>
<tr>
<td>4</td>
<td>0.257819</td>
<td>1.62</td>
<td>0.150701</td>
<td>2.10</td>
</tr>
<tr>
<td>5</td>
<td>0.098093</td>
<td>1.39</td>
<td>0.037882</td>
<td>1.99</td>
</tr>
<tr>
<td>6</td>
<td>0.042202</td>
<td>1.22</td>
<td>0.009436</td>
<td>2.01</td>
</tr>
<tr>
<td>7</td>
<td>0.019559</td>
<td>1.11</td>
<td>0.002362</td>
<td>2.00</td>
</tr>
</tbody>
</table>

Finally, we would like to remark that we must use the reference elements \([-1, 1]^n\), but not \([0, 1]^n\). When we use the latter reference elements, a straightforward Taylor expansion would lose one order, compared to that from the first case. The difference is that in the first case, the Taylor expansions are done at the barycentric-center of each element, while at one corner point of each element for the second case. To show this phenomenon, we list in Table 8 the nodal errors...
and the rates for bilinear elements when using the reference element \([0, 1]^2\). We can see the first rate of convergence this time is one order lower than that listed in Table 1.

**References**


