SUPERCONVERGENCE OF HIGH ORDER FINITE DIFFERENCE SCHEMES BASED ON VARIATIONAL FORMULATION FOR ELLIPTIC EQUATIONS *

HAO LI† AND XIANGXIONG ZHANG †

Abstract. The classical continuous finite element method with Lagrangian Q^k basis reduces to a finite difference scheme when all the integrals are replaced by the $(k+1)\times(k+1)$ Gauss-Lobatto quadrature. We prove that this finite difference scheme is $(k+2)$-th order accurate in the discrete 2-norm for an elliptic equation with Dirichlet boundary conditions, which is a superconvergence result of function values. We also give a convenient implementation for the case $k = 2$, which is a simple fourth order accurate elliptic solver on a rectangular domain.

Key words. Superconvergence, high order accurate discrete Laplacian, elliptic equations, finite difference scheme based on variational formulation, Gauss-Lobatto quadrature.

AMS subject classifications. 65N30, 65N15, 65N06

1. Introduction.

1.1. Motivation. In this paper we consider solving a two-dimensional elliptic equation with smooth coefficients on a rectangular domain by high order finite difference schemes, which are constructed via using suitable quadrature in the classical continuous finite element method on a rectangular mesh. Consider the following model problem as an example: a variable coefficient Poisson equation $-\nabla\cdot(a(\mathbf{x})\nabla u) = f, a(\mathbf{x}) > 0$ on a square domain $\Omega = (0,1)\times(0,1)$ with homogeneous Dirichlet boundary conditions. The variational form is to find $u \in H_0^1(\Omega) = \{v \in H^1(\Omega) : v|_{\partial\Omega} = 0\}$ satisfying

$$A(u, v) = (f, v), \quad \forall v \in H_0^1(\Omega),$$

where $A(u, v) = \iint_\Omega a\nabla u \cdot \nabla v dxdy$, $(f, v) = \iint_\Omega f v dxdy$. Let h be the mesh size of an uniform rectangular mesh and $V_0^h \subseteq H_0^1(\Omega)$ be the continuous finite element space consisting of piecewise Q^k polynomials (i.e., tensor product of piecewise polynomials of degree k), then the C^0-Q^k finite element solution is defined as $u_h \in V_0^h$ satisfying

$$A(u_h, v_h) = (f, v_h), \quad \forall v_h \in V_0^h. \tag{1.1}$$

Standard error estimates of (1.1) are $\|u - u_h\|_1 \le Ch^k\|u\|_{k+1}$ and $\|u - u_h\|_0 \le Ch^{k+1}\|u\|_{k+1}$ where $\|\cdot\|_k$ denotes $H^k(\Omega)$-norm, see [5]. For $k \ge 2$, $\mathcal{O}(h^{k+1})$ superconvergence for the gradient at Gauss quadrature points and $\mathcal{O}(h^{k+2})$ superconvergence for functions values at Gauss-Lobatto quadrature points were proven for one-dimensional case in [11, 2, 1] and for two-dimensional case in [8, 17, 4, 14].

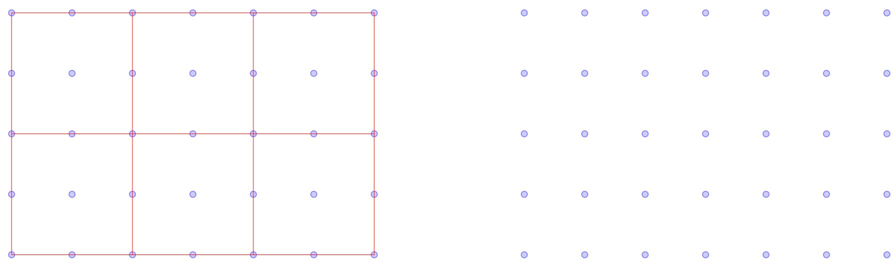
When implementing the scheme (1.1), integrals are usually approximated by quadrature. The most convenient implementation is to use $(k + 1) \times (k + 1)$ Gauss-Lobatto quadrature because they not only are superconvergence points but also can define all the degree of freedoms of Lagrangian Q^k basis. See Figure 1 for the case $k = 2$. Such a quadrature scheme can be denoted as finding $u_h \in V_0^h$ satisfying

$$A_h(u_h, v_h) = \langle f, v_h\rangle_h, \quad \forall v_h \in V_0^h, \tag{1.2}$$

*Research is supported by the NSF grants DMS-1522593 and DMS-1913120.

†Department of Mathematics, Purdue University, 150 N. University Street, West Lafayette, IN 47907-2067 (li2497@purdue.edu, zhan1966@purdue.edu).

where $A_h(u_h, v_h)$ and $\langle f, v_h \rangle_h$ denote using tensor product of $(k+1)$-point Gauss-Lobatto quadrature for integrals $A(u_h, v_h)$ and (f, v_h) respectively.



(a) The quadrature points and a FEM mesh

(b) The corresponding finite difference grid

FIG. 1. *An illustration of Lagrangian Q^2 element and the 3×3 Gauss-Lobatto quadrature.*

It is well known that many classical finite difference schemes are exactly finite element methods with specific quadrature scheme, see [5]. We will write scheme (1.2) as an exact finite difference type scheme in Section 7 for $k = 2$. Such a finite difference scheme not only provides an efficient and also convenient way for assembling the stiffness matrix especially for a variable coefficient problem, but also with has advantages inherited from the variational formulation, such as symmetry of stiffness matrix and easiness of handling boundary conditions in high order schemes. This is the variational approach to construct a high order accurate finite difference scheme .

Classical quadrature error estimates imply that standard finite element error estimates still hold for (1.2), see [7, 5]. The focus of this paper is to prove that the superconvergence of function values at Gauss-Lobatto points still holds. To be more specific, for Dirichlet type boundary conditions, we will show that (1.2) with $k \geq 2$ is a $(k+2)$-th order accurate finite difference scheme in the discrete 2-norm under suitable smoothness assumptions on the exact solution and the coefficients.

In this paper, the main motivation to study superconvergence is to use it for constructing $(k+2)$-th order accurate finite difference schemes. For such a task, superconvergence points should define all degree of freedoms over the whole computational domain including boundary points. For high order finite element methods, this seems possible only on quite structured meshes such as rectangular meshes for a rectangular domain and equilateral triangles for a hexagonal domain, even though there are numerous superconvergence results for interior cells in unstructured meshes.

1.2. Related work and difficulty in using standard tools. To illustrate our perspectives and difficulties, we focus on the case $k = 2$ in the following. For computing the bilinear form in the scheme (1.1), another convenient implementation is to replace the smooth coefficient $a(x, y)$ by a piecewise Q^2 polynomial $a_I(x, y)$ obtained by interpolating $a(x, y)$ at the quadrature points in each cell shown in Figure 1. Then one can compute the integrals in the bilinear form exactly since the integrand is a polynomial. Superconvergence of function values for such an approximated coefficient scheme was proven in [13] and the proof can be easily extended to higher order polynomials and three-dimensional cases. This result might seem surprising since interpolation error $a(x, y) - a_I(x, y)$ is of third order. On the other hand, all the tools used in [13] are standard in the literature.

From a practical point of view, (1.2) is more interesting since it gives a genuine finite difference scheme. It is straightforward to use standard tools in the literature for showing superconvergence still holds for accurate enough quadrature. Even though the 3×3 Gauss-Lobatto quadrature is fourth order accurate, the standard quadrature error estimates cannot be used directly to establish the fourth order accuracy of (1.2), as will be explained in detail in Remark 3.8 in Section 3.2.

We can also rewrite (1.2) for $k = 2$ as a finite difference scheme but its local truncation error is only second order as will be shown in Section 7.4. The phenomenon that truncation errors have lower orders was named *supraconvergence* in the literature. The second order truncation error makes it difficult to establish the fourth order accuracy following any traditional finite difference analysis approaches.

To construct high order finite difference schemes from variational formulation, we can also consider finite element method with P^2 basis on a regular triangular mesh in which two adjacent triangles form a rectangle [18]. Superconvergence of function values in C^0-P^2 finite element method at the three vertices and three edge centers can be proven [4, 17]. See also [10]. Even though the quadrature using only three edge centers is third order accurate, error cancellations happen on two adjacent triangles forming a rectangle, thus fourth order accuracy of the corresponding finite difference scheme is still possible. However, extensions to construct higher order finite difference schemes are much more difficult.

1.3. Contributions and organization of the paper. The main contribution is to give the proof of the $(k+2)$-th order accuracy of (1.2) with $k \geq 2$, which is an easy construction of high order finite difference schemes for variable coefficient problems. An important step is to obtain desired sharp quadrature estimate for the bilinear form, for which it is necessary to count in quadrature error cancellations between neighboring cells. Conventional quadrature estimating tools such as the Bramble-Hilbert Lemma only give the sharp estimate on each cell thus cannot be used directly. A key technique in this paper is to apply the Bramble-Hilbert Lemma after integration by parts on proper interpolation polynomials to allow error cancellations.

The paper is organized as follows. In Section 2, we introduce our notations and assumptions. In Section 3, standard quadrature estimates are reviewed. Superconvergence of bilinear forms with quadrature is shown in Section 4. Then we prove the main result for homogeneous Dirichlet boundary conditions in Section 5 and for nonhomogeneous Dirichlet boundary conditions in Section 6. Section 7 provides a simple finite difference implementation of (1.2). Section 8 contains numerical tests. Concluding remarks are given in Section 9.

2. Notations and assumptions.

2.1. Notations and basic tools. We will use the same notations as in [13]:

- We only consider a rectangular domain $\Omega = (0,1) \times (0,1)$ with its boundary denoted as $\partial\Omega$.
- Only for convenience, we assume Ω_h is an uniform rectangular mesh for $\bar{\Omega}$ and $e = [x_e - h, x_e + h] \times [y_e - h, y_e + h]$ denotes any cell in Ω_h with cell center (x_e, y_e). The assumption of an uniform mesh is not essential to the discussion of superconvergence. All superconvergence results in this paper can be easily extended to continuous finite element method with Q^k element on a quasi-uniform rectangular mesh, but not on a generic quadrilateral mesh or any curved mesh.
- $Q^k(e) = \left\{ p(x,y) = \sum\limits_{i=0}^{k} \sum\limits_{j=0}^{k} p_{ij} x^i y^j, (x,y) \in e \right\}$ is the set of tensor product of

polynomials of degree k on a cell e.

- $V^h = \{p(x,y) \in C^0(\Omega_h) : p|_e \in Q^k(e), \quad \forall e \in \Omega_h\}$ denotes the continuous piecewise Q^k finite element space on Ω_h.
- $V_0^h = \{v_h \in V^h : v_h = 0 \quad \text{on} \quad \partial\Omega\}$.
- The norm and seminorms for $W^{k,p}(\Omega)$ and $1 \le p < +\infty$, with standard modification for $p = +\infty$:

$$\|u\|_{k,p,\Omega} = \left(\sum_{i+j\le k} \iint_\Omega |\partial_x^i \partial_y^j u(x,y)|^p dxdy \right)^{1/p},$$

$$|u|_{k,p,\Omega} = \left(\sum_{i+j=k} \iint_\Omega |\partial_x^i \partial_y^j u(x,y)|^p dxdy \right)^{1/p},$$

$$[u]_{k,p,\Omega} = \left(\iint_\Omega |\partial_x^k u(x,y)|^p dxdy + \iint_\Omega |\partial_y^k u(x,y)|^p dxdy \right)^{1/p}.$$

Notice that $[u]_{k+1,p,\Omega} = 0$ if u is a Q^k polynomial.

- For simplicity, sometimes we may use $\|u\|_{k,\Omega}$, $|u|_{k,\Omega}$ and $[u]_{k,\Omega}$ denote norm and seminorms for $H^k(\Omega) = W^{k,2}(\Omega)$.
- When there is no confusion, Ω may be dropped in the norm and seminorms, e.g., $\|u\|_k = \|u\|_{k,2,\Omega}$.
- **For any $v_h \in V^h$, $1 \le p < +\infty$ and $k \ge 1$, we will abuse the notation to denote the broken Sobolev norm and seminorms by the following symbols**

$$\|v_h\|_{k,p,\Omega} := \left(\sum_e \|v_h\|_{k,p,e}^p \right)^{\frac{1}{p}}, \quad |v_h|_{k,p,\Omega} := \left(\sum_e |v_h|_{k,p,e}^p \right)^{\frac{1}{p}}, \quad [v_h]_{k,p,\Omega} := \left(\sum_e [v_h]_{k,p,e}^p \right)^{\frac{1}{p}}.$$

- Let $Z_{0,e}$ denote the set of $(k+1) \times (k+1)$ Gauss-Lobatto points on a cell e.
- $Z_0 = \bigcup_e Z_{0,e}$ denotes all Gauss-Lobatto points in the mesh Ω_h.
- Let $\|u\|_{2,Z_0}$ and $\|u\|_{\infty,Z_0}$ denote the discrete 2-norm and the maximum norm over Z_0 respectively:

$$\|u\|_{2,Z_0} = \left[h^2 \sum_{(x,y)\in Z_0} |u(x,y)|^2 \right]^{\frac{1}{2}}, \quad \|u\|_{\infty,Z_0} = \max_{(x,y)\in Z_0} |u(x,y)|.$$

- For a continuous function $f(x,y)$, let $f_I(x,y)$ denote its piecewise Q^k Lagrange interpolant at $Z_{0,e}$ on each cell e, i.e., $f_I \in V^h$ satisfies:

$$f(x,y) = f_I(x,y), \quad \forall (x,y) \in Z_0.$$

- $P^k(t)$ denotes the set of polynomial of degree k of variable t.
- $(f,v)_e$ denotes the inner product in $L^2(e)$ and (f,v) denotes the inner product in $L^2(\Omega)$:

$$(f,v)_e = \iint_e fv\,dxdy, \quad (f,v) = \iint_\Omega fv\,dxdy = \sum_e (f,v)_e.$$

- $\langle f, v\rangle_{e,h}$ denotes the approximation to $(f, v)_e$ by using $(k+1)\times(k+1)$-point Gauss Lobatto quadrature with $k \geq 2$ for integration over cell e.
- $\langle f, v\rangle_h$ denotes the approximation to (f, v) by using $(k+1)\times(k+1)$-point Gauss Lobatto quadrature with $k \geq 2$ for integration over each cell e.
- $\hat{K} = [-1, 1] \times [-1, 1]$ denotes a reference cell.
- For $f(x, y)$ defined on e, consider $\hat{f}(s, t) = f(sh + x_e, th + y_e)$ defined on $\hat{K}$. Let $\hat{f}_I$ denote the Q^k Lagrange interpolation of $\hat{f}$ at the $(k+1)\times(k+1)$ Gauss Lobatto quadrature points on $\hat{K}$.
- $(\hat{f}, \hat{v})_{\hat{K}} = \iint_{\hat{K}} \hat{f}\hat{v}\, dsdt$.
- $\langle \hat{f}, \hat{v}\rangle_{\hat{K}}$ denotes the approximation to $(\hat{f}, \hat{v})_{\hat{K}}$ by using $(k+1)\times(k+1)$-point Gauss-Lobatto quadrature.
- On the reference cell $\hat{K}$, for convenience we use the superscript h over the ds or dt to denote we use $(k+1)$-point Gauss-Lobatto quadrature on the corresponding variable. For example,

$$\iint_{\hat{K}} \hat{f} d^h s dt = \int_{-1}^{1} [w_1 \hat{f}(-1, t) + w_{k+1}\hat{f}(1, t) + \sum_{i=2}^{k} w_i \hat{f}(x_i, t)]dt.$$

  Since $(\hat{f}\hat{v})_I$ coincides with $\hat{f}\hat{v}$ at the quadrature points, we have

$$\iint_{\hat{K}} (\hat{f}\hat{v})_I dxdy = \iint_{\hat{K}} (\hat{f}\hat{v})_I d^h x d^h y = \iint_{\hat{K}} \hat{f}\hat{v} d^h x d^h y = \langle \hat{f}, \hat{v}\rangle_{\hat{K}}.$$

The following are commonly used tools and facts:

- For two-dimensional problems,

$$h^{k-2/p}|v|_{k,p,e} = |\hat{v}|_{k,p,\hat{K}}, \quad h^{k-2/p}[v]_{k,p,e} = [\hat{v}]_{k,p,\hat{K}}, \quad 1 \leq p \leq \infty.$$

- Inverse estimates for polynomials:

$$\|v_h\|_{k+1,e} \leq Ch^{-1}\|v_h\|_{k,e}, \quad \forall v_h \in V^h, k \geq 0. \tag{2.1}$$

- Sobolev's embedding in two and three dimensions: $H^2(\hat{K}) \hookrightarrow C^0(\hat{K})$.
- The embedding implies

$$\|\hat{f}\|_{0,\infty,\hat{K}} \leq C\|\hat{f}\|_{k,2,\hat{K}}, \quad \forall \hat{f} \in H^k(\hat{K}), k \geq 2,$$

$$\|\hat{f}\|_{1,\infty,\hat{K}} \leq C\|\hat{f}\|_{k+1,2,\hat{K}}, \quad \forall \hat{f} \in H^{k+1}(\hat{K}), k \geq 2.$$

- Cauchy-Schwarz inequalities in two dimensions:

$$\sum_e \|u\|_{k,e}\|v\|_{k,e} \leq \left(\sum_e \|u\|_{k,e}^2\right)^{\frac{1}{2}} \left(\sum_e \|v\|_{k,e}^2\right)^{\frac{1}{2}}, \quad \|u\|_{k,1,e} = \mathcal{O}(h)\|u\|_{k,2,e}.$$

- Poincaré inequality: let $\bar{u}$ be the average of $u \in H^1(\Omega)$ on Ω, then

$$|u - \bar{u}|_{0,p,\Omega} \leq C|\nabla u|_{0,p,\Omega}, \quad p \geq 1.$$

  If $\bar{u}$ is the average of $u \in H^1(e)$ on a cell e, we have

$$|u - \bar{u}|_{0,p,e} \leq Ch|\nabla u|_{0,p,e}, \quad p \geq 1.$$

- For $k \geq 2$, the $(k+1)\times(k+1)$ Gauss-Lobatto quadrature is exact for integration of polynomials of degree $2k-1 \geq k+1$ on $\hat{K}$.
- Define the projection operator $\hat{\Pi}_1 : \hat{u} \in L^1(\hat{K}) \to \hat{\Pi}_1 \hat{u} \in Q^1(\hat{K})$ by

$$\iint_{\hat{K}} (\hat{\Pi}_1 \hat{u}) w ds dt = \iint_{\hat{K}} \hat{u} w ds dt, \forall w \in Q^1(\hat{K}). \tag{2.2}$$

  Notice that all degree of freedoms of $\hat{\Pi}_1 \hat{u}$ can be represented as a linear combination of $\iint_{\hat{K}} \hat{u}(s,t)p(s,t)dsdt$ for $p(s,t) = 1, s, t, st$, thus the $H^1(\hat{K})$ (or $H^2(\hat{K})$) norm of $\hat{\Pi}_1 \hat{u}$ are determined by $\iint_{\hat{K}} \hat{u}(s,t)p(s,t)dsdt$. By Cauchy-Schwarz inequality $|\iint_{\hat{K}} \hat{u}(s,t)\hat{p}(s,t)dsdt| \leq \|\hat{u}\|_{0,2,\hat{K}} \|\hat{p}\|_{0,2,\hat{K}} \leq C\|\hat{u}\|_{0,2,\hat{K}}$, we have $\|\Pi_1 \hat{u}\|_{1,2,\hat{K}} \leq C\|\hat{u}\|_{0,2,\hat{K}}$, which means $\hat{\Pi}_1$ is a continuous linear mapping from $L^2(\hat{K})$ to $H^1(\hat{K})$. By a similar argument, one can show $\hat{\Pi}_1$ is a continuous linear mapping from $L^2(\hat{K})$ to $H^2(\hat{K})$.

2.2. Coercivity and elliptic regularity. We consider the elliptic variational problem of finding $u \in H_0^1(\Omega)$ to satisfy

$$A(u,v) := \iint_\Omega (\nabla v^T \mathbf{a} \nabla u + \mathbf{b}\nabla u v + cuv)\, dxdy = (f,v), \forall v \in H_0^1(\Omega), \tag{2.3}$$

where $\mathbf{a} = \begin{pmatrix} a^{11} & a^{12} \\ a^{21} & a^{22} \end{pmatrix}$ is real symmetric positive definite and $\mathbf{b} = [b^1 \quad b^2]$. Assume the coefficients $\mathbf{a}$, $\mathbf{b}$ and c are smooth with uniform upper bounds, thus $A(u,v) \leq C\|u\|_1 \|v\|_1$ for any $u, v \in H_0^1(\Omega)$. We denote $\lambda_{\mathbf{a}}$ as the smallest eigenvalues of $\mathbf{a}$. Assume $\lambda_{\mathbf{a}}$ has a positive lower bound and $\nabla \cdot \mathbf{b} \leq 2c$, so that coercivity of the bilinear form can be easily achieved. Since

$$(\mathbf{b}\cdot\nabla u, v) = \int_{\partial\Omega} uv\mathbf{b}\cdot\mathbf{n}ds - (\nabla\cdot(v\mathbf{b}), u) = \int_{\partial\Omega} uv\mathbf{b}\cdot\mathbf{n}ds - (\mathbf{b}\cdot\nabla v, u) - (v\nabla\cdot\mathbf{b}, u),$$

we have

$$2(\mathbf{b}\cdot\nabla v, v) + 2(cv,v) = \int_{\partial\Omega} v^2 \mathbf{b}\cdot\mathbf{n}ds + ((2c - \nabla\cdot\mathbf{b})v, v) \geq 0, \quad \forall v \in H_0^1(\Omega). \tag{2.4}$$

By the equivalence of two norms $|\cdot|_1$ and $\|\cdot\|_1$ for the space $H_0^1(\Omega)$ (see [5]), we conclude that the bilinear form $A(u,v) = (\mathbf{a}\nabla u, \nabla v) + (\mathbf{b}\cdot\nabla u, v) + (cu,v)$ satisfies coercivity $A(v,v) \geq C\|v\|_1$ for any $v \in H_0^1(\Omega)$.

The coercivity can also be achieved if we assume $|\mathbf{b}| < 4\lambda_{\mathbf{a}} c$. By Young's inequality

$$|(\mathbf{b}\cdot\nabla v, v)| \leq \iint_\Omega \frac{|\mathbf{b}\cdot\nabla v|^2}{4c} + c|v|^2 dxdy \leq \left(\frac{|\mathbf{b}|^2}{4c}\nabla v, \nabla v\right) + (cv,v),$$

we have

$$A(v,v) \geq (\mathbf{a}\nabla v, \nabla v) + (cv,v) - |(\mathbf{b}\cdot\nabla v, v)| \geq \left((\lambda_{\mathbf{a}} - \frac{|\mathbf{b}|^2}{4c})\nabla v, \nabla v\right) > 0, \quad \forall v \in H_0^1(\Omega). \tag{2.5}$$

Let A^* be the dual operator of A, i.e., $A^*(u,v) = A(v,u)$. We need to assume the elliptic regularity holds for the dual problem of (2.3) :

$$w \in H_0^1(\Omega), A^*(w,v) = (f,v), \quad \forall v \in H_0^1(\Omega) \Longrightarrow \|w\|_2 \leq C\|f\|_0, \tag{2.6}$$

where C is independent of w and f. See [16, 9] for the elliptic regularity with Lipschitz continuous coefficients on a Lipschitz domain.

3. Quadrature error estimates. In the following, we will use $\hat{}$ for a function to emphasize the function is defined on or transformed to the reference cell $\hat{K} = [-1,1]\times[-1,1]$ from a mesh cell.

3.1. Standard estimates. The Bramble-Hilbert Lemma for Q^k polynomials can be stated as follows, see Exercise 3.1.1 and Theorem 4.1.3 in [6]:

THEOREM 3.1. *If a continuous linear mapping $\hat{\Pi}: H^{k+1}(\hat{K}) \to H^{k+1}(\hat{K})$ satisfies $\hat{\Pi}\hat{v} = \hat{v}$ for any $\hat{v} \in Q^k(\hat{K})$, then*

$$\|\hat{u} - \hat{\Pi}\hat{u}\|_{k+1,\hat{K}} \leq C[\hat{u}]_{k+1,\hat{K}}, \quad \forall \hat{u} \in H^{k+1}(\hat{K}). \tag{3.1}$$

Thus if $l(\cdot)$ is a continuous linear form on the space $H^{k+1}(\hat{K})$ satisfying $l(\hat{v}) = 0, \forall \hat{v} \in Q^k(\hat{K})$, then

$$|l(\hat{u})| \leq C\|l\|'_{k+1,\hat{K}}[\hat{u}]_{k+1,\hat{K}}, \quad \forall \hat{u} \in H^{k+1}(\hat{K}),$$

where $\|l\|'_{k+1,\hat{K}}$ is the norm in the dual space of $H^{k+1}(\hat{K})$.

By applying Bramble-Hilbert Lemma, we have the following standard quadrature estimates. See Theorem 2.3 and Theorem 2.4 in [13] for the detailed proof.

THEOREM 3.2. *For a sufficiently smooth function $a(x,y) \in H^{2k}(e)$ and $k \geq 2$, let m is an integer satisfying $k \leq m \leq 2k$, we have*

$$\iint_e a(x,y)dxdy - \iint_e a_I(x,y)dxdy = \mathcal{O}(h^{m+1})[a]_{m,e} = \mathcal{O}(h^{m+2})[a]_{m,\infty,e}.$$

THEOREM 3.3. *If $f \in H^{k+2}(\Omega)$ with $k \geq 2$, then*

$$(f, v_h) - \langle f, v_h\rangle_h = \mathcal{O}(h^{k+2})\|f\|_{k+2}\|v_h\|_2, \quad \forall v_h \in V^h.$$

REMARK 3.4. *By the Theorem 3.1, on the reference cell $\hat{K}$, for $a(x,y) \in H^{k+2}(e)$ and $k \geq 2$, we have*

$$\iint_{\hat{K}} \hat{a}(s,t) - \hat{a}_I(s,t)dsdt \leq C[\hat{a}]_{k+2,\hat{K}} \leq C[\hat{a}]_{k+2,\infty,\hat{K}}, \tag{3.2}$$

and

$$\|\hat{a} - \hat{a}_I\|_{k+1,\hat{K}} \leq C[\hat{a}]_{k+1,\hat{K}}. \tag{3.3}$$

The following two results are also standard estimates obtained by applying the Bramble-Hilbert Lemma.

LEMMA 3.5. *If $f \in H^2(\Omega)$ or $f \in V^h$, we have $(f, v_h) - \langle f, v_h\rangle_h = \mathcal{O}(h^2)|f|_2\|v_h\|_0, \quad \forall v_h \in V^h$.*

Proof. For simplicity, we ignore the subscript in v_h. Let $E(f)$ denote the quadrature error for integrating $f(x,y)$ on e. Let $\hat{E}(\hat{f})$ denote the quadrature error for integrating $\hat{f}(s,t) = f(x_e + sh, y_e + th)$ on the reference cell $\hat{K}$. Due to the embedding $H^2(\hat{K}) \hookrightarrow C^0(\hat{K})$, we have

$$|\hat{E}(\hat{f}\hat{v})| \leq C|\hat{f}\hat{v}|_{0,\infty,\hat{K}} \leq C|\hat{f}|_{0,\infty,\hat{K}}|\hat{v}|_{0,\infty,\hat{K}} \leq C\|\hat{f}\|_{2,\hat{K}}\|\hat{v}\|_{0,\hat{K}}.$$

Thus the mapping $\hat{f} \to E(\hat{f}\hat{v})$ is a continuous linear form on $H^2(\hat{K})$ and its norm is bounded by $C\|\hat{v}\|_{0,\hat{K}}$. If $\hat{f} \in Q^1(\hat{K})$, then we have $\hat{E}(\hat{f}\hat{v}) = 0$. By the Bramble-Hilbert Lemma Theorem 3.1 on this continuous linear form, we get

$$|\hat{E}(\hat{f}\hat{v})| \leq C[\hat{f}]_{2,\hat{K}}\|\hat{v}\|_{0,\hat{K}}.$$

So on a cell e, we get

$$E(fv) = h^2\hat{E}(\hat{f}\hat{v}) \leq Ch^2[\hat{f}]_{2,\hat{K}}\|\hat{v}\|_{0,\hat{K}} \leq Ch^2|f|_{2,e}\|v\|_{0,e}. \tag{3.4}$$

Summing over all elements and use Cauchy-Schwarz inequality, we get the desired result. ∎

THEOREM 3.6. *Assume all coefficients of* (2.3) *are in* $W^{2,\infty}(\Omega)$. *We have*

$$A(z_h, v_h) - A_h(z_h, v_h) = \mathcal{O}(h)\|v_h\|_2\|z_h\|_1, \quad \forall v_h, z_h \in V^h.$$

Proof. Following the same arguments as in the proof of Lemma 3.4, we have

$$E(fv) \leq Ch^2|f|_{2,e}\|v\|_{0,e}, \forall f, v \in V^h.$$

Let $f = a^{11}(v_h)_x$ and $v = (z_h)_x$ in the estimate above, we get

$$\begin{aligned}
&|(a^{11}(z_h)_x, (v_h)_x) - \langle a^{11}(z_h)_x, (v_h)_x\rangle_h| \leq Ch^2\|a^{11}(v_h)_x\|_2\|(z_h)_x\|_0 \\
\leq &Ch^2\|a^{11}\|_{2,\infty}\|v_h\|_3|z_h|_1 \leq Ch\|a^{11}\|_{2,\infty}\|v_h\|_2|z_h|_1,
\end{aligned}$$

where the inverse estimate (2.1) is used in the last inequality. Similarly, we have

$$\begin{aligned}
(a^{12}(z_h)_x, (v_h)_y) - \langle a^{12}(z_h)_x, (v_h)_y\rangle_h &= Ch\|a^{12}\|_{2,\infty}\|v_h\|_2|z_h|_1, \\
(a^{22}(z_h)_y, (v_h)_y) - \langle a^{22}(z_h)_y, (v_h)_y\rangle_h &= Ch\|a^{22}\|_{2,\infty}\|v_h\|_2|z_h|_1, \\
(b^1(z_h)_x, v_h) - \langle b^1(z_h)_x, v_h\rangle_h &= Ch\|b^1\|_{2,\infty}\|v_h\|_2|z_h|_0, \\
(b^2(z_h)_y, v_h) - \langle b^2(z_h)_y, v_h\rangle_h &= Ch\|b^2\|_{2,\infty}\|v_h\|_2|z_h|_0, \\
(cz_h, v_h) - \langle cz_h, v_h\rangle_h &= Ch\|c\|_{2,\infty}\|v_h\|_1|z_h|_0,
\end{aligned}$$

which implies

$$A(z_h, v_h) - A_h(z_h, v_h) = \mathcal{O}(h)\|v_h\|_2\|z_h\|_1.$$ ∎

3.2. A refined consistency error. In this subsection, we will show how to establish the desired consistency error estimate for smooth enough coefficients:

$$A(u, v_h) - A_h(u, v_h) = \begin{cases} \mathcal{O}(h^{k+2})\|u\|_{k+3}\|v_h\|_2, & \forall v_h \in V_0^h \\ \mathcal{O}(h^{k+\frac{3}{2}})\|u\|_{k+3}\|v_h\|_2, & \forall v_h \in V^h \end{cases}.$$

THEOREM 3.7. *Assume* $a(x,y) \in W^{k+2,\infty}(\Omega)$, $u \in H^{k+3}(\Omega)$, $k \geq 2$, *then*

$$(a\partial_x u, \partial_x v_h) - \langle a\partial_x u, \partial_x v_h\rangle_h = \begin{cases} \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty}\|u\|_{k+3}\|v_h\|_2, & \forall v_h \in V_0^h, \quad (3.5a) \\ \mathcal{O}(h^{k+\frac{3}{2}})\|a\|_{k+2,\infty}\|u\|_{k+3}\|v_h\|_2, & \forall v_h \in V^h, \quad (3.5b) \end{cases}$$

$$(a\partial_x u, \partial_y v_h) - \langle a\partial_x u, \partial_y v_h\rangle_h = \begin{cases} \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty}\|u\|_{k+3}\|v_h\|_2, & \forall v_h \in V_0^h, \quad (3.6a) \\ \mathcal{O}(h^{k+\frac{3}{2}})\|a\|_{k+2,\infty}\|u\|_{k+3}\|v_h\|_2, & \forall v_h \in V^h, \quad (3.6b) \end{cases}$$

$$(a\partial_x u, v_h) - \langle a\partial_x u, v_h\rangle_h = \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty}\|u\|_{k+3}\|v_h\|_2, \quad \forall v_h \in V_0^h, \tag{3.7}$$

$$(au, v_h) - \langle au, v_h\rangle_h = \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty}\|u\|_{k+2}\|v_h\|_2, \quad \forall v_h \in V_0^h. \tag{3.8}$$

REMARK 3.8. *We emphasize that Theorem 3.7 cannot be proven by applying the Bramble-Hilbert Lemma directly. Consider the constant coefficient case $a(x,y) \equiv 1$ and $k = 2$ as an example,*

$$(\partial_x u, \partial_x v_h) - \langle \partial_x u, \partial_x v_h\rangle_h = \sum_e \left(\iint_e u_x (v_h)_x dxdy - \iint_e u_x (v_h)_x d^h x d^h y \right).$$

Since the 3×3 Gauss-Lobatto quadrature is exact for integrating Q^3 polynomials, by Theorem 3.1 we have

$$\left| \iint_e u_x (v_h)_x dxdy - \iint_e u_x (v_h)_x d^h x d^h y \right| = \left| \iint_{\hat K} \hat u_s (\hat v_h)_s dsdt - \iint_{\hat K} \hat u_s (\hat v_h)_s d^h s d^h t \right| \le C[\hat u_s (\hat v_h)_s]_{4,\hat K}.$$

Notice that $\hat v_h$ is Q^2 thus $(\hat v_h)_{stt}$ does not vanish and $[(\hat v_h)_s]_{4,\hat K} \le C|\hat v_h|_{3,\hat K}$. So by Bramble-Hilbert Lemma for Q^k polynomials, we can only get

$$\iint_e u_x (v_h)_x dxdy - \iint_e u_x (v_h)_x d^h x d^h y = \mathcal{O}(h^4)\|u\|_{5,e}\|v_h\|_{3,e}.$$

Thus by Cauchy-Schwarz inequality after summing over e, we only have

$$(\partial_x u, \partial_x v_h) - \langle \partial_x u, \partial_x v_h\rangle_h = \mathcal{O}(h^4)\|u\|_5\|v_h\|_3.$$

In order to get the desired estimate involving only the broken H^2-norm of v_h, we will take advantage of error cancellations between neighboring cells through integration by parts.

Proof. For simplicity, we ignore the subscript $_h$ of v_h in this proof and all the following v are in V^h which are Q^k polynomials in each cell. First, by Theorem 3.3, we easily obtain (3.7) and (3.8):

$$(au_x, v) - \langle au_x, v\rangle_h = \mathcal{O}(h^{k+2})\|au_x\|_{k+2}\|v\|_2 = \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty}\|u\|_{k+3}\|v\|_2,$$

$$(au, v) - \langle au, v\rangle_h = \mathcal{O}(h^{k+2})\|au\|_{k+2}\|v\|_2 = \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty}\|u\|_{k+2}\|v\|_2.$$

We will only discuss $(au_x, v_x) - \langle au_x, v_x\rangle_h$ and the same discussion also applies to derive (3.6a) and (3.6b).

Since we have

$$(au_x, v_x) - \langle au_x, v_x\rangle_h = \sum_e \left(\iint_e au_x v_x dxdy - \iint_e au_x v_x d^h x d^h y \right)$$

$$= \sum_e \left(\iint_{\hat K} \hat a \hat u_s \hat v_s dsdt - \iint_{\hat K} \hat a \hat u_s \hat v_s d^h s d^h t \right) = \sum_e \left(\iint_{\hat K} \hat a \hat u_s \hat v_s dsdt - \iint_{\hat K} (\hat a \hat u_s)_I \hat v_s d^h s d^h t \right),$$

where we use the fact $\hat a \hat u_s \hat v_s = (\hat a \hat u_s)_I \hat v_s$ on the Gauss-Lobatto quadrature points. For fixed t, $(\hat a \hat u_s)_I \hat v_s$ is a polynomial of degree $2k-1$ w.r.t. variable s, thus the $(k+1)$-point Gauss-Lobatto quadrature is exact for its s-integration, i.e.,

$$\iint_{\hat K} (\hat a \hat u_s)_I \hat v_s d^h s d^h t = \iint_{\hat K} (\hat a \hat u_s)_I \hat v_s ds d^h t.$$

To estimate the quadrature error we introduce some intermediate values then do interpretation by parts,

$$\iint_{\hat{K}} \hat{a}\hat{u}_s\hat{v}_s dsdt - \iint_{\hat{K}} (\hat{a}\hat{u}_s)_I \hat{v}_s d^h s d^h t \tag{3.9}$$

$$= \iint_{\hat{K}} \hat{a}\hat{u}_s\hat{v}_s dsdt - \iint_{\hat{K}} (\hat{a}\hat{u}_s)_I \hat{v}_s dsdt + \iint_{\hat{K}} (\hat{a}\hat{u}_s)_I \hat{v}_s dsdt - \iint_{\hat{K}} (\hat{a}\hat{u}_s)_I \hat{v}_s dsd^h t \tag{3.10}$$

$$= \iint_{\hat{K}} \left[\hat{a}\hat{u}_s - (\hat{a}\hat{u}_s)_I\right] \hat{v}_s dsdt + \left(\iint_{\hat{K}} \left[(\hat{a}\hat{u}_s)_I\right]_s \hat{v} dsd^h t - \iint_{\hat{K}} \left[(\hat{a}\hat{u}_s)_I\right]_s \hat{v} dsdt \right) \tag{3.11}$$

$$+ \left(\int_{-1}^{1} (\hat{a}\hat{u}_s)_I \hat{v} dt \Big|_{s=-1}^{s=1} - \int_{-1}^{1} (\hat{a}\hat{u}_s)_I \hat{v} d^h t \Big|_{s=-1}^{s=1} \right) = I + II + III. \tag{3.12}$$

For the first term in (3.12), let $\overline{\hat{v}_s}$ be the cell average of $\hat{v}_s$ on $\hat{K}$, then

$$I = \iint_{\hat{K}} \left(\hat{a}\hat{u}_s - (\hat{a}\hat{u}_s)_I\right) \overline{\hat{v}_s} dsdt + \iint_{\hat{K}} \left(\hat{a}\hat{u}_s - (\hat{a}\hat{u}_s)_I\right) (\hat{v}_s - \overline{\hat{v}_s}) dsdt.$$

By (3.2) we have

$$\left| \iint_{\hat{K}} \left(\hat{a}\hat{u}_s - (\hat{a}\hat{u}_s)_I\right) \overline{\hat{v}_s} dsdt \right| \le C[\hat{a}\hat{u}_s]_{k+2,\hat{K}} \left|\overline{\hat{v}_s}\right| = \mathcal{O}(h^{k+2}) \|\hat{a}\|_{k+2,\infty,e} \|\hat{u}\|_{k+3,e} \|\hat{v}\|_{1,e}.$$

By Cauchy-Schwarz inequality, the Bramble-Hilbert Lemma on interpolation error and Poincaré inequality, we have

$$\left| \iint_{\hat{K}} \left(\hat{a}\hat{u}_s - (\hat{a}\hat{u}_s)_I\right) (\hat{v}_s - \overline{\hat{v}_s}) dsdt \right| \le |\hat{a}\hat{u}_s - (\hat{a}\hat{u}_s)_I|_{0,\hat{K}} |\hat{v}_s - \overline{\hat{v}_s}|_{0,\hat{K}}$$

$$\le C[\hat{a}\hat{u}_s]_{k+1,\hat{K}} |\hat{v}|_{2,\hat{K}} = \mathcal{O}(h^{k+2}) \|a\|_{k+1,\infty,e} \|u\|_{k+2,e} \|v\|_{2,e}.$$

Thus we have

$$I = \mathcal{O}(h^{k+2}) \|a\|_{k+2,\infty,e} \|u\|_{k+3,e} \|v\|_{2,e}.$$

For the second term in (3.12), we can estimate it the same way as in the proof of Theorem 2.4. in [13]. For each $\hat{v} \in Q^k(\hat{K})$ we can define a linear form on $H^k(\hat{K})$ as

$$\hat{E}_{\hat{v}}(\hat{f}) = \iint_{\hat{K}} (\hat{F}_I)_s \hat{v} dsdt - \iint_{\hat{K}} (\hat{F}_I)_s \hat{v} dsd^h t,$$

where $\hat{F}$ is an antiderivative of $\hat{f}$ w.r.t. variable s. Due to the linearity of interpolation operator and differentiating operation, $\hat{E}_{\hat{v}}$ is well defined. By the embedding $H^2(\hat{K}) \hookrightarrow C^0(\hat{K})$, we have

$$\hat{E}_{\hat{v}}(\hat{f}) \le C\|\hat{F}\|_{0,\infty,\hat{K}} \|\hat{v}\|_{0,\infty,\hat{K}} \le C\|\hat{f}\|_{0,\infty,\hat{K}} \|\hat{v}\|_{0,\infty,\hat{K}} \le C\|\hat{f}\|_{2,\hat{K}} \|\hat{v}\|_{0,\hat{K}} \le C\|\hat{f}\|_{k,\hat{K}} \|\hat{v}\|_{0,\hat{K}},$$

where we use the fact that all the norms on $Q^k(\hat{K})$ are equivalent to derive the first inequality. The above inequalities imply that the mapping $\hat{E}_{\hat{v}}$ is a continuous linear form on $H^k(\hat{K})$. With projection Π_1 defined in (2.2), we have

$$\hat{E}_{\hat{v}}(\hat{f}) = \hat{E}_{\hat{v}-\Pi_1\hat{v}}(\hat{f}) + \hat{E}_{\Pi_1\hat{v}}(\hat{f}), \quad \forall \hat{v} \in Q^k(\hat{K}).$$

Notice that $\hat{F}$ by definition is an antiderivative of $\hat{f}$ w.r.t. only variable s. If $\hat{f} \in Q^{k-1}(\hat{K})$, then $\hat{F}_I$ is a polynomial of degree only $k-1$ w.r.t. to variable t thus $(\hat{F}_I)_s \in Q^{k-1}(\hat{K})$. The quadrature is exact for polynomials of degree $2k-1$, thus $Q^{k-1}(\hat{K}) \subset \ker \hat{E}_{\hat{v}-\Pi_1\hat{v}}$. So by the Bramble-Hilbert Lemma, we get

$$\hat{E}_{\hat{v}-\Pi_1\hat{v}}(\hat{f}) \le C[f]_{k,\hat{K}} \|\hat{v}-\Pi_1\hat{v}\|_{0,\hat{K}} \le C[f]_{k,\hat{K}} |\hat{v}|_{2,\hat{K}},$$

and we also have

$$\hat{E}_{\Pi_1\hat{v}}(\hat{f}) = \iint_{\hat{K}} (\hat{F}_I)_s \Pi_1\hat{v} ds dt - \iint_{\hat{K}} (\hat{F}_I)_s \Pi_1 \hat{v} ds d^h t = 0.$$

Thus we have

$$\iint_{\hat{K}} [(\hat{a}\hat{u}_s)_I]_s \, \hat{v} ds d^h t - \iint_{\hat{K}} [(\hat{a}\hat{u}_s)_I]_s \, \hat{v} ds dt = -\hat{E}_{\hat{v}}((\hat{a}\hat{u}_s)_s) = -\hat{E}_{\hat{v}-\Pi_1\hat{v}}((\hat{a}\hat{u}_s)_s)$$
$$\le C[(\hat{a}\hat{u}_s)_s]_{k,\hat{K}} |\hat{v}_h|_{2,\hat{K}} \le C|\hat{a}\hat{u}_s|_{k+1,\hat{K}} |\hat{v}|_{2,\hat{K}} = \mathcal{O}(h^{k+2}) \|a\|_{k+1,\infty,e} \|u\|_{k+2,e} |v|_{2,e}$$

Now we only need to discuss the line integral term. Let L_2 and L_4 denote the left and right boundary of Ω and let l_2^e and l_4^e denote the left and right edge of element e or $l_2^{\hat{K}}$ and $l_4^{\hat{K}}$ for $\hat{K}$. Since $(\hat{a}\hat{u}_s)_I\hat{v}$ mapped back to e will be $\frac{1}{h}(au_x)_I v$ which is continuous across l_2^e and l_4^e, after summing over all elements e, the line integrals along the inner edges are canceled out and only the line integrals on L_2 and L_4 remain.

For a cell e adjacent to L_2, consider its reference cell $\hat{K}$, and define a linear form $\hat{E}(\hat{f}) = \int_{-1}^{1} \hat{f}(-1,t)dt - \int_{-1}^{1} \hat{f}(-1,t)d^h t$, then we have

$$\hat{E}(\hat{f}\hat{v}) \le C|\hat{f}|_{0,\infty,l_2^{\hat{K}}} |\hat{v}|_{0,\infty,l_2^{\hat{K}}} \le C\|\hat{f}\|_{2,l_2^{\hat{K}}} \|\hat{v}\|_{0,l_2^{\hat{K}}},$$

which means that the mapping $\hat{f} \to \hat{E}(\hat{f}\hat{v})$ is continuous with operator norm less than $C\|\hat{v}\|_{0,l_2^{\hat{K}}}$ for some C. Clearly we have

$$\hat{E}(\hat{f}\hat{v}) = \hat{E}(\hat{f}\Pi_1\hat{v}) + \hat{E}(\hat{f}(\hat{v}-\Pi_1\hat{v})).$$

By the Theorem 3.1 we get

$$\hat{E}((\hat{a}\hat{u}_s)_I(\hat{v}-\Pi_1\hat{v})) \le C[(\hat{a}\hat{u}_s)_I]_{k,l_2^{\hat{K}}} [\hat{v}]_{2,l_2^{\hat{K}}} \le C(|\hat{a}\hat{u}_s - (\hat{a}\hat{u}_s)_I|_{k,l_2^{\hat{K}}} + |\hat{a}\hat{u}_s|_{k,l_2^{\hat{K}}})[\hat{v}]_{2,l_2^{\hat{K}}}$$
$$\le (|\hat{a}\hat{u}_s|_{k+1,l_2^{\hat{K}}} + |\hat{a}\hat{u}_s|_{k,l_2^{\hat{K}}})[\hat{v}]_{2,l_2^{\hat{K}}} = \mathcal{O}(h^{k+2})\|a\|_{k+1,\infty,l_2^e} \|u\|_{k+2,l_2^e} [v]_{2,l_2^e},$$

where the first inequality comes from the accuracy of the $(k+1)$-point Gauss-Lobatto quadrature rule, i.e. $\hat{E}(\hat{f}) = 0, \forall \hat{f} \in Q^{2k-1}(\hat{K})$. The $(k+1)$-point Gauss-Lobatto quadrature rule also gives

$$\hat{E}((\hat{a}\hat{u}_s)_I \Pi_1\hat{v}) = 0.$$

For the third term in (3.12), we sum them up over all the elements. Then for the line integral along L_2

$$\sum_{e \cap L_2 \neq \emptyset} \int_{-1}^{1} (\hat{a}\hat{u}_s)_I(-1,t)\hat{v}(-1,t)dt - \sum_{e \cap L_2 \neq \emptyset} \int_{-1}^{1} (\hat{a}\hat{u}_s)_I(-1,t)\hat{v}(-1,t)d^h t$$
$$= \sum_{e \cap L_2 \neq \emptyset} \hat{E}((\hat{a}\hat{u}_s)_I\hat{v}) = \sum_{e \cap L_2 \neq \emptyset} \mathcal{O}(h^{k+2})\|a\|_{k+1,\infty,l_2^e} \|u\|_{k+2,l_2^e} |v|_{2,l_2^e}.$$

Let s_α and ω_α ($\alpha = 1, 2, \cdots, k+2$) denote the quadrature points and weights in $(k+2)$-point Gauss-Lobatto quadrature rule for $s \in [-1, 1]$. Since $\hat{v}_{tt}^2(s,t) \in Q^{2k}(\hat{K})$, $(k+2)$-point Gauss-Lobatto quadrature is exact for s-integration thus

$$\int_{-1}^{1}\int_{-1}^{1} \hat{v}_{tt}^2(s,t)dsdt = \sum_{\alpha=1}^{k+2} \omega_\alpha \int_{-1}^{1} \hat{v}_{tt}^2(s_\alpha, t)dt,$$

which implies

$$\int_{-1}^{1} \hat{v}_{tt}^2(\pm 1, t)dt \leq C \int_{-1}^{1}\int_{-1}^{1} \hat{v}_{tt}^2(s,t)dsdt, \tag{3.13}$$

thus

$$h^{\frac{1}{2}}|v|_{2,l_2^e} \leq C[v]_{2,e}.$$

By Cauchy-Schwarz inequality and trace inequality, we have

$$\begin{aligned}
&\sum_{e\cap L_2\neq\emptyset} \left(\int_{-1}^{1} (\hat{a}\hat{u}_s)_I \hat{v} dt\Big|_{s=-1}^{s=1} - \int_{-1}^{1} (\hat{a}\hat{u}_s)_I \hat{v} d^h t\Big|_{s=-1}^{s=1} \right) \\
= &\sum_{e\cap L_2\neq\emptyset} \mathcal{O}(h^{k+2})\|a\|_{k+1,\infty,l_2^e}\|u\|_{k+2,l_2^e}|v|_{2,l_2^e} \\
= &\sum_{e\cap L_2\neq\emptyset} \mathcal{O}(h^{k+\frac{3}{2}})\|a\|_{k+1,\infty,l_2^e}\|u\|_{k+2,l_2^e}|v|_{2,e} = \mathcal{O}(h^{k+\frac{3}{2}})\|a\|_{k+1,\infty,\Omega}\|u\|_{k+2,L_2}|v|_{2,\Omega} \\
= &\mathcal{O}(h^{k+\frac{3}{2}})\|a\|_{k+1,\infty,\Omega}\|u\|_{k+3,\Omega}|v|_{2,\Omega}.
\end{aligned}$$

Combine all the estimates above, we get (3.5b). Since the $\frac{1}{2}$ order loss is only due to the line integral along the boundary $\partial\Omega$. If $v \in V_0^h$, $v_{yy} = 0$ on L_2 and L_4 so we have (3.5a). ◻

4. Superconvergence of bilinear forms.

The M-type projection in [3, 4] is a very convenient tool for discussing the superconvergence of function values. Let u_p be the M-type Q^k projection of the smooth exact solution u and its definition will be given in the following subsection. To establish the superconvergence of the original finite element method (1.1) for a generic elliptic problem (2.3) with smooth coefficients, one can show the following superconvergence of bilinear forms, see [4, 14] (see also [13] for a detailed proof):

$$A(u - u_p, v_h) = \begin{cases} \mathcal{O}(h^{k+2})\|u\|_{k+3}\|v_h\|_2, & \forall v_h \in V_0^h, \\ \mathcal{O}(h^{k+\frac{3}{2}})\|u\|_{k+3}\|v_h\|_2, & \forall v_h \in V^h. \end{cases}$$

In this section we will show the superconvergence of the bilinear form A_h:

$$A_h(u - u_p, v_h) = \begin{cases} \mathcal{O}(h^{k+2})\|u\|_{k+3}\|v_h\|_2, & \forall v_h \in V_0^h, \quad (4.1a) \\ \mathcal{O}(h^{k+\frac{3}{2}})\|u\|_{k+3}\|v_h\|_2, & \forall v_h \in V^h. \quad (4.1b) \end{cases}$$

4.1. Definition of M-type projection.

We first recall the definition of M-type projection. More detailed definition can also be found in [13]. Legendre polynomials on the reference interval $[-1, 1]$ are given as

$$l_k(t) = \frac{1}{2^k k!}\frac{d^k}{dt^k}(t^2-1)^k : l_0(t) = 1, l_1(t) = t, l_2(t) = \frac{1}{2}(3t^2-1), \cdots,$$

which are L^2-orthogonal to one another. Define their antiderivatives as M-type polynomials:

$$M_{k+1}(t) = \frac{1}{2^k k!}\frac{d^{k-1}}{dt^{k-1}}(t^2-1)^k : M_0(t) = 1, M_1(t) = t, M_2(t) = \frac{1}{2}(t^2-1), M_3(t) = \frac{1}{2}(t^3-t), \cdots.$$

which satisfy the following properties:

- If $j - i \neq 0, \pm 2$, then $M_i(t) \perp M_j(t)$, i.e., $\int_{-1}^{1} M_i(t)M_j(t)dt = 0$.
- Roots of $M_k(t)$ are the k-point Gauss-Lobatto quadrature points for $[-1, 1]$.

Since Legendre polynomials form a complete orthogonal basis for $L^2([-1, 1])$, for any $\hat{f}(t) \in H^1([-1, 1])$, its derivative $\hat{f}'(t)$ can be expressed as Fourier-Legendre series

$$\hat{f}'(t) = \sum_{j=0}^{\infty} \hat{b}_{j+1} l_j(t), \quad \hat{b}_{j+1} = (j + \frac{1}{2}) \int_{-1}^{1} \hat{f}'(t) l_j(t) dt.$$

The one-dimensional M-type projection is defined as $\hat{f}_k(t) = \sum_{j=0}^{k} \hat{b}_j M_j(t)$, where $\hat{b}_0 = \frac{\hat{f}(1)+\hat{f}(-1)}{2}$ is determined by $\hat{b}_1 = \frac{\hat{f}(1)-\hat{f}(-1)}{2}$ so that $\hat{f}_k(\pm 1) = \hat{f}(\pm 1)$. We have $\hat{f}(t) = \lim_{k\to\infty} \hat{f}_k(t) = \sum_{j=0}^{\infty} \hat{b}_j M_j(t)$. The remainder $\hat{R}[\hat{f}]_k(t)$ of one-dimensional M-type projection is

$$\hat{R}[\hat{f}]_k(t) = \hat{f}(t) - \hat{f}_k(t) = \sum_{j=k+1}^{\infty} \hat{b}_j M_j(t).$$

For a function $\hat{f}(s, t) \in H^2(\hat{K})$ on the reference cell $\hat{K} = [-1, 1] \times [-1, 1]$, its two-dimensional M-type expansion is given as

$$\hat{f}(s, t) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \hat{b}_{i,j} M_i(s) M_j(t),$$

where

$$\begin{aligned}
\hat{b}_{0,0} &= \frac{1}{4}[\hat{f}(-1, -1) + \hat{f}(-1, 1) + \hat{f}(1, -1) + \hat{f}(1, 1)],\\
\hat{b}_{0,j}, \hat{b}_{1,j} &= \frac{2j-1}{4} \int_{-1}^{1} [\hat{f}_t(1, t) \pm \hat{f}_t(-1, t)] l_{j-1}(t) dt, \quad j \geq 1,\\
\hat{b}_{i,0}, \hat{b}_{i,1} &= \frac{2i-1}{4} \int_{-1}^{1} [\hat{f}_s(s, 1) \pm \hat{f}_s(s, -1)] l_{i-1}(s) ds, \quad i \geq 1,\\
\hat{b}_{i,j} &= \frac{(2i-1)(2j-1)}{4} \iint_{\hat{K}} \hat{f}_{st}(s, t) l_{i-1}(s) l_{j-1}(t) ds dt, \quad i, j \geq 1.
\end{aligned}$$

The M-type Q^k projection of $\hat{f}$ on $\hat{K}$ and its remainder are defined as

$$\hat{f}_{k,k}(s, t) = \sum_{i=0}^{k} \sum_{j=0}^{k} \hat{b}_{i,j} M_i(s) M_j(t), \quad \hat{R}[\hat{f}]_{k,k}(s, t) = \hat{f}(s, t) - \hat{f}_{k,k}(s, t).$$

The M-type Q^k projection is equivalent to the point-line-plane interpolation used in [15, 14]. See Theorem 3.1 in [13] for the proof of the following fact:

THEOREM 4.1. *For $k \geq 2$, the M-type Q^k projection is equivalent to the Q^k point-line-plane projection Π defined as follows:*

1. $\Pi\hat{u} = \hat{u}$ *at four corners of* $\hat{K} = [-1,1]\times[-1,1]$.
2. $\Pi\hat{u} - \hat{u}$ *is orthogonal to polynomials of degree* $k-2$ *on each edge of* $\hat{K}$.
3. $\Pi\hat{u} - \hat{u}$ *is orthogonal to any* $\hat{v} \in Q^{k-2}(\hat{K})$ *on* $\hat{K}$.

For $f(x,y)$ on $e = [x_e - h, x_e + h] \times [y_e - h, y_e + h]$, let $\hat{f}(s,t) = f(sh + x_e, th + y_e)$ then the M-type Q^k projection of f on e and its remainder are defined as

$$f_{k,k}(x,y) = \hat{f}_{k,k}\Big(\frac{x-x_e}{h}, \frac{y-y_e}{h}\Big), \quad R[f]_{k,k}(x,y) = f(x,y) - f_{k,k}(x,y).$$

Now consider a function $u(x,y) \in H^{k+2}(\Omega)$, let $u_p(x,y)$ denote its piecewise M-type Q^k projection on each element e in the mesh Ω_h. The first two properties in Theorem 4.1 imply that $u_p(x,y)$ on each edge of e is uniquely determined by $u(x,y)$ along that edge. So $u_p(x,y)$ is a piecewise continuous Q^k polynomial on Ω_h.

M-type projection has the following properties. See Theorem 3.2, Lemma 3.1 and Lemma 3.2 in [13] for the proof.

THEOREM 4.2. *For* $k \geq 2$,

$$\|u - u_p\|_{2,Z_0} = \mathcal{O}(h^{k+2})\|u\|_{k+2}, \quad \forall u \in H^{k+2}(\Omega).$$

$$\|u - u_p\|_{\infty,Z_0} = \mathcal{O}(h^{k+2})\|u\|_{k+2,\infty}, \quad \forall u \in W^{k+2,\infty}(\Omega).$$

LEMMA 4.3. *For* $\hat{f} \in H^{k+1}(\hat{K})$, $k \geq 2$,

1. $|\hat{R}[\hat{f}]_{k,k}|_{0,\infty,\hat{K}} \leq C[\hat{f}]_{k+1,\hat{K}}, \quad |\partial_s \hat{R}[\hat{f}]_{k,k}|_{0,\infty,\hat{K}} \leq C[\hat{f}]_{k+1,\hat{K}}$.
2. $\hat{R}[\hat{f}]_{k+1,k+1} - \hat{R}[\hat{f}]_{k,k} = M_{k+1}(t)\sum_{i=0}^{k}\hat{b}_{i,k+1}M_i(s) + M_{k+1}(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t)$.
3. $|\hat{b}_{i,k+1}| \leq C_k|\hat{f}|_{k+1,2,\hat{K}}, |\hat{b}_{k+1,i}| \leq C_k|\hat{f}|_{k+1,2,\hat{K}}, \quad 0 \leq i \leq k+1$.
4. *If* $\hat{f} \in H^{k+2}(\hat{K})$, *then* $|\hat{b}_{i,k+1}| \leq C_k|\hat{f}|_{k+2,2,\hat{K}}, \quad 1 \leq i \leq k+1$.

4.2. Estimates of M-type projection with quadrature.

LEMMA 4.4. *Assume* $\hat{f}(s,t) \in H^{k+3}(\hat{K})$, $k \geq 2$,

$$\langle \hat{R}[\hat{f}]_{k+1,k+1} - \hat{R}[\hat{f}]_{k,k}, 1\rangle_{\hat{K}} = 0, \quad |\langle \partial_s \hat{R}[\hat{f}]_{k+1,k+1}, 1\rangle_{\hat{K}}| \leq C|\hat{f}|_{k+3,\hat{K}}.$$

Proof. First, we have

$$\langle \hat{R}[\hat{f}]_{k+1,k+1} - \hat{R}[\hat{f}]_{k,k}, 1\rangle_{\hat{K}} = \langle M_{k+1}(t)\sum_{i=0}^{k}\hat{b}_{i,k+1}M_i(s) + M_{k+1}(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t), 1\rangle_{\hat{K}} = 0$$

due to the fact that roots of $M_{k+1}(t)$ are the $(k+1)$-point Gauss-Lobatto quadrature points for $[-1,1]$.

We have

$$\begin{aligned}
&\langle \partial_s \hat{R}[\hat{f}]_{k+1,k+1}, 1\rangle_{\hat{K}} \\
=&\langle \partial_s \hat{R}[\hat{f}]_{k+2,k+2}, 1\rangle_{\hat{K}} - \langle \partial_s(\hat{R}[\hat{f}]_{k+2,k+2} - \hat{R}[\hat{f}]_{k+1,k+1}), 1\rangle_{\hat{K}} \\
=&\langle \partial_s \hat{R}[\hat{f}]_{k+2,k+2}, 1\rangle_{\hat{K}} - \langle M_{k+2}(t)\sum_{i=0}^{k+1}\hat{b}_{i,k+2}M_i'(s) + M_{k+2}'(s)\sum_{j=0}^{k+2}\hat{b}_{k+2,j}M_j(t), 1\rangle_{\hat{K}} \\
=&\langle \partial_s \hat{R}[\hat{f}]_{k+2,k+2}, 1\rangle_{\hat{K}} - \langle M_{k+2}(t)\sum_{i=0}^{k}\hat{b}_{i+1,k+2}l_i(s), 1\rangle_{\hat{K}} + \langle l_{k+1}(s)\sum_{j=0}^{k+2}\hat{b}_{k+2,j}M_j(t), 1\rangle_{\hat{K}}.
\end{aligned}$$

Then by Lemma 4.3,

$$|\langle \partial_s \hat{R}[\hat{f}]_{k+2,k+2}, 1\rangle_{\hat{K}}| \le C|\hat{f}|_{k+3,\hat{K}}. \qquad \square$$

Notice that we have $\langle l_{k+1}(s) \sum_{j=0}^{k+2} \hat{b}_{k+2,j} M_j(t), 1\rangle_{\hat{K}} = 0$ since the $(k+1)$-point Gauss-Lobatto quadrature for s-integration is exact and $l_{k+1}(s)$ is orthogonal to 1. Lemma 4.3 implies $|\hat{b}_{i+1,k+2}| \le C[\hat{f}]_{k+3,\hat{K}}$ for $i \ge 0$, thus we have

$$|\langle M_{k+2}(t) \sum_{i=0}^{k} \hat{b}_{i+1,k+2} l_i(s), 1\rangle_{\hat{K}}| \le C[\hat{f}]_{k+3,\hat{K}}.$$

LEMMA 4.5. *Assume $a(x,y) \in W^{k,\infty}(\Omega)$, $u(x,y) \in H^{k+3}(\Omega)$ and $k \ge 2$. Then*

$$\langle a(u-u_p)_x, (v_h)_x\rangle_h = \mathcal{O}(h^{k+2})\|a\|_{2,\infty}\|u\|_{k+3}\|v_h\|_2, \quad \forall v_h \in V^h.$$

Proof. As before, we ignore the subscript of v_h for simplicity. We have

$$\langle a(u-u_p)_x, v_x\rangle_h = \sum_e \langle a(u-u_p)_x, v_x\rangle_{e,h},$$

and on each cell e,

$$\begin{aligned} \langle a(u-u_p)_x, v_x\rangle_{e,h} &= \langle (R[u]_{k,k})_x, av_x\rangle_{e,h} = \langle (\hat{R}[\hat{u}]_{k,k})_s, \hat{a}\hat{v}_s\rangle_{\hat{K}} \\ &= \langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\hat{v}_s\rangle_{\hat{K}} + \langle (\hat{R}[\hat{u}]_{k,k} - \hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\hat{v}_s\rangle_{\hat{K}}. \end{aligned} \tag{4.2}$$

For the first term in (4.2), we have

$$\langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\hat{v}_s\rangle_{\hat{K}} = \langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\overline{\hat{v}_s}\rangle_{\hat{K}} + \langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}(\hat{v}_s - \overline{\hat{v}_s})\rangle_{\hat{K}}.$$

By Lemma 4.4,

$$\langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \overline{\hat{a}}\,\overline{\hat{v}_s}\rangle_{\hat{K}} \le C|\hat{a}|_{0,\infty}|\hat{u}|_{k+3,\hat{K}}|\hat{v}|_{1,\hat{K}}.$$

By Lemma 4.3,

$$|(\hat{R}[\hat{u}]_{k+1,k+1})_s|_{0,\infty,\hat{K}} \le C[\hat{u}]_{k+2,\hat{K}}.$$

By Bramble-Hilbert Lemma Theorem 3.1 we have

$$\begin{aligned} &\langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\overline{\hat{v}_s}\rangle_{\hat{K}} = \langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \overline{\hat{a}}\,\overline{\hat{v}_s}\rangle_{\hat{K}} + \langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, (\hat{a}-\overline{\hat{a}})\overline{\hat{v}_s}\rangle_{\hat{K}} \\ &\le C(|\hat{a}|_{0,\infty}|\hat{u}|_{k+3,\hat{K}}|\hat{v}|_{1,\hat{K}} + |\hat{a}-\overline{\hat{a}}|_{0,\infty}|\hat{u}|_{k+2,\hat{K}}|\hat{v}|_{1,\hat{K}}) \\ &\le C(|\hat{a}|_{0,\infty}|\hat{u}|_{k+3,\hat{K}}|\hat{v}|_{1,\hat{K}} + |\hat{a}|_{1,\infty}|\hat{u}|_{k+2,\hat{K}}|\hat{v}|_{1,\hat{K}}) = \mathcal{O}(h^{k+2})\|a\|_{1,\infty,e}\|u\|_{k+3,e}\|v\|_{1,e}, \end{aligned}$$

and

$$\begin{aligned} &\langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}(\hat{v}_s - \overline{\hat{v}_s})\rangle_{\hat{K}} \le C[\hat{u}]_{k+2,2,\hat{K}}|\hat{a}|_{0,\infty,\hat{K}}|\hat{v}_s - \overline{\hat{v}_s}|_{0,\infty,\hat{K}} \\ &\le C[\hat{u}]_{k+2,2,\hat{K}}|\hat{a}|_{0,\infty,\hat{K}}|\hat{v}_s - \overline{\hat{v}_s}|_{0,2,\hat{K}} = \mathcal{O}(h^{k+2})[u]_{k+2,2,e}|a|_{0,\infty,e}|v|_{2,2,e}. \end{aligned}$$

Thus,

$$\langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\hat{v}_s\rangle_{\hat{K}} = \mathcal{O}(h^{k+2})\|a\|_{1,\infty,e}|u|_{k+3,2,e}\|v\|_{2,e}. \tag{4.3}$$

For the second term in (4.2), we have

$$\begin{aligned}
&\langle(\hat{R}[\hat{u}]_{k,k}-\hat{R}[\hat{u}]_{k+1,k+1})_s,\hat{a}\hat{v}_s\rangle_{\hat{K}}\\
=&-\langle(M_{k+1}(t)\sum_{i=0}^{k}\hat{b}_{i,k+1}M_i(s)+M_{k+1}(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t))_s,\hat{a}\hat{v}_s\rangle_{\hat{K}}\\
=&-\langle M_{k+1}(t)\sum_{i=0}^{k-1}\hat{b}_{i+1,k+1}l_i(s)+l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\hat{a}\hat{v}_s\rangle_{\hat{K}}\\
=&-\langle M_{k+1}(t)\sum_{i=0}^{k-1}\hat{b}_{i+1,k+1}l_i(s),\hat{a}\hat{v}_s\rangle_{\hat{K}}-\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\hat{a}\hat{v}_s\rangle_{\hat{K}}.
\end{aligned}\tag{4.4}$$

Since $M_{k+1}(t)$ vanishes at $(k+1)$ Gauss-Lobatto points, we have

$$\langle M_{k+1}(t)\sum_{i=0}^{k-1}\hat{b}_{i+1,3}l_i(s),\hat{a}\hat{v}_s\rangle_{\hat{K}}=0.$$

For the second term in (4.4),

$$\begin{aligned}
&\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\hat{a}\hat{v}_s\rangle_{\hat{K}}=\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\hat{a}\overline{\hat{v}_s}\rangle_{\hat{K}}+\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\hat{a}(\hat{v}_s-\overline{\hat{v}_s})\rangle_{\hat{K}}\\
=&\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),(\hat{a}-\hat{\Pi}_1\hat{a})\overline{\hat{v}_s}\rangle_{\hat{K}}+\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),(\hat{\Pi}_1\hat{a})\overline{\hat{v}_s}\rangle_{\hat{K}}\\
&+\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),(\hat{a}-\overline{\hat{a}})(\hat{v}_s-\overline{\hat{v}_s})\rangle_{\hat{K}}+\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\overline{\hat{a}}(\hat{v}_s-\overline{\hat{v}}_s)\rangle_{\hat{K}}\\
=&\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),(\hat{a}-\hat{\Pi}_1\hat{a})\overline{\hat{v}}_s\rangle_{\hat{K}}+\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),(\hat{a}-\overline{\hat{a}})(\hat{v}_s-\overline{\hat{v}}_s)\rangle_{\hat{K}},
\end{aligned}$$

where the last step is due to the facts that $(\hat{\Pi}_1\hat{a})\overline{\overline{\hat{v}_s}}$ and $\overline{\hat{a}}(\hat{v}_s-\overline{\hat{v}}_s)$ are polynomials of degree at most $k-1$ with respect to variable s, the $(k+1)$-point Gauss-Lobatto quadrature on s-integration is exact for polynomial of degree $2k-1$, and $l_k(s)$ is orthogonal to polynomials of lower degree. With Lemma 4.3, we have

$$\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\hat{a}\hat{v}_s\rangle_{\hat{K}}\leq C|\hat{u}|_{k+1,2,\hat{K}}(|\hat{a}|_{2,\infty}|\hat{v}|_{1,\hat{K}}+|\hat{a}|_{1,\infty}|\hat{v}|_{2,\hat{K}})=\mathcal{O}(h^{k+2})\|a\|_{2,\infty}\|u\|_{k+1,e}\|v\|_{2,e}.\tag{4.5}$$

Combined with (4.3), we have proved the estimate. ∎

LEMMA 4.6. *Assume $a(x,y)\in W^{2,\infty}(\Omega)$, $u(x,y)\in H^{k+2}(\Omega)$ and $k\geq 2$. Then*

$$\langle a(u-u_p),v_h\rangle_h=\mathcal{O}(h^{k+2})\|a\|_{2,\infty}\|u\|_{k+2}\|v_h\|_2,\quad\forall v_h\in V^h.$$

Proof. As before, we ignore the subscript of v_h for simplicity and

$$\langle a(u-u_p),v\rangle_h=\sum_e\langle a(u-u_p),v\rangle_{e,h}.$$

On each cell e we have

$$\langle a(u-u_p), v\rangle_{e,h} = \langle R[u]_{k,k}, av\rangle_{e,h} = h^2\langle \hat{R}[\hat{u}]_{k,k}, \hat{a}\hat{v}\rangle_{\hat{K}} = h^2\langle \hat{R}[\hat{u}]_{k,k}, \hat{a}\hat{v}-\overline{\hat{a}\hat{v}}\rangle_{\hat{K}} + h^2\langle \hat{R}[\hat{u}]_{k,k}, \overline{\hat{a}\hat{v}}\rangle_{\hat{K}}. \tag{4.6}$$

For the first term in (4.6), due to the embedding $H^2(\hat{K}) \hookrightarrow C^0(\hat{K})$, Bramble-Hilbert Lemma Theorem 3.1 and Lemma 4.3, we have

$$\begin{aligned}
&h^2\langle \hat{R}[\hat{u}]_{k,k}, \hat{a}\hat{v}-\overline{\hat{a}\hat{v}}\rangle_{\hat{K}} \le Ch^2|R[\hat{u}]_{k,k}|_\infty |\hat{a}\hat{v}-\overline{\hat{a}\hat{v}}|_\infty \le Ch^2|\hat{u}|_{k+1,\hat{K}}\|\hat{a}\hat{v}-\overline{\hat{a}\hat{v}}\|_{2,\hat{K}}\\
&\le Ch^2|\hat{u}|_{k+1,\hat{K}}(\|\hat{a}\hat{v}-\overline{\hat{a}\hat{v}}\|_{L^2(\hat{K})} + |\hat{a}\hat{v}|_{1,\hat{K}} + |\hat{a}\hat{v}|_{2,\hat{K}})\\
&\le Ch^2|\hat{u}|_{k+1,\hat{K}}(|\hat{a}\hat{v}|_{1,\hat{K}} + |\hat{a}\hat{v}|_{2,\hat{K}}) = \mathcal{O}(h^{k+2})\|a\|_{2,\infty,e}\|u\|_{k+1,e}\|v\|_{2,e}.
\end{aligned}$$

For the second term in (4.6), we have

$$h^2\langle \hat{R}[\hat{u}]_{k+1,k+1}, \overline{\hat{a}\hat{v}}\rangle_{\hat{K}} = h^2\langle \hat{R}[\hat{u}]_{k+1,k+1}, \overline{\hat{a}\hat{v}}\rangle_{\hat{K}} - h^2\langle \hat{R}[\hat{u}]_{k+1,k+1} - \hat{R}[\hat{u}]_{k,k}, \overline{\hat{a}\hat{v}}\rangle_{\hat{K}}.$$

By Lemma 4.3 and Lemma 4.4 we have

$$h^2\langle \hat{R}[\hat{u}]_{k+1,k+1}, \overline{\hat{a}\hat{v}}\rangle_{\hat{K}} \le Ch^2|\hat{u}|_{k+2,\hat{K}}|\hat{a}\hat{v}|_{0,\hat{K}} = \mathcal{O}(h^{k+2})\|a\|_{0,\infty,e}\|u\|_{k+2,e}\|v\|_{0,e},$$

and

$$h^2\langle \hat{R}[\hat{u}]_{k+1,k+1} - \hat{R}[\hat{u}]_{k,k}, \overline{\hat{a}\hat{v}}\rangle_{\hat{K}} = 0.$$

Thus, we have $\langle a(u-u_p), v_h\rangle_h = \mathcal{O}(h^{k+2})\|a\|_{2,\infty}\|u\|_{k+2}\|v_h\|_2$. ∎

LEMMA 4.7. *Assume $a\in W^{2,\infty}(\Omega)$, $u\in H^{k+3}(\Omega)$ and $k\ge 2$. Then*

$$\langle a(u-u_p)_x, v_h\rangle_h = \mathcal{O}(h^{k+2})\|a\|_{2,\infty}\|u\|_{k+3}\|v_h\|_2, \quad \forall v_h \in V^h.$$

Proof. As before, we ignore the subscript in v_h and we have

$$\langle a(u-u_p)_x, v\rangle_h = \sum_e \langle a(u-u_p)_x, v\rangle_{e,h}.$$

On each cell e, we have

$$\begin{aligned}
\langle a(u-u_p)_x, v\rangle_{e,h} &= \langle (R[u]_{k,k})_x, av\rangle_{e,h} = h\langle (\hat{R}[\hat{u}]_{k,k})_s, \hat{a}\hat{v}\rangle_{\hat{K}}\\
&= h\langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\hat{v}\rangle_{\hat{K}} - h\langle (\hat{R}[\hat{u}]_{k+1,k+1} - \hat{R}[\hat{u}]_{k,k})_s, \hat{a}\hat{v}\rangle_{\hat{K}}.
\end{aligned} \tag{4.7}$$

For the first term in (4.7), we have

$$\langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\hat{v}\rangle_{\hat{K}} \le \langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \overline{\hat{a}\hat{v}}\rangle_{\hat{K}} + \langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\hat{v}-\overline{\hat{a}\hat{v}}\rangle_{\hat{K}}$$

Due to Lemma 4.4,

$$h\langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \overline{\hat{a}\hat{v}}\rangle_{\hat{K}} \le Ch\|a\|_{0,\infty}|u|_{k+3,\hat{K}}\|v\|_{0,\hat{K}} = \mathcal{O}(h^{k+2})\|a\|_{0,\infty}\|u\|_{k+3,e}\|v\|_{0,e},$$

and by the same arguments as in the proof of Lemma 4.6 we have

$$\begin{aligned}
&h\langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\hat{v}-\overline{\hat{a}\hat{v}}\rangle_{\hat{K}} \le Ch|(R[\hat{u}]_{k+1,k+1})_s|_\infty |\hat{a}\hat{v}-\overline{\hat{a}\hat{v}}|_\infty \le Ch|\hat{u}|_{k+2,\hat{K}}\|\hat{a}\hat{v}-\overline{\hat{a}\hat{v}}\|_{2,\hat{K}}\\
&\le Ch|\hat{u}|_{k+2,\hat{K}}(\|\hat{a}\hat{v}-\overline{\hat{a}\hat{v}}\|_{L^2(\hat{K})} + |\hat{a}\hat{v}|_{1,\hat{K}} + |\hat{a}\hat{v}|_{2,\hat{K}}) \le Ch|\hat{u}|_{k+2,\hat{K}}(|\hat{a}\hat{v}|_{1,\hat{K}} + |\hat{a}\hat{v}|_{2,\hat{K}}) = \mathcal{O}(h^{k+2})\|a\|_{2,\infty}\|u\|_{k+2,e}\|v\|_{2,e}.
\end{aligned}$$

Thus

$$h\langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\hat{v}\rangle_{\hat{K}} = \mathcal{O}(h^{k+2})\|a\|_{2,\infty}\|u\|_{k+3,e}\|v\|_{2,e}. \tag{4.8}$$

For the second term in (4.7), we have

$$
\begin{aligned}
&\langle(\hat{R}[\hat{u}]_{k+1,k+1}-\hat{R}[\hat{u}]_{k,k})_s,\hat{a}\hat{v}\rangle_{\hat{K}}\\
=&\langle(M_{k+1}(t)\sum_{i=0}^{k}\hat{b}_{i,k+1}M_i(s)+M_{k+1}(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t))_s,\hat{a}\hat{v}\rangle_{\hat{K}}\\
=&\langle M_{k+1}(t)\sum_{i=0}^{k-1}\hat{b}_{i+1,k+1}l_i(s)+l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\hat{a}\hat{v}\rangle_{\hat{K}}\\
=&\langle M_{k+1}(t)\sum_{i=0}^{k-1}\hat{b}_{i+1,k+1}l_i(s),\hat{a}\hat{v}\rangle_{\hat{K}}+\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\hat{a}\hat{v}\rangle_{\hat{K}}\\
=&\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\hat{a}\hat{v}\rangle_{\hat{K}},
\end{aligned}
$$

where the last step is due to that $M_{k+1}(t)$ vanishes at $(k+1)$ Gauss-Lobatto points. Then

$$
\begin{aligned}
&\langle(\hat{R}[\hat{u}]_{k,k}-\hat{R}[\hat{u}]_{k+1,k+1})_s,\hat{a}\hat{v}\rangle_{\hat{K}}=\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\hat{a}\hat{v}\rangle_{\hat{K}}\\
=&\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\hat{a}\hat{v}-\hat{\Pi}_1(\hat{a}\hat{v})\rangle_{\hat{K}}+\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\hat{\Pi}_1(\hat{a}\hat{v})\rangle_{\hat{K}}\\
=&\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\hat{a}\hat{v}-\hat{\Pi}_1(\hat{a}\hat{v})\rangle_{\hat{K}},
\end{aligned}
$$

where the last step is due to the facts that $\hat{\Pi}_1(\hat{a}\hat{v})$ is a linear function in s thus the $(k+1)$-point Gauss-Lobatto quadrature on s-variable is exact, and $l_k(s)$ is orthogonal to linear functions.

By Lemma 4.3 and Theorem 3.1, we have

$$
\begin{aligned}
&\langle(\hat{R}[\hat{u}]_{k,k}-\hat{R}[\hat{u}]_{k+1,k+1})_s,\hat{a}\hat{v}\rangle_{\hat{K}}=\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t),\hat{a}\hat{v}-\hat{\Pi}_1(\hat{a}\hat{v})\rangle_{\hat{K}}\\
\leq&C|u|_{k+1,\hat{K}}|\hat{a}\hat{v}|_{2,\hat{K}}\leq C|u|_{k+1,\hat{K}}(|\hat{a}|_{2,\infty,\hat{K}}|\hat{v}|_{0,\hat{K}}+|\hat{a}|_{1,\infty,\hat{K}}|\hat{v}|_{1,\hat{K}}+|\hat{a}|_{0,\infty}|\hat{v}|_{2,\hat{K}})
\end{aligned}
$$

Thus

$$
h\langle(\hat{R}[\hat{u}]_{k,k}-\hat{R}[\hat{u}]_{k+1,k+1})_s,\hat{a}\hat{v}\rangle_{\hat{K}}=\mathcal{O}(h^{k+2})\|a\|_{2,\infty}\|u\|_{k+1,e}\|v\|_{2,e}. \tag{4.9}
$$

By (4.8) and (4.9) and sum up over all the cells, we get the desired estimate. ∎

LEMMA 4.8. *Assume $a(x,y)\in W^{4,\infty}(\Omega)$, $u(x,y)\in H^{k+3}(\Omega)$ and $k\geq 2$. Then*

$$
\langle a(u-u_p)_x,(v_h)_y\rangle_h=\begin{cases}\mathcal{O}(h^{k+\frac{3}{2}})\|a\|_{k+2,\infty}\|u\|_{k+3}\|v_h\|_2, & \forall v_h\in V^h, \quad (4.10a)\\ \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty}\|u\|_{k+3}\|v_h\|_2, & \forall v_h\in V_0^h. \quad (4.10b)\end{cases}
$$

Proof. We ignore the subscript in v_h and we have

$$
\langle a(u-u_p)_x,v_y\rangle_h=\sum_e\langle a(u-u_p)_x,v_y\rangle_{e,h},
$$

and on each cell e

$$\begin{aligned}\langle a(u-u_p)_x, v_y\rangle_{e,h} &= \langle (R[u]_{k,k})_x, av_y\rangle_{e,h} = \langle (\hat{R}[\hat{u}]_{k,k})_s, \hat{a}\hat{v}_t\rangle_{\hat{K}} \\ &=\langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\hat{v}_t\rangle_{\hat{K}} + \langle (\hat{R}[\hat{u}]_{k,k} - \hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\hat{v}_t\rangle_{\hat{K}}. \qquad (4.11)\end{aligned}$$

By the same arguments as in the proof of Lemma 4.5, we have

$$\langle (\hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\hat{v}_t\rangle_{\hat{K}} = \mathcal{O}(h^{k+2})\|a\|_{1,\infty}|u|_{k+3,2,e}\|v\|_{2,e}, \qquad (4.12)$$

and

$$\langle (\hat{R}[\hat{u}]_{k,k} - \hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\hat{v}_t\rangle_{\hat{K}} = -\langle l_k(s)\sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t), \hat{a}\hat{v}_t\rangle_{\hat{K}}.$$

For simplicity, we define

$$\hat{b}_{k+1}(t) := \sum_{j=0}^{k+1}\hat{b}_{k+1,j}M_j(t).$$

then by the third and fourth estimates in Lemma 4.3, we have

$$\begin{aligned}|\hat{b}_{k+1}(t)| &\le C\sum_{j=0}^{k+1}|\hat{b}_{k+1,j}| \le C|\hat{u}|_{k+1,\hat{K}}, \\ |\hat{b}^{(m)}_{k+1}(t)| &\le C\sum_{j=m}^{k+1}|\hat{b}_{k+1,j}| \le C|\hat{u}|_{k+2,\hat{K}}, \quad 1\le m,\end{aligned}$$

where $\hat{b}^{(m)}_{k+1}(t)$ is the m-th derivative of $\hat{b}_{k+1}(t)$. We use the same technique in the proof of Theorem 3.7 and we let $l_k = l_k(s)$, $b_{k+1} = b_{k+1}(t)$ in the following,

$$\begin{aligned}&\langle (\hat{R}[\hat{u}]_{k,k} - \hat{R}[\hat{u}]_{k+1,k+1})_s, \hat{a}\hat{v}_t\rangle_{\hat{K}} = -\langle l_k(s)\hat{b}_{k+1}(t), \hat{a}\hat{v}_t\rangle_{\hat{K}} \\ &= -\iint_{\hat{K}} l_k(s)\hat{b}_{k+1}(t)\hat{a}\hat{v}_t d^h s d^h t = -\iint_{\hat{K}} (l_k\hat{b}_{k+1}\hat{a})_I\hat{v}_t d^h s d^h t \\ &= -\iint_{\hat{K}} (l_k\hat{b}_{k+1}\hat{a})_I\hat{v}_t d^h s d^h t + \iint_{\hat{K}} l_k\hat{b}_{k+1}\hat{a}\hat{v}_t dsdt - \iint_{\hat{K}} l_k\hat{b}_{k+1}\hat{a}\hat{v}_t dsdt,\end{aligned}$$

and

$$\begin{aligned}&-\iint_{\hat{K}} (l_k\hat{b}_{k+1}\hat{a})_I\hat{v}_t d^h s d^h t + \iint_{\hat{K}} l_k\hat{b}_{k+1}\hat{a}\hat{v}_t dsdt \\ &= \iint_{\hat{K}} \left[l_k\hat{b}_{k+1}\hat{a} - (l_k\hat{b}_{k+1}\hat{a})_I\right]\hat{v}_t dsdt + \iint_{\hat{K}} (l_k\hat{b}_{k+1}\hat{a})_I\hat{v}_t dsdt - \iint_{\hat{K}} (l_k\hat{b}_{k+1}\hat{a})_I\hat{v}_t d^h sdt \\ &= \iint_{\hat{K}} \left[l_k\hat{b}_{k+1}\hat{a} - (l_k\hat{b}_{k+1}\hat{a})_I\right]\hat{v}_t dsdt + \iint_{\hat{K}} \partial_t(l_k\hat{b}_{k+1}\hat{a})_I\hat{v} d^h sdt - \iint_{\hat{K}} \partial_t(l_k\hat{b}_{k+1}\hat{a})_I\hat{v} dsdt \\ &+ \left(\int_{-1}^{1}(l_k\hat{b}_{k+1}\hat{a})_I\hat{v}ds\Big|_{t=-1}^{t=1} - \int_{-1}^{1}(l_k\hat{b}_{k+1}\hat{a})_I\hat{v}d^h s\Big|_{t=-1}^{t=1}\right) = I + II + III.\end{aligned}$$

∎

After integration by parts with respect to the variable s, we have

$$\iint_{\hat{K}} l_k(s)\hat{b}_{k+1}(t)\hat{a}\hat{v}_t dsdt = -\iint_{\hat{K}} M_{k+1}(s)\hat{b}_{k+1}(t)(\hat{a}_s\hat{v}_t + \hat{a}\hat{v}_{st})dsdt,$$

which is exactly the same integral estimated in the proof of Lemma 3.7 in [13]. By the same proof of Lemma 3.7 in [13], after summing over all elements, we have the estimate for the term $\iint_{\hat K} l_k(s)\hat b_{k+1}(t)\hat a\hat v_t dsdt$:

$$\sum_e \iint_{\hat K} l_k(s)\hat b_{k+1}(t)\hat a\hat v_t dsdt = \begin{cases}\mathcal{O}(h^{k+\frac{3}{2}})\|a\|_{k+2,\infty}\|u\|_{k+3}\|v\|_2, & \forall v\in V^h,\\ \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty}\|u\|_{k+3}\|v\|_2, & \forall v\in V_0^h.\end{cases}$$

Then we can do similar estimation as in Theorem 3.7 for I, II, III separately.

For term I, by Theorem 3.1 and the estimate (3.2), we have

$$\begin{aligned}
&\iint_{\hat K}\left[l_k\hat b_{k+1}\hat a-(l_k\hat b_{k+1}\hat a)_I\right]\hat v_t dsdt\\
=&\iint_{\hat K}\left[l_k\hat b_{k+1}\hat a-(l_k\hat b_{k+1}\hat a)_I\right]\overline{\hat v_t}dsdt+\iint_{\hat K}\left[l_k\hat b_{k+1}\hat a-(l_k\hat b_{k+1}\hat a)_I\right](\hat v_t-\overline{\hat v_t})dsdt\\
\le& C\left[l_k\hat b_{k+1}\hat a\right]_{k+2,\hat K}|\hat v|_{1,\hat K}+C\left[l_k\hat b_{k+1}\hat a\right]_{k+1,\hat K}|\hat v|_{2,\hat K}\\
\le& C\left(\sum_{m=2}^{k+2}|\hat a|_{m,\infty,\hat K}\max_{t\in[-1,1]}|\hat b_{k+1}(t)|\right)|\hat v|_{1,\hat K}+C\left(\sum_{m=0}^{k+2}|\hat a|_{m,\infty,\hat K}\max_{t\in[-1,1]}|\hat b_{k+1}^{(k+2-m)}(t)|\right)|\hat v|_{1,\hat K}\\
&+C\left(\sum_{m=1}^{k+1}|\hat a|_{m,\infty,\hat K}\max_{t\in[-1,1]}|\hat b_{k+1}(t)|\right)|\hat v|_{2,\hat K}+C\left(\sum_{m=0}^{k+1}|\hat a|_{m,\infty,\hat K}\max_{t\in[-1,1]}|\hat b_{k+1}^{(k+1-m)}(t)|\right)|\hat v|_{2,\hat K}\\
=&\mathcal{O}(h^{k+2})\|a\|_{k+2,\infty}\|u\|_{k+2,e}\|v\|_{2,e}.
\end{aligned}$$

For term II, as in the proof of Theorem 3.7, we define the linear form as

$$\hat E_{\hat v}(\hat f)=\iint_{\hat K}(\hat F_I)_t\hat v dsdt-\iint_{\hat K}(\hat F_I)_t\hat v d^h sdt,$$

for each $\hat v\in Q^k(\hat K)$ and $\hat F$ is an antiderivative of $\hat f$ w.r.t. variable t. We can easily see that $\hat E_{\hat v}$ is well defined and $\hat E_{\hat v}$ is a continuous linear form on $H^k(\hat K)$. With projection $\hat\Pi_1$ defined in (2.2), we have

$$\hat E_{\hat v}(\hat f)=\hat E_{\hat v-\hat\Pi_1\hat v}(\hat f)+\hat E_{\hat\Pi_1\hat v}(\hat f),\quad \forall \hat v\in Q^k(\hat K).$$

Since $Q^{k-1}(\hat K)\subset\ker\hat E_{\hat v-\hat\Pi_1\hat v}$ thus

$$\hat E_{\hat v-\hat\Pi_1\hat v}(\hat f)\le C[f]_{k,\hat K}\|\hat v-\hat\Pi_1\hat v\|_{0,\hat K}\le C[f]_{k,\hat K}|\hat v|_{2,\hat K}$$

and

$$\hat E_{\hat\Pi_1\hat v}(\hat f)=\iint_{\hat K}(\hat F_I)_t\hat\Pi_1\hat v dsdt-\iint_{\hat K}(\hat F_I)_t\hat\Pi_1\hat v d^h sdt=0.$$

Thus we have

$$\begin{aligned}
&\iint_{\hat K}\partial_t(l_k\hat b_{k+1}\hat a)_I\hat v d^h sdt-\iint_{\hat K}\partial_t(l_k\hat b_{k+1}\hat a)_I\hat v dsdt=-\hat E_{\hat v}((l_k\hat b_{k+1}\hat a)_t)\\
=&-\hat E_{\hat v-\Pi_1\hat v}((l_k\hat b_{k+1}\hat a)_t)\le C[(l_k\hat b_{k+1}\hat a)_t]_{k,\hat K}|\hat v_h|_{2,\hat K}=\mathcal{O}(h^{k+2})\|a\|_{k+1,\infty,e}\|u\|_{k+2,e}|v|_{2,e}.
\end{aligned}$$

Now we only need to discuss term III. Let L_1 and L_3 denote the top and bottom boundaries of Ω and let l_1^e, l_3^e denote the top and bottom edges of element e (and $l_1^{\hat K}$

and $l_3^{\hat{K}}$ for $\hat{K}$). Notice that after mapping back to the cell e we have

$$b_{k+1}(y_e+h) = \hat{b}_{k+1}(1) = \sum_{j=0}^{k+1} \hat{b}_{k+1,j} M_j(1) = \hat{b}_{k+1,0} + \hat{b}_{k+1,1}$$

$$= (k+\frac{1}{2})\int_{-1}^{1} \partial_s \hat{u}(s,1) l_k(s) ds = (k+\frac{1}{2})\int_{x_e-h}^{x_e+h} \partial_x u(x, y_e+h) l_k(\frac{x-x_e}{h}) dx,$$

and similarly we get $b_{k+1}(y_e-h) = \hat{b}_{k+1}(-1) = (k+\frac{1}{2})\int_{x_e-h}^{x_e+h} \partial_x u(x,y_e-h) l_k(\frac{x-x_e}{h})dx$. Thus the term $l(\frac{x-x_e}{h})b_{k+1}(y)av$ is continuous across the top and bottom edges of cells. Therefore, if summing over all elements e, the line integral on the inner edges are cancelled out. So after summing over all elements, the line integral reduces to two line integrals along L_1 and L_3. We only need to discuss one of them. For a cell e adjacent to L_1, consider its reference cell $\hat{K}$ and define linear form $\hat{E}(\hat{f}) = \int_{-1}^{1} \hat{f}(s,1)ds - \int_{-1}^{1} \hat{f}(s,1)d^h s$, then we have

$$\hat{E}(\hat{f}\hat{v}) \le C|\hat{f}|_{0,\infty,l_1^{\hat{K}}} |\hat{v}|_{0,\infty,l_1^{\hat{K}}} \le C\|\hat{f}\|_{2,l_1^{\hat{K}}} \|\hat{v}\|_{0,l_1^{\hat{K}}},$$

thus the mapping $\hat{f} \to \hat{E}(\hat{f}\hat{v})$ is continuous with operator norm less than $C\|\hat{v}\|_{0,l_1^{\hat{K}}}$ for some C. Since $\hat{E}((\hat{a}\hat{u}_s)_I \hat{\Pi}_1 \hat{v}) = 0$ we have

$$\sum_{e\cap L_1 \neq \emptyset} \int_{-1}^{1} (l_k \hat{b}_{k+1}\hat{a})_I \hat{v} ds - \int_{-1}^{1} (l_k \hat{b}_{k+1}\hat{a})_I \hat{v} d^h s$$

$$= \sum_{e\cap L_1 \neq \emptyset} \hat{E}((l_k \hat{b}_{k+1}\hat{a})_I \hat{v}) = \sum_{e\cap L_1 \neq \emptyset} \hat{E}((l_k \hat{b}_{k+1}\hat{a})_I (\hat{v} - \hat{\Pi}_1 \hat{v})) \le \sum_{e\cap L_1 \neq \emptyset} C[(l_k \hat{b}_{k+1}\hat{a})_I]_{k,l_1^{\hat{K}}} [\hat{v}]_{2,l_1^{\hat{K}}}$$

$$\le \sum_{e\cap L_1 \neq \emptyset} C(|l_k \hat{b}_{k+1}\hat{a} - (l_k \hat{b}_{k+1}\hat{a})_I|_{k,l_1^{\hat{K}}} + |l_k \hat{b}_{k+1}\hat{a}|_{k,l_1^{\hat{K}}})[\hat{v}]_{2,l_1^{\hat{K}}}$$

$$\le \sum_{e\cap L_1 \neq \emptyset} (|l_k \hat{b}_{k+1}\hat{a}|_{k+1,l_1^{\hat{K}}} + |l_k \hat{b}_{k+1}\hat{a}|_{k,l_1^{\hat{K}}})[\hat{v}]_{2,l_1^{\hat{K}}} \le \sum_{e\cap L_1 \neq \emptyset} C\|\hat{a}\|_{k,\infty,\hat{K}} |\hat{b}_{k+1}(1)| [\hat{v}]_{2,l_1^{\hat{K}}}, \qquad \blacksquare$$

where the first inequality is derived from $\hat{E}(\hat{f}(\hat{v} - \hat{\Pi}_1 \hat{v})) = 0, \forall \hat{f} \in Q^{k-1}(\hat{K})$ and Theorem 3.1.

Since $l_k(t) = \frac{1}{2^k k!}\frac{d^k}{dt^k}(t^2-1)^k$, after integration by parts k times,

$$\hat{b}_{k+1}(1) = (k+\frac{1}{2})\int_{-1}^{1} \partial_s u(s,1) l_k(s) dx = (-1)^k (k+\frac{1}{2})\int_{-1}^{1} \partial_s^{k+1} u(s,1) L(s) ds,$$

where $L(s)$ is a polynomial of degree $2k$ by taking antiderivatives of $l_k(s)$ k times. Then by Cauchy-Schwarz inequality we have

$$\hat{b}_{k+1}(1) \le C\left(\int_{-1}^{1} |\partial_s^{k+1}\hat{u}(s,1)|^2 ds\right)^{\frac{1}{2}} \le Ch^{k+\frac{1}{2}} |u|_{k+1,l_1^e}.$$

By (3.13), we get $|\hat{v}|_{2,l_1^{\hat{K}}} = h^{\frac{3}{2}} |\hat{v}|_{2,l_1^e} \le Ch|v|_{2,e}$. Thus we have

$$\sum_{e\cap L_1 \neq \emptyset} \int_{-1}^{1} (l_k \hat{b}_{k+1}\hat{a})_I \hat{v} ds - \int_{-1}^{1} (l_k \hat{b}_{k+1}\hat{a})_I \hat{v} d^h s \le \sum_{e\cap L_1 \neq \emptyset} C\|\hat{a}\|_{k,\infty,\hat{K}} |\hat{b}_{k+1}(1)| |\hat{v}|_{2,l_1^{\hat{K}}}$$

$$= \mathcal{O}(h^{k+\frac{3}{2}}) \sum_{e\cap L_1 \neq \emptyset} \|a\|_{k,\infty} |u|_{k+1,l_1^e} |v|_{2,e} = \mathcal{O}(h^{k+\frac{3}{2}}) \|a\|_{k,\infty} |u|_{k+1,L_1} \|v\|_{2,\Omega} = \mathcal{O}(h^{k+\frac{3}{2}}) \|a\|_{k,\infty} \|u\|_{k+2,\Omega} \|v\|_{2,\Omega}, \qquad \blacksquare$$

where the trace inequality $\|u\|_{k+1,\partial\Omega} \le C\|u\|_{k+2,\Omega}$ is used.

Combine all the estimates above, we get (4.10a). Since the $\frac{1}{2}$ order loss is only due to the line integral along L_1 and L_3, on which $v_{xx} = 0$ if $v \in V_0^h$, we get (4.10b).□

By all the discussions in this subsection, we have proven (4.1a) and (4.1b).

5. Homogeneous Dirichlet Boundary Conditions.

5.1. V^h-ellipticity. In order to discuss the scheme (1.2), we need to show A_h satisfies V^h-ellipticity

$$\forall v_h \in V_0^h, \quad C\|v_h\|_1^2 \le A_h(v_h, v_h). \tag{5.1}$$

We first consider the V_h-ellipticity for the case $\mathbf{b} \equiv 0$.

LEMMA 5.1. *Assume the coefficients in* (2.3) *satisfy that* $\mathbf{b} \equiv 0$*, both* $c(x, y)$ *and the eigenvalues of* $\mathbf{a}(x, y)$ *have a uniform upper bound and a uniform positive lower bound, then there exist two constants* $C_1, C_2 > 0$ *independent of mesh size* h *such that*

$$\forall v_h \in V_0^h, \quad C_1\|v_h\|_1^2 \le A_h(v_h, v_h) \le C_2\|v_h\|_1^2.$$

Proof. Let $Z_{0,\hat{K}}$ denote the set of $(k+1) \times (k+1)$ Gauss-Lobatto points on the reference cell $\hat{K}$. First we notice that the set $Z_{0,\hat{K}}$ is a $Q^k(\hat{K})$-unisolvent subset. Since the Gauss-Lobatto quadrature weights are strictly positive, we have

$$\forall \hat{p} \in Q^k(\hat{K}), \sum_{i=1}^{2} \langle \partial_i \hat{p}, \partial_i \hat{p} \rangle_{\hat{K}} = 0 \Longrightarrow \partial_i \hat{p} = 0 \text{ at quadrature points},$$

where $i = 1, 2$ represents the spatial derivative on variable x_i respectively. Since $\partial_i \hat{p} \in Q^k(\hat{K})$ and it vanishes on a $Q^k(\hat{K})$-unisolvent subset, we have $\partial_i \hat{p} \equiv 0$. As a consequence, $\sqrt{\sum_{i=1}^n \langle \partial_i \hat{p}, \partial_i \hat{p} \rangle_h}$ defines a norm over the quotient space $Q^k(\hat{K})/Q^0(\hat{K})$. Since that $|\cdot|_{1,\hat{K}}$ is also a norm over the same quotient space, by the equivalence of norms over a finite dimensional space, we have

$$\forall \hat{p} \in Q^k(\hat{K}), \quad C_1|\hat{p}|_{1,\hat{K}}^2 \le \sum_{i=1}^n \langle \partial_i \hat{p}, \partial_i \hat{p} \rangle_{\hat{K}} \le C_2|\hat{p}|_{1,\hat{K}}^2.$$

On the reference cell $\hat{K}$, by the assumption on the coefficients, we have

$$C_1|\hat{v}_h|_{1,\hat{K}}^2 \le C_1 \sum_{i}^{n} \langle \partial_i \hat{v}_h, \partial_i \hat{v}_h \rangle_{\hat{K}} \le \sum_{i,j=1}^{n} \left(\langle \hat{a}_{ij} \partial_i \hat{v}_h, \partial_j \hat{v}_h \rangle_{\hat{K}} + \langle \hat{c} \hat{v}_h, \hat{v}_h \rangle_{\hat{K}} \right) \le C_2 \|\hat{v}_h\|_{1,\hat{K}}^2$$

Mapping these back to the original cell e and summing over all elements, by the equivalence of two norms $|\cdot|_1$ and $\|\cdot\|_1$ for the space $H_0^1(\Omega) \supset V_0^h$ [5], we get $C_1\|v_h\|_1^2 \le A_h(v_h, v_h) \le C_2\|v_h\|_1^2$. □

For discussing V_h-ellipticity when $\mathbf{b}$ is nonzero, by Young's inequality we have

$$|\langle \mathbf{b} \cdot \nabla v_h, v_h \rangle_h| \le \sum_e \iint_e \frac{(\mathbf{b} \cdot \nabla v_h)^2}{4c} + c|v_h|^2 d^h x d^h y \le \langle \frac{|\mathbf{b}|^2}{4c} \nabla v_h, \nabla v_h \rangle_h + \langle c v_h, v_h \rangle_h.$$

Thus we have

$$\langle \mathbf{a} \nabla v_h, \nabla v_h \rangle_h + \langle \mathbf{b} \cdot \nabla v_h, v_h \rangle_h + \langle c v_h, v_h \rangle_h \ge \langle \lambda_{\mathbf{a}} \nabla v_h, \nabla v_h \rangle_h - \langle \frac{|\mathbf{b}|^2}{4c} \nabla v_h, \nabla v_h \rangle_h,$$

where $\lambda_{\mathbf{a}}$ is smallest eigenvalue of $\mathbf{a}$. Then we have the following Lemma

LEMMA 5.2. *Assume $4\lambda_{\mathbf{a}} c > |\mathbf{b}|^2$, then there exists a constant $C > 0$ independent of mesh size h such that*

$$\forall v_h \in V_0^h, \quad A_h(v_h, v_h) \geq C\|v_h\|_1^2.$$

5.2. Standard estimates for the dual problem. In order to apply the Aubin-Nitsche duality argument for establishing superconvergence of function values, we need certain estimates on a proper dual problem. Define $\theta_h := u_h - u_p$. Then we consider the dual problem: find $w \in H_0^1(\Omega)$ satisfying

$$A^*(w, v) = (\theta_h, v), \quad \forall v \in H_0^1(\Omega), \tag{5.2}$$

where $A^*(\cdot,\cdot)$ is the adjoint bilinear form of $A(\cdot,\cdot)$ such that

$$A^*(u, v) = A(v, u) = (\mathbf{a}\nabla v, \nabla u) + (\mathbf{b} \cdot \nabla v, u) + (cv, u).$$

Let $w_h \in V_0^h$ be the solution to

$$A_h^*(w_h, v_h) = (\theta_h, v_h), \quad \forall v_h \in V_0^h. \tag{5.3}$$

Notice that the right hand side of (5.3) is different from the right hand side of the scheme (1.2).

We need the following standard estimates on w_h for the dual problem.

THEOREM 5.3. *Assume all coefficients in (2.3) are in $W^{2,\infty}(\Omega)$. Let w be defined in (5.2), w_h be defined in (5.3), and $\theta_h = u_h - u_p$. Assume elliptic regularity (2.6) and V^h ellipticity holds, we have*

$$\|w - w_h\|_1 \leq Ch\|w\|_2,$$

$$\|w_h\|_2 \leq C\|\theta_h\|_0.$$

Proof. By V^h ellipticity, we have $C_1\|w_h - v_h\|_1^2 \leq A_h^*(w_h - v_h, w_h - v_h)$. By the definition of the dual problem, we have

$$A_h^*(w_h, w_h - v_h) = (\theta_h, w_h - v_h) = A^*(w, w_h - v_h), \quad \forall v_h \in V_0^h.$$

Thus for any $v_h \in V_0^h$, by Theorem 3.6, we have

$$\begin{aligned} &C_1\|w_h - v_h\|_1^2 \leq A_h^*(w_h - v_h, w_h - v_h) \\ =&A^*(w - v_h, w_h - v_h) + [A_h^*(w_h, w_h - v_h) - A^*(w, w_h - v_h)] + [A^*(v_h, w_h - v_h) - A_h^*(v_h, w_h - v_h)] \\ =&A^*(w - v_h, w_h - v_h) + [A(w_h - v_h, v_h) - A_h(w_h - v_h, v_h)] \\ \leq&C\|w - v_h\|_1\|w_h - v_h\|_1 + Ch\|v_h\|_2\|w_h - v_h\|_1. \end{aligned}$$

∎

Thus

$$\|w - w_h\|_1 \leq \|w - v_h\|_1 + \|w_h - v_h\|_1 \leq C\|w - v_h\|_1 + Ch\|v_h\|_2. \tag{5.4}$$

Now consider $\Pi_1 w \in V_0^h$ where Π_1 is the piecewise Q^1 projection and its definition on each cell is defined through (2.2) on the reference cell. By the Bramble Hilbert Lemma Theorem 3.1 on the projection error, we have

$$\|w - \Pi_1 w\|_1 \leq Ch\|w\|_2, \quad \|w - \Pi_1 w\|_2 \leq C\|w\|_2, \tag{5.5}$$

thus $\|\Pi_1 w\|_2 \leq \|w\|_2 + \|w - \Pi_1 w\|_2 \leq C\|w\|_2$. By setting $v_h = \Pi_1 w$, from (5.4) we have

$$\|w - w_h\|_1 \leq C\|w - \Pi_1 w\|_1 + Ch\|\Pi_1 w\|_2 \leq Ch\|w\|_2. \tag{5.6}$$

By the inverse estimate on the piecewise polynomial $w_h - \Pi_1 w$, we get

$$\|w_h\|_2 \leq \|w_h - \Pi_1 w\|_2 + \|\Pi_1 w - w\|_2 + \|w\|_2 \leq Ch^{-1}\|w_h - \Pi_1 w\|_1 + C\|w\|_2. \tag{5.7}$$

By (5.5) and (5.6), we also have

$$\|w_h - \Pi_1 w\|_1 \leq \|w - \Pi_1 w\|_1 + \|w - w_h\|_1 \leq Ch\|w\|_2. \tag{5.8}$$

With (5.7), (5.8) and the elliptic regularity $\|w\|_2 \leq C\|\theta_h\|_0$, we get □

$$\|w_h\|_2 \leq C\|w\|_2 \leq C\|\theta_h\|_0.$$

5.3. Superconvergence of function values.

THEOREM 5.4. *Assume $a_{ij}, b_i, c \in W^{k+2,\infty}(\Omega)$ and $u(x,y) \in H^{k+3}(\Omega)$, $f(x,y) \in H^{k+2}(\Omega)$ with $k \geq 2$. Assume elliptic regularity (2.6) and V^h ellipticity holds. Then u_h, the numerical solution from scheme (1.2), is a $(k+2)$-th order accurate approximation to the exact solution u in the discrete 2-norm over all the $(k+1) \times (k+1)$ Gauss-Lobatto points:*

$$\|u_h - u\|_{2,Z_0} = \mathcal{O}(h^{k+2})(\|u\|_{k+3,\Omega} + \|f\|_{k+2,\Omega}).$$

Proof. By Theorem 3.7 and Theorem 3.3, for any $v_h \in V_0^h$,

$$\begin{aligned} & A_h(u - u_h, v_h) = [A(u, v_h) - A_h(u_h, v_h)] + [A_h(u, v_h) - A(u, v_h)] \\ = \; & A(u, v_h) - A_h(u_h, v_h) + \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty}\|u\|_{k+3}\|v_h\|_2 \\ = \; & [(f, v_h) - \langle f, v_h\rangle_h] + \mathcal{O}(h^{k+2})\|u\|_{k+3}\|v_h\|_2 = \mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2})\|v_h\|_2. \end{aligned}$$

Let $\theta_h = u_h - u_p$, then $\theta_h \in V_0^h$ due to the properties of the M-type projection. So by (4.1a) and Theorem 5.3, we get

$$\begin{aligned} \|\theta_h\|_0^2 = (\theta_h, \theta_h) = A_h(\theta_h, w_h) = A_h(u_h - u, w_h) + A_h(u - u_p, w_h) \\ = A_h(u - u_p, w_h) + \mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2})\|w_h\|_2 \\ = \mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2})\|w_h\|_2 = \mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2})\|\theta_h\|_0, \end{aligned}$$

thus

$$\|u_h - u_p\|_0 = \|\theta_h\|_0 = \mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2}).$$

Finally, by the equivalence of the discrete 2-norm on Z_0 and the $L^2(\Omega)$ norm in finite-dimensional space V^h and Theorem 4.2, we obtain

$$\begin{aligned} \|u_h - u\|_{2,Z_0} \leq \|u_h - u_p\|_{2,Z_0} + \|u_p - u\|_{2,Z_0} \leq C\|u_h - u_p\|_0 + \|u_p - u\|_{2,Z_0} \\ = \mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2}). \end{aligned}$$ □

REMARK 5.5. *To extend the discussions to Neumann type boundary conditions, due to (4.1b) and Theorem 3.7, one can only prove $(k+\frac{3}{2})$-th order accuracy:*

$$\|u_h - u\|_{2,Z_0} = \mathcal{O}(h^{k+\frac{3}{2}})(\|u\|_{k+3} + \|f\|_{k+2}).$$

On the other hand, for solving a general elliptic equation, only $\mathcal{O}(h^{k+\frac{3}{2}})$ superconvergence at all Lobatto point can be proven for Neumann boundary conditions even for the full finite element scheme (1.1), see [4].

REMARK 5.6. *All key discussions can be extended to three-dimensional cases. For instance, M-type expansion has been used for discussing superconvergence for the three-dimensional case [4]. The most useful technique in Section 3.2 to obtain desired consistency error estimate is to derive error cancellations between neighboring cells through integration by parts on suitable interpolation polynomials, which still seems possible on rectangular meshes in three dimensions.*

6. Nonhomogeneous Dirichlet Boundary Conditions.

We consider a two-dimensional elliptic problem on $\Omega = (0,1)^2$ with nonhomogeneous Dirichlet boundary condition,

$$\begin{aligned} -\nabla\cdot(\mathbf{a}\nabla u) + \mathbf{b}\cdot\nabla u + cu &= f \text{ on } \Omega \\ u &= g \text{ on } \partial\Omega. \end{aligned} \tag{6.1}$$

Assume there is a function $\bar{g} \in H^1(\Omega)$ as a smooth extension of g so that $\bar{g}|_{\partial\Omega} = g$. The variational form is to find $\tilde{u} = u - \bar{g} \in H_0^1(\Omega)$ satisfying

$$A(\tilde{u}, v) = (f, v) - A(\bar{g}, v), \quad \forall v \in H_0^1(\Omega). \tag{6.2}$$

In practice, $\bar{g}$ is not used explicitly. By abusing notations, the most convenient implementation is to consider

$$g(x,y) = \begin{cases} 0, & \text{if} \quad (x,y) \in (0,1)\times(0,1), \\ g(x,y), & \text{if} \quad (x,y) \in \partial\Omega, \end{cases}$$

and $g_I \in V^h$ which is defined as the Q^k Lagrange interpolation at $(k+1)\times(k+1)$ Gauss-Lobatto points for each cell on Ω of $g(x,y)$. Namely, $g_I \in V^h$ is the piecewise P^k interpolation of g along the boundary grid points and $g_I = 0$ at the interior grid points. The numerical scheme is to find $\tilde{u}_h \in V_0^h$, s.t.

$$A_h(\tilde{u}_h, v_h) = \langle f, v_h\rangle_h - A_h(g_I, v_h), \quad \forall v_h \in V_0^h. \tag{6.3}$$

Then $u_h = \tilde{u}_h + g_I$ will be our numerical solution for (6.1). Notice that (6.3) is not a straightforward approximation to (6.2) since $\bar{g}$ is never used. Assuming elliptic regularity and V^h ellipticity hold, we will show that $u_h - u$ is of $(k+2)$-th order in the discrete 2-norm over all $(k+1)\times(k+1)$ Gauss-Lobatto points.

6.1. An auxiliary scheme.

In order to discuss the superconvergence of (6.3), we need to prove the superconvergence of an auxiliary scheme. Notice that we discuss the auxiliary scheme only for proving the accuracy of (6.3). In practice one should not implement the auxiliary scheme since (6.3) is a much more convenient implementation with the same accuracy.

Let $\bar{g}_p \in V^h$ be the piecewise M-type Q^k projection of the smooth extension function $\bar{g}$, and define $g_p \in V^h$ as $g_p = \bar{g}_p$ on $\partial\Omega$ and $g_p = 0$ at all the inner grids. The auxiliary scheme is to find $\tilde{u}_h^* \in V_0^h$ satisfying

$$A_h(\tilde{u}_h^*, v_h) = \langle f, v_h\rangle_h - A_h(g_p, v_h), \quad \forall v_h \in V_0^h, \tag{6.4}$$

Then $u_h^* = \tilde{u}_h^* + g_p$ is the numerical solution for problem (6.2). Define $\theta_h = u_h^* - u_p$, then by Theorem 4.1 we have $\theta_h \in V_0^h$. Following Section 5.2, define the following dual problem: find $w \in H_0^1(\Omega)$ satisfying

$$A^*(w, v) = (\theta_h, v), \quad \forall v \in H_0^1(\Omega). \tag{6.5}$$

Let $w_h \in V_0^h$ be the solution to

$$A_h^*(w_h, v_h) = (\theta_h, v_h), \quad \forall v_h \in V_0^h. \tag{6.6}$$

Notice that the dual problem has homogeneous Dirichlet boundary conditions. By Theorem 3.7, Theorem 3.3, for any $v_h \in V_0^h$,

$$\begin{aligned} & A_h(u - u_h^*, v_h) = [A(u, v_h) - A_h(u_h^*, v_h)] + [A_h(u, v_h) - A(u, v_h)] \\ = \; & A(u, v_h) - A_h(u_h^*, v_h) + \mathcal{O}(h^{k+2})\|a\|_{k+2,\infty}\|u\|_{k+3}\|v_h\|_2 \\ = \; & [(f, v_h) - \langle f, v_h\rangle_h] + \mathcal{O}(h^{k+2})\|u\|_{k+3}\|v_h\|_2 = \mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2})\|v_h\|_2. \end{aligned}$$

By (4.1a) and Theorem 5.3, we get

$$\begin{aligned} & \|\theta_h\|_0^2 = (\theta_h, \theta_h) = A_h(\theta_h, w_h) = A_h(u_h^* - u, w_h) + A_h(u - u_p, w_h) \\ = & A_h(u - u_p, w_h) + \mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2})\|w_h\|_2 \\ = & \mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2})\|w_h\|_2 = \mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2})\|\theta_h\|_0, \end{aligned}$$

thus $\|u_h^* - u_p\|_0 = \|\theta_h\|_0 = \mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2})$. So Theorem 5.4 still holds for the auxiliary scheme (6.4):

$$\|u_h^* - u\|_{2,Z_0} = \mathcal{O}(h^{k+2})(\|u\|_{k+3} + \|f\|_{k+2}). \tag{6.7}$$

6.2. The main result. In order to extend Theorem 5.4 to (6.3), we only need to prove

$$\|u_h - u_h^*\|_0 = \mathcal{O}(h^{k+2}).$$

The difference between (6.4) and (6.3) is

$$A_h(\tilde{u}_h^* - \tilde{u}_h, v_h) = A_h(g_I - g_p, v_h), \quad \forall v_h \in V_0^h. \tag{6.8}$$

We need the following Lemma.

LEMMA 6.1. *Assuming $u \in H^{k+4}(\Omega)$ for $k \geq 2$, with g_I and g_p being defined as in this Section, then we have*

$$A_h(g_I - g_p, v_h) = \mathcal{O}(h^{k+2})\|u\|_{k+4,\Omega}\|v_h\|_{2,\Omega}, \quad \forall v_h \in V_0^h. \tag{6.9}$$

Proof. For simplicity, we ignore the subscript $_h$ of v_h in this proof and all the following v are in V^h.

Notice that $g_I - g_p \equiv 0$ in interior cells. Thus we only consider cells adjacent to $\partial\Omega$. Let L_1, L_2, L_3 and L_4 denote the top, left, bottom and right boundary edges of $\bar{\Omega} = [0, 1] \times [0, 1]$ respectively. Without loss of generality, we consider cell $e = [x_e - h, x_e + h] \times [y_e - h, y_e + h]$ adjacent to the left boundary L_2, i.e., $x_e - h = 0$. Let l_1^e, l_2^e, l_3^e and l_4^e denote the top, left, bottom and right boundary edges of e respectively.

On $l_2 \subset L_2$, Let $\phi_{ij}(x, y), i, j = 0, 1, \dots, k$, be Lagrange basis functions on edge l_2^e for the $(k + 1) \times (k + 1)$ Gauss-Lobatto points in cell e. Then $g_I - g_p =$

$\sum_{i,j=0}^{k} \lambda_{ij}\phi_{ij}(x,y)$ and $|\lambda_{ij}| \leq \|g_I - g_p\|_{\infty,Z_0}$. Due to Sobolev's embedding, we have $u \in W^{k+2,\infty}(\Omega)$. By Theorem 4.2, we have

$$\|g_I - g_p\|_{\infty,Z_0} \leq \|u - u_p\|_{\infty,Z_0} = \mathcal{O}(h^{k+2})\|u\|_{k+2,\infty,\Omega} = \mathcal{O}(h^{k+2})\|u\|_{k+4,\Omega}.$$

Thus we get $\forall v \in V_0^h$,

$$\langle a(g_I - g_p)_x, v_x\rangle_e = \langle a \sum_{i,j=0}^{k} \lambda_{ij}\phi_{ij}(x,y)_x, v_x\rangle_e \leq C\|a\|_{\infty,\Omega} \max_{i,j} |\lambda_{ij}||\langle \sum_{i,j=0}^{k} \phi_{ij}(x,y)_x, v_x\rangle_e|.$$

Since for polynomials on $\hat{K}$ all the norm are equivalent, we have

$$|\langle \sum_{i,j=0}^{k} \phi_{ij}(x,y)_x, v_x\rangle_e| = |\langle \sum_{i,j=0}^{k} \hat{\phi}_{ij}(s,t)_s, \hat{v}_s\rangle_{\hat{K}}| \leq C|\hat{v}_s|_{\infty,\hat{K}} \leq C|v|_{1,\hat{K}} = C|v|_{1,e},$$

which implies

$$\langle a(g_I - g_p)_x, v_x\rangle_h \leq C\|a\|_{\infty,\Omega} \sum_e \max_{i,j} |\lambda_{ij}||v|_{1,e} = \mathcal{O}(h^{k+2})\|a\|_{\infty,\Omega}\|u\|_{k+4,\Omega}\|v\|_{2,\Omega}$$

Similarly, for any $v \in V_0^h$, we have

$$\begin{aligned}
\langle a(g_I - g_p)_y, v_y\rangle_h =& \mathcal{O}(h^{k+2})\|a\|_\infty\|u\|_{k+4}\|v\|_2,\\
\langle a(g_I - g_p)_x, v_y\rangle_h =& \mathcal{O}(h^{k+2})\|a\|_\infty\|u\|_{k+4}\|v\|_2,\\
\langle \mathbf{b}\cdot\nabla(g_I - g_p), v\rangle_h =& \mathcal{O}(h^{k+2})\|\mathbf{b}\|_\infty\|u\|_{k+4}\|v\|_2,\\
\langle c(g_I - g_p), v\rangle_h =& \mathcal{O}(h^{k+2})\|c\|_\infty\|u\|_{k+4}\|v\|_2.
\end{aligned}$$

Thus we conclude that

$$A_h(g_I - g_p, v_h) = \mathcal{O}(h^{k+2})\|u\|_{k+4}\|v_h\|_2, \quad \forall v_h \in V_0^h.$$

∎

By (6.8) and Lemma 6.1, we have

$$A_h(\tilde{u}_h^* - \tilde{u}_h, v_h) = \mathcal{O}(h^{k+2})\|u\|_{k+4}\|v_h\|_2, \quad \forall v_h \in V_0^h. \tag{6.10}$$

Let $\theta_h = \tilde{u}_h^* - \tilde{u}_h \in V_0^h$. Following Section 5.2, define the following dual problem: find $w \in H_0^1(\Omega)$ satisfying

$$A^*(w, v) = (\theta_h, v), \quad \forall v \in H_0^1(\Omega). \tag{6.11}$$

Let $w_h \in V_0^h$ be the solution to

$$A_h^*(w_h, v_h) = (\theta_h, v_h), \quad \forall v_h \in V_0^h. \tag{6.12}$$

By (6.10) and Theorem 5.3, we get

$$\|\theta_h\|_0^2 = (\theta_h, \theta_h) = A_h^*(w_h, \theta_h) = A_h(\tilde{u}_h^* - \tilde{u}_h, w_h) = \mathcal{O}(h^{k+2})\|u\|_{k+4}\|w_h\|_2 = \mathcal{O}(h^{k+2})\|u\|_{k+4}\|\theta_h\|_0,$$

thus $\|\tilde{u}_h^* - \tilde{u}_h\|_0 = \|\theta_h\|_0 = \mathcal{O}(h^{k+2})\|u\|_{k+4}$. By equivalence of norms for polynomials, we have

$$\|\tilde{u}_h^* - \tilde{u}_h\|_{2,Z_0} \leq C\|\tilde{u}_h^* - \tilde{u}_h\|_0 = \mathcal{O}(h^{k+2})\|u\|_{k+4,\Omega}. \tag{6.13}$$

Notice that both $\tilde{u}_h$ and $\tilde{u}_h^*$ are constant zero along $\partial\Omega$, and $u_h|_{\partial\Omega} = g_I$ is the Lagrangian interpolation of g along $\partial\Omega$. With (6.7), we have proven the following main result.

THEOREM 6.2. *Assume elliptic regularity* (2.6) *and V^h ellipticity holds. For a nonhomogeneous Dirichlet boundary problem* (6.1)*, with suitable smoothness assumptions for $k \geq 2$, $a_{ij}, b_i, c \in W^{k+2,\infty}(\Omega)$, the exact solution of* (6.2) *$u(x,y) = \tilde{u} + \bar{g} \in H^{k+4}(\Omega)$ and $f(x,y) \in H^{k+2}(\Omega)$, the numerical solution u_h by scheme* (6.3) *is a $(k+2)$-th order accurate approximation to u in the discrete 2-norm over all the $(k+1)\times(k+1)$ Gauss-Lobatto points:*

$$\|u_h - u\|_{2,Z_0} = \mathcal{O}(h^{k+2})(\|u\|_{k+4} + \|f\|_{k+2}).$$

7. Finite difference implementation.

In this section we present the finite difference implementation of the scheme (6.3) for the case $k = 2$ on a uniform mesh. The finite difference implementation of the nonhomogeneous Dirichlet boundary value problem is based on a homogeneous Neumann boundary value problem, which will be discussed first. We demonstrate how it is derived for the one-dimensional case then give the two-dimensional implementation. It provides efficient assembling of the stiffness matrix and one can easily implement it in MATLAB. Implementations for higher order elements or quasi-uniform meshes can be similarly derived, even though it will no longer be a conventional finite difference scheme on a uniform grid.

7.1. One-dimensional case.

Consider a homogeneous Neumann boundary value problem $-(au')' = f$ on $[0,1], u'(0) = 0, u'(1) = 0$, and its variational form is to seek $u \in H^1([0,1])$ satisfying

$$(au', v') = (f, v), \quad \forall v \in H^1([0,1]). \tag{7.1}$$

Consider a uniform mesh $x_i = ih$, $i = 0, 1, \ldots, n+1$, $h = \frac{1}{n+1}$. Assume n is odd and let $N = \frac{n+1}{2}$. Define intervals $I_k = [x_{2k}, x_{2k+2}]$ for $k = 0, \ldots, N-1$ as a finite element mesh for P^2 basis. Define

$$V^h = \{v \in C^0([0,1]) : v|_{I_k} \in P^2(I_k), k = 0, \ldots, N-1\}.$$

Let $\{v_i\}_{i=0}^{n+1} \subset V^h$ be a basis of V^h such that $v_i(x_j) = \delta_{ij}$, $i, j = 0, 1, \ldots, n+1$. With 3-point Gauss-Lobatto quadrature, the C^0-P^2 finite element method for (7.1) is to seek $u_h \in V^h$ satisfying

$$\langle a u_h', v_i' \rangle_h = \langle f, v_i \rangle_h, \quad i = 0, 1, \ldots, n+1. \tag{7.2}$$

Let $u_j = u_h(x_j)$, $a_j = a(x_j)$ and $f_j = f(x_j)$ then $u_h(x) = \sum_{j=0}^{n+1} u_j v_j(x)$. We have

$$\sum_{j=0}^{n+1} u_j \langle a v_j', v_i' \rangle_h = \langle a u_h', v_j' \rangle_h = \langle f, v_i \rangle_h = \sum_{j=0}^{n+1} f_j \langle v_j, v_i \rangle_h, \quad i = 0, 1, \ldots, n+1.$$

The matrix form of this scheme is $\bar{S}\bar{\mathbf{u}} = \bar{M}\bar{\mathbf{f}}$, where

$$\bar{\mathbf{u}} = \left[u_0, u_1, \ldots, u_n, u_{n+1}\right]^T, \quad \bar{\mathbf{f}} = \left[f_0, f_1, \ldots, f_n, f_{n+1}\right]^T,$$

the stiffness matrix $\bar{S}$ is has size $(n+2)\times(n+2)$ with (i,j)-th entry as $\langle a v_i', v_j' \rangle_h$, and the lumped mass matrix M is a $(n+2)\times(n+2)$ diagonal matrix with diagonal entries $h\left(\frac{1}{3}, \frac{4}{3}, \frac{2}{3}, \frac{4}{3}, \frac{2}{3}, \ldots, \frac{2}{3}, \frac{4}{3}, \frac{1}{3}\right)$.

Next we derive an explicit representation of the matrix $\bar{S}$. Since basis functions $v_i \in V^h$ and $u_h(x)$ are not C^1 at the knots x_{2k} $(k = 1, 2, \ldots, N-1)$, their derivatives at the knots are double valued. We will use superscripts + and − to denote derivatives obtained from the right and from the left respectively, e.g., v'^{+}_{2k} and v'^{-}_{2k+2} denote the derivatives of v_{2k} and v_{2k+2} respectively in the interval $I_k = [x_{2k}, x_{2k+2}]$. Then in the interval $I_k = [x_{2k}, x_{2k+2}]$ we have the following representation of derivatives

$$\begin{bmatrix} v'^{+}_{2k}(x) \\ v'_{2k+1}(x) \\ v'^{-}_{2k+2}(x) \end{bmatrix} = \frac{1}{2h} \begin{bmatrix} -3 & 4 & -1 \\ -1 & 0 & 1 \\ 1 & -4 & 3 \end{bmatrix} \begin{bmatrix} v_{2k}(x) \\ v_{2k+1}(x) \\ v_{2k+2}(x) \end{bmatrix}. \tag{7.3}$$

By abusing notations, we use $(v_i)'_{2k}$ to denote the average of two derivatives of v_i at the knots x_{2k}:

$$(v_i)'_{2k} = \frac{1}{2}[(v'_i)^{-}_{2k} + (v'_i)^{+}_{2k}].$$

Let $[v_i]$ denote the difference between the right derivative and left derivative:

$$[v'_i]_0 = [v'_i]_{n+2} = 0, \quad [v'_i]_{2k} := (v'_i)^{+}_{2k} - (v'_i)^{-}_{2k}, \quad k = 1, 2, \ldots, N-1.$$

Then at the knots, we have

$$(v'_i)^{-}_{2k}(v'_j)^{-}_{2k} + (v'_i)^{+}_{2k}(v'_j)^{+}_{2k} = 2(v'_i)_{2k}(v'_j)_{2k} + \frac{1}{2}[v_i]_{2k}[v_j]_{2k}. \tag{7.4}$$

We also have

$$\langle av'_j, v'_i\rangle_{I_{2k}} = h\left[\frac{1}{3}a_{2k}(v'_j)^{+}_{2k}(v'_i)^{+}_{2k} + \frac{4}{3}a_{2k+1}(v'_j)_{2k+1}(v'_i)_{2k+1} + \frac{1}{3}a_{2k+2}(v'_j)^{-}_{2k+2}(v'_i)^{-}_{2k+2}\right]. \tag{7.5}$$

Let $\mathbf{v}_i$ denote a column vector of size $n+2$ consisting of grid point values of $v_i(x)$. Plugging (7.4) into (7.5), with (7.3), we get

$$\langle av'_j, v'_i\rangle_h = \sum_{k=0}^{N-1}\langle av'_j, v'_i\rangle_{I_{2k}} = \frac{1}{h}\mathbf{v}_i^T(D^TWAD + E^TWAE)\mathbf{v}_j,$$

where A is a diagonal matrix with diagonal entries $a_0, a_1, \ldots, a_n, a_{n+1}$, and

$$W = diag\left(\tfrac{1}{3}, \tfrac{4}{3}, \tfrac{2}{3}, \tfrac{4}{3}, \tfrac{2}{3}, \ldots, \tfrac{2}{3}, \tfrac{4}{3}, \tfrac{1}{3}\right)_{(n+2)\times(n+2)},$$

$$D = \frac{1}{2}\begin{pmatrix} -3 & 4 & -1 & & & & & & & \\ -1 & 0 & 1 & & & & & & & \\ \frac{1}{2} & -2 & 0 & 2 & -\frac{1}{2} & & & & & \\ & & -1 & 0 & 1 & & & & & \\ & & \frac{1}{2} & -2 & 0 & 2 & -\frac{1}{2} & & & \\ & & & & -1 & 0 & 1 & & & \\ & & & & & \ddots & \ddots & \ddots & & \\ & & & & & -1 & 0 & 1 & & \\ & & & & & \frac{1}{2} & -2 & 0 & 2 & -\frac{1}{2} \\ & & & & & & & -1 & 0 & 1 \\ & & & & & & & 1 & -4 & 3 \end{pmatrix}_{(n+2)\times(n+2)}, E = \frac{1}{2}\begin{pmatrix} 0 & 0 & 0 & & & & & & & \\ 0 & 0 & 0 & & & & & & & \\ -\frac{1}{2} & 2 & -3 & 2 & -\frac{1}{2} & & & & & \\ & & 0 & 0 & 0 & & & & & \\ & & -\frac{1}{2} & 2 & -3 & 2 & -\frac{1}{2} & & & \\ & & & & 0 & 0 & 0 & & & \\ & & & & & \ddots & \ddots & \ddots & & \\ & & & & & 0 & 0 & 0 & & \\ & & & & & -\frac{1}{2} & 2 & -3 & 2 & -\frac{1}{2} \\ & & & & & & & 0 & 0 & 0 \\ & & & & & & & 0 & 0 & 0 \end{pmatrix}_{(n+2)\times(n+2)}.$$

Since $\{v_i\}_{i=0}^{n}$ are the Lagrangian basis for V^h, we have

$$\bar{S} = \frac{1}{h}(D^TWAD + E^TWAE). \tag{7.6}$$

Now consider the one-dimensional Dirichlet boundary value problem:

$$-(au')' = f \text{ on } [0,1],$$
$$u(0) = \sigma_1, \quad u(1) = \sigma_2.$$

Consider the same mesh as above and define

$$V_0^h = \{v \in C^0([0,1]) : v|_{I_k} \in P^2(I_k), k = 0, \dots, N-1; v(0) = v(1) = 0\}.$$

Then $\{v_i\}_{i=1}^n \subset V^h$ is a basis of V_0^h for $\{v_i\}_{i=0}^{n+1}$ defined above. The one-dimensional version of (6.3) is to seek $u_h \in V_0^h$ satisfying

$$\begin{aligned} \langle a u_h', v_i' \rangle_h &= \langle f, v_i \rangle_h - \langle a g_I', v_i' \rangle_h, \quad i = 1, 2, \dots, n, \\ g_I(x) &= \sigma_0 v_0(x) + \sigma_1 v_{n+1}(x). \end{aligned} \tag{7.7}$$

Notice that we can obtain (7.7) by simply setting $u_h(0) = \sigma_0$ and $u_h(1) = \sigma_1$ in (7.2). So the finite difference implementation of (7.7) is given as follows:

1. Assemble the $(n+2) \times (n+2)$ stiffness matrix $\bar{S}$ for homogeneous Neumann problem as in (7.6).
2. Let S denote the $n \times n$ submatrix $\bar{S}(2 : n+1, 2 : n+1)$, i.e., $[\bar{S}_{ij}]$ for $i, j = 2, \cdots, n+1$.
3. Let $\mathbf{l}$ denote the $n \times 1$ submatrix $\bar{S}(2 : n+1, 1)$ and $\mathbf{r}$ denote the $n \times 1$ submatrix $\bar{S}(2 : n+1, n+2)$, which correspond to $v_0(x)$ and $v_{n+1}(x)$.
4. Let $\mathbf{u} = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix}^T$ and $\mathbf{f} = \begin{bmatrix} f_1 & f_2 & \cdots & f_n \end{bmatrix}^T$. Define $\mathbf{w} = \left[\frac{4}{3}, \frac{2}{3}, \frac{4}{3}, \frac{2}{3}, \dots, \frac{2}{3}, \frac{4}{3}\right]$ as a column vector of size n. The scheme (7.7) can be implemented as
$$S\mathbf{u} = h\mathbf{w}^T\mathbf{f} - \sigma_0 \mathbf{l} - \sigma_1 \mathbf{r}.$$

7.2. Notations and tools for the two-dimensional case. We will need two operators:

- Kronecker product of two matrices: if A is $m \times n$ and B is $p \times q$, then $A \otimes B$ is $mp \times nq$ give by
$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \vdots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{pmatrix}.$$
- For a $m \times n$ matrix X, $vec(X)$ denotes the vectorization of the matrix X by rearranging X into a vector column by column.

The following properties will be used:

1. $(A \otimes B)(C \otimes D) = AC \otimes BD$.
2. $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$.
3. $(B^T \otimes A)vec(X) = vec(AXB)$.
4. $(A \otimes B)^T = A^T \otimes B^T$.

Consider a uniform grid (x_i, y_j) for a rectangular domain $\bar{\Omega} = [0,1] \times [0,1]$ where $x_i = ih_x$, $i = 0, 1, \dots, n_x+1$, $h_x = \frac{1}{n_x+1}$ and $y_j = jh_y$, $j = 0, 1, \dots, n_y+1$, $h_y = \frac{1}{n_y+1}$.

Assume n_x and n_y are odd and let $N_x = \frac{n_x+1}{2}$ and $N_y = \frac{n_y+1}{2}$. We consider rectangular cells $e_{kl} = [x_{2k}, x_{2k+2}] \times [y_{2l}, y_{2l+2}]$ for $k = 0, \dots, N_x - 1$ and $l = 0, \dots, N_y - 1$ as a finite element mesh for Q^2 basis. Define

$$V^h = \{v \in C^0(\Omega) : v|_{e_{kl}} \in Q^2(e_{kl}), k = 0, \dots, N_x - 1, l = 0, \dots, N_y - 1\},$$

$$V_0^h = \{v \in C^0(\Omega) : v|_{e_{kl}} \in Q^2(e_{kl}), k = 0, \dots, N_x - 1, l = 0, \dots, N_y - 1; v|_{\partial\Omega} \equiv 0\}.$$

For the coefficients $\mathbf{a}(x,y) = \begin{pmatrix} a^{11} & a^{12} \\ a^{21} & a^{22} \end{pmatrix}$, $\mathbf{b} = [b^1 \quad b^2]$ and c in the elliptic operator (2.3), consider their grid point values in the following form:

$$A^{kl} = \begin{pmatrix} a_{00} & a_{01} & \cdots & a_{0,n_x+1} \\ a_{10} & a_{11} & \cdots & a_{1,n_x+1} \\ \vdots & \vdots & & \vdots \\ a_{n_y+1,0} & a_{n_y+1,1} & \cdots & a_{n_y+1,,n_x+1} \end{pmatrix}_{(n_y+2)\times(n_x+2)}, \quad a_{ij} = a^{kl}(x_j, y_i), \quad k,l = 1,2,$$

$$B^{m} = \begin{pmatrix} b_{00} & b_{01} & \cdots & b_{0,n_x+1} \\ b_{10} & b_{11} & \cdots & b_{1,n_x+1} \\ \vdots & \vdots & & \vdots \\ b_{n_y+1,0} & b_{n_y+1,1} & \cdots & b_{n_y+1,n_x+1} \end{pmatrix}_{(n_y+2)\times(n_x+2)}, \quad b_{ij} = b^{m}(x_j, y_i), \quad m = 1,2,$$

$$C = \begin{pmatrix} c_{00} & c_{01} & \cdots & c_{0,n_x+1} \\ c_{10} & c_{11} & \cdots & c_{1,n_x+1} \\ \vdots & \vdots & & \vdots \\ c_{n_y+1,0} & c_{n_y+1,1} & \cdots & c_{n_y+1,n_x+1} \end{pmatrix}_{(n_y+2)\times(n_x+2)}, \quad c_{ij} = c(x_j, y_i).$$

Let $diag(\mathbf{x})$ denote a diagonal matrix with the vector $\mathbf{x}$ as diagonal entries and define

$$\bar{W}_x = diag\left(\tfrac{1}{3}, \tfrac{4}{3}, \tfrac{2}{3}, \tfrac{4}{3}, \tfrac{2}{3}, \ldots, \tfrac{2}{3}, \tfrac{4}{3}, \tfrac{1}{3}\right)_{(n_x+2)\times(n_x+2)},$$

$$\bar{W}_y = diag\left(\tfrac{1}{3}, \tfrac{4}{3}, \tfrac{2}{3}, \tfrac{4}{3}, \tfrac{2}{3}, \ldots, \tfrac{2}{3}, \tfrac{4}{3}, \tfrac{1}{3}\right)_{(n_y+2)\times(n_y+2)},$$

$$W_x = diag\left(\tfrac{4}{3}, \tfrac{2}{3}, \tfrac{4}{3}, \tfrac{2}{3}, \ldots, \tfrac{2}{3}, \tfrac{4}{3}\right)_{n_x\times n_x}, W_y = diag\left(\tfrac{4}{3}, \tfrac{2}{3}, \tfrac{4}{3}, \tfrac{2}{3}, \ldots, \tfrac{2}{3}, \tfrac{4}{3}\right)_{n_y\times n_y}.$$

Let $s = x$ or y, we define the D and E matrices with dimension $(n_s+2)\times(n_s+2)$ for each variable:

$$D_s = \frac{1}{2}\begin{pmatrix} -3 & 4 & -1 & & & & & & & & \\ -1 & 0 & 1 & & & & & & & & \\ \frac{1}{2} & -2 & 0 & 2 & -\frac{1}{2} & & & & & & \\ & & -1 & 0 & 1 & & & & & & \\ & & \frac{1}{2} & -2 & 0 & 2 & -\frac{1}{2} & & & & \\ & & & & -1 & 0 & 1 & & & & \\ & & & & & \ddots & \ddots & \ddots & & & \\ & & & & & & -1 & 0 & 1 & & \\ & & & & & & \frac{1}{2} & -2 & 0 & 2 & -\frac{1}{2} \\ & & & & & & & & -1 & 0 & 1 \\ & & & & & & & & 1 & -4 & 3 \end{pmatrix}, \quad E_s = \frac{1}{2}\begin{pmatrix} 0 & 0 & 0 & & & & & & & & \\ 0 & 0 & 0 & & & & & & & & \\ -\frac{1}{2} & 2 & -3 & 2 & -\frac{1}{2} & & & & & & \\ & & 0 & 0 & 0 & & & & & & \\ & & -\frac{1}{2} & 2 & -3 & 2 & -\frac{1}{2} & & & & \\ & & & & 0 & 0 & 0 & & & & \\ & & & & & \ddots & \ddots & \ddots & & & \\ & & & & & & 0 & 0 & 0 & & \\ & & & & & & -\frac{1}{2} & 2 & -3 & 2 & -\frac{1}{2} \\ & & & & & & & & 0 & 0 & 0 \\ & & & & & & & & 0 & 0 & 0 \end{pmatrix}.$$

Define an inflation operator $Infl : \mathbb{R}^{n_y\times n_x} \longrightarrow \mathbb{R}^{(n_y+2)\times(n_x+2)}$ by adding zeros:

$$Infl(U) = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & U & \vdots \\ 0 & \cdots & 0 \end{pmatrix}_{(n_y+2)\times(n_x+2)}$$

and its matrix representation is given as $\tilde{I}_x \otimes \tilde{I}_y$ where

$$\tilde{I}_x = \begin{pmatrix} \mathbf{0} \\ I_{n_x\times n_x} \\ \mathbf{0} \end{pmatrix}_{(n_x+2)\times n_x}, \tilde{I}_y = \begin{pmatrix} \mathbf{0} \\ I_{n_y\times n_y} \\ \mathbf{0} \end{pmatrix}_{(n_y+2)\times n_y}.$$

Its adjoint is a restriction operator $Res : \mathbb{R}^{(n_y+2)\times(n_x+2)} \longrightarrow \mathbb{R}^{n_y\times n_x}$ as

$$Res(X) = X(2 : n_y + 1, 2 : n_x + 1) \quad , \forall X \in \mathbb{R}^{(n_y+2)\times(n_x+2)},$$

and its matrix representation is $\tilde{I}_x^T \otimes \tilde{I}_y^T$.

7.3. Two-dimensional case. For $\bar{\Omega} = [0, 1]^2$ we first consider an elliptic equation with homogeneous Neumann boundary condition:

$$-\nabla \cdot (\mathbf{a}\nabla u) + \mathbf{b}\nabla u + cu = f \text{ on } \Omega, \tag{7.8}$$

$$\mathbf{a}\nabla u \cdot \mathbf{n} = 0 \text{ on } \partial\Omega. \tag{7.9}$$

The variational form is to find $u \in H^1(\Omega)$ satisfying

$$A(u, v) = (f, v), \quad \forall v \in H^1(\Omega). \tag{7.10}$$

The C^0-Q^2 finite element method with 3×3 Gauss-Lobatto quadrature is to find $u_h \in V^h$ satisfying

$$\langle \mathbf{a}\nabla u_h, \nabla v_h\rangle_h + \langle \mathbf{b}\nabla u_h, v_h\rangle_h + \langle cu_h, v_h\rangle_h = \langle f, v_h\rangle_h, \quad \forall v_h \in V^h, \tag{7.11}$$

Let $\bar{U}$ be a $(n_y + 2) \times (n_x + 2)$ matrix such that its (j, i)-th entry is $\bar{U}(j, i) = u_h(x_{i-1}, y_{j-1})$, $i = 1, \dots, n_x + 2$, $j = 1, \dots, n_y + 2$. Let $\bar{F}$ be a $(n_y + 2) \times (n_x + 2)$ matrix such that its (j, i)-th entry is $\bar{F}(j, i) = f(x_{i-1}, y_{j-1})$. Then the matrix form of (7.11) is

$$\bar{S}vec(\bar{U}) = \bar{M}vec(\bar{F}), \quad \bar{M} = h_x h_y \bar{W}_x \otimes \bar{W}_y, \quad \bar{S} = \sum_{k,l=1}^{2} S_a^{kl} + \sum_{m=1}^{2} S_b^m + S_c, \tag{7.12}$$

where

$$S_a^{11} = \frac{h_y}{h_x}(D_x^T \otimes I_y)diag(vec(\bar{W}_y A^{11}\bar{W}_x))(D_x \otimes I_y) + \frac{h_y}{h_x}(E_x^T \otimes I_y)diag(vec(\bar{W}_y A^{11}\bar{W}_x))(E_x \otimes I_y),$$
$$S_a^{12} = (D_x^T \otimes I_y)diag(vec(\bar{W}_y A^{12}\bar{W}_x))(I_x \otimes D_y) + (E_x^T \otimes I_y)diag(vec(\bar{W}_y A^{12}\bar{W}_x))(I_x \otimes E_y),$$
$$S_a^{21} = (I_x \otimes D_y^T)diag(vec(\bar{W}_y A^{21}\bar{W}_x))(D_x \otimes I_y) + (I_x \otimes E_y^T)diag(vec(\bar{W}_y A^{21}\bar{W}_x))(E_x \otimes I_y),$$
$$S_a^{22} = \frac{h_x}{h_y}(I_x \otimes D_y^T)diag(vec(\bar{W}_y A^{22}\bar{W}_x))(I_x \otimes D_y) + \frac{h_x}{h_y}(I_x \otimes E_y^T)diag(vec(\bar{W}_y A^{22}\bar{W}_x))(I_x \otimes E_y),$$
$$S_b^1 = h_y diag(vec(\bar{W}_y B^1\bar{W}_x))(D_x \otimes I_y), \quad S_b^2 = h_x diag(vec(\bar{W}_y B^2\bar{W}_x))(I_x \otimes D_y),$$
$$S_c = h_x h_y diag(vec(\bar{W}_y C\bar{W}_x).$$

Now consider the scheme (6.3) for nonhomogeneous Dirichlet boundary conditions. Its numerical solution can be represented as a matrix U of size $ny \times nx$ with (j, i)-entry $U(j, i) = u_h(x_i, y_j)$ for $i = 1, \cdots, nx; j = 1, \cdots, ny$. Similar to the one-dimensional case, its stiffness matrix can be obtained as the submatrix of $\bar{S}$ in (7.12). Let $\bar{G}$ be a $(n_y + 2)$ by $(n_x + 2)$ matrix with (j, i)-th entry as $\bar{G}(j, i) = g(x_{i-1}, y_{j-1})$, where

$$g(x, y) = \begin{cases} 0, & \text{if} \quad (x, y) \in (0, 1) \times (0, 1), \\ g(x, y), & \text{if} \quad (x, y) \in \partial\Omega. \end{cases}$$

In particular, $\bar{G}(j + 1, i + 1) = 0$ for $j = 1, \dots, n_y$, $i = 1, \dots, n_x$. Let F be a matrix of size $ny \times nx$ with (j, i)-entry as $F(j, i) = f(x_i, y_j)$ for $i = 1, \cdots, nx; j = 1, \cdots, ny$. Then the scheme (6.3) becomes

$$(\tilde{I}_x^T \otimes \tilde{I}_y^T)\bar{S}(\tilde{I}_x \otimes \tilde{I}_y)vec(U) = (W_x \otimes W_y)vec(F) - (\tilde{I}_x^T \otimes \tilde{I}_y^T)\bar{S}vec(\bar{G}). \tag{7.13}$$

Even though the stiffness matrix is given as $S = (\tilde{I}_x^T \otimes \tilde{I}_y^T)\bar{S}(\tilde{I}_x \otimes \tilde{I}_y)$, S should be implemented as a linear operator in iterative linear system solvers. For example, the matrix vector multiplication $(\tilde{I}_x^T \otimes \tilde{I}_y^T)S_a^{11}(\tilde{I}_x \otimes \tilde{I}_y)vec(U)$ is equivalent to the following linear operator from $\mathbb{R}^{ny\times nx}$ to $\mathbb{R}^{ny\times nx}$:

$$\frac{h_y}{h_x}\tilde{I}_y^T\left\{I_y\left([\bar{W}_yA^{11}\bar{W}_x]\circ[I_y(\tilde{I}_yU\tilde{I}_x^T)D_x^T]\right)D_x + I_y\left([\bar{W}_yA^{11}\bar{W}_x]\circ[I_y(\tilde{I}_yU\tilde{I}_x^T)E_x^T]\right)E_x\right\}\tilde{I}_x,$$

where $\circ$ is the Hadamard product (i.e., entrywise multiplication).

7.4. The Laplacian case. For one-dimensional constant coefficient case with homogeneous Dirichlet boundary condition, the scheme can be written as a classical finite difference scheme $H\mathbf{u} = \mathbf{f}$ with

$$H = M^{-1}S = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & & & & & & \\ -2 & \frac{7}{2} & -2 & \frac{1}{4} & & & & \\ & -1 & 2 & -1 & & & & \\ & \frac{1}{4} & -2 & \frac{7}{2} & -2 & \frac{1}{4} & & \\ & & & -1 & 2 & -1 & & \\ & & & & \ddots & \ddots & & \\ & & & & \frac{1}{4} & -2 & \frac{7}{2} & -2 \\ & & & & & & -1 & 2 \end{pmatrix}$$

In other words, if x_i is a cell center, the scheme is

$$\frac{-u_{i-1} + 2u_i - u_{i+1}}{h^2} = f_i,$$

and if x_i is a knot away from the boundary, the scheme is

$$\frac{u_{i-2} - 8u_{i-1} + 14u_i - 8u_{i+1} + u_{i+2}}{4h^2} = f_i.$$

It is straightforward to verify that the local truncation error is only second order.

For the two-dimensional Laplacian case homogeneous Dirichlet boundary condition, the scheme can be rewritten as

$$(H_x \otimes I_y) + (I_x \otimes H_y)vec(U) = vec(F),$$

where H_x and H_y are the same H matrix above with size $n_x \times n_x$ and $n_y \times n_y$ respectively. The inverse of $(H_x \otimes I_y) + (I_x \otimes H_y)$ can be efficiently constructed via the eigen-decomposition of small matrices H_x and H_y:

1. Compute eigen-decomposition of $H_x = T_x\Lambda_xT_x^{-1}$ and $H_y = T_y\Lambda_yT_y^{-1}$.
2. The properties of Kronecker product imply that

   $$(H_x \otimes I_y) + (I_x \otimes H_y) = (T_x \otimes T_y)(\Lambda_x \otimes I_y + I_x \otimes \Lambda_y)(T_x^{-1} \otimes T_y^{-1}),$$

   thus

   $$[(H_x \otimes I_y) + (I_x \otimes H_y)]^{-1} = (T_x \otimes T_y)(\Lambda_x \otimes I_y + I_x \otimes \Lambda_y)^{-1}(T_x^{-1} \otimes T_y^{-1}).$$

3. It is nontrivial to determine whether H is diagonalizable. In all our numerical tests, H has no repeated eigenvalues. So if assuming Λ_x and Λ_y are diagonal matrices, the matrix vector multiplication $[(H_x \otimes I_y) + (I_x \otimes H_y)]^{-1}vec(F)$ can be implemented as a linear operator on F:

   $$T_y([T_y^{-1}F(T_x^{-1})^T]./\Lambda)T_x^T, \tag{7.14}$$

   where Λ is a $n_y \times n_x$ matrix with (i,j)-th entry as $\Lambda(i,j) = \Lambda_y(i,i) + \Lambda_x(j,j)$ and ./ denotes entry-wise division for two matrices of the same size.

For the 3D Laplacian, the matrix can be represented as $H_x \otimes I_y \otimes I_z + I_x \otimes H_y \otimes I_z + I_x \otimes I_y \otimes H_z$ thus can be efficiently inverted through eigen-decomposition of small matrices H_x, H_y and H_z as well.

Since the eigen-decomposition of small matrices H_x and H_y can be precomputed, and (7.14) costs only $\mathcal{O}(n^3)$ for a 2D problem on a mesh size $n \times n$, in practice (7.14) can be used as a simple preconditioner in conjugate gradient solvers for the following linear system equivalent to (7.13):

$$(W_x^{-1} \otimes W_y^{-1})(\tilde{I}_x^T \otimes \tilde{I}_y^T)\bar{S}(\tilde{I}_x \otimes \tilde{I}_y)vec(U) = vec(F) - (W_x^{-1} \otimes W_y^{-1})(\tilde{I}_x^T \otimes \tilde{I}_y^T)\bar{S}vec(G),$$

even though the multigrid method as reviewed in [19] is the optimal solver in terms of computational complexity.

8. Numerical results. In this section we show a few numerical tests verifying the accuracy of the scheme (6.3) for $k = 2$ implemented as a finite difference scheme on a uniform grid. We first consider the following two dimensional elliptic equation:

$$-\nabla \cdot (\mathbf{a}\nabla u) + \mathbf{b} \cdot \nabla u + cu = f \quad \text{on } [0,1] \times [0,2] \tag{8.1}$$

where $\mathbf{a} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$, $a_{11} = 10 + 30y^5 + x\cos y + y$, $a_{12} = a_{21} = 2 + 0.5(\sin(\pi x) + x^3)(\sin(\pi y) + y^3) + \cos(x^4 + y^3)$, $a_{22} = 10 + x^5$, $\mathbf{b} = \mathbf{0}$, $c = 1 + x^4 y^3$, with an exact solution

$$u(x, y) = 0.1(\sin(\pi x) + x^3)(\sin(\pi y) + y^3) + \cos(x^4 + y^3).$$

The errors at grid points are listed in Table 1 for purely Dirichlet boundary condition and Table 2 for purely Neumann boundary condition. We observe fourth order accuracy in the discrete 2-norm for both tests, even though only $\mathcal{O}(h^{3.5})$ can be proven for Neumann boundary condition as discussed in Remark 5.5. Regarding the maximum norm of the superconvergence of the function values at Gauss-Lobatto points, one can only prove $\mathcal{O}(h^3 \log h)$ even for the full finite element scheme (1.1) since discrete Green's function is used, see [4].

TABLE 1
A 2D elliptic equation with Dirichlet boundary conditions. The first column is the number of regular cells in a finite element mesh. The second column is the number of grid points in a finite difference implementation, i.e., number of degree of freedoms.

FEM Mesh	FD Grid	l^2 error	order	l^∞ error	order
2×4	3×7	3.94E-2	-	7.15E-2	-
4×8	7×15	1.23E-2	1.67	3.28E-2	1.12
8×16	15×31	1.46E-3	3.08	5.42E-3	2.60
16×32	31×63	1.14E-4	3.68	3.96E-4	3.78
32×64	63×127	7.75E-6	3.88	2.62E-5	3.92
64×128	127×255	5.02E-7	3.95	1.73E-6	3.92
128×256	255×511	3.23E-8	3.96	1.13E-7	3.94

Next we consider a three-dimensional problem $-\Delta u = f$ with homogeneous Dirichlet boundary conditions on a cube $[0,1]^3$ with the following exact solution

$$u(x, y, z) = \sin(\pi x)\sin(2\pi y)\sin(3\pi z) + (x - x^3)(y^2 - y^4)(z - z^2).$$

See Table 3 for the performance of the finite difference scheme. There is no essential difficulty to extend the proof to three dimensions, even though it is not

TABLE 2
A 2D elliptic equation with Neumann boundary conditions.

FEM Mesh	FD Grid	l^2 error	order	l^∞ error	order
2×4	5×9	1.38E0	-	2.27E0	-
4×8	9×17	1.46E-1	3.24	2.52E-1	3.17
8×16	17×33	7.49E-3	4.28	1.64E-2	3.94
16×32	33×65	4.31E-4	4.12	1.02E-3	4.01
32×64	65×129	2.61E-5	4.04	7.47E-5	3.78

very straightforward. Nonetheless we observe that the scheme is indeed fourth order accurate. The linear system is solved by the eigenvector method shown in Section 7.4. The discrete 2-norm over the set of all grid points Z_0 is defined as $\|u\|_{2,Z_0} = \left[h^3 \sum_{(x,y,z)\in Z_0} |u(x,y,z)|^2\right]^{\frac{1}{2}}$.

TABLE 3
$-\Delta u = f$ in 3D with homogeneous Dirichlet boundary condition.

Finite Difference Grid	l^2 error	order	l^∞ error	order
$7 \times 7 \times 7$	1.51E-2	-	4.87E-2	-
$15 \times 15 \times 15$	9.23E-4	4.04	3.12E-3	3.96
$31 \times 31 \times 31$	5.68E-5	4.02	1.95E-4	4.00
$63 \times 63 \times 63$	3.54E-6	4.01	1.22E-5	4.00
$127 \times 127 \times 127$	2.21E-7	4.00	7.59E-7	4.00

Last we consider (8.1) with convection term and the coefficients $\mathbf{b}$ is incompressible $\nabla \cdot \mathbf{b} = 0$: $\mathbf{a} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$, $a_{11} = 100 + 30y^5 + x\cos y + y$, $a_{12} = a_{21} = 2 + 0.5(\sin(\pi x) + x^3)(\sin(\pi y) + y^3) + \cos(x^4 + y^3)$, $a_{22} = 100 + x^5$, $\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$, $b_1 = \psi_y$, $b_2 = -\psi_x$, $\psi = x\exp(x^2 + y)$, $c = 1 + x^4y^3$, with an exact solution

$$u(x,y) = 0.1(\sin(\pi x) + x^3)(\sin(\pi y) + y^3) + \cos(x^4 + y^3).$$

The errors at grid points are listed in Table 4 for Dirichlet boundary conditions.

TABLE 4
A 2D elliptic equation with convection term and Dirichlet boundary conditions.

FEM Mesh	FD Grid	l^2 error	order	l^∞ error	order
2×4	3×7	1.26E-1	-	2.71E-1	-
4×8	7×15	2.85E-2	2.15	9.70E-2	1.48
8×16	15×31	1.89E-3	3.92	7.25E-3	3.74
16×32	31×63	1.17E-4	4.01	4.01E-4	4.17
32×64	63×127	7.41E-6	3.98	2.54E-5	3.98

9. Concluding remarks.

In this paper we have proven the superconvergence of function values in the simplest finite difference implementation of C^0-Q^k finite element method for elliptic equations. In particular, for the case $k = 2$ the scheme (6.3) can be easily implemented as a fourth order accurate finite difference scheme as shown in

Section 7. It provides only only an convenient approach for constructing fourth order accurate finite difference schemes but also the most efficient implementation of C^0-Q^k finite element method without losing superconvergence of function values. In a follow up paper [12], we will show that discrete maximum principle can be proven for the scheme (6.3) in the case $k = 2$ when solving a variable coefficient Poisson equation.

REFERENCES

[1] M. Bakker, *A note on C^0 Galerkin methods for two-point boundary problems*, Numerische Mathematik, 38 (1982), pp. 447–453.

[2] C. Chen, *Superconvergent points of Galerkin's method for two point boundary value problems*, Numerical Mathematics A Journal of Chinese Universities, 1 (1979), pp. 73–79.

[3] C. Chen, *Superconvergence of finite element solutions and their derivatives*, Numerical Mathematics A Journal of Chinese Universities, 3 (1981), pp. 118–125.

[4] C. Chen, *Structure theory of superconvergence of finite elements (In Chinese)*, Hunan Science and Technology Press, Changsha, 2001.

[5] P. G. Ciarlet, *Basic error estimates for elliptic problems*, Handbook of Numerical Analysis, 2 (1991), pp. 17–351.

[6] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, Society for Industrial and Applied Mathematics, 2002.

[7] P. G. Ciarlet and P.-A. Raviart, *The combined effect of curved boundaries and numerical integration in isoparametric finite element methods*, in The mathematical foundations of the finite element method with applications to partial differential equations, Elsevier, 1972, pp. 409–474.

[8] J. Douglas, T. Dupont, and M. F. Wheeler, *An L^∞ estimate and a superconvergence result for a Galerkin method for elliptic equations based on tensor products of piecewise polynomials*, (1974).

[9] P. Grisvard, *Elliptic problems in nonsmooth domains*, vol. 69, SIAM, 2011.

[10] Y. Huang and J. Xu, *Superconvergence of quadratic finite elements on mildly structured grids*, Mathematics of computation, 77 (2008), pp. 1253–1268.

[11] P. Lesaint and M. Zlamal, *Superconvergence of the gradient of finite element solutions*, RAIRO. Analyse numérique, 13 (1979), pp. 139–166.

[12] H. Li and X. Zhang, *On the monotonicity and discrete maximum principle of the finite difference implementation of C^0-Q^2 finite element method*, in preparation, (2019).

[13] H. Li and X. Zhang, *Superconvergence of C^0-Q^k finite element method for elliptic equations with approximated coefficients*, arXiv preprint arXiv:1902.00945, (2019).

[14] Q. Lin and N. Yan, *Construction and Analysis for Efficient Finite Element Method (In Chinese)*, Hebei University Press, 1996.

[15] Q. Lin, N. Yan, and A. Zhou, *A rectangle test for interpolated finite elements*, in Proc. Sys. Sci. and Sys. Eng.(Hong Kong), Great Wall Culture Publ. Co, 1991, pp. 217–229.

[16] G. Savaré, *Regularity results for elliptic equations in lipschitz domains*, Journal of Functional Analysis, 152 (1998), pp. 176–201.

[17] L. Wahlbin, *Superconvergence in Galerkin finite element methods*, Springer, 2006.

[18] J. Whiteman, *Lagrangian finite element and finite difference methods for poisson problems*, in Numerische Behandlung von Differentialgleichungen, Springer, 1975, pp. 331–355.

[19] J. Xu and L. Zikatanov, *Algebraic multigrid methods*, Acta Numerica, 26 (2017), pp. 591–721.