A MONOTONE Q^1 FINITE ELEMENT METHOD FOR ANISOTROPIC ELLIPTIC EQUATIONS

HAO LI * AND XIANGXIONG ZHANG †

Abstract. We construct a monotone continuous Q^1 finite element method on the uniform mesh for the anisotropic diffusion problem with a diagonally dominant diffusion coefficient matrix. The monotonicity implies the discrete maximum principle. Convergence of the new scheme is rigorously proven. On quadrilateral meshes, the matrix coefficient conditions translate into specific a mesh constraint.

Key words. Inverse positivity, Q^1 finite element method, monotonicity, discrete maximum principle, anisotropic diffusion

AMS subject classifications. 65N30, 65N15, 65N12

1. Introduction.

1.1. Monotonicity and discrete maximum principle.

Consider solving the following elliptic equation on $\Omega = (0,1)^2$ with Dirichlet boundary conditions:

$$
\begin{aligned}
\mathcal{L}u \equiv -\nabla\cdot(\mathbf{a}\nabla u) + cu &= f \quad \text{on} \quad \Omega,\\
u &= g \quad \text{on} \quad \partial\Omega,
\end{aligned}
\tag{1.1}
$$

where the diffusion matrix $\mathbf{a}(\mathbf{x}) \in \mathbb{R}^{2\times 2}$, $c(\mathbf{x})$, $f(\mathbf{x})$ and $g(\mathbf{x})$ are sufficiently smooth functions over $\bar{\Omega}$ or $\partial\Omega$. We assume that $\forall \mathbf{x} \in \Omega$, $\mathbf{a}(\mathbf{x})$ is symmetric and uniformly positive definite on Ω. In the literature, (1.1) is called a heterogeneous anisotropic diffusion problem when the eigenvalues of $\mathbf{a}(\mathbf{x})$ are unequal and vary over on Ω. For a smooth function $u \in C^2(\Omega)\cap C(\bar{\Omega})$, a maximum principle holds [7]:

$$
\mathcal{L}u \le 0 \quad \text{on} \quad \Omega \quad \Longrightarrow \quad \max_{\bar{\Omega}} u \le \max\left\{0, \max_{\partial\Omega} u\right\}.
$$

In particular,

$$
\mathcal{L}u = 0 \text{ in } \Omega \Longrightarrow |u(x_1,x_2)| \le \max_{\partial\Omega}|u|, \quad \forall (x_1,x_2)\in\Omega. \tag{1.2}
$$

For simplicity, we only consider the homogeneous Dirichlet boundary condition, i.e. $g = 0$. The anisotropic diffusion problem (1.1) arises from various areas of science and engineering, including plasma physics, Lagrangian hydrodynamics, and image processing. To avoid spurious oscillations or non-physical numerical solution, it is desired to have numerical schemes to satisfy (1.2) in the discrete sense. We are interested in a linear approximation to $\mathcal{L}$ which can be represented as a matrix L_h. The matrix L_h is called monotone if its inverse only has nonnegative entries, i.e., $L_h^{-1} \ge 0$. Monotonicity of the scheme is a sufficient condition for the discrete maximum principle and has various applications espeically for parabolic problems, see [1, 33, 14, 9, 31, 21, 5, 6, 22, 21, 13, 16].

*lihao199408@gmail.com,

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

1.2. Monotone schemes for anisotropic diffusion equations. Monotone (or positive-type in some literature) numerical methods for problem (1.1) have received considerable attention, e.g., see [11, 17, 18, 19, 20, 25, 34, 30, 12, 2, 27]. The major efforts of studying linear monotone schemes take advantage of M-matrix (see [29] for the definition), either by showing the coefficient matrix is M-matrix directly or the coefficient matrix can be factorized into product of M-matrices. In the following, we call a numerical scheme satisfying *M-matrix property* if the corresponding coefficient matrix is a M-matrix.

By factorizing the stiffness matrix into a product of M-matrices, the monotonocity can still be ensured. For a nine-point scheme on a two-dimensional quadrilateral grid, the matrix condition for monotonicity with specific splitting strategy in [28] aligns with the Lorenz's condition presented in [23, 14]. The difference is in [23, 14], only the existence of the factorization was proved while in [28] the authors found the exact matrix factorization.

In [26], it is proved that a monotone finite difference scheme exists for any linear second-order elliptic problem on fine enough uniform mesh and a finite difference method with fixed stencil for all the problems satisfying the M-matrix property does not exist. With nonnegative directional splittings, [32, 8, 27] propose to construct finite difference schemes for elliptic operators in the nondivergence form and divergence form. Particularly in [27], it is shown that a monotone scheme satisfying the M-matrix property can be constructed for continuous diffusion matrix for sufficiently fine mesh and sufficiently large finite difference stencil.

In [17], for the P^1 finite elements in two and three dimensions, the author generalized the well known non-obtuse angle condition for anisotropic diffusion problem in the sense to have the dihedral angles of all mesh elements, measured in a metric depending on $\mathbf{a}(\mathbf{x})$, be non-obtuse. It reduces to the non-obtuse angle condition for isotropic diffusion matrices when $\mathbf{a}(\mathbf{x}) = \alpha(\mathbf{x})\mathbb{I}$. The formulation was also utilized in [17] for the construction of the so called M-uniform meshes on which the numerical scheme is monotone. The approach to show monotonicity in [17] is to write the global matrix as the sum of local contributions. In [10], the Delaunay condition is extended to anisotropic diffusion problems through a refined analysis studying the whole stiffness matrix for the two-dimensional situation. The analysis of [17] was extended to the anisotropic diffusion–convection–reaction problems in [24].

For the Q^1 finite elements, research on monotonicity has predominantly been focused on meshes whose cells are rectangular blocks. For the two-dimensional Poisson equation, it was noted in [3] that the M-matrix property is violated when the aspect ratio, i.e. the ratio between the length of the longer edge and the shorter edge of the cell, becomes excessively large. Then the discrete maximum principle is not guaranteed.

1.3. Contributions and organization of the paper. It is well known that the second-order accurate linear schemes, such as mixed finite element and multipoint flux approximation, do not always satisfy monotonicity for distorted meshes or with high anisotropy ratio. In this paper, we construct a monotone Q^1 finite element method for solving the equation (1.1), which is second-order accurate for function values.

To analyze the monotonicity of the stiffness matrix, we approximate integrals with a specific quadrature rule, particularly, the linear combination of the trapezoid rule and midpoint rule. We demonstrate that a continuous Q^1 finite element method with the specific quadrature rule, when applied to the anisotropic diffusion problem on a uniform mesh, ensures monotonicity for the problem with a diagonally dominant diffusion coefficient matrix. The method is linear, second-order accurate. The convergence of the function values for this method is also proven. The coefficient constraints become mesh constraints when this Q^1 finite element method is used on general quadrilateral meshes.

The paper is organized as follows. In Section 2, we introduce the notations and review standard quadrature estimates. In Section 3, we derive the Q^1 scheme for anisotropic diffusion equation with Dirichlet boundary condition and derive the coefficient constraints for the stiffness matrix to be an M-matrix. In Section 4, we prove the convergence of function values. In Section 5, we discuss the extension to general quadrilateral meshes. Numerical results are given in Section 6.

2. Preliminaries.

2.1. Notation and tools. We list the tools and notation as follows.

- For the problem dimension d, though we only consider the case $d = 2$, sometimes we keep the general notation d to illustrate how the results are influenced by the dimension.
- For the Q^1 finite element space, i.e., tensor product of linear polynomials, the local space is defined on a reference cell $\hat{K}$, e.g., $\hat{K} = [0,1]^2$. Then, the finite element space on a physical mesh cell e is given by the reference map from $\hat{K}$ to e. The reference element $\hat{K}$ is as Figure 1.

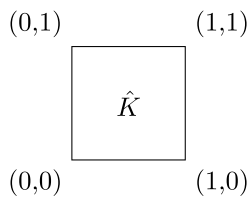

FIG. 1. *The reference element.*

On a reference element $\hat{K}$, we have the Lagrangian basis $\hat{\phi}_{0,0}$, $\hat{\phi}_{0,1}$, $\hat{\phi}_{1,1}, \hat{\phi}_{1,0}$ as

$$\hat{\phi}_{0,0} = (1-\hat{x}_1)(1-\hat{x}_2), \quad \hat{\phi}_{0,1} = (1-\hat{x}_1)\hat{x}_2, \quad \hat{\phi}_{1,1} = \hat{x}_1\hat{x}_2, \quad \hat{\phi}_{1,0} = \hat{x}_1(1-\hat{x}_2). \tag{2.1}$$

- We will use ^ for a function to emphasize the function is defined on or transformed to the reference element $\hat{K}$ from a physical mesh element.
- For a quadrilateral element e, we assume $\mathbf{F}_e$ is the bilinear mapping such that $\mathbf{F}_e(\hat{K}) = e$. Let $\mathbf{c}_{i,j}$, $i,j = 0,1$ be the vertices of the quadrilateral element e. The mapping $\boldsymbol{F}_e$ can be written as

$$\mathbf{F}_e = \sum_{\ell=0}^{1}\sum_{m=0}^{1} \mathbf{c}_{\ell,m}\hat{\phi}_{\ell,m}.$$

- $Q^1(\hat{K}) = \left\{ p(\mathbf{x}) = \sum_{i=0}^{1}\sum_{j=0}^{1} p_{ij}\hat{\phi}_{i,j}(\hat{\mathbf{x}}), \quad \hat{\mathbf{x}} \in \hat{K} \right\}$ is the set of Q^1 polynomials on the reference element $\hat{K}$.
- $Q^1(e) = \left\{ v_h \in H^1(e) : v_h \circ \boldsymbol{F}_j \in Q_1(\hat{K}) \right\}$ is the set of Q^1 polynomials on an element e.

- $V^h = \left\{p(\mathbf{x}) \in H^1(\Omega_h) : p\big|_e \in Q^1(e), \quad \forall e \in \Omega_h\right\}$ denotes the continuous Q^1 finite element space on Ω_h.
- $V_0^h = \left\{v_h \in V^h : v_h = 0 \quad \text{on} \quad \partial\Omega\right\}$
- Let $(f,v)_e$ denote the inner product in $L^2(e)$ and (f,v) denote the inner product in $L^2(\Omega)$:

$$(f,v)_e = \int_e fv\,d\mathbf{x}, \quad (f,v) = \int_\Omega fv\,d\mathbf{x} = \sum_e (f,v)_e.$$

- Let $\langle f,v\rangle_{e,h}$ denote the approximation to $(f,v)_e$ by the mixed quadrature defined in (2.7) over element e with some specified quadrature parameter and $\langle f,v\rangle_h$ denotes the approximation to (f,v) by

$$\langle f,v\rangle_h = \sum_e \langle f,v\rangle_{e,h}.$$

- Let $E(f)$ denote the quadrature error for integrating $f(\hat{\mathbf{x}})$ on element e. Let $\hat{E}(\hat{f})$ denote the quadrature error for integrating $\hat{f}(\hat{\mathbf{x}}) = f(\mathbf{F}_e(\hat{\mathbf{x}}))$ on the reference element $\hat{K}$. Then $E(f) = h^d\hat{E}(\hat{f})$ on uniform rectangular mesh with mesh size h.
- The norm and semi-norms for $W^{k,p}(\Omega)$ and $1 \le p < +\infty$, with standard modification for $p = +\infty$:

$$\|u\|_{k,p,\Omega} = \left(\sum_{i+j\le k} \iint_\Omega \left|\partial_{x_1}^i \partial_{x_2}^j u(x_1,x_2)\right|^p d\mathbf{x}\right)^{1/p},$$

$$|u|_{k,p,\Omega} = \left(\sum_{i+j=k} \iint_\Omega \left|\partial_{x_1}^i \partial_{x_2}^j u(x_1,x_2)\right|^p d\mathbf{x}\right)^{1/p},$$

$$[u]_{k,p,\Omega} = \left(\iint_\Omega \left|\partial_{x_1}^k u(x_1,x_2)\right|^p d\mathbf{x} + \iint_\Omega \left|\partial_{x_2}^k u(x_1,x_2)\right|^p d\mathbf{x}\right)^{1/p}.$$

- In the special case where $\omega = \Omega$, we drop the subscript, i.e. $(\cdot,\cdot) := (\cdot,\cdot)_\Omega$ and $\|\cdot\| := \|\cdot\|_\Omega$.
- 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 semi-norms by the following symbols

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

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

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

- For simplicity, sometimes we may use $\|u\|_{k,\Omega}, |u|_{k,\Omega}$ and $[u]_{k,\Omega}$ denote norm and semi-norms for $H^k(\Omega) = W^{k,2}(\Omega)$. When there is no confusion, Ω may be dropped in the norm and semi-norms, e.g., $\|u\|_k := \|u\|_{k,\Omega}$.

- Inverse estimates for polynomials:

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

- Elliptic regularity holds for the problem (3.1):

$$\|u\|_2 \leq C\|f\|_0$$

- Let Ω_h is a finite element mesh for Ω. For each element $e \in \Omega_h$, we denote $\bar{\mathbf{a}}_e = (\bar{a}_e^{ij})$ as an approximation to the average of $\mathbf{a}$ on element e, i.e. $\bar{a}_e^{ij} = \frac{1}{meas(e)} \int_e a^{ij} d\mathbf{x}$. Specifically, we choose $\bar{a}_e^{ij}$ as the the function value of a^{ij} at the center of element e. Then we define piece-wise constant function $\bar{\mathbf{a}}$ as

$$\bar{\mathbf{a}}(\mathbf{x}) = \bar{\mathbf{a}}_e, \quad \text{for } \mathbf{x} \in e.$$

- 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

$$\int_{\hat{K}} \left(\hat{\Pi}_1\hat{u}\right) \hat{w} d\hat{\mathbf{x}} = \int_{\hat{K}} \hat{u}\hat{w} d\hat{\mathbf{x}}, \quad \forall \hat{w} \in Q^1(\hat{K}). \tag{2.2}$$

  Observe that all degrees of freedom of $\hat{\Pi}_1\hat{u}$ can be expressed as a linear combination of $\int_{\hat{K}} \hat{u}\hat{p} d\hat{\mathbf{x}}$ where $\hat{p}(\mathbf{x})$ takes the forms $1, \hat{x}_1, \hat{x}_2$, and $\hat{x}_1\hat{x}_2$. This implies that the $H^1(\hat{K})$ (or $H^2(\hat{K})$) norm of $\hat{\Pi}_1\hat{u}$ is dictated by $\int_{\hat{K}} \hat{u}\hat{p} d\hat{\mathbf{x}}$. Utilizing the Cauchy-Schwartz inequality, we deduce:

$$\left| \int_{\hat{K}} \hat{u}\hat{p} d\hat{\mathbf{x}} \right| \leq \|\hat{u}\|_{0,2,\hat{K}} \|\hat{p}\|_{0,2,\hat{K}} \leq C\|\hat{u}\|_{0,2,\hat{K}}$$

  From which it follows that:

$$\|\Pi_1\hat{u}\|_{1,2,\hat{K}} \leq C\|\hat{u}\|_{0,2,\hat{K}}$$

  This establishes that $\hat{\Pi}_1$ acts as a continuous linear mapping from $L^2(\hat{K})$ to $H^1(\hat{K})$. Similarly, by extending this argument, we can also demonstrate that $\hat{\Pi}_1$ is a continuous linear mapping from $L^2(\hat{K})$ to $H^2(\hat{K})$.
- We denote all the the vertices of Ω_h inside Ω by $\mathbf{x}_j, j = 1, \ldots, N_h$. We denote nodal basis functions in V_h by $\varphi_i, i = 1, \ldots N_h$, which are continuous in Ω, linear in each element e and

$$\varphi_i(\mathbf{x}_i) = 1, \quad \varphi_i(\mathbf{x}_j) = 0, \quad j \neq i.$$

2.2. Mixed quadrature. To analyze and impose the monotonicity of the stiffness matrix, we will use numerical quadrature rules to approximate integrals. As we will see, the choice of quadrature rules can significantly affect the monotonicity of the numerical schemes.

For a one-dimensional integral of function f over the interval $[0, 1]$, we can approximate $\int_0^1 f(\hat{x})d\hat{x}$ using either the trapezoid rule, given by $\frac{f(0)+f(1)}{2}$, or the midpoint rule, $f\left(\frac{1}{2}\right)$. Both quadrature offer second-order accuracy. We will use the linear combination of these two kinds of quadrature as follows:

$$\begin{aligned} \int_0^1 f(\hat{x})d\hat{x} \simeq& \lambda \frac{f(0)+f(1)}{2} + (1-\lambda) f\left(\frac{1}{2}\right) \\ =& \hat{\omega}_1 f(\hat{\xi}_1) + \hat{\omega}_2 f(\hat{\xi}_2) + \hat{\omega}_3 f(\hat{\xi}_1), \end{aligned} \tag{2.3}$$

where λ is a parameter to be determined and

$$\hat{\omega}_1 = \frac{\lambda}{2}, \quad \hat{\omega}_2 = 1-\lambda, \quad \hat{\omega}_3 = \frac{\lambda}{2}, \quad \hat{\xi}_1 = 0, \quad \hat{\xi}_2 = \frac{1}{2}, \quad \hat{\xi}_3 = 1. \tag{2.4}$$

When $\lambda = 1$, the mixed quadrature recovers the trapezoid rule and when $\lambda = 0$ the mixed quadrature recovers the midpoint rule.

To approximate integration on square $\hat{K}$, we may use the mixed quadrature (2.3) with different parameters λ^1 and λ^2 for different dimension x_1 and x_2 respectively. By Fubini's theorem,

$$\begin{aligned} &\int_{\hat{K}} f(\hat{\mathbf{x}}) d\hat{\mathbf{x}} = \int_0^1 \int_0^1 f(\hat{\mathbf{x}}) d\hat{\mathbf{x}} = \int_0^1 \left(\int_0^1 f\left(\hat{x}_1, \hat{x}_2\right) d\hat{x}_2 \right) d\hat{x}_1 \\ &\simeq \int_0^1 \left(\sum_{q=1}^{3} \hat{\omega}_q^2 f\left(\hat{x}_1, \hat{\xi}_q\right) \right) d\hat{x}_1 \simeq \sum_{p=1}^{r+1} \hat{\omega}_p^1 \left(\sum_{q=1}^{r+1} \hat{\omega}_q^2 f\left(\hat{\xi}_p, \hat{\xi}_q\right) \right) = \sum_{p=1}^{3} \sum_{q=1}^{3} \hat{\omega}_p^1 \hat{\omega}_q^2 f\left(\hat{\xi}_p, \hat{\xi}_q\right), \end{aligned} \tag{2.5}$$

where ω_i^j are just ω_i while replacing λ with λ^j in (2.4) for $i = 1, 2, 3,\ j = 1, 2$.

On the reference element $\hat{K}$, for convenience, to denote the above quadrature for integral approximation with parameter $\boldsymbol{\lambda} = \left(\lambda^1, \lambda^2\right)$, we will use the following notation

$$\int_{\hat{K}} \hat{f}(\hat{\mathbf{x}}) d^h_{\boldsymbol{\lambda}} \hat{\mathbf{x}} := \sum_{p=1}^{3} \sum_{q=1}^{3} \hat{\omega}_p^1 \hat{\omega}_q^2 f\left(\hat{\xi}_p, \hat{\xi}_q\right). \tag{2.6}$$

Given the quadrature parameter $\boldsymbol{\lambda}_e = \left(\lambda_e^1, \lambda_e^2\right)$, the quadrature approximation to $\int_e f(\mathbf{x}) d\mathbf{x}$ is denoted as

$$\int_e f(\mathbf{x}) d^h_{\boldsymbol{\lambda}_e} \mathbf{x} := \int_{\hat{K}} f \circ \mathbf{F}_e(\hat{\mathbf{x}}) d^h_{\boldsymbol{\lambda}_e} \hat{\mathbf{x}}. \tag{2.7}$$

Then we define the quadrature approximation over the entire domain Ω as

$$\int_\Omega f d^h_{\boldsymbol{\lambda}_\Omega} \mathbf{x} := \sum_{e \in \Omega_h} \int_e f d^h_{\boldsymbol{\lambda}_e} \mathbf{x}, \tag{2.8}$$

where $\boldsymbol{\lambda}_\Omega = \left(\boldsymbol{\lambda}_e\right)_{e \in \Omega_h}$ can be viewed as a vector-valued piece-wise constant function, with values $\boldsymbol{\lambda}_e$ that differ across elements.

As a particular instance, $\int_\Omega f d^h_1 \mathbf{x}$ denote the case $\boldsymbol{\lambda}_e = (1, 0)$ for all $e \in \Omega_h$, i.e. the integral on each element are approximated by the trapezoid rule in all directions.

2.3. Quadrature error 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 [4]:

THEOREM 2.1. *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*

$$\|\hat{u} - \hat{\Pi}\hat{u}\|_{k+1,\hat{K}} \leq C[\hat{u}]_{k+1,\hat{K}}, \quad \forall \hat{u} \in H^{k+1}(\hat{K}). \tag{2.9}$$

Therefore 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 quadrature estimates.

LEMMA 2.2. *For a sufficiently smooth function $a \in H^2(e)$, we have*

$$\int_e a d\mathbf{x} - \int_e a d^h\mathbf{x} = \mathcal{O}\left(h^{2+\frac{d}{2}}\right)[a]_{2,e} = \mathcal{O}\left(h^{2+d}\right)[a]_{2,\infty,e} \tag{2.10}$$

$$\int_e a d\mathbf{x} - \int_e \bar{a}_e d\mathbf{x} = \mathcal{O}\left(h^{2+\frac{d}{2}}\right)[a]_{2,e} = \mathcal{O}\left(h^{2+d}\right)[a]_{2,\infty,e} \tag{2.11}$$

Proof. For any $\hat{f} \in H^2(\hat{K})$, since quadrature are represented by point values, with the Sobolev's embedding we have

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

Therefore $\hat{E}(\cdot)$ is a continuous linear form on $H^2(\hat{K})$ and $\hat{E}(\hat{f}) = 0$ if $\hat{f} \in Q^1(\hat{K})$. Then the Bramble-Hilbert lemma implies

$$|E(a)| = h^d|\hat{E}(\hat{a})| \leq Ch^d[\hat{a}]_{2,2,\hat{K}} = \mathcal{O}\left(h^{2+\frac{d}{2}}\right)[a]_{2,2,e} = \mathcal{O}\left(h^{2+d}\right)[a]_{2,\infty,e}$$

LEMMA 2.3. *If $f \in H^2(\Omega)$, $\forall v_h \in V^h$, we have*

$$(f, v_h) - \langle f, v_h\rangle_h = \mathcal{O}\left(h^2\right)\|f\|_2\|v_h\|_1.$$

Proof. Applying Theorem 2.1, on element e, with $\frac{\partial^2 \hat{v}_h}{\partial^2 \hat{x}_i}$ vanish, we obtain: ∎

$$\begin{aligned}
E(fv) = h^d\hat{E}(\hat{f}\hat{v}_h) &\leq Ch^d[\hat{f}\hat{v}_h]_{2,2,\hat{K}} \\
\leq &Ch^d\left(|\hat{f}|_{2,2,\hat{K}}|\hat{v}_h|_{0,\infty,\hat{K}} + |\hat{f}|_{1,2,\hat{K}}|\hat{v}_h|_{1,\infty,\hat{K}}\right) \\
\leq &Ch^d\left(|\hat{f}|_{2,2,\hat{K}}|\hat{v}_h|_{0,2,\hat{K}} + |\hat{f}|_{1,2,\hat{K}}|\hat{v}_h|_{1,2,\hat{K}}\right) \\
\leq &Ch^2\left(|f|_{2,2,e}|v_h|_{0,2,e} + |f|_{1,2,e}|v_h|_{1,2,e}\right) = \mathcal{O}\left(h^2\right)\|f\|_{2,e}\|v_h\|_{1,e}.
\end{aligned}$$

By sum the above result over all elements of Ω_h, then we conclude with

$$(f, v_h) - \langle f, v_h\rangle_h = \mathcal{O}\left(h^2\right)\|f\|_2\|v_h\|_1.$$

LEMMA 2.4. *If $u \in H^3(e)$, for $i, j = 1, 2$, then $\forall v_h$,*

$$\int_e u_{x_i}(v_h)_{x_j} d\mathbf{x} - \int u_{x_i}(v_h)_{x_j} d^h_{\boldsymbol{\lambda}_e}\mathbf{x} = \mathcal{O}\left(h^2\right)\|u\|_{3,e}\|v_h\|_{2,e}.$$

Proof. Applying Theorem 2.1, we obtain:

$$\begin{aligned}
E(u_{x_i}(v_h)_{x_j}) &= h^{d-2}\hat{E}(\hat{u}_{\hat{x}_i}(\hat{v}_h)_{\hat{x}_j}) \leq Ch^{d-2}[\hat{u}_{\hat{x}_i}(\hat{v}_h)_{\hat{x}_j}]_{2,2,\hat{K}} \\
\leq &Ch^{d-2}\left(|\hat{u}_{\hat{x}_i}|_{2,2,\hat{K}}|(\hat{v}_h)_{\hat{x}_j}|_{0,\infty,\hat{K}} + |\hat{u}_{\hat{x}_i}|_{1,2,\hat{K}}|(\hat{v}_h)_{\hat{x}_j}|_{1,\infty,\hat{K}} + |\hat{u}_{\hat{x}_i}|_{0,2,\hat{K}}|(\hat{v}_h)_{\hat{x}_j}|_{2,\infty,\hat{K}}\right) \\
\leq &Ch^{d-2}\left(|\hat{u}_{\hat{x}_i}|_{2,2,\hat{K}}|(\hat{v}_h)_{\hat{x}_j}|_{0,2,\hat{K}} + |\hat{u}_{\hat{x}_i}|_{1,2,\hat{K}}|(\hat{v}_h)_{\hat{x}_j}|_{1,2,\hat{K}} + |\hat{u}_{\hat{x}_i}|_{0,2,\hat{K}}|(\hat{v}_h)_{\hat{x}_j}|_{2,2,\hat{K}}\right) \\
\leq &Ch^{d-2}\left(|\hat{u}|_{3,2,\hat{K}}|\hat{v}_h|_{1,2,\hat{K}} + |\hat{u}|_{2,2,\hat{K}}|\hat{v}_h|_{2,2,\hat{K}}\right).
\end{aligned}$$

where the second last inequality is implied by the equivalence of norms over $Q^1(\hat{K})$ and in the last inequality we use the fact that the third derivative of Q^1 polynomial vanish.

Therefore,

$$E(u_{x_i}(v_h)_{x_j}) \le Ch^2\left(|u|_{3,2,e}|v_h|_{1,2,e} + |u|_{2,2,e}|v_h|_{2,2,e}\right) = \mathcal{O}\left(h^2\right)\|u\|_{3,e}\|v_h\|_{2,e}. \qquad \square$$

LEMMA 2.5. *If $f \in H^2(\Omega)$ or $f \in V^h$, $\forall v_h$, we have*

$$(f, v_h) - \langle f, v_h\rangle_h = \mathcal{O}(h)\|f\|_2\|v_h\|_0.$$

Proof. As in the proof of Lemma 2.3, we have

$$E(fv) = \mathcal{O}\left(h^2\right)\|f\|_{2,e}\|v_h\|_{1,e}. \qquad \square$$

By applying the inverse estimate to polynomial v_h, we have

$$E(fv) = \mathcal{O}(h)\|f\|_{2,e}\|v_h\|_{0,e}.$$

Summing the previous result across all elements in Ω_h, we conclude:

$$(f, v_h) - \langle f, v_h\rangle_h = \mathcal{O}(h)\|f\|_2\|v_h\|_0.$$

3. The Q^1 finite element method and its monotonicity.

In this section, we give a derivation of the Q^1 finite element scheme and then discuss its monotonicity.

3.1. Derivation of the scheme.

The variational form of (1.1) is to find $u \in H_0^1(\Omega)$ satisfying

$$\mathcal{A}(u, v) = (f, v), \quad \forall v \in H_0^1(\Omega), \tag{3.1}$$

where $\mathcal{A}(u,v) = \int_\Omega \mathbf{a}\nabla u \cdot \nabla v d\mathbf{x} + \int_\Omega cuvd\mathbf{x}$, $(f,v) = \int_\Omega fvd\mathbf{x}$.

Let $V_0^h \subseteq H_0^1(\Omega)$ be the continuous finite element space consisting of piece-wise Q^1 polynomials. To have a second-order monotone method, we first approximate the matrix coefficients $\mathbf{a} = (a^{ij}(\mathbf{x}))$ by either its average $\frac{1}{meas(e)}\int_e \mathbf{a}d\mathbf{x}$ or its middle point value on each element e. The approximation is denoted by $\bar{\mathbf{a}}_e$. Then we get the modified bilinear form

$$\bar{\mathcal{A}}(u, v) = \int_\Omega \bar{\mathbf{a}}\nabla u \cdot \nabla v d\mathbf{x} + \int_\Omega cuvd\mathbf{x},$$

where $\bar{\mathbf{a}} = (\bar{\mathbf{a}}_e)_{e\in\Omega_h}$. In practice, we take $\bar{\mathbf{a}}_e$ to be the middle point value of $\bar{\mathbf{a}}$ on element e for smooth enough $\mathbf{a}$ and fine enough mesh.

By approximating integrals in $\bar{\mathcal{A}}(u_h, v_h)$ with quadrature specified in (2.8), along with designated quadrature parameter $\boldsymbol{\lambda}_\Omega$, we derive the following numerical scheme: find $u_h \in V_0^h$ satisfying

$$\mathcal{A}_h(u_h, v_h) = \langle f, v_h\rangle_h, \quad \forall v_h \in V_0^h, \tag{3.2}$$

where the approximated bilinear form is defined as

$$\mathcal{A}_h(u_h, v_h) := \int_\Omega \bar{\mathbf{a}}\nabla u_h \cdot \nabla v_h d^h_{\boldsymbol{\lambda}}\mathbf{x} + \int_\Omega cu_h v_h d^h_1\mathbf{x} \tag{3.3}$$

and the right hand side is

$$\langle f, v_h\rangle_h := \int_\Omega f v_h d^h_1\mathbf{x}. \tag{3.4}$$

Of course, the quadrature parameter $\boldsymbol{\lambda} = (\lambda^1, \lambda^2)$ on each element need to be determined for the quadrature (2.7).

It is not obvious that the numerical solution u_h is an accurate approximation of the exact solution u as $\bar{\mathbf{a}}$ varies depending on the mesh.

3.2. Monotonicity. Let $A = (\mathcal{A}_h(\nabla\varphi_i, \nabla\varphi_j))$ be the stiffness matrix of our Q^1 scheme (3.2) for equation (1.1). To have the monotonicity, we enforce the stiffness matrix A to be a M-matrix. We are interested in conditions for A to be an M-matrix. Recall a sufficient condition for M-matrix, see condition C_{10} in [29]:

LEMMA 3.1. *For a real irreducible square matrix A with positive diagonal entries and non-positive off-diagonal entries, A is a nonsingular M-matrix if all the row sums of A are non-negative and at least one row sum is positive.*

Then we have the following result on the uniform rectangular mesh.

THEOREM 3.2. *Assume $\forall e \in \Omega_h$, $|\bar{a}_e^{12}| \le \min\{\bar{a}_e^{11}, \bar{a}_e^{22}\}$. Then for the Q^1 scheme given by (3.2) for the elliptic equation (1.1) on uniform rectangular mesh, the stiffness matrix is a M-matrix, provided the quadrature parameters for each element e are chosen as:*

$$\lambda_e^1, \lambda_e^2 \in \left(\frac{|\bar{a}_e^{11} - \bar{a}_e^{22}|}{\bar{a}_e^{11} + \bar{a}_e^{22}}, 1 - \frac{2|\bar{a}_e^{12}|}{\bar{a}_e^{11} + \bar{a}_e^{22}} \right]. \tag{3.5}$$

When $|\bar{a}_e^{12}| = \min\{\bar{a}_e^{11}, \bar{a}_e^{22}\}$, (3.5) means we take λ_e^1, λ_e^2 to be the upper bound of the interval, i.e. $1 - \frac{2|\bar{a}_e^{12}|}{\bar{a}_e^{11}+\bar{a}_e^{22}}$.

Proof. First, we consider the following quadrature approximation results on the reference element $\hat{K}$. With quadrature (2.6) and quadrature parameter $\boldsymbol{\lambda}_e = (\lambda_e^1, \lambda_e^2)$, we have

$$\begin{aligned}
\langle \bar{\mathbf{a}}\nabla\phi_{0,0}, \nabla\phi_{0,1}\rangle_h = \langle \bar{\mathbf{a}}\nabla\phi_{1,1}, \nabla\phi_{1,0}\rangle_h &= -\frac{1}{4}(\lambda_e^2\bar{a}_e^{11} + \lambda_e^1\bar{a}_e^{22}) + \frac{1}{4}(\bar{a}_e^{11} - \bar{a}_e^{22}),\\
\langle \bar{\mathbf{a}}\nabla\phi_{0,0}, \nabla\phi_{1,0}\rangle_h = \langle \bar{\mathbf{a}}\nabla\phi_{0,1}, \nabla\phi_{1,1}\rangle_h &= -\frac{1}{4}(\lambda_e^2\bar{a}_e^{11} + \lambda_e^1\bar{a}_e^{22}) + \frac{1}{4}(\bar{a}_e^{22} - \bar{a}_e^{11}),\\
\langle \bar{\mathbf{a}}\nabla\phi_{0,0}, \nabla\phi_{1,1}\rangle_h &= -\frac{1}{4}\left((1-\lambda_e^2)\bar{a}_e^{11} + (1-\lambda_e^1)\bar{a}_e^{22}\right) - \frac{1}{2}\bar{a}_e^{12},\\
\langle \bar{\mathbf{a}}\nabla\phi_{0,1}, \nabla\phi_{1,0}\rangle_h &= -\frac{1}{4}\left((1-\lambda_e^2)\bar{a}_e^{11} + (1-\lambda_e^1)\bar{a}_e^{22}\right) + \frac{1}{2}\bar{a}_e^{12}.
\end{aligned}$$

With (3.5) and the assumption $|\bar{a}_e^{12}| \le \min\{\bar{a}_e^{11}, \bar{a}_e^{22}\}$, we have

$$\begin{aligned}
\langle \bar{\mathbf{a}}\nabla\phi_{0,0}, \nabla\phi_{0,1}\rangle_h = \langle \bar{\mathbf{a}}\nabla\phi_{1,1}, \nabla\phi_{1,0}\rangle_h &\in \left[\frac{1}{2}\left(|\bar{a}_e^{12}| - \bar{a}_e^{22}\right), \frac{1}{4}(\bar{a}_e^{11} - \bar{a}_e^{22} - |\bar{a}_e^{11} - \bar{a}_e^{22}|\right),\\
\langle \bar{\mathbf{a}}\nabla\phi_{0,0}, \nabla\phi_{1,0}\rangle_h = \langle \bar{\mathbf{a}}\nabla\phi_{0,1}, \nabla\phi_{1,1}\rangle_h &\in \left[\frac{1}{2}\left(|\bar{a}_e^{12}| - \bar{a}_e^{11}\right), \frac{1}{4}(\bar{a}_e^{22} - \bar{a}_e^{11} - |\bar{a}_e^{11} - \bar{a}_e^{22}|\right),\\
\langle \bar{\mathbf{a}}\nabla\phi_{0,0}, \nabla\phi_{1,1}\rangle_h &\in \left(-\frac{1}{2}(\min\{\bar{a}_e^{11}, \bar{a}_e^{22}\} - \bar{a}_e^{12}), -\frac{1}{2}(|\bar{a}_e^{12}| + \bar{a}_e^{12})\right],\\
\langle \bar{\mathbf{a}}\nabla\phi_{0,1}, \nabla\phi_{1,0}\rangle_h &\in \left(-\frac{1}{2}(\min\{\bar{a}_e^{11}, \bar{a}_e^{22}\} + \bar{a}_e^{12}), -\frac{1}{2}(|\bar{a}_e^{12}| - \bar{a}_e^{12})\right],
\end{aligned} \tag{3.6}$$

which are all non-positive. Again, when $|\bar{a}_e^{12}| = \min\{\bar{a}_e^{11}, \bar{a}_e^{22}\}$, we will take the above values as the bound of the closed side of the interval.

Given $j \in \{1, \ldots, N_h\}$, consider the corresponding node x_j. Obviously, if both x_i

and x_j are vertices of the same elements e,

$$
\begin{aligned}
A_{ij} &= \mathcal{A}_h(\varphi_j, \varphi_i) \\
&= \sum_{e\in\Omega_h} \int_e \bar{\mathbf{a}}\nabla\varphi_j \cdot \nabla\varphi_i d^h_{\boldsymbol{\lambda}_e}\mathbf{x} + \int_e c\varphi_j\varphi_i d^h_1\mathbf{x} \\
&= \sum_{e\in\Omega_h} \int_{\hat{K}} \bar{\mathbf{a}}\hat{\nabla}\hat{\varphi}_j \cdot \hat{\nabla}\hat{\varphi}_i d^h_{\boldsymbol{\lambda}_e}\hat{\mathbf{x}} + \int_{\hat{K}} \hat{c}\hat{\varphi}_j\hat{\varphi}_i d^h_1\hat{\mathbf{x}} \\
&= \sum_{i,j\in e} \int_{\hat{K}} \bar{\mathbf{a}}\hat{\nabla}\hat{\varphi}_j \cdot \hat{\nabla}\hat{\varphi}_i d^h_{\boldsymbol{\lambda}_e}\hat{\mathbf{x}} + \int_{\hat{K}} \hat{c}\hat{\varphi}_j\hat{\varphi}_i d^h_1\hat{\mathbf{x}}
\end{aligned} \tag{3.7}
$$

where $\sum_{i,j\in e}$ means summation over all elements e containing both vertices i and j.

Notice that $\int_{\hat{K}} \hat{c}\hat{\varphi}_j\hat{\varphi}_i d^h_1\hat{\mathbf{x}}$ vanish if $i \neq j$ and $\int_{\hat{K}} \bar{\mathbf{a}}\hat{\nabla}\hat{\varphi}_j \cdot \hat{\nabla}\hat{\varphi}_i d^h_{\boldsymbol{\lambda}_e}\hat{\mathbf{x}}$ aligns with one of the values in (3.6) depending on their relative positions. Therefore, for $i \neq j$, with (3.5) and the assumption $|\bar{a}^{12}_e| \leq \min\{\bar{a}^{11}_e, \bar{a}^{22}_e\}$ we have

$$
A_{ij} = \sum_{i,j\in e} \int_{\hat{K}} \bar{\mathbf{a}}\hat{\nabla}\hat{\varphi}_j \cdot \hat{\nabla}\hat{\varphi}_i d^h_{\boldsymbol{\lambda}_e}\hat{\mathbf{x}} \leq 0. \tag{3.8}
$$

If $\mathbf{x}_i$ has no neighboring node on the boundary, then the i-th row sum of A is non-negative:

$$
\sum_j A_{ij} = \sum_{j=0}^{N_h} \mathcal{A}_h(\varphi_j, \varphi_i) = \mathcal{A}_h(1, \varphi_i) = Cc_i \geq 0,
$$

where C is a certain positive number and $c_i = c(\mathbf{x}_i) \geq 0$. Therefore, $A_{ii} \geq \sum_{j\neq i} |A_{ij}|$.

When $\mathbf{x}_i$ has a neighboring node on the boundary, we do have $A_{ii} \geq \sum_{j\neq i} |A_{ij}|$. When $\mathbf{x}_i$ has two neighboring node on the boundary, based on (3.6), in the stencil of x_i, one of the corresponding coefficients of the two neighboring nodes on the boundary must be negative, and it is not in $A_{i,:}$, then $\sum_j A_{ij} > 0$, i.e. $A_{ii} > \sum_{j\neq i} |A_{ij}|$.

Therefore, we conclude the proof. ∎

REMARK 1. *For each element e, the choice in (3.5) make $\lambda^1_e, \lambda^2_e > 0$, which implies the V^h-ellipticity of the bilinear form (3.3) discussed in Section 4.2. Therefore, we can assure of V^h-ellipticity and the stiffness matrix being an M-matrix simultaneously.*

REMARK 2. *The constraint on the coefficient, $|\bar{a}^{12}_e| \leq \min\{\bar{a}^{11}_e, \bar{a}^{22}_e\}$, aligns with the condition for rendering the stiffness matrix as an M-matrix in the seven-point stencil control volume method with optimal optimal monotonicity region in the case of homogeneous medium and uniform mesh in [28]. In [27], the authors show that a three-by-three stencil can be used to construct monotone finite difference schemes under the assumption $|a^{12}| < \min\{a^{11}, a^{22}\}$.*

4. Convergence of the Q^1 finite element method with mixed quadrature. In this section, we prove the second-order accuracy of the scheme (3.2) on uniform rectangular mesh. For convenience, in this section, we may drop the subscript h in a test function $v_h \in V^h$. When there is no confusion, we may also drop $d\mathbf{x}$ or $d\hat{\mathbf{x}}$ in a integral.

4.1. Approximation error estimate of bilinear forms. In this subsection, we estimate the approximation error of $\mathcal{A}_h(u, v)$ to $\mathcal{A}(u, v)$.

THEOREM 4.1. *Assume $a^{ij}, c \in W^{2,\infty}(\Omega)$ for $i,j = 1,2$ and $u \in H^3(\Omega)$, then $\forall v \in V^h$, on element e, we have*

$$\int_e (\mathbf{a}\nabla u)\cdot\nabla v d\mathbf{x} - \int_e (\bar{\mathbf{a}}_e \nabla u)\cdot \nabla v d^h_{\boldsymbol{\lambda}_e}\mathbf{x} = \mathcal{O}(h^2)\|u\|_{3,e}\|v\|_{2,e}, \tag{4.1}$$

$$\int_e cuv d\mathbf{x} - \int_e cuv d^h_1\mathbf{x} = \mathcal{O}\left(h^2\right)\|u\|_{2,e}\|v\|_{2,e}. \tag{4.2}$$

Proof. For $k,l = 1,2$ and function $a \in W^{2,\infty}(e)$, we have

$$\begin{aligned}
&\int_e a u_{x_k} v_{x_l} d\mathbf{x} - \int_e \bar{a}_e u_{x_k} v_{x_l} d^h_{\boldsymbol{\lambda}_e}\mathbf{x}\\
=&\int_e (a-\bar{a}_e) u_{x_k} v_{x_l} d\mathbf{x} + \bar{a}_e\left(\int_e u_{x_k} v_{x_l} d\mathbf{x} - \int_e u_{x_k} v_{x_l} d^h_{\boldsymbol{\lambda}_e}\mathbf{x}\right)\\
=&\int_e (a-\bar{a}_e) u_{x_k} v_{x_l} d\mathbf{x} + \bar{a}_e E(u_{x_k} v_{x_l}).
\end{aligned} \tag{4.3}$$

For the first term,

$$\begin{aligned}
&\int_e (a-\bar{a}_e) u_{x_k} v_{x_l} d\mathbf{x}\\
=&\int_e (a-\bar{a}_e)(u_{x_k} v_{x_l} - \overline{u_{x_k} v_{x_l}}) d\mathbf{x} + \int_e (a-\bar{a}_e)\overline{u_{x_k} v_{x_l}} d\mathbf{x}\\
\leq&\|a-\bar{a}_e\|_{0,\infty,e}\|u_{x_k} v_{x_l} - \overline{u_{x_k} v_{x_l}}\|_{0,1,e} + \frac{1}{meas(e)}\int_e (a-\bar{a}_e) d\mathbf{x}\int_e u_{x_k} v_{x_l} d\mathbf{x}.
\end{aligned} \tag{4.4}$$

By Poincare inequality and Cauchy-Schwartz inequality, we have

$$\begin{aligned}
&\|a-\bar{a}_e\|_{0,\infty,e}\|u_{x_k} v_{x_l} - \overline{u_{x_k} v_{x_l}}\|_{0,1,e}\\
=&\mathcal{O}(h^2)\|a\|_{1,\infty,e}\left\|\nabla\left(u_{x_k} v_{x_l}\right)\right\|_{0,1,e} = \mathcal{O}(h^2)\|u\|_{2,e}\|v\|_{2,e}.
\end{aligned} \tag{4.5}$$

By Lemma 2.2 and Cauchy-Schwartz inequality

$$\begin{aligned}
&\frac{1}{meas(e)}\int_e (a-\bar{a}_e) d\mathbf{x}\int_e u_{x_k} v_{x_l} d\mathbf{x}\\
=&\frac{h^{2+d}}{meas(e)}[a]_{2,\infty,e}\|u_{x_k}\|_{0,e}\|v_{x_l}\|_{0,e} = \mathcal{O}\left(h^2\right)\|u\|_{1,e}\|v\|_{1,e}
\end{aligned} \tag{4.6}$$

where in the last equation $meas(e) = \mathcal{O}(h^d)$ is also used. Therefore, we have the estimate of the first term of (4.3):

$$\int_e (a-\bar{a}_e) u_{x_k} v_{x_l} d\mathbf{x} = \mathcal{O}(h^2)\|a\|_{2,\infty,e}\|u\|_{2,e}\|v\|_{2,e}. \tag{4.7}$$

For the second term of (4.3), by Lemma 2.4, we obtain

$$\int_e \bar{a}_e u_{x_k} v_{x_l} d\mathbf{x} - \int_e \bar{a}_e u_{x_l} v_{x_l} d^h_{\boldsymbol{\lambda}_e}\mathbf{x} = \mathcal{O}(h^2)\|a\|_{0,\infty,e}\|u\|_{3,e}\|v\|_{2,e}, \tag{4.8}$$

which together with (4.7) imply the estimate of (4.3):

$$\int_e a u_{x_k} v_{x_l} d\mathbf{x} - \int_e \bar{a}_e u_{x_k} v_{x_l} d^h_{\boldsymbol{\lambda}_e}\mathbf{x} = \mathcal{O}(h^2)\|a\|_{2,\infty,e}\|u\|_{3,e}\|v\|_{2,e}. \tag{4.9}$$

Therefore, we have

$$\int_e (\mathbf{a}(\mathbf{x})\nabla u)\cdot\nabla v d\mathbf{x} - \int_e (\bar{\mathbf{a}}(\mathbf{x})\cdot\nabla u)\,\nabla v d^h_{\boldsymbol{\lambda}_e}\mathbf{x} = \mathcal{O}\left(h^2\right)\|\mathbf{a}\|_{2,\infty,e}\|u\|_{3,e}\|v\|_{2,e}. \tag{4.10}$$

Similarly we have

$$\int_e cuv d\mathbf{x} - \int_e cuv d^h_1\mathbf{x} = \mathcal{O}\left(h^2\right)\|c\|_{2,\infty,e}\|u\|_{2,e}\|v\|_{2,e}. \tag{4.11}$$

□

We also have

Lemma 4.2. *Assume $a^{ij}, c \in W^{2,\infty}(\Omega)$ for $i,j = 1,2$. We have*

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

Proof. By Theorem 4.1 and noticing that the third derivative of Q^1 polynomial vanish, we have

$$\int_e (\mathbf{a}\nabla v_h)\cdot\nabla w_h d\mathbf{x} - \int_e (\bar{\mathbf{a}}_e\nabla v_h)\cdot\nabla w_h d^h_{\boldsymbol{\lambda}_e}\mathbf{x} = \mathcal{O}(h^2)\|v_h\|_{2,e}\|w_h\|_{2,e}, \tag{4.12}$$

$$\int_e cv_h w_h d\mathbf{x} - \int_e cv_h w_h d^h_1\mathbf{x} = \mathcal{O}\left(h^2\right)\|v_h\|_{2,e}\|w_h\|_{2,e}. \tag{4.13}$$

By applying the inverse estimate to polynomial z_h, we get

$$\int_e (\mathbf{a}\nabla v_h)\cdot\nabla w_h d\mathbf{x} - \int_e (\bar{\mathbf{a}}_e\nabla v_h)\cdot\nabla w_h d^h_{\boldsymbol{\lambda}_e}\mathbf{x} = \mathcal{O}(h)\|v_h\|_{2,e}\|w_h\|_{1,e}, \tag{4.14}$$

$$\int_e cv_h w_h d\mathbf{x} - \int_e cv_h w_h d^h_1\mathbf{x} = \mathcal{O}\left(h\right)\|v_h\|_{2,e}\|w_h\|_{1,e}. \tag{4.15}$$

Then by summing over all the elements we get prove the Lemma. □

4.2. V^h-ellipticity and the dual problem. In order to prove the convergence results of the scheme (3.2), we need A_h satisfies V^h-ellipticity:

$$\forall v_h \in V_0^h, \quad C\left\|v_h\right\|_1^2 \le A_h\left(v_h, v_h\right). \tag{4.16}$$

By following the proof of Lemma 5.1 in [15], we have

Lemma 4.3. *Assume the eigenvalues of $\mathbf{a}$ have a uniform positive lower bound and a uniform upper bound and c have a upper bound. If there exists lower bound $\lambda_0 > 0$ such that $\forall e \in \Omega_h$, the quadrature parameter $\lambda_e^1, \lambda_e^2 > \lambda_0$, then there are two constants $C_1, C_2 > 0$ independent of mesh size h such that*

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

Proof. For element e, at first we map all the functions to the reference element $\hat{K}$. Let $Z_{0,\hat{K}}$ denote the set of vertices on the reference element $\hat{K}$. We notice that the set $Z_{0,\hat{K}}$ is a $Q^1(\hat{K})$-unisolvent subset. Since the weights of trapezoid rule are strictly positive, we have

$$\forall \hat{p} \in Q^1(\hat{K}), \quad \sum_{i=1}^{2}\int_{\hat{K}} \hat{p}_{\hat{x}^i}^2 d^h_1\hat{\mathbf{x}} = 0 \Longrightarrow \hat{p}_{\hat{x}^i} = 0 \text{ at } Z_{0,\hat{K}},$$

where $i=1,2$. As a consequence, $\sum_{i=1}^{2}\int_{\hat K}\hat p_{\hat x^i}^2 d_1^h\hat{\mathbf{x}}$ defines a norm over the quotient space $Q^1(\hat K)/Q^0(\hat K)$. 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

$$\forall \hat p\in Q^1(\hat K),\quad C_1|\hat p|_{1,\hat K}^2\le\sum_{i=1}^{2}\int_{\hat K}\hat p_{\hat x^i}^2 d_1^h\hat{\mathbf{x}}\le C_2|\hat p|_{1,\hat K}^2$$

As the quadrature parameter $\lambda_e^1,\lambda_e^2\ge\lambda_0\ge 0$, we have

$$C_1\left|\hat v_h\right|_{1,\hat K}^2\le C_1\sum_{i=1}^{2}\int_{\hat K}(\hat v_h)_{\hat x_i}^2 d_1^h\hat{\mathbf{x}}\le\int_{\hat K}(\bar{\mathbf{a}}_e^{ij}\nabla\hat v_h)\cdot\nabla\hat v_h d_{\boldsymbol{\lambda_e}}^h\hat{\mathbf{x}}+\int_{\hat K}\hat c\hat v_h^2 d_1^h\hat{\mathbf{x}}\le C_2\left\|\hat v_h\right\|_{1,\hat K}^2.$$

Mapping these back to the original element e and summing over all elements, by the equivalence of two norms $|\cdot|_1$ and $\|\cdot\|_1$ for the space $H_0^1(\Omega)\supset V_0^h$, we get the conclusion. ∎

In the following part, we assume the assumption of Lemma 4.3 is fulfilled, i.e. the V^h-ellipticity holds.

In order to apply the Aubin-Nitsche duality argument for establishing convergence of function values, we need certain estimates on a proper dual problem.

Define $\theta:=u-u_h$ and consider the dual problem: find $w\in H_0^1(\Omega)$ satisfying

$$A^*(w,v)=(\theta,v)\,,\quad\forall v\in H_0^1(\Omega), \tag{4.17}$$

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

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

Although here the bilinear form we considered is symmetric i.e. $A(\cdot,\cdot)=A^*(\cdot,\cdot)$, we still use $A^*(\cdot,\cdot)$ for abstractness.

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

$$A_h^*\left(w_h,v_h\right)=\left(\theta,v_h\right),\quad\forall v_h\in V_0^h. \tag{4.18}$$

Notice that the right hand side of (4.18) is different from the right hand side of the scheme (3.2).

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

Lemma 4.4. *Assume $a^{ij},c\in W^{2,\infty}(\Omega)$ and $u\in H^3(\Omega),f\in H^2(\Omega)$. Let w be defined in (4.17), w_h be defined in (4.18). With elliptic regularity and V^h-ellipticity hold, we have*

$$\begin{aligned}\|w-w_h\|_1\le& Ch\|w\|_2\\ \|w_h\|_2\le& C\left\|\theta\right\|_0.\end{aligned} \tag{4.19}$$

Proof. By V^h-ellipticity, we have $C_1\left\|w_h-v_h\right\|_1^2\le A_h^*\left(w_h-v_h,w_h-v_h\right)$. By the definition of the dual problem (4.17), we have

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

Therefore $\forall v_h\in V_0^h$, by Lemma 4.2, we have

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

which implies

$$\|w-w_h\|_1 \le \|w-v_h\|_1 + \|w_h-v_h\|_1 \le C\,\|w-v_h\|_1 + Ch\,\|v_h\|_2 . \tag{4.20}$$

Now consider $\Pi_1 w \in V_0^h$ where Π_1 is the piece-wise Q^1 projection and its definition on each element is defined through (2.2) on the reference element. By Theorem 2.1 on the projection error, we have

$$\|w-\Pi_1 w\|_1 \le Ch\|w\|_2, \quad \|w-\Pi_1 w\|_2 \le C\|w\|_2, \tag{4.21}$$

which implies

$$\|\Pi_1 w\|_2 \le \|w\|_2 + \|w-\Pi_1 w\|_2 \le C\|w\|_2. \tag{4.22}$$

By setting $v_h = \Pi_1 w$, using (4.20), (4.21) and (4.22), we have

$$\|w-w_h\|_1 \le C\,\|w-\Pi_1 w\|_1 + Ch\,\|\Pi_1 w\|_2 \le Ch\|w\|_2. \tag{4.23}$$

By (4.21) and (4.23), we also have

$$\|w_h-\Pi_1 w\|_1 \le \|w-\Pi_1 w\|_1 + \|w-w_h\|_1 \le Ch\|w\|_2. \tag{4.24}$$

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

$$\|w_h\|_2 \le \|w_h-\Pi_1 w\|_2 + \|\Pi_1 w - w\|_2 + \|w\|_2 \le Ch^{-1}\,\|w_h-\Pi_1 w\|_1 + C\|w\|_2. \tag{4.25}$$ □

With (4.24), (4.25) and the elliptic regularity $\|w\|_2 \le C\,\|\theta\|_0$, we get

$$\|w_h\|_2 \le C\|w\|_2 \le C\,\|\theta\|_0 .$$

4.3. Convergence results. In this section, we initially establish the error estimate for $\|u-u_h\|_{1,\Omega}$. Subsequently, we demonstrate that the Q^1 finite element method, as given by (3.2), achieves second-order accuracy for function values.

We have the estimate of the error $\|u-u_h\|_{1,\Omega}$ as follows:

THEOREM 4.5. *Assume $a^{ij}, c \in W^{2,\infty}(\Omega)$ and $u \in H^2(\Omega), f \in H^2(\Omega)$. With elliptic regularity and V^h-ellipticity hold, we have*

$$\|u-u_h\|_{1,\Omega} = \mathcal{O}(h)\left(\|u\|_{2,\Omega} + \|f\|_{2,\Omega}\right).$$

Proof. By the First Strang Lemma,

$$\begin{aligned}\|u-u_h\|_{1,\Omega} \le & C\left(\inf_{v_h\in V^h}\left\{\|u-v_h\|_{1,\Omega} + \sup_{w_h\in V_h}\frac{|\mathcal{A}(v_h,w_h)-\mathcal{A}_h(v_h,w_h)|}{\|w_h\|_{1,\Omega}}\right\} + \right. \\ & \left. + \sup_{w_h\in V^h}\frac{|\langle f,w_h\rangle_h - (f,w_h)|}{\|w_h\|_{1,\Omega}}\right).\end{aligned} \tag{4.26}$$

By Lemma 4.2, we have:

$$\frac{|\mathcal{A}(v_h,w_h)-\mathcal{A}_h(v_h,w_h)|}{\|w_h\|_{1,\Omega}} = \frac{\mathcal{O}(h)\|v_h\|_{2,\Omega}\|w_h\|_{1,\Omega}}{\|w_h\|_{1,\Omega}} = \mathcal{O}(h)\|v_h\|_{2,\Omega}.$$

By Lemma 2.3, we have

$$\sup_{w_h\in V^h}\frac{|\langle f,w_h\rangle_h-(f,w_h)|}{\|w_h\|_{1,\Omega}}=\frac{\mathcal{O}(h^2)\|f\|_{2,\Omega}\|w_h\|_{1,\Omega}}{\|w_h\|_{1,\Omega}}=\mathcal{O}(h^2)\|f\|_{2,\Omega}.$$

By the approximation property of piece-wise Q^1 polynomials,

$$\|u-u_h\|_{1,\Omega}=\mathcal{O}(h)(\|u\|_{2,\Omega}+|f\|_{2,\Omega}).$$

□

In the following part we prove the Aubin-Nitsche Lemma up to the quadrature error for establishing convergence of function values.

THEOREM 4.6. *Assume $a^{ij},c\in W^{2,\infty}(\Omega)$ and $u(\mathbf{x})\in H^3(\Omega), f\in H^2(\Omega)$. Assume V^h ellipticity holds. Then the numerical solution from scheme (3.2) u_h is a 2-th order accurate approximation to the exact solution u:*

$$\|u_h-u\|_{0,\Omega}=\mathcal{O}\left(h^2\right)(\|u\|_{2,\Omega}+\|f\|_{2,\Omega}).$$

Proof. With $\theta=u-u_h\in H_0^1(\Omega)$, we have

$$\|\theta\|_0^2=(\theta,\theta)=A(\theta,w)=A(u-u_h,w_h)+A(u-u_h,w-w_h) \tag{4.27}$$

For the first term (4.27), by Lemma 4.1, we have

$$\begin{aligned} A(u-u_h,w_h)&=[A(u,w_h)-A_h(u_h,w_h)]+[A_h(u_h,w_h)-A(u_h,w_h)]\\ &=(f,w_h)-\langle f,w_h\rangle_h+\mathcal{O}\left(h^2\right)\|u_h\|_3\|w_h\|_2\\ &=\mathcal{O}\left(h^2\right)\|f\|_2\|w_h\|_1+\mathcal{O}\left(h^2\right)\|u_h\|_2\|w_h\|_2\\ &=\mathcal{O}\left(h^2\right)(\|f\|_2+\|u_h\|_2)\|\theta\|_0, \end{aligned} \tag{4.28}$$

where in the second last equation Lemma 2.3 and the fact the third derivative of Q^1 polynomials vanish are used. As the estimate of $\|w_h\|_2$ and $\|w\|_2$ in the proof of Lemma 4.4, we have

$$\begin{aligned} \|u_h\|_2&\le\|u_h-\Pi_1u\|_2+\|\Pi_1u-u\|_2+\|u\|_2\le Ch^{-1}\|u_h-\Pi_1u\|_1+C\|u\|_2\\ &\le Ch^{-1}(\|u-\Pi_1u\|_1+\|u-u_h\|_1)+C\|u\|_2\\ &\le Ch^{-1}\|u-u_h\|_1+C\|u\|_2\\ &\le C(\|u\|_2+\|f\|_2), \end{aligned} \tag{4.29}$$

where Theorem 4.5 is used in the last inequality. Therefore, we have

$$A(u-u_h,w_h)=\mathcal{O}\left(h^2\right)(\|f\|_2+\|u\|_2)\|\theta\|_0. \tag{4.30}$$

For the second term (4.27), by continuity of the bilinear form and Lemma 4.4, we have

$$\begin{aligned} A(u-u_h,w-w_h)&\le C\|u-u_h\|_1\|w-w_h\|_1\le Ch\|u-u_h\|_1\|w\|_2\\ &\le Ch\|u-u_h\|_1\|\theta\|_0=\mathcal{O}\left(h^2\right)(\|f\|_2+\|u\|_2)\|\theta\|_0. \end{aligned} \tag{4.31}$$

Therefore, by (4.27), (4.28) and (4.31), we have

$$\|\theta\|_0=\mathcal{O}\left(h^2\right)(\|f\|_2+\|u\|_2). \tag{4.32}$$

□

REMARK 3. *Similar convergence results for the Q^1 method on general quasi-uniform quadrilateral meshes can be established via the same proof procedure in this section.*

5. Extension to general quadrilateral meshes. For a quadrilateral element e as in Fig. 2, let $\boldsymbol{F}_e$ the mapping such that $\boldsymbol{F}_e(\hat{K}) = e$.

For $\varphi \in V_0^h$, by definition $\hat{\varphi} = \varphi|_e \circ \boldsymbol{F}_e \in Q^1(\hat{K})$. According to the chain rule, we have

$$\nabla\varphi \circ \boldsymbol{F}_e = DF_e^{T-1}\hat{\nabla}\hat{\varphi}$$

where $\varphi \circ \boldsymbol{F}_e = \hat{\varphi}, \nabla = \left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}\right)^T$, $\hat{\nabla} = \left(\frac{\partial}{\partial \hat{x}_1}, \frac{\partial}{\partial \hat{x}_2}\right)^T$.

Therefore, we have

$$\int_e \mathbf{a}\nabla u_h \cdot \nabla v_h d\mathbf{x} = \int_{\hat{K}} \left(DF_e^{-1}\hat{\mathbf{a}}DF_e^{T-1}\hat{\nabla}\hat{u}_h\right) \cdot \hat{\nabla}\hat{v}_h \left|J_e\right| d\hat{\mathbf{x}} \tag{5.1}$$

In the case of regular meshes with mesh size h, the matrix $DF_e^{-1}\hat{\mathbf{a}}DF_e^{*-1} = \frac{1}{h^2}\hat{\mathbf{a}}$ and $J_e = h^2$.

Approximate (5.1) by the mixed quadrature (2.6) with parameter $\boldsymbol{\lambda} = (\lambda^1, \lambda^2)$, i.e.

$$\int_e (\mathbf{a}\nabla u_h) \cdot \nabla v_h \ \mathrm{d}\mathbf{x} \approx \int_{\hat{K}} \left(\tilde{\mathbf{a}}\hat{\nabla}\hat{u}_h\right) \cdot \hat{\nabla}\hat{v}_h d^h_{\boldsymbol{\lambda}}\hat{\mathbf{x}} \tag{5.2}$$

where $\tilde{\mathbf{a}} = \left(\left|J_e\right| DF_e^{-1}\hat{\mathbf{a}}DF_e^{T-1}\right)(\frac{1}{2}, \frac{1}{2})$.

As in Fig. 2, denote

$$\overrightarrow{\mathbf{c}_0} = \mathbf{c}_{0,1} - \mathbf{c}_{0,0}, \quad \overrightarrow{\mathbf{c}_1} = \mathbf{c}_{1,0} - \mathbf{c}_{0,0}, \quad \overrightarrow{\mathbf{c}_2} = \mathbf{c}_{1,1} - \mathbf{c}_{1,0}, \quad \overrightarrow{\mathbf{c}_3} = \mathbf{c}_{1,1} - \mathbf{c}_{0,1}$$

and

$$\overrightarrow{\mathbf{c}_i} = (c_i^1, c_i^2)^T,\ i = 0, 1, 2, 3, \quad DF_h = DF(\frac{1}{2}, \frac{1}{2}), \quad J_{e,h} = \left|J_e\right|(\frac{1}{2}, \frac{1}{2}), \quad \bar{\mathbf{a}}_e = \hat{\mathbf{a}}_e(\frac{1}{2}, \frac{1}{2}),$$

then we have

$$DF_h = \frac{1}{2}\begin{pmatrix} c_1^1 + c_3^1 & c_0^1 + c_2^1 \\ c_1^2 + c_3^2 & c_0^2 + c_2^2 \end{pmatrix}, \quad DF_h^{-1} = \frac{1}{2det(DF_h)}\begin{pmatrix} c_0^2 + c_2^2 & -c_0^1 - c_2^1 \\ -c_1^2 - c_3^2 & c_1^1 + c_3^1 \end{pmatrix},$$

$$\tilde{\mathbf{a}} = J_{e,h}DF_h^{-1}\bar{\mathbf{a}}_e DF_h^{T-1} = \begin{pmatrix} \tilde{a}_e^{11} & \tilde{a}_e^{12} \\ \tilde{a}_e^{12} & \tilde{a}_e^{22} \end{pmatrix}. \tag{5.3}$$

To make the stiffness matrix a M-matrix, by Theorem 3.2, the following is a sufficient condition:

$$\left|\tilde{a}_e^{12}\right| \leq \min\{\tilde{a}_e^{11}, \tilde{a}_e^{22}\}. \tag{5.4}$$

While we have

$$\begin{aligned} \tilde{a}^{11} =& det(\bar{\mathbf{a}}_e)C\begin{pmatrix} c_0^2 + c_2^2 & -c_0^1 - c_2^1 \end{pmatrix}\begin{pmatrix} \bar{a}^{11} & \bar{a}^{12} \\ \bar{a}^{12} & \bar{a}^{22} \end{pmatrix}\begin{pmatrix} c_0^2 + c_2^2 \\ -c_0^1 - c_2^1 \end{pmatrix} \\ =& det(\bar{\mathbf{a}}_e)C\begin{pmatrix} c_0^1 + c_2^1 & c_0^2 + c_2^2 \end{pmatrix}\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} \bar{a}^{11} & \bar{a}^{12} \\ \bar{a}^{12} & \bar{a}^{22} \end{pmatrix}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} c_0^1 + c_2^1 \\ c_0^2 + c_2^2 \end{pmatrix} \\ =& C\begin{pmatrix} c_0^1 + c_2^1 & c_0^2 + c_2^2 \end{pmatrix}\begin{pmatrix} \bar{a}^{22} & -\bar{a}^{12} \\ -\bar{a}^{12} & \bar{a}^{11} \end{pmatrix}\begin{pmatrix} c_0^1 + c_2^1 \\ c_0^2 + c_2^2 \end{pmatrix} \\ =& C\left(\overrightarrow{\mathbf{c}_0} + \overrightarrow{\mathbf{c}_2}\right)^T \bar{\mathbf{a}}_e^{-1}\left(\overrightarrow{\mathbf{c}_0} + \overrightarrow{\mathbf{c}_2}\right), \end{aligned} \tag{5.5}$$

and similarly

$$\tilde{a}^{12} = -C\left(\overrightarrow{\mathbf{c}_0}+\overrightarrow{\mathbf{c}_2}\right)^T \bar{\mathbf{a}}_e^{-1}\left(\overrightarrow{\mathbf{c}_1}+\overrightarrow{\mathbf{c}_3}\right), \quad \tilde{a}^{22} = C\left(\overrightarrow{\mathbf{c}_1}+\overrightarrow{\mathbf{c}_3}\right)^T \bar{\mathbf{a}}_e^{-1}\left(\overrightarrow{\mathbf{c}_1}+\overrightarrow{\mathbf{c}_3}\right), \tag{5.6}$$

with $C = \frac{J_{e,h}}{4det(DF_h)^2 det(\bar{\mathbf{a}}_e)}$.

By $\overrightarrow{\mathbf{c}_1}+\overrightarrow{\mathbf{c}_2}-\overrightarrow{\mathbf{c}_3}-\overrightarrow{\mathbf{c}_0}=\overrightarrow{0}$, (5.4) is equivalent to

$$\begin{aligned}
&\left(\overrightarrow{\mathbf{c}_0}+\overrightarrow{\mathbf{c}_2}\right)^T \bar{\mathbf{a}}_e^{-1}\left(\overrightarrow{\mathbf{c}_0}+\overrightarrow{\mathbf{c}_3}\right)>0, \quad \left(\overrightarrow{\mathbf{c}_0}+\overrightarrow{\mathbf{c}_2}\right)^T \bar{\mathbf{a}}_e^{-1}\left(\overrightarrow{\mathbf{c}_0}-\overrightarrow{\mathbf{c}_1}\right)>0,\\
&\left(\overrightarrow{\mathbf{c}_1}+\overrightarrow{\mathbf{c}_3}\right)^T \bar{\mathbf{a}}_e^{-1}\left(\overrightarrow{\mathbf{c}_0}+\overrightarrow{\mathbf{c}_3}\right)>0, \quad \left(\overrightarrow{\mathbf{c}_1}+\overrightarrow{\mathbf{c}_3}\right)^T \bar{\mathbf{a}}_e^{-1}\left(\overrightarrow{\mathbf{c}_1}-\overrightarrow{\mathbf{c}_0}\right)>0.
\end{aligned} \tag{5.7}$$

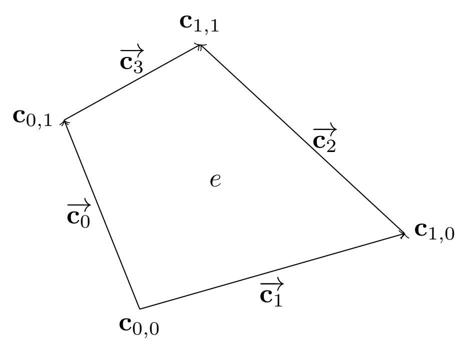

FIG. 2. *A quadrilateral element e.*

THEOREM 5.1. *If the quadrilateral mesh fulfill the condition (5.4) with $\tilde{\mathbf{a}}$ defined in (5.3) or the mesh condition (5.7), then the stiffness matrix of the linear Q^1 finite element scheme (3.2) for solving BVP (1.1) is an M-matrix.*

REMARK 4. *If the diffusion operator degenerate to Laplacian, i.e. $\mathbf{a} = \alpha(\mathbf{x})I$. A sufficient condition for (5.7) is that both diagonals of the quadrilateral element bisect each angle, resulting in two non-obtuse angles for each vertex.*

REMARK 5. *By adopting some anisotropic mesh adaptation strategy where an anisotropic mesh is generated as an M-uniform mesh or a uniform mesh in the metric specified by the diffusion matrix $\mathbf{a}$. The method (3.2) for any anistropic problem possibly can be monotone on that anisotropic mesh.*

If we consider rectangular meshes, for simplicity we assume

$$\mathbf{c}_{0,0} = (0,0), \quad \mathbf{c}_{1,0} = (h_1, 0), \quad \mathbf{c}_{0,1} = (0, h_2), \quad \mathbf{c}_{1,1} = (h_1, h_2).$$

Then we have

$$\tilde{\mathbf{a}} = \begin{pmatrix} \frac{h_2}{h_1}\bar{a}^{11} & \bar{a}^{12} \\ \bar{a}^{12} & \frac{h_1}{h_2}\bar{a}^{22} \end{pmatrix}$$

and (5.4) becomes

$$|\bar{a}_e^{12}| \le \min\{\frac{h_2}{h_1}\bar{a}_e^{11}, \frac{h_1}{h_2}\bar{a}_e^{22}\}. \tag{5.8}$$

Recall that $\sqrt{\bar{a}_e^{11}\bar{a}_e^{22}} \geq |\bar{a}_e^{12}|$, taking $\frac{h_1}{h_2} = \sqrt{\frac{\bar{a}_e^{11}}{\bar{a}_e^{22}}}$ will guarantee (5.8). Therefore, if the rectangular mesh is deployed with aspect ratio $\sqrt{\frac{\bar{a}_e^{11}}{\bar{a}_e^{22}}}$, then the stiffness matrix of the Q^1 method (3.2) is a M-matrix.

If the elliptic coefficient $\mathbf{a}$ is constant on the whole domain Ω, when the rectangular mesh are fine enough, there must exist rectangular mesh with aspect ratio approximatly $\sqrt{\frac{\bar{a}_e^{11}}{\bar{a}_e^{22}}}$ such that the stiffness matrix of scheme (3.2) solve the BVP (1.1) is an M-matrix.

6. Numerical experiment. In this section, we show an accuracy test verifying the proved order of accuracy of the scheme (3.2) on uniform meshes. We consider the following two dimensional elliptic equation:

$$-\nabla \cdot (\mathbf{a}\nabla u) + cu = f \quad \text{on } [0,\pi]^2 \tag{6.1}$$

where $\mathbf{a} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$, $a_{11} = a_{12} = a_{21} = 1 + 10x_2^2 + x_1 \cos x_2 + x_2$, $a_{22} = 2 + 10x_2^2 + x_1 \cos x_2 + x_2$, with an exact solution

$$u(x_1, x_2) = -\sin {x_1}^2 \sin x_2 \cos x_2.$$

The errors at grid points are listed in Table 1. We observe the desired second order accuracy in the discrete 2-norm and infinity norm for the function values.

Table 1
A 2D elliptic equation with Dirichlet boundary conditions. The first column is the number of elements in a finite element mesh. The second column is the number of degree of freedoms.

FEM Mesh	DoF	l^2 error	order	l^∞ error	order
4×4	3^2	4.41E-1	-	3.48E-1	-
8×8	7^2	7.20E-2	2.61	5.93E-2	2.55
16×16	15^2	1.65E-2	2.13	1.39E-2	2.09
32×32	31^2	4.03E-3	2.03	3.45E-3	2.02
64×64	63^2	1.00E-3	2.01	8.61E-4	2.00

7. Conclusion. We constructed a linear monotone Q^1 finite element method for anistropic diffusion problem (1.1). On uniform meshes, when the diffusion matrix is diagonally dominant, the M-matrix property is guaranteed thus monotonicity is achieved. When this Q^1 finite element method is deployed on a general quadrilateral mesh, we get a local mesh constraint.

REFERENCES

[1] J. Bramble, B. Hubbard, and V. Thomée, *Convergence estimates for essentially positive type discrete dirichlet problems*, Mathematics of Computation, 23 (1969), pp. 695–709.
[2] C. Cancès, M. Cathala, and C. Le Potier, *Monotone corrections for generic cell-centered finite volume approximations of anisotropic diffusion equations*, Numerische Mathematik, 125 (2013), pp. 387–417.
[3] I. Christie and C. Hall, *The maximum principle for bilinear elements*, Internat. J. Numer. Methods Engrg., 20 (1984), pp. 549–553.
[4] P. G. Ciarlet, *The finite element method for elliptic problems*, Classics in applied mathematics, 40 (2002), pp. 1–511.

[5] L. J. Cross and X. Zhang, *Monotonicity of Q^3 spectral element method for discrete Laplacian*, 2023, https://arxiv.org/abs/2010.07282.

[6] L. J. Cross and X. Zhang, *On the monotonicity of Q^2 spectral element method for Laplacian on quasi-uniform rectangular meshes*, to appear in Communications in Computational Physics, (2023).

[7] L. C. Evans, *Partial Differential Equations*, vol. 019 of Graduate Studies in Mathematics, American Mathematical Society, 2 ed., 2010.

[8] D. Greenspan and P. Jain, *On non negative difference analogues of elliptic differential equations*, Journal of the Franklin Institute, 279 (1965), pp. 360–365.

[9] J. Hu and X. Zhang, *Positivity-preserving and energy-dissipative finite difference schemes for the Fokker–Planck and Keller–Segel equations*, IMA Journal of Numerical Analysis, 43 (2022), pp. 1450–1484.

[10] W. Huang, *Discrete maximum principle and a delaunay-type mesh condition for linear finite element approximations of two-dimensional anisotropic diffusion problems*, Numerical Mathematics: Theory, Methods and Applications, 4 (2011), pp. 319–334.

[11] D. Kuzmin, M. J. Shashkov, and D. Svyatskiy, *A constrained finite element method satisfying the discrete maximum principle for anisotropic diffusion problems*, J. Comput. Phys., 228 (2009), pp. 3448–3463.

[12] C. Le Potier, *A nonlinear finite volume scheme satisfying maximum and minimum principles for diffusion operators*, International Journal on Finite Volumes, (2009), pp. 1–20.

[13] H. Li, S. Xie, and X. Zhang, *A high order accurate bound-preserving compact finite difference scheme for scalar convection diffusion equations*, SIAM Journal on Numerical Analysis, 56 (2018), pp. 3308–3345.

[14] H. Li and X. Zhang, *On the monotonicity and discrete maximum principle of the finite difference implementation of C^ 0 C 0-Q^ 2 Q 2 finite element method*, Numerische Mathematik, 145 (2020), pp. 437–472.

[15] H. Li and X. Zhang, *Superconvergence of high order finite difference schemes based on variational formulation for elliptic equations*, Journal of Scientific Computing, 82 (2020), pp. 1–39.

[16] H. Li and X. Zhang, *A high order accurate bound-preserving compact finite difference scheme for two-dimensional incompressible flow*, Communications on Applied Mathematics and Computation, (2023), pp. 1–29.

[17] X. Li and W. Huang, *An anisotropic mesh adaptation method for the finite element solution of heterogeneous anisotropic diffusion problems*, Journal of Computational Physics, 229 (2010), pp. 8072–8094.

[18] X. Li, D. Svyatskiy, and M. Shashkov, *Mesh adaptation and discrete maximum principle for 2d anisotropic diffusion problems*, tech. report, Technical Report LA-UR 10-01227, Los Alamos National Laboratory, Los Alamos, NM, 2007.

[19] K. Lipnikov, M. Shashkov, D. Svyatskiy, and Y. Vassilevski, *Monotone finite volume schemes for diffusion equations on unstructured triangular and shape-regular polygonal meshes*, Journal of Computational Physics, 227 (2007), pp. 492–512.

[20] R. Liska and M. Shashkov, *Enforcing the discrete maximum principle for linear finite element solutions of second-order elliptic problems*, Commun. Comput. Phys., 3 (2008), pp. 852–877.

[21] C. Liu, Y. Gao, and X. Zhang, *Structure preserving schemes for fokker-planck equations of irreversible processes*, to appear in Journal of Scientific Computing, (2023).

[22] C. Liu and X. Zhang, *A positivity-preserving implicit-explicit scheme with high order polynomial basis for compressible Navier–Stokes equations*, Journal of Computational Physics, 493 (2023), p. 112496.

[23] J. Lorenz, *Zur inversmonotonie diskreter probleme*, Numer. Math., 27 (1977), pp. 227–238.

[24] C. Lu, W. Huang, and J. Qiu, *Maximum principle in linear finite element approximations of anisotropic diffusion–convection–reaction problems*, Numerische Mathematik, 127 (2014), pp. 515–537.

[25] M. J. Mlacnik and L. J. Durlofsky, *Unstructured grid optimization for improved monotonicity of discrete solutions of elliptic equations with highly anisotropic coefficients*, Journal of Computational Physics, 216 (2006), pp. 337–361.

[26] T. S. Motzkin and W. Wasow, *On the approximation of linear elliptic differential equations by difference equations with positive coefficients*, Journal of Mathematics and Physics, 31 (1952), pp. 253–259.

[27] C. Ngo and W. Huang, *Monotone finite difference schemes for anisotropic diffusion problems via nonnegative directional splittings*, Communications in Computational Physics, 19 (2016), pp. 473–495.

[28] J. M. Nordbotten, I. Aavatsmark, and G. Eigestad, *Monotonicity of control volume methods*, Numerische Mathematik, 106 (2007), pp. 255–288.
[29] R. J. Plemmons, *M-matrix characterizations. I-nonsingular M-matrices*, Numer. Anal. Appl., 18 (1977), pp. 175–188.
[30] P. Sharma and G. W. Hammett, *Preserving monotonicity in anisotropic diffusion*, Journal of Computational Physics, 227 (2007), pp. 123–142.
[31] J. Shen and X. Zhang, *Discrete maximum principle of a high order finite difference scheme for a generalized Allen–Cahn equation*, Communications in Mathematical Sciences, 20 (2022), pp. 1409–1436.
[32] J. Weickert et al., *Anisotropic diffusion in image processing*, vol. 1, Teubner Stuttgart, 1998.
[33] J. Xu and L. Zikatanov, *A monotone finite element scheme for convection-diffusion equations*, Math. Comp., 68 (1999), pp. 1429–1446.
[34] G. Yuan and Z. Sheng, *Monotone finite volume schemes for diffusion equations on polygonal meshes*, Journal of computational physics, 227 (2008), pp. 6288–6312.