

Maximum principle in linear finite element approximations of anisotropic diffusion–convection–reaction problems

Changna Lu · Weizhang Huang · Jianxian Qiu

Received: 12 January 2012 / Revised: 3 March 2013 / Published online: 6 November 2013
© Springer-Verlag Berlin Heidelberg 2013

Abstract A mesh condition is developed for linear finite element approximations of anisotropic diffusion–convection–reaction problems to satisfy a discrete maximum principle. Loosely speaking, the condition requires that the mesh be simplicial and $\mathcal{O}(\|\mathbf{b}\|_\infty h + \|c\|_\infty h^2)$ -nonobtuse when the dihedral angles are measured in the metric specified by the inverse of the diffusion matrix, where h denotes the mesh size and \mathbf{b} and c are the coefficients of the convection and reaction terms. In two dimensions, the condition can be replaced by a weaker mesh condition (an $\mathcal{O}(\|\mathbf{b}\|_\infty h + \|c\|_\infty h^2)$ perturbation of a generalized Delaunay condition). These results include many existing mesh conditions as special cases. Numerical results are presented to verify the theoretical findings.

Mathematics Subject Classification 65N30 · 65N50

The work was supported in part by the National Science Foundation (USA) under Grant DMS-1115118 and the Natural Science Foundation of China under Grants 10931004 and 40906048. The authors would like to thank the anonymous referees for their valuable comments and suggestions.

C. Lu (✉)
College of Math and Statistics, Nanjing University of Information Science and Technology,
Nanjing 210044, Jiangsu, China
e-mail: luchangna@nuist.edu.cn

W. Huang
Department of Mathematics, The University of Kansas, Lawrence, KS 66045, USA
e-mail: huang@math.ku.edu

J. Qiu
School of Mathematical Sciences, Xiamen University, Xiamen 361005, Fujian, China
e-mail: jxqiu@xmu.edu.cn

1 Introduction

We are concerned with the linear finite element (FEM) solution of the anisotropic diffusion equation

$$-\nabla \cdot (\mathbb{D} \nabla u) + \mathbf{b} \cdot \nabla u + c u = f, \quad \text{in } \Omega \quad (1)$$

subject to the Dirichlet boundary condition

$$u = g, \quad \text{on } \partial\Omega \quad (2)$$

where $\Omega \subset \mathbb{R}^d$ ($d \geq 1$) is a connected polyhedron and $\mathbb{D} = \mathbb{D}(\mathbf{x}) \in \mathbb{R}^{d \times d}$ (the diffusion matrix), $\mathbf{b} = \mathbf{b}(\mathbf{x}) \in \mathbb{R}^d$, $c = c(\mathbf{x})$, $f = f(\mathbf{x})$, and $g = g(\mathbf{x})$ are given, sufficiently smooth functions defined on Ω . We assume that for any $\mathbf{x} \in \Omega$, $\mathbb{D}(\mathbf{x})$ is symmetric and strictly positive definite and functions \mathbf{b} and c satisfy

$$c(\mathbf{x}) - \frac{1}{2} \nabla \cdot \mathbf{b}(\mathbf{x}) \geq 0, \quad c(\mathbf{x}) \geq 0, \quad \forall \mathbf{x} \in \Omega. \quad (3)$$

It is known (e.g., see [8]) that the solution of the boundary value problem (BVP) (1) and (2) satisfies the maximum principle.

The numerical solution of BVP (1) and (2) has attracted considerable attention from scientists and engineers. The BVP is a prototype model for anisotropic diffusion problems which arise in various fields such as plasma physics [10, 11, 28], petroleum reservoir simulation [1, 26], and image processing [27, 32]. Moreover, it has been amply demonstrated that a standard numerical method, such as a finite element, a finite difference, or a finite volume method, does not necessarily satisfy a discrete maximum principle (DMP) and may produce unphysical solutions that typically contain spurious oscillations, undershoots, and overshoots. Furthermore, designing a numerical scheme to preserve the maximum principle is an important research topic in its own right. As a matter of fact, considerable work has been done in the past to develop numerical schemes to satisfy DMP; e.g., see [2–4, 6, 16, 17, 19, 21, 29–31, 33] for isotropic diffusion problems ($\mathbb{D} = \alpha(\mathbf{x})I$ with $\alpha(\mathbf{x})$ being a scalar function) and [7, 10, 11, 14, 18, 20, 22–26, 28] for anisotropic diffusion problems. In particular, it is shown in [6] that the linear FEM satisfies DMP when the mesh is simplicial and satisfies the so-called non-obtuse angle condition which requires that the dihedral angles of all mesh elements be non-obtuse. In two dimensions the condition can be replaced by a weaker condition (the Delaunay condition) which requires that the sum of any pair of angles opposite a common edge is less than or equal to π [30]. Similar results have been obtained recently for anisotropic diffusion problems in [14, 22].

It is pointed out that most of the existing work has been concerned with problems without convection terms. For continuous problems, it is known (e.g., see [8]) that convection terms have no effect on the satisfaction of the maximum principle by the solution. However, the situation is different for discrete schemes. The main difficulty comes from the fact that discrete convection terms typically do not vanish at an interior maximum point and the entries of the corresponding matrix can be both positive and negative. A few researchers have tried to address the issue for isotropic diffusion

problems. For example, Xu and Zikatanov [33] employ a special number treatment for convection terms so that they have no effect on the DMP satisfaction by the discrete solution. Burman and Ern [3] propose a nonlinear stabilized Galerkin approximation of the Laplace operator which satisfies DMP on arbitrary meshes and for arbitrary space dimension without resorting to the non-obtuse angle condition. They prove that the result can extend to diffusion–convection–reaction problems with constant diffusion coefficient when the mesh is locally quasi-uniform. More recently, Wang and Zhang [31] study quasilinear isotropic diffusion–convection–reaction problems and show that linear finite element approximations satisfy DMP when the mesh is $\mathcal{O}(\|\mathbf{b}\|_\infty h + \|c\|_\infty h^2)$ -acute (i.e., the dihedral angles of all mesh elements are less than or equal to $\frac{\pi}{2} - \gamma_1 \|\mathbf{b}\|_\infty h - \gamma_2 \|c\|_\infty h^2$ for some positive constants γ_1 and γ_2). On the other hand, no work has been done for anisotropic diffusion–convection–reaction problems.

The objective of this paper is to develop a mesh condition for linear finite element approximations of anisotropic diffusion–convection–reaction problems (1) and (2) in any dimension to satisfy a discrete maximum principle. We shall use the approach of [22] to show the stiffness matrix associated with the linear finite element discretization to be an M -matrix and have non-negative row sums, with the focus on the treatments of the convection and reaction terms. We shall also investigate the two dimensional case where a weaker sufficient condition can be developed.

The paper is organized as follows. A linear finite element discretization for BVP (1) and (2) is introduced in Sect. 2. In Sect. 3 geometric properties of the gradient of linear basis functions are studied. A general mesh condition valid in any dimension and a specific and weaker condition in two dimensions are developed in Sect. 4, followed by numerical results in Sect. 5. Finally, Sect. 6 contains conclusions.

2 Linear finite element formulation

We consider the linear finite element solution of BVP (1) and (2). Assume that an affine family of simplicial meshes $\{\mathcal{T}_h\}$ is given for Ω . Let

$$U_g = \{v \in H^1(\Omega) \mid v|_{\partial\Omega} = g\}.$$

Denote by $U_{g^h}^h$ the linear finite element space associated with mesh \mathcal{T}_h , where g^h is a piecewise linear approximation to g on the boundary. A linear finite element approximation $u^h \in U_{g^h}^h$ to BVP (1) and (2) is defined by

$$\begin{aligned} & \int_{\Omega} (\nabla v^h)^T \mathbb{D} \nabla u^h dx + \int_{\Omega} v^h (\mathbf{b} \cdot \nabla u^h) dx + \int_{\Omega} c u^h v^h dx \\ & = \int_{\Omega} f v^h dx, \quad \forall v^h \in U_0^h. \end{aligned} \tag{4}$$

The above equation can be rewritten as

$$\begin{aligned} & \sum_{K \in \mathcal{T}_h} |K| (\nabla v^h)^T \mathbb{D}_K \nabla u^h + \sum_{K \in \mathcal{T}_h} \int_K v^h (\mathbf{b} \cdot \nabla u^h) dx \\ & + \sum_{K \in \mathcal{T}_h} \int_K c u^h v^h dx = \sum_{K \in \mathcal{T}_h} \int_K f v^h dx, \quad \forall v^h \in U_0^h \end{aligned} \tag{5}$$

where $|K|$ is the volume of element K and \mathbb{D}_K is the integral average of \mathbb{D} over K , viz.,

$$\mathbb{D}_K = \frac{1}{|K|} \int_K \mathbb{D} dx. \tag{6}$$

Scheme (5) can be expressed in a matrix form. Denote the numbers of the elements, vertices, and interior vertices of mesh \mathcal{T}_h by N , N_v , and N_{vi} , respectively. Assume that the vertices are ordered in such a way that the first N_{vi} vertices are the interior vertices. Then U_0^h and u^h can be expressed as

$$U_0^h = \text{span}\{\phi_1, \dots, \phi_{N_{vi}}\}, \tag{7}$$

$$u^h = \sum_{j=1}^{N_{vi}} u_j \phi_j + \sum_{j=N_{vi}+1}^{N_v} u_j \phi_j, \tag{8}$$

where ϕ_j denotes the linear basis function associated with the j th vertex, \mathbf{a}_j . The boundary condition (2) is approximated by

$$u_j = g(\mathbf{a}_j), \quad j = N_{vi} + 1, \dots, N_v. \tag{9}$$

Substituting (8) into (5), taking $v^h = \phi_j$ ($j = 1, \dots, N_{vi}$), and combining the resulting equations with (9), we obtain the linear algebraic system

$$A \mathbf{u} = \mathbf{f}, \tag{10}$$

where $\mathbf{u} = (u_1, \dots, u_{N_{vi}}, u_{N_{vi}+1}, \dots, u_{N_v})^T$, $\mathbf{f} = (f_1, \dots, f_{N_{vi}}, g_{N_{vi}+1}, \dots, g_{N_v})^T$,

$$A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & I \end{bmatrix}, \tag{11}$$

and I is the identity matrix of size $(N_v - N_{vi})$. The entries of the stiffness matrix A and the right-hand-side vector f are given by

$$a_{ij} = \sum_{K \in \mathcal{T}_h} |K| (\nabla \phi_i)^T \mathbb{D}_K \nabla \phi_j + \sum_{K \in \mathcal{T}_h} \int_K \phi_i (\mathbf{b} \cdot \nabla \phi_j) dx + \sum_{K \in \mathcal{T}_h} \int_K c \phi_j \phi_i dx, \quad i = 1, \dots, N_{vi}, \quad j = 1, \dots, N_v, \tag{12}$$

$$f_i = \sum_{K \in \mathcal{T}_h} \int_K f \phi_i dx, \quad i = 1, \dots, N_{vi}. \tag{13}$$

In the following sections we shall investigate under what condition on the mesh the solution of (5) satisfies a maximum principle. A key to this investigation is to understand geometric properties of the gradient of linear basis functions which are to be described in the next section.

3 Geometric properties of the gradient of linear basis functions

Let K be an arbitrary simplex with vertices $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_{d+1}$. Denote the face opposite to vertex \mathbf{a}_i (i.e. the face not having \mathbf{a}_i as its vertex) by S_i and its unit inward (pointing to \mathbf{a}_i) normal by \mathbf{n}_i . The distance (or height) from vertex \mathbf{a}_i to face S_i is denoted by h_i . The result of the following lemma exists in literature; e.g., see [2, 19, 33].

Lemma 1 *For any simplex $K \in \mathbb{R}^d$, the gradient of linear basis function ϕ_i associated any vertex \mathbf{a}_i ($i = 1, \dots, d + 1$) is given by*

$$\nabla \phi_i = \frac{\mathbf{n}_i}{h_i}. \tag{14}$$

It is remarked that Brandts et al. [2] have obtained the same result using the so-called \mathbf{q} -vectors defined through the edge matrix of elements. Specifically, they show that \mathbf{q}_i , a \mathbf{q} -vector associated with face S_i , is an inward normal to S_i , has the length $1/h_i$, and is equal to $\nabla \phi_i$; i.e.,

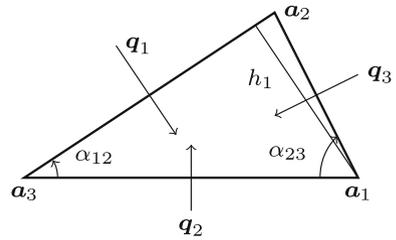
$$\mathbf{q}_i = \frac{1}{h_i} \mathbf{n}_i = \nabla \phi_i, \quad i = 1, \dots, d + 1. \tag{15}$$

These \mathbf{q} -vectors will be used frequently in the remaining of the paper.

The next property of gradient of linear basis functions is related to the diffusion term in stiffness matrix (12) for the case $\mathbb{D}_K = I$. Denote the dihedral angle between any two faces S_i and S_j ($i \neq j$) by α_{ij} . It can be calculated as the supplement of the angle between the inward normals to the faces, i.e.,

$$\cos(\alpha_{ij}) = -\mathbf{n}_i \cdot \mathbf{n}_j = -\frac{\mathbf{q}_i \cdot \mathbf{q}_j}{\|\mathbf{q}_i\| \cdot \|\mathbf{q}_j\|}, \quad i \neq j. \tag{16}$$

Fig. 1 A sketch of unit inward normals, dihedral angles, and heights of element K



(In fact, (16) is often used as the definition of the dihedral angle.) A sketch of the q -vectors, dihedral angles, and heights of an element are shown in Fig. 1.

The result of the following lemma is also known in literature; e.g., see [2, 9, 14].

Lemma 2 For any simplex $K \in \mathbb{R}^d$, we have

$$|K|(\nabla\phi_i)^T \nabla\phi_j = -\frac{|K|}{h_i h_j} \cos(\alpha_{ij}), \quad i \neq j. \tag{17}$$

It reduces to

$$|K|(\nabla\phi_i)^T \nabla\phi_j = -\frac{1}{2} \cot(\alpha_{ij}), \quad i \neq j \tag{18}$$

in two dimensions.

Proof Equation (17) follows from Lemma 1 and (16).

In two dimensions, K is a triangle. Consider the case with $i = 1$ and $j = 2$. From Fig. 1, we have

$$|K| = \frac{h_1}{2} \|a_2 - a_3\| = \frac{h_1 h_2}{2 \sin(\alpha_{12})}.$$

Combining this result and (17) gives (18). □

We now study the diffusion term $|K|(\nabla\phi_i)^T \mathbb{D}_K \nabla\phi_j$ for general symmetric and positive definite matrix \mathbb{D}_K using Lemma 2. Define

$$G_K(x) = \mathbb{D}_K^{-\frac{1}{2}} x : K \rightarrow \tilde{K}, \tag{19}$$

where $\tilde{K} = G(K)$. Obviously, \tilde{K} is also a simplex in \mathbb{R}^d . For any vertex a_i , we denote the corresponding vertex, face, height, and q -vector of \tilde{K} by $\tilde{a}_i, \tilde{S}_i, \tilde{h}_i$, and \tilde{q}_i , respectively. We have

$$\begin{cases} \tilde{a}_i = \mathbb{D}_K^{-\frac{1}{2}} a_i, & \tilde{S}_i = \mathbb{D}_K^{-\frac{1}{2}} S_i, & |\tilde{K}| = \det(\mathbb{D}_K)^{-\frac{1}{2}} |K|, \\ \tilde{q}_i = \mathbb{D}_K^{\frac{1}{2}} q_i, & \tilde{h}_i = \|q_i\|_{\mathbb{D}_K}^{-1}, \end{cases} \tag{20}$$

where $\|\cdot\|_{\mathbb{D}_K}$ denotes the distance measured in the metric \mathbb{D}_K . The derivations of the first three relations are trivial. To derive the last two, we first notice that

$$\phi_i(\mathbf{x}) = \phi_i\left(\mathbb{D}_K^{\frac{1}{2}}\tilde{\mathbf{x}}\right) = \tilde{\phi}_i(\tilde{\mathbf{x}}).$$

Then from (15) we have

$$\tilde{\mathbf{q}}_i = \tilde{\nabla}\tilde{\phi}_i = \mathbb{D}_K^{\frac{1}{2}}\nabla\phi_i = \mathbb{D}_K^{\frac{1}{2}}\mathbf{q}_i,$$

which gives the second last relation in (20). The last relation is obtained by taking the norm of the above equation.

To obtain the relation between h_i and \tilde{h}_i , we rewrite the last relation in (20) as

$$\tilde{h}_i = \frac{1}{\sqrt{(\mathbf{q}_i)^T \mathbb{D}_K \mathbf{q}_i}},$$

from which we obtain

$$\frac{h_i}{\sqrt{\lambda_{max}(\mathbb{D}_K)}} \leq \tilde{h}_i \leq \frac{h_i}{\sqrt{\lambda_{min}(\mathbb{D}_K)}}, \tag{21}$$

where $\lambda_{max}(\mathbb{D}_K)$ and $\lambda_{min}(\mathbb{D}_K)$ denote the maximum and minimum eigenvalues of \mathbb{D}_K , respectively.

Denote the dihedral angle between faces \tilde{S}_i and \tilde{S}_j by $\alpha_{ij, \mathbb{D}_K^{-1}}$. Since $\tilde{S}_i = \mathbb{D}_K^{-\frac{1}{2}}S_i$ and $\tilde{S}_j = \mathbb{D}_K^{-\frac{1}{2}}S_j$, it can also be viewed as the dihedral angle between S_i and S_j measured in the metric \mathbb{D}_K^{-1} . Moreover, from (16) we see that the angle can be calculated by

$$\cos\left(\alpha_{ij, \mathbb{D}_K^{-1}}\right) = -\frac{\tilde{\mathbf{q}}_i \cdot \tilde{\mathbf{q}}_j}{\|\tilde{\mathbf{q}}_i\| \cdot \|\tilde{\mathbf{q}}_j\|} = -\frac{\mathbf{q}_i^T \mathbb{D}_K \mathbf{q}_j}{\|\mathbf{q}_i\|_{\mathbb{D}_K} \|\mathbf{q}_j\|_{\mathbb{D}_K}}. \tag{22}$$

We now go back to the quantity $|K|(\nabla\phi_i)^T \mathbb{D}_K \nabla\phi_j$. Notice that

$$|K|(\nabla\phi_i)^T \mathbb{D}_K \nabla\phi_j = \det(\mathbb{D}_K)^{\frac{1}{2}} |\tilde{K}| (\tilde{\mathbf{q}}_i)^T \tilde{\mathbf{q}}_j.$$

Applying Lemma 2 to \tilde{K} and using relations (20), we have the following lemma.

Lemma 3 *For any simplex $K \in \mathbb{R}^d$ and any symmetric and positive definite matrix \mathbb{D}_K , we have*

$$|K|(\nabla\phi_i)^T \mathbb{D}_K \nabla\phi_j = -\frac{|\tilde{K}| \det(\mathbb{D}_K)^{\frac{1}{2}}}{\tilde{h}_i \tilde{h}_j} \cos\left(\alpha_{ij, \mathbb{D}_K^{-1}}\right), \quad i \neq j. \tag{23}$$

It reduces to

$$|K|(\nabla\phi_i)^T \mathbb{D}_K \nabla\phi_j = -\frac{\det(\mathbb{D}_K)^{\frac{1}{2}}}{2} \cot\left(\alpha_{ij, \mathbb{D}_K^{-1}}\right), \quad i \neq j \tag{24}$$

in two dimensions.

4 Mesh conditions for DMP satisfaction

In this section we study the mesh conditions under which the linear finite element scheme (5) satisfies DMP. The main conclusions are given in Theorems 1 and 2.

Theorem 1 *If the mesh satisfies*

$$\frac{h_i^K}{\lambda_{\min}(\mathbb{D}_K)} \cdot \frac{\|\mathbf{b}\|_{\infty, K}}{(d+1)} + \frac{h_i^K h_j^K}{\lambda_{\min}(\mathbb{D}_K)} \cdot \frac{\|c\|_{\infty, K}}{(d+1)(d+2)} \leq \cos\left(\alpha_{ij, \mathbb{D}_K^{-1}}\right),$$

$$i, j = 1, \dots, d+1, \quad i \neq j, \quad \forall K \in \mathcal{T}_h \tag{25}$$

where $\|\mathbf{b}\|_{\infty, K} = \max_{\mathbf{x} \in K} \|\mathbf{b}(\mathbf{x})\|$, $\|c\|_{\infty, K} = \max_{\mathbf{x} \in K} c(\mathbf{x})$, and h_i^K 's and $\alpha_{ij, \mathbb{D}_K^{-1}}$'s are the heights and dihedral angles of element K , respectively, then the linear finite element scheme (5) for BVP (1) and (2) satisfies DMP.

Proof Following [22] we prove this theorem by showing that stiffness matrix A defined in (11) and (12) has non-negative row sums and is an M-matrix.¹ From Stoyan [29, Theorem 1], this implies that scheme (5) satisfies DMP.

- (1) We first show that matrix A has non-negative row sums. Notice that we only need to show that the first N_{vi} row sums are non-negative. Using the fact $\sum_{j=1}^{N_v} \phi_j(\mathbf{x}) = 1$ and the assumption $c \geq 0$ (cf. (3)), from (12) we have, for $i = 1, \dots, N_{vi}$,

$$\begin{aligned} \sum_{j=1}^{N_v} a_{ij} &= \sum_{K \in \mathcal{T}_h} |K| (\nabla\phi_i)^T \mathbb{D}_K \nabla \left(\sum_{j=1}^{N_v} \phi_j \right) \\ &\quad + \sum_{K \in \mathcal{T}_h} \int_K \phi_i \left(\mathbf{b} \cdot \nabla \left(\sum_{j=1}^{N_v} \phi_j \right) \right) dx \\ &\quad + \sum_{K \in \mathcal{T}_h} \int_K c \phi_i \left(\sum_{j=1}^{N_v} \phi_j \right) dx \end{aligned}$$

¹ Matrix A is called an M-matrix if it is a Z-matrix (see (27) and (28) below) and satisfies $A^{-1} \geq 0$ (i.e., all entries of its inverse are nonnegative).

$$\begin{aligned}
 &= \sum_{K \in \mathcal{T}_h} \int_K c \phi_i dx \\
 &\geq 0.
 \end{aligned}
 \tag{26}$$

(2) Next, we show that A is a Z-matrix; i.e.,

$$a_{ij} \leq 0, \quad \forall i \neq j, \quad i = 1, \dots, N_{vi}, \quad j = 1, \dots, N_v \tag{27}$$

$$a_{ii} \geq 0, \quad i = 1, \dots, N_{vi}. \tag{28}$$

Recall from Ciarlet [5, Page 201] that

$$\int_{K \in \omega_i} \phi_i dx = \frac{|K|}{d+1}, \quad \int_{K \in \omega_i \cap \omega_j} \phi_i \phi_j dx = \frac{|K|}{(d+1)(d+2)}, \tag{29}$$

where ω_i and ω_j are the element patches associated with vertices \mathbf{a}_i and \mathbf{a}_j , respectively. We have

$$\begin{aligned}
 a_{ij} &= \sum_{K \in \omega_i \cap \omega_j} \left(|K| (\nabla \phi_i)^T \mathbb{D}_K \nabla \phi_j + \int_K \phi_i (\mathbf{b} \cdot \nabla \phi_j) dx + \int_K c \phi_i \phi_j dx \right) \quad (\text{from (12)}) \\
 &\leq \sum_{K \in \omega_i \cap \omega_j} \left(|K| (\nabla \phi_i)^T \mathbb{D}_K \nabla \phi_j + \frac{1}{h_j^K} \int_K \phi_i |\mathbf{b} \cdot \mathbf{n}_j^K| dx + \int_K c \phi_i \phi_j dx \right) \quad (\text{Lemma 1}) \\
 &\leq \sum_{K \in \omega_i \cap \omega_j} \left(|K| (\nabla \phi_i)^T \mathbb{D}_K \nabla \phi_j + \frac{|K| \|\mathbf{b}\|_{\infty, K}}{h_j^K (d+1)} + \frac{|K| \|c\|_{\infty, K}}{(d+1)(d+2)} \right) \quad (\text{from (29)}) \\
 &= \sum_{K \in \omega_i \cap \omega_j} \left(-\frac{|K|}{\tilde{h}_i^K \tilde{h}_j^K} \cos(\alpha_{ij, \mathbb{D}_K^{-1}}) + \frac{|K| \|\mathbf{b}\|_{\infty, K}}{h_j^K (d+1)} + \frac{|K| \|c\|_{\infty, K}}{(d+1)(d+2)} \right) \quad (\text{Lemma 3}) \\
 &= \sum_{K \in \omega_i \cap \omega_j} \frac{|K|}{\tilde{h}_i^K \tilde{h}_j^K} \left(-\cos(\alpha_{ij, \mathbb{D}_K^{-1}}) + \frac{\tilde{h}_i^K \tilde{h}_j^K \|\mathbf{b}\|_{\infty, K}}{h_j^K (d+1)} + \frac{\tilde{h}_i^K \tilde{h}_j^K \|c\|_{\infty, K}}{(d+1)(d+2)} \right) \\
 &\leq \sum_{K \in \omega_i \cap \omega_j} \frac{|K|}{\tilde{h}_i^K \tilde{h}_j^K} \left(-\cos(\alpha_{ij, \mathbb{D}_K^{-1}}) + \frac{h_i^K \|\mathbf{b}\|_{\infty, K}}{\lambda_{\min}(\mathbb{D}_K)(d+1)} \right. \\
 &\quad \left. + \frac{h_i^K h_j^K \|c\|_{\infty, K}}{\lambda_{\min}(\mathbb{D}_K)(d+1)(d+2)} \right) \quad (\text{from (21)}).
 \end{aligned}$$

Combining this with (25) implies (27).

On the other hand, for $i = 1, \dots, N_{vi}$,

$$\begin{aligned} a_{ii} &= \sum_{K \in \mathcal{T}_h} |K| (\nabla \phi_i)^T \mathbb{D}_K \nabla \phi_i + \int_{\Omega} \phi_i (\mathbf{b} \cdot \nabla \phi_i) dx + \int_{\Omega} c \phi_i^2 dx \quad (\text{from (12)}) \\ &\geq \int_{\Omega} \phi_i (\mathbf{b} \cdot \nabla \phi_i) dx + \int_{\Omega} c \phi_i^2 dx \\ &= \int_{\Omega} \left(c - \frac{1}{2} \nabla \cdot \mathbf{b} \right) \phi_i^2 dx \quad (\text{Gauss' divergence thm}). \end{aligned}$$

The assumption (3) implies that $a_{ii} \geq 0$. Thus, the stiffness matrix A is a Z-matrix.

- (3) We now show that A_{11} , the northwest block of matrix A , is an M-matrix. This is done by showing A_{11} is positive definite. For any vector $\mathbf{v} = (v_1, v_2, \dots, v_{N_{vi}})^T$, we define $v^h = \sum_{i=1}^{N_{vi}} v_i \phi_i \in U_0$. Notice that ∇v^h is constant on K . As in the proof for $a_{ii} \geq 0$, from (12) we have

$$\begin{aligned} \mathbf{v}^T A_{11} \mathbf{v} &= \sum_{K \in \mathcal{T}_h} |K| (\nabla v^h)^T \mathbb{D}_K \nabla v^h + \int_{\Omega} v^h (\mathbf{b} \cdot \nabla v^h) dx + \int_{\Omega} c (v^h)^2 dx \\ &\geq \sum_{K \in \mathcal{T}_h} |K| (\nabla v^h)^T \mathbb{D}_K \nabla v^h + \int_{\Omega} \left(c - \frac{1}{2} \nabla \cdot \mathbf{b} \right) (v^h)^2 dx \geq 0. \end{aligned}$$

Moreover, from the above inequality, $\mathbf{v}^T A_{11} \mathbf{v} = 0$ implies $v^h = \text{constant}$, which in turn implies $v^h = 0$ due to the fact that $v^h \in U_0$. From these, we know that A_{11} is positive definite. Since A_{11} is a Z-matrix, so it is an M-matrix.

- (4) Finally, we show matrix A is an M-matrix by showing the inverse of A is positive. From (11), the inverse of A is given by

$$A^{-1} = \begin{bmatrix} A_{11}^{-1} & -A_{11}^{-1} A_{12} \\ 0 & I \end{bmatrix}. \tag{30}$$

Using the fact that $A_{11}^{-1} \geq 0$ and $A_{12} \leq 0$, then $A^{-1} \geq 0$, which, together with the fact that A is a Z-matrix, implies that A is an M-matrix. □

Remark 1 Loosely speaking, (25) requires

$$\cos \left(\alpha_{ij, \mathbb{D}_K^{-1}} \right) \geq \mathcal{O}(h \|\mathbf{b}\|_{\infty}) + \mathcal{O}(h^2 \|\mathbf{c}\|_{\infty}) \tag{31}$$

or

$$0 < \alpha_{ij, \mathbb{D}_K^{-1}} \leq \frac{\pi}{2} - \mathcal{O}(h \|\mathbf{b}\|_{\infty}) - \mathcal{O}(h^2 \|\mathbf{c}\|_{\infty}) \tag{32}$$

for all dihedral angles, where $h = \max_{K \in \mathcal{T}_h} h_K$ is the maximum element size. In other words, if the mesh is $\mathcal{O}(h)$ -acute in the metric \mathbb{D}^{-1} for the case $\mathbf{b} \neq 0$ or $\mathcal{O}(h^2)$ -acute

in the metric \mathbb{D}^{-1} for the case $\mathbf{b} \equiv 0$ and $c \neq 0$, then the linear finite element solution of (1) and (2) satisfies a DMP. \square

Remark 2 When no convection and reaction terms are involved, condition (25) reduces to the nonobtuse angle condition of [6] and the anisotropic nonobtuse angle condition of [22] for isotropic and anisotropic diffusion problems, respectively. Moreover, the condition is consistent with the DMP conditions obtained by Brandts et al. [2] and Wang and Zhang [31] for isotropic diffusion–reaction or diffusion–convection–reaction problems. Thus, (25) can be viewed as a generalization of those existing results to anisotropic diffusion–convection–reaction problems.

Remark 3 Condition (25) can be rewritten into a more friendly form to mesh generation. Indeed, combining (22) with (25) we have

$$\begin{aligned} & \frac{h_i^K}{\lambda_{\min}(\mathbb{D}_K)} \cdot \frac{\|\mathbf{b}\|_{\infty,K}}{(d+1)} + \frac{h_i^K h_j^K}{\lambda_{\min}(\mathbb{D}_K)} \cdot \frac{\|c\|_{\infty,K}}{(d+1)(d+2)} \\ & \leq - \frac{\mathbf{q}_i^T \mathbb{D}_K \mathbf{q}_j}{\sqrt{\mathbf{q}_i^T \mathbb{D}_K \mathbf{q}_i} \sqrt{\mathbf{q}_j^T \mathbb{D}_K \mathbf{q}_j}}. \end{aligned} \tag{33}$$

Denote the reference element and its \mathbf{q} -vectors by \hat{K} and $\hat{\mathbf{q}}_i$ ($i = 1, \dots, d+1$), respectively. Let F_K and F'_K be the affine mapping from \hat{K} to K and its Jacobian matrix. As for (20), it is not difficult to show that

$$\mathbf{q}_i = (F'_K)^{-1} \hat{\mathbf{q}}_i, \quad i = 1, \dots, d+1.$$

Inserting this into (33) leads to

$$\begin{aligned} & \frac{h_i^K}{\lambda_{\min}(\mathbb{D}_K)} \cdot \frac{\|\mathbf{b}\|_{\infty,K}}{(d+1)} + \frac{h_i^K h_j^K}{\lambda_{\min}(\mathbb{D}_K)} \cdot \frac{\|c\|_{\infty,K}}{(d+1)(d+2)} \\ & \leq - \frac{\hat{\mathbf{q}}_i^T (F'_K)^{-T} \mathbb{D}_K (F'_K)^{-1} \hat{\mathbf{q}}_j}{(\hat{\mathbf{q}}_i^T (F'_K)^{-T} \mathbb{D}_K (F'_K)^{-1} \hat{\mathbf{q}}_i)^{\frac{1}{2}} (\hat{\mathbf{q}}_j^T (F'_K)^{-T} \mathbb{D}_K (F'_K)^{-1} \hat{\mathbf{q}}_j)^{\frac{1}{2}}}. \end{aligned} \tag{34}$$

We now consider so-called M -uniform meshes for a given tensor $M = M(\mathbf{x})$ which is assumed to be a symmetric and positive definite $d \times d$ matrix for any $\mathbf{x} \in \bar{\Omega} = \Omega \cup \partial\Omega$. These meshes are approximately uniform in the metric specified by M . They are defined (e.g., see [13, 15]) as meshes satisfying

$$(F'_K)^{-T} M_K^{-1} (F'_K)^{-1} = \left(\frac{\sigma_h}{N}\right)^{-\frac{2}{d}} I, \quad \forall K \in \mathcal{T}_h \tag{35}$$

where

$$M_K = \frac{1}{|K|} \int_K M(\mathbf{x}) d\mathbf{x}, \quad \sigma_h = \sum_K |K| \sqrt{\det(M_K)},$$

and $\det(M_K)$ is the determinant of M_K . M -uniform meshes (or more practically, almost M -uniform meshes) can be generated using a variety of techniques including blue refinement, directional refinement, Delaunay-type triangulation, front advancing, bubble packing, local refinement and modification, and variational mesh generation; see references in [22].

When the metric tensor is chosen as $M_K = \mathbb{D}_K^{-1}$ or, more generally,

$$M_K = \theta_K \mathbb{D}_K^{-1}, \tag{36}$$

where θ_K is a scalar, piecewise constant function, the corresponding M -uniform meshes will be referred to as \mathbb{D}^{-1} -uniform meshes. For those meshes, (35) becomes

$$(F'_K)^{-T} \mathbb{D}_K (F'_K)^{-1} = \theta_K \left(\frac{\sigma_h}{N}\right)^{-\frac{2}{d}} I, \quad \forall K \in \mathcal{T}_h.$$

Inserting this into (34), we get

$$\frac{h_i^K}{\lambda_{\min}(\mathbb{D}_K)} \cdot \frac{\|\mathbf{b}\|_{\infty, K}}{(d+1)} + \frac{h_i^K h_j^K}{\lambda_{\min}(\mathbb{D}_K)} \cdot \frac{\|c\|_{\infty, K}}{(d+1)(d+2)} \leq \cos(\hat{\alpha}_{ij}),$$

where $\hat{\alpha}_{ij}$ is a dihedral angle of \hat{K} . If the reference element \hat{K} is chosen as a regular d -dimensional simplex, we have $\cos(\hat{\alpha}_{ij}) = 1/d$, and the above inequality becomes

$$\frac{h_i^K}{\lambda_{\min}(\mathbb{D}_K)} \cdot \frac{\|\mathbf{b}\|_{\infty, K}}{(d+1)} + \frac{h_i^K h_j^K}{\lambda_{\min}(\mathbb{D}_K)} \cdot \frac{\|c\|_{\infty, K}}{(d+1)(d+2)} \leq \frac{1}{d}. \tag{37}$$

This inequality holds if the maximum element size satisfies

$$h \|\mathbf{b}\|_{\infty} + \frac{1}{d+2} h^2 \|c\|_{\infty} \leq \frac{d+1}{d} \min_{\mathbf{x} \in \Omega \cup \partial\Omega} \lambda_{\min}(\mathbb{D}(\mathbf{x})). \tag{38}$$

It should be emphasized that (38) is generally too conservative to be useful in practical computation. However, it does show that *when the reference element is chosen as a regular d -dimensional simplex, a sufficiently fine (with h satisfying (38)), \mathbb{D}^{-1} -uniform mesh satisfies the condition (25). In other words, a mesh satisfying (25) can be obtained by refining a \mathbb{D}^{-1} -uniform mesh.* □

It is known that the acute or nonobtuse angle condition can be replaced by the weaker, so-called Delaunay condition in two dimensions for a linear finite element solution to satisfy a DMP; e.g., see Strang and Fix [30] for the anisotropic diffusion case and Huang [14] for the anisotropic diffusion case. In the current situation with convection and reaction terms, a similar weaker condition can also be obtained in two dimensions. The argument is almost the same as that of Theorem 1 except that Step (2) of the proof needs to be fine-tuned. Let e_{ij} be the edge connecting vertices \mathbf{a}_i and

\mathbf{a}_j ($i = 1, \dots, N_{vi}$, $j = 1, \dots, N$, $i \neq j$). Denote the two elements sharing $e_{i,j}$ by K and K' . From Step (2), we have

$$\begin{aligned}
 a_{ij} &\leq |K|(\nabla\phi_i|_K)^T \mathbb{D}_K \nabla\phi_j|_K + \frac{|K| \|\mathbf{b}\|_{\infty,K}}{h_j^K(d+1)} + \frac{|K| \|c\|_{\infty,K}}{(d+1)(d+2)} \\
 &\quad + |K'|(\nabla\phi_i|_{K'})^T \mathbb{D}_{K'} \nabla\phi_j|_{K'} + \frac{|K'| \|\mathbf{b}\|_{\infty,K'}}{h_j^{K'}(d+1)} + \frac{|K'| \|c\|_{\infty,K'}}{(d+1)(d+2)} \\
 &= -\frac{\det(\mathbb{D}_K)^{\frac{1}{2}}}{2} \cot(\alpha_{ij,\mathbb{D}_K^{-1}}) - \frac{\det(\mathbb{D}_{K'})^{\frac{1}{2}}}{2} \cot(\alpha_{ij,\mathbb{D}_{K'}^{-1}}) \\
 &\quad + \frac{|K| \|\mathbf{b}\|_{\infty,K}}{h_j^K(d+1)} + \frac{|K| \|c\|_{\infty,K}}{(d+1)(d+2)} + \frac{|K'| \|\mathbf{b}\|_{\infty,K'}}{h_j^{K'}(d+1)} + \frac{|K'| \|c\|_{\infty,K'}}{(d+1)(d+2)}, \quad (\text{Lemma 3})
 \end{aligned}$$

where $\alpha_{ij,\mathbb{D}_K^{-1}}$ and $\alpha_{ij,\mathbb{D}_{K'}^{-1}}$ are the angles of K and K' , respectively, that face the common edge e_{ij} . From this we can conclude that the linear finite element solution in 2D satisfies a DMP if the mesh satisfies

$$\begin{aligned}
 &\frac{|K| \|\mathbf{b}\|_{\infty,K}}{h_j^K(d+1)} + \frac{|K| \|c\|_{\infty,K}}{(d+1)(d+2)} + \frac{|K'| \|\mathbf{b}\|_{\infty,K'}}{h_j^{K'}(d+1)} + \frac{|K'| \|c\|_{\infty,K'}}{(d+1)(d+2)} \\
 &\leq \frac{\det(\mathbb{D}_K)^{\frac{1}{2}}}{2} \cot(\alpha_{ij,\mathbb{D}_K^{-1}}) + \frac{\det(\mathbb{D}_{K'})^{\frac{1}{2}}}{2} \cot(\alpha_{ij,\mathbb{D}_{K'}^{-1}}) \tag{39}
 \end{aligned}$$

for all internal edges. Following [14], we can rewrite the above inequality as

$$\begin{aligned}
 0 &< \frac{1}{2} \left[\alpha_{ij,\mathbb{D}_K^{-1}} + \alpha_{ij,\mathbb{D}_{K'}^{-1}} + \arccot\left(\sqrt{\frac{\det(\mathbb{D}_{K'})}{\det(\mathbb{D}_K)}} \cot(\alpha_{ij,\mathbb{D}_{K'}^{-1}}) - \frac{2 C(K, K', j)}{\sqrt{\det(\mathbb{D}_K)}}\right) \right] \\
 &\quad + \arccot\left(\sqrt{\frac{\det(\mathbb{D}_K)}{\det(\mathbb{D}_{K'})}} \cot(\alpha_{ij,\mathbb{D}_K^{-1}}) - \frac{2 C(K, K', j)}{\sqrt{\det(\mathbb{D}_{K'})}}\right) \leq \pi, \tag{40}
 \end{aligned}$$

where

$$\begin{aligned}
 C(K, K', j) &= \frac{|K| \|\mathbf{b}\|_{\infty,K}}{h_j^K(d+1)} + \frac{|K| \|c\|_{\infty,K}}{(d+1)(d+2)} + \frac{|K'| \|\mathbf{b}\|_{\infty,K'}}{h_j^{K'}(d+1)} \\
 &\quad + \frac{|K'| \|c\|_{\infty,K'}}{(d+1)(d+2)}. \tag{41}
 \end{aligned}$$

The following theorem summarizes the above analysis.

Theorem 2 *If (40) holds for all internal edges of the simplicial mesh \mathcal{T}_h , then the linear finite element scheme (5) for BVP (1) and (2) in two dimensions satisfies DMP.*

Loosely speaking, (40) can be written as

$$\begin{aligned}
 0 &< \frac{1}{2} \left[\alpha_{ij, \mathbb{D}_K^{-1}} + \alpha_{ij, \mathbb{D}_{K'}^{-1}} + \operatorname{arccot} \left(\sqrt{\frac{\det(\mathbb{D}_{K'})}{\det(\mathbb{D}_K)}} \cot(\alpha_{ij, \mathbb{D}_{K'}^{-1}}) \right) \right. \\
 &\quad \left. + \operatorname{arccot} \left(\sqrt{\frac{\det(\mathbb{D}_K)}{\det(\mathbb{D}_{K'})}} \cot(\alpha_{ij, \mathbb{D}_K^{-1}}) \right) \right] \\
 &\leq \pi - \mathcal{O}(h \|\mathbf{b}\|_\infty) - \mathcal{O}(h^2 \|c\|_\infty).
 \end{aligned}
 \tag{42}$$

Remark 4 For the case where $\mathbb{D} = I$, $\mathbf{b} = 0$, and $c = 0$, it is easy to see that (42) reduces to the Delaunay condition: $\alpha_{ij, K} + \alpha_{ij, K'} \leq \pi$. Moreover, for the case without convection and reaction terms, (42) gives the Delaunay-type mesh condition obtained by Huang [14] for two dimensional anisotropic diffusion problems. \square

5 Numerical examples

In this section we present numerical results obtained for four 2D examples to verify the mesh condition (25) and (42). In all but Example 4, the convection vector \mathbf{b} is taken as a constant vector with equal, positive x and y components, i.e., $\mathbf{b} = \|\mathbf{b}\|_\infty(1, 1)^T$.

Example 1 The first example is in the form of (1) and (2), and the coefficients are given as

$$\begin{aligned}
 c &= 0, \quad f = 0, \quad g(x, 0) = g(16, y) = 0, \\
 g(0, y) &= \begin{cases} 0.5y, & \text{for } 0 \leq y < 2 \\ 1, & \text{for } 2 \leq y \leq 16 \end{cases} \\
 g(x, 16) &= \begin{cases} 1, & \text{for } 0 \leq x \leq 14 \\ 8 - 0.5x, & \text{for } 14 < x \leq 16. \end{cases}
 \end{aligned}$$

For this example, the diffusion matrix is taken as the identity matrix, i.e., $\mathbb{D} = I$. This is an isotropic homogeneous diffusion problem. Note that the example satisfies the maximum principle and its solution stays between 0 and 1.

An acute-type mesh is used in the computation. Such a mesh is obtained by partitioning each square element of a uniform mesh into eight triangles with acute angles; see Fig. 2. The maximum angle of the mesh is 0.49π and thus condition (25) holds when the mesh size is sufficiently small.

Figure 3 shows the contours of the linear finite element solutions obtained for $N = 9,800$ and $N = 20,000$. (N is the number of elements.) There are no undershoot nor overshoot for $N = 20,000$ whereas both undershoots and overshoots occur for the case with $N = 9800$. In Fig. 4a, $-u_{min}$ is shown as functions of the number of elements N . From the figure one can see that $-u_{min}$ decreases as the mesh is refined and the decrease rate is about quadratic initially and then exponential near $N = 10,000$.

Fig. 2 A typical mesh ($N = 200$) used for Example 1

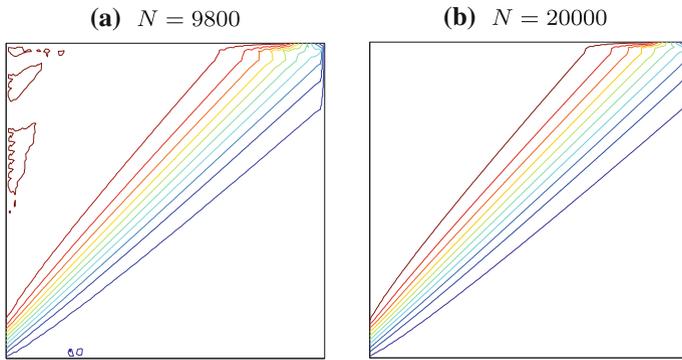
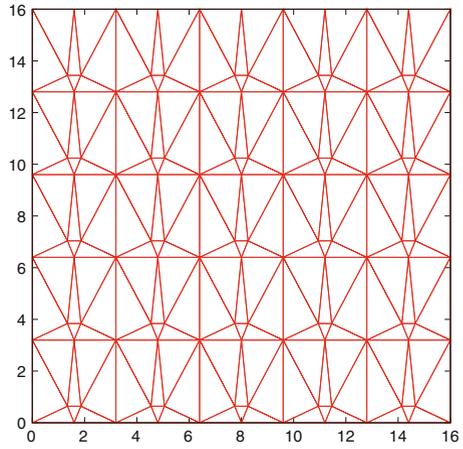


Fig. 3 Contours of the linear finite element solutions for Example 1

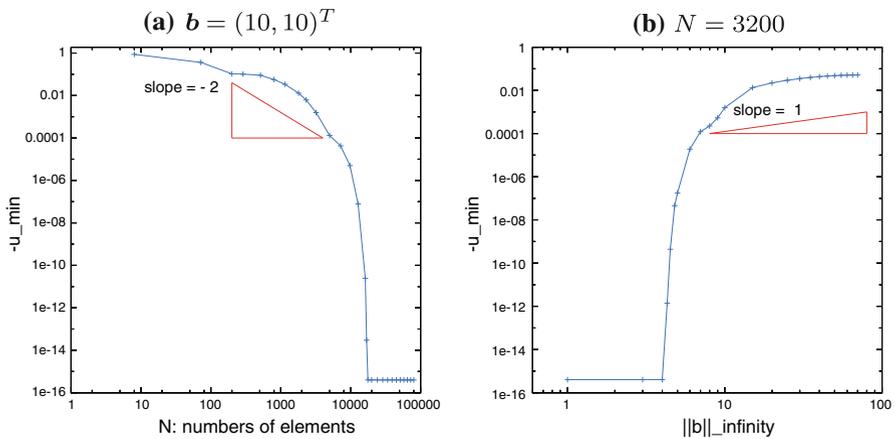


Fig. 4 The undershoot, $-u_{min}$, is plotted as a function of the number of elements N in **(a)** and as a function of $\|b\|_\infty$ in **(b)** for Example 1

Moreover, $-u_{min}$ becomes zero (more precisely, at the level of roundoff error) after around $N = 17,000$. This is consistent with Theorem 1 which states that there are no undershoot nor overshoot when the mesh size is sufficiently small.

To further verify Theorem 1, we fix the number of elements at $N = 3,200$ and let $\|\mathbf{b}\|_\infty$ vary. Quantity $-u_{min}$ is plotted in Fig. 4b as a function of $\|\mathbf{b}\|_\infty$. From the figure, we can see that there is no undershoot until $\|\mathbf{b}\|_\infty \approx 4$. Then $-u_{min}$ increases exponentially until $\|\mathbf{b}\|_\infty \approx 20$ where the increase rate is about linear as $\|\mathbf{b}\|_\infty$ increases.

Finally, it is pointed out that a similar behavior can be observed for the overshoot. The results are omitted here to save space. □

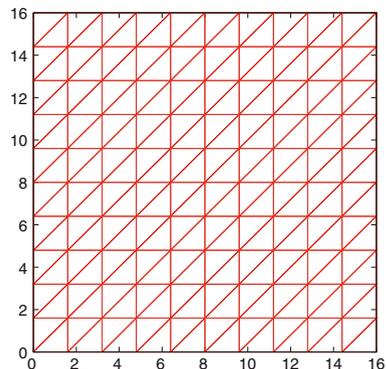
Example 2 In the second example, BVP (1) and (2) with all the coefficients being the same with Example 1 except the diffusion matrix is used. The diffusion matrix is taken as

$$\mathbb{D}(x, y) = \begin{pmatrix} 500.5 & 499.5 \\ 499.5 & 500.5 \end{pmatrix}.$$

This matrix represents a homogeneous but highly anisotropic diffusion process. A mesh with right triangle elements (see Fig. 5) is used for this example. Such a mesh is obtained by dividing each square element of a uniform mesh into two right triangular elements. Although each element of the mesh is a right triangle (in the Euclidean sense), the maximum angle is 0.49π when measured in metric \mathbb{D}^{-1} . Thus, the mesh is of acute-type in the metric and condition (25) can be satisfied if the mesh size is sufficiently small.

Contours of linear finite element solutions are shown in Fig. 6 while the undershoot is plotted as functions of N and $\|\mathbf{b}\|_\infty$ in Fig. 7. From these results we can observe a similar behavior of the undershoot and overshoot as in Example 1, i.e., they occur only for relatively coarse meshes or relatively large $\|\mathbf{b}\|_\infty$. The behavior is consistent with Theorem 1. □

Fig. 5 A typical mesh ($N = 200$) used for Example 2 (with anisotropic \mathbb{D})



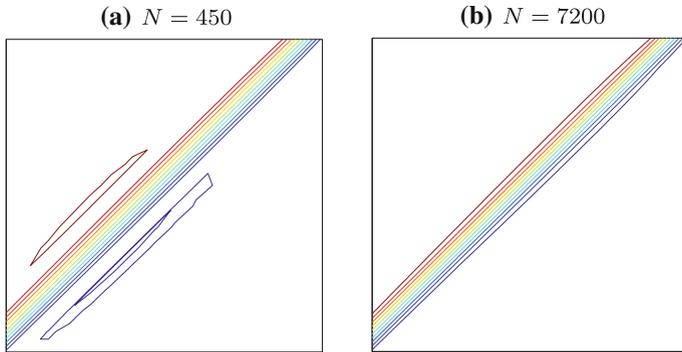


Fig. 6 Contours of linear finite element solutions for Example 2

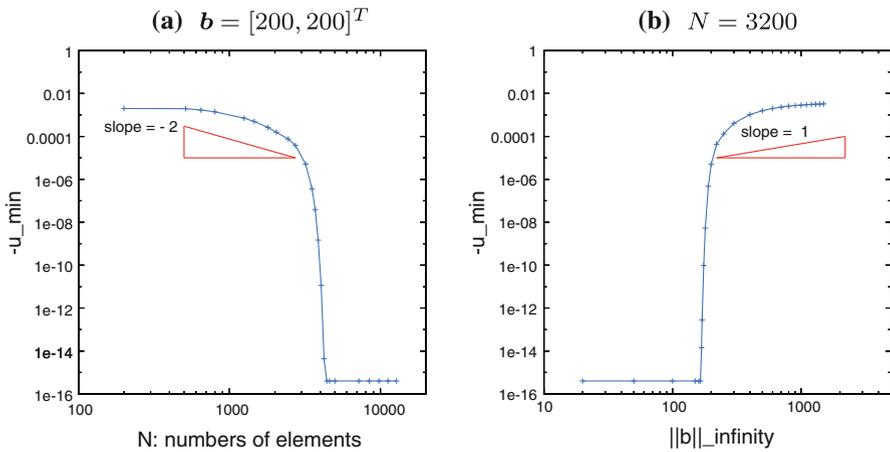


Fig. 7 The undershoot, $-u_{min}$, is plotted as a function of the number of elements N in (a) and as a function of $\|b\|_\infty$ in (b) for Example 2

Example 3 In this example, the same BVP (1) and (2) with Example 1 except the diffusion matrix is used for this example. The diffusion matrix is taken as

$$\mathbb{D}(x, y) = \begin{pmatrix} 50 & 12 \\ 12 & 50 \end{pmatrix}.$$

This diffusion matrix has a weaker anisotropy than that in the previous example.

The mesh used for Example 1 (see Fig. 2) is also used for this example. Recall that the mesh is acute in the Euclidean sense. When measured in metric \mathbb{D}^{-1} , however, the maximum angle of the mesh is 0.55π and the maximum sum of any pair of angles opposite a common edge is 0.97π . Thus, the mesh will satisfy (42) but not (25) when its size is sufficiently small.

Contours of numerical solutions are shown in Fig. 8 while the undershoot is plotted as functions of N and $\|b\|_\infty$ in Fig. 9. A similar behavior of the undershoot and overshoot can be observed as for the two previous examples.

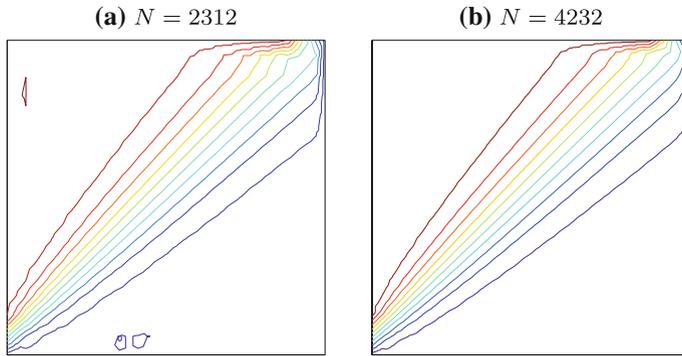


Fig. 8 Contours of linear finite element solutions for Example 3

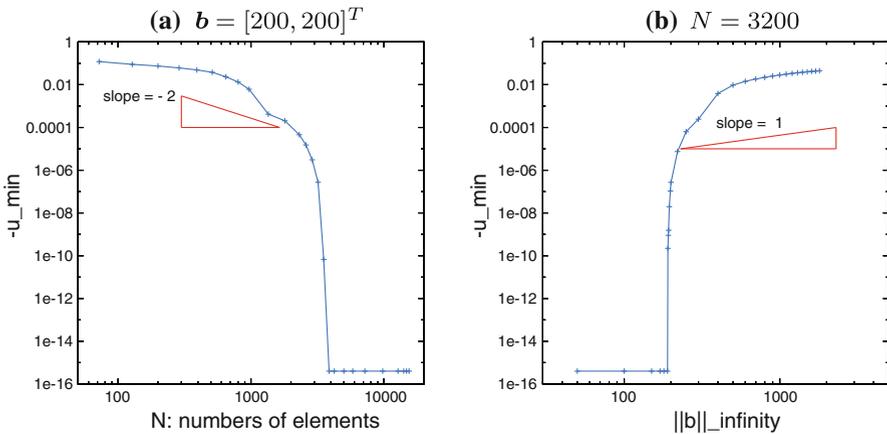


Fig. 9 The undershoot, $-u_{min}$, is plotted as a function of the number of elements N in (a) and as a function of $\|b\|_\infty$ in (b) for Example 3

In particular, for $N \geq 4,000$ there is no undershoot or overshoot in the numerical solutions for the case with $b = [200, 200]^T$. This example shows that condition (42) is weaker than condition (25). \square

Example 4 In the last example, we consider BVP (1) and (2) on domain $\Omega = [0, 1]^2 \setminus [\frac{4}{9}, \frac{5}{9}]^2$ with

$$\begin{cases} b = [5000(0.5 - y), 5000(x - 0.5)]^T, & c = 100, & f = 0, \\ g = 0 \text{ on } \partial\Omega_{out}, & g = 2 \text{ on } \partial\Omega_{in}, \end{cases} \tag{43}$$

where $\partial\Omega_{out}$ and $\partial\Omega_{in}$ are the outer and inner boundaries of Ω , respectively. The diffusion matrix is taken as

$$\mathbb{D}(x, y) = \begin{pmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{pmatrix} \begin{pmatrix} 1000 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \cos \alpha & \sin \alpha \\ -\sin \alpha & \cos \alpha \end{pmatrix},$$

where $\alpha = \pi \sin(x) \cos(y)$. This diffusion matrix is anisotropic and heterogeneous. The BVP satisfies the maximum principle and its solution stays between 0 and 2.

We use two sets of \mathbb{D}^{-1} -uniform meshes (cf. Remark 3) for this example. The first set (referred to as M_{DMP} meshes) is M -uniform meshes with M defined in (36) and $\theta_K = 1$. Notice that this set of meshes is completely determined by the diffusion matrix \mathbb{D} . The other set of meshes (referred to as $M_{DMP+adap}$ meshes) is M -uniform meshes with M defined in the form (36) and θ_K determined by minimizing an interpolation error estimate. M has the expression [22, equation (55)] as

$$M_K = \left(1 + \frac{1}{\alpha_h} B_K \right)^{\frac{1}{2}} \sqrt{\det(\mathbb{D}_K)} \mathbb{D}_K^{-1}, \tag{44}$$

where

$$B_K = \det(\mathbb{D}_K)^{-\frac{1}{2}} \|\mathbb{D}_K^{-1}\| \cdot \frac{1}{|K|} \int_K \|\mathbb{D}_K |H(u)|\|^2 dx,$$

$$\alpha_h = \left(\frac{1}{|\Omega|} \sum_K |K| \sqrt{B_K} \right)^2,$$

$\|\cdot\|$ is the l_2 matrix norm, $|H(u)| = \sqrt{H(u)^2}$, and $H(u)$ denotes the Hessian of the exact solution u . In the computation, the integral is calculated with a Gaussian quadrature rule and the Hessian is replaced by approximations obtained with a Hessian recovery technique [22] which employs piecewise quadratic polynomials fitting in least-squares sense to nodal values of the currently available computed solution.

An iterative procedure is used for solving this example. It involves three basic steps, solving the BVP on the current mesh, computing the metric tensor, and generating a new mesh. In our computation, each run is stopped after 10 iterations. We have found that there is very little improvement in the computed solution after 10 iterations. A new mesh is generated using the computer code BAMG (bidimensional anisotropic mesh generator) developed by Hecht [12] based on Delaunay-type triangulation. The code allows the user to supply his/her own metric tensor defined on a background mesh.

Two typical M_{DMP} and $M_{DMP+adap}$ meshes for this example are shown in Fig. 10. The quantity $-u_{min}$ is plotted as a function of the number of mesh elements in Fig. 11. One can see that undershoots occur for relatively coarser meshes but not for fine meshes. This is consistent with Remark 3 which shows that sufficiently fine \mathbb{D}^{-1} -uniform meshes satisfy the mesh condition (25). It should be pointed out that the maximum element size of the finest mesh for the considered range of N in Fig. 13 is $h \approx 0.058$, which is much larger than $h = 0.0003$ required by (38) for the current example. This indicates that (38) is quite conservative in estimating h for a mesh satisfying condition (25). From Fig. 13, one may also notice that $M_{DMP+adap}$ meshes lead to larger undershoots than M_{DMP} meshes. Figure 12 shows that the former has a larger maximum element size than the latter. There are no clear explanations why this should happen. We recall that M_{DMP} meshes are completely determined by \mathbb{D} whereas

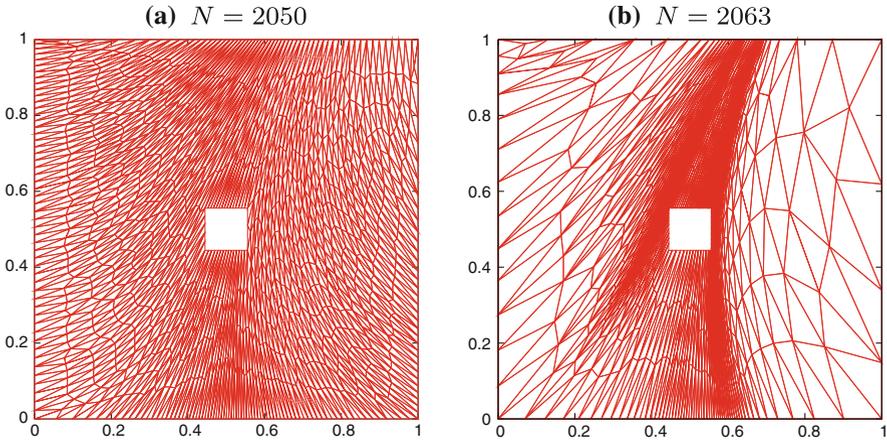


Fig. 10 Two typical **a** M_{DMP} and **b** $M_{DMP+adap}$ meshes obtained for Example 4

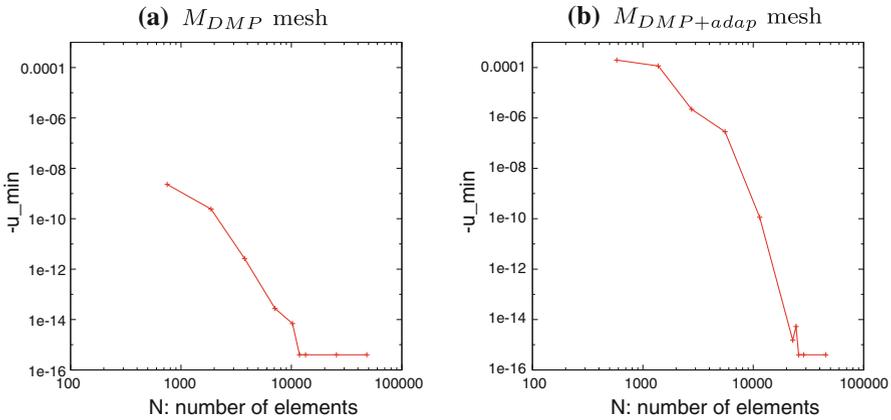


Fig. 11 The undershoot, $-u_{min}$, is plotted as a function of the number of elements N for **a** M_{DMP} and **b** $M_{DMP+adap}$ meshes for Example 4

$M_{DMP+adap}$ meshes are determined by \mathbb{D} in element shape and by minimization of interpolation error in element size. These two sets of meshes serve different purposes and it seems that it can go either way.

Finally, contours of linear finite element solutions are shown in Figs. 13 and 14. Once again, one can see that for both sets of meshes, undershoots occur for a relatively coarse mesh and vanish for a finer mesh. □

6 Conclusions

In the previous sections we have developed a mesh condition (25) under which the linear finite element solution defined in (5) for the general anisotropic diffusion problem (1) and (2) involving convection and reaction terms to satisfy a discrete maximum

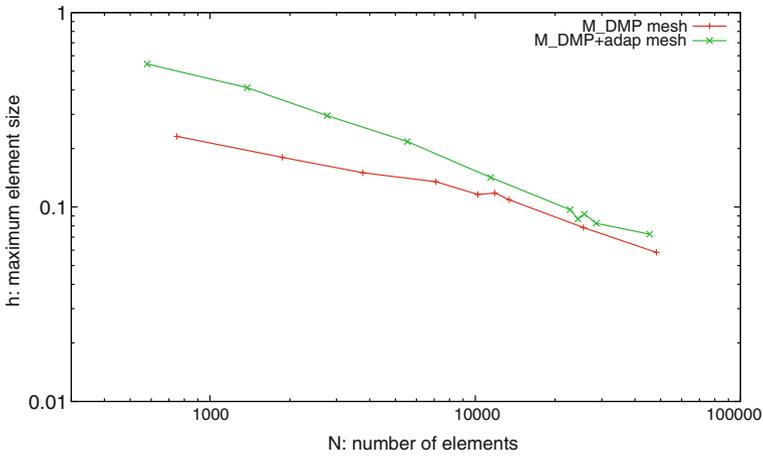


Fig. 12 The maximum element size is plotted as a function of the number of elements for both M_{DMP} and $M_{DMP+adap}$ meshes for Example 4

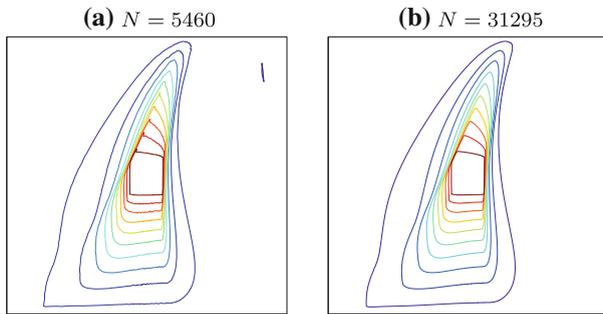


Fig. 13 Contours of linear finite element solutions obtained with M_{DMP} meshes for Example 4

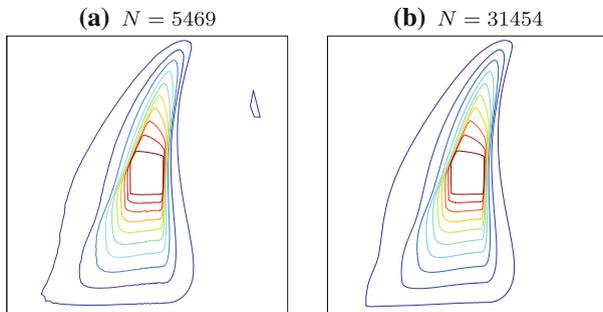


Fig. 14 Contours of linear finite element solutions obtained with $M_{DMP+adap}$ meshes for Example 4

principle. Loosely speaking, the condition requires that the dihedral angles of all elements of the mesh be $\mathcal{O}(\|\mathbf{b}\|_\infty h + \|c\|_\infty h^2)$ – acute when they are measured in the metric specific by the inverse of the coefficient matrix, where \mathbf{b} and c are the coeffi-

icients of the convection and reaction terms, respectively. Moreover, we have shown that in two dimensions a weaker condition, (40) or (42)—an $\mathcal{O}(\|\mathbf{b}\|_\infty h + \|c\|_\infty h^2)$ perturbation of the generalized Delaunay condition developed in [14], is sufficient for the linear finite element solution to satisfy a discrete maximum principle. Finally, it is worth pointing out that many existing mesh conditions such as those developed in Ciarlet and Raviart [6] (for isotropic diffusion without convection terms), Strang and Fix [30] (for 2D isotropic diffusion without convection terms), Wang and Zhang [31] (for isotropic diffusion with convection and reaction terms), Li and Huang [22] (for anisotropic diffusion without convection and reaction terms), and Huang [14] (for 2D anisotropic diffusion without convection and reaction terms) are special cases of mesh condition (25) or (40).

References

1. Aavatsmark, I., Barkve, T., Bøe, Ø., Mannseth, T.: Discretization on unstructured grids for inhomogeneous, anisotropic media. I. Derivation of the methods. *SIAM J. Sci. Comput.* **19**, 1700–1716 (1998, electronic)
2. Brandts, J., Korotov, S., Křížek, M.: The discrete maximum principle for linear simplicial finite element approximations of a reaction–diffusion problem. *Linear Algebra Appl.* **429**, 2344–2357 (2008)
3. Burman, E., Ern, A.: Discrete maximum principle for Galerkin approximations of the Laplace operator on arbitrary meshes. *C. R. Acad. Sci. Paris Ser. I* **338**, 641–646 (2004)
4. Ciarlet, P.G.: Discrete maximum principle for finite difference operators. *Aequ. Math.* **4**, 338–352 (1970)
5. Ciarlet, P.G.: *The Finite Element Method for Elliptic Problems*. North-Holland, Amsterdam (1978)
6. Ciarlet, P.G., Raviart, P.-A.: Maximum principle and uniform convergence for the finite element method. *Comput. Meth. Appl. Mech. Eng.* **2**, 17–31 (1973)
7. Drăgănescu, A., Dupont, T.F., Scott, L.R.: Failure of the discrete maximum principle for an elliptic finite element problem. *Math. Comput.* **74**, 1–23 (2004)
8. Evans, L.C.: *Partial Differential Equations*. Graduate Studies in Mathematics, vol. 19 American Mathematical Society, Providence (1998)
9. Forsyth, P.A.: A control-volume, finite-element method for local mesh refinement in thermal reservoir simulation. *SPE Reserv. Eng.* **5**, 561–566 (1990). (Paper SPE 18415)
10. Günther, S., Lackner, K.: A mixed implicit–explicit finite difference scheme for heat transport in magnetised plasmas. *J. Comput. Phys.* **228**, 282–293 (2009)
11. Günther, S., Yu, Q., Kruger, J., Lackner, K.: Modelling of heat transport in magnetised plasmas using non-aligned coordinates. *J. Comput. Phys.* **209**, 354–370 (2005)
12. Hecht, F.: Bidimensional anisotropic mesh generator. Technical report, INRIA, Rocquencourt (1997). <http://www.ann.jussieu.fr/hecht/ftp/bang>
13. Huang, W.: Mathematical principles of anisotropic mesh adaptation. *Comm. Comput. Phys.* **1**, 276–310 (2006)
14. Huang, W.: Discrete maximum principle and a delaunay-type mesh condition for linear finite element approximations of two-dimensional anisotropic diffusion problems. *Numer. Math. Theory Meth. Appl.* **4**, 319–334 (2011). (arXiv:1008.0562v1)
15. Huang, W., Russell, R.D.: *Adaptive Moving Mesh Methods*. Springer, New York (2011)
16. Karátson, J., Korotov, S.: An algebraic discrete maximum principle in Hilbert space with applications to nonlinear cooperative elliptic systems. *SIAM J. Numer. Anal.* **47**, 2518–2549 (2009)
17. Karátson, J., Korotov, S., Křížek, M.: On discrete maximum principles for nonlinear elliptic problems. *Math. Comput. Sim.* **76**, 99–108 (2007)
18. Kuzmin, D., Shashkov, M.J., Svyatskiy, D.: A constrained finite element method satisfying the discrete maximum principle for anisotropic diffusion problems. *J. Comput. Phys.* **228**, 3448–3463 (2009)
19. Křížek, M., Lin, Q.: On diagonal dominance of stiffness matrices in 3D. *East West J. Numer. Math.* **3**, 59–69 (1995)

20. Le Potier, C.: A nonlinear finite volume scheme satisfying maximum and minimum principles for diffusion operators. *Int. J. Finite Vol.* **6**, 20 (2009)
21. Letniowski, F.W.: Three-dimensional Delaunay triangulations for finite element approximations to a second-order diffusion operator. *SIAM J. Sci. Stat. Comput.* **13**, 765–770 (1992)
22. Li, X.P., Huang, W.: An anisotropic mesh adaptation method for the finite element solution of heterogeneous anisotropic diffusion problems. *J. Comput. Phys.* **229**, 8072–8094 (2010). (arXiv:1003.4530v2)
23. Li, X.P., Svyatskiy, D., Shashkov, M.: Mesh adaptation and discrete maximum principle for 2D anisotropic diffusion problems. Technical Report LA-UR 10–01227, Los Alamos National Laboratory, Los Alamos, NM (2007)
24. Lipnikov, K., Shashkov, M., Svyatskiy, D., Vassilevski, Yu.: Monotone finite volume schemes for diffusion equations on unstructured triangular and shape-regular polygonal meshes. *J. Comput. Phys.* **227**, 492–512 (2007)
25. Liska, R., Shashkov, M.: Enforcing the discrete maximum principle for linear finite element solutions of second-order elliptic problems. *Comm. Comput. Phys.* **3**, 852–877 (2008)
26. Mlacnik, M.J., Durlafsky, L.J.: Unstructured grid optimization for improved monotonicity of discrete solutions of elliptic equations with highly anisotropic coefficients. *J. Comput. Phys.* **216**, 337–361 (2006)
27. Perona, P., Malik, J.: Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. Pattern Anal. Mach. Intel.* **12**, 629–639 (1990)
28. Sharma, P., Hammett, G.W.: Preserving monotonicity in anisotropic diffusion. *J. Comput. Phys.* **227**, 123–142 (2007)
29. Stoyan, G.: On maximum principles for monotone matrices. *Linear Algebra Appl.* **78**, 147–161 (1986)
30. Strang, G., Fix, G.J.: *An Analysis of the Finite Element Method*. Prentice Hall, Englewood Cliffs (1973)
31. Wang, J., Zhang, R.: Maximum principle for P1-conforming finite element approximations of quasi-linear second order elliptic equations. (2011). (arXiv:1105.1466v3)
32. Weickert, J.: *Anisotropic Diffusion in Image Processing*. Teubner-Verlag, Stuttgart (1998)
33. Xu, J., Zikatanov, L.: A monotone finite element scheme for convection–diffusion equations. *Math. Comput.* **69**, 1429–1446 (1999)