Maximum Principle for the Finite Element Solution
of Time-Dependent Anisotropic Diffusion
Problems

Xianping Li, 1 Weizhang Huang 2

1 Department of Mathematics, The University of Central Arkansas, Conway, Arkansas 72034
2 Department of Mathematics, The University of Kansas, Lawrence, Kansas 66045

Received 26 September 2012; accepted 28 February 2013
Published online 14 April 2013 in Wiley Online Library (wileyonlinelibrary.com).
DOI 10.1002/num.21784

Preservation of the maximum principle is studied for the combination of the linear finite element method
in space and the \( \theta \)-method in time for solving time-dependent anisotropic diffusion problems. It is shown
that the numerical solution satisfies a discrete maximum principle when all element angles of the mesh mea-
sured in the metric specified by the inverse of the diffusion matrix are nonobtuse, and the time step size is
bounded below and above by bounds proportional essentially to the square of the maximal element diameter.
The lower bound requirement can be removed when a lumped mass matrix is used. In two dimensions, the
mesh and time step conditions can be replaced by weaker Delaunay-type conditions. Numerical results are
presented to verify the theoretical findings. © 2013 Wiley Periodicals, Inc. Numer Methods Partial Differential
Eq 29: 1963–1985, 2013

Keywords: finite element; time-dependent; anisotropic diffusion; maximum principle

I. INTRODUCTION

We are concerned with the linear finite element solution of the initial-boundary value problem
(IBVP) of a linear diffusion equation,

\[
\begin{align*}
\frac{\partial u}{\partial t} - \nabla \cdot (D \nabla u) &= f(x, t), & \text{in } \Omega_T = \Omega \times (0, T] \\
u(x, t) &= g(x, t), & \text{on } \partial \Omega \times [0, T] \\
u(x, 0) &= u_0(x), & \text{in } \Omega \times \{t = 0\}
\end{align*}
\]

(1)

where \( \Omega \subset \mathbb{R}^d \) (\( d \geq 1 \)) is a connected polygonal or polyhedral domain, \( T > 0 \) is a fixed time,
\( f(x, t), \ g(x, t) \), and \( u_0(x) \) are given functions, and \( D \) is the diffusion matrix. We assume that
\[ D = D(x) \] is a general symmetric and strictly positive definite matrix-valued function on \( \Omega_T \). It includes both isotropic and anisotropic diffusion as special examples. In the former case, \( D \) takes the form \( \alpha(x) I \), where \( I \) is the \( d \times d \) identity matrix and \( \alpha = \alpha(x) \) is a scalar function. In the latter case, on the other hand, \( D \) has not-all-equal eigenvalues at least on a certain portion of \( \Omega_T \). Note that we consider only time-independent \( D \) in this work. In principle, the procedure used in this work can also apply to the time-dependent situation. For that situation, however, different meshes are needed for different time steps and the numerical solution has to be interpolated between these meshes. Then, a conservative interpolation scheme must be used in order for the underlying scheme to preserve the maximum principle, nonnegativity, or monotonicity. The development of conservative interpolation schemes and their use for unstructured meshes is an interesting research topic in its own right (e.g., see Ref. [1]) and beyond the scope of the current study. To avoid this possible complexity, we restrict our attention to the time-independent diffusion matrix in this work.

Anisotropic diffusion problems arise from various areas of science and engineering including plasma physics [2–7], petroleum reservoir simulation [8–12], and image processing [13–18]. IBVP (1) is a prototype of those anisotropic diffusion problems. It satisfies the maximum principle

\[
\max_{(x,t) \in \Omega_T} v(x,t) = \max \left\{ 0, \max_{(x,t) \in \partial \Omega_T} v(x,t) \right\}, \quad \forall v \text{ satisfying } v_t - \nabla \cdot (D \nabla v) \leq 0 \text{ in } \Omega_T \tag{2}
\]

where \( \partial \Omega_T \) denotes the parabolic boundary (i.e., \( \partial \Omega \times \{ 0 < t \leq T \} \cup \Omega \times \{ t = 0 \} \)). When a standard numerical method such as a finite element or a finite difference method is used to solve this problem, the numerical solution may violate the maximum principle and contain spurious oscillations. It is of practical and theoretical importance to study when a numerical solution satisfies a discrete maximum principle (DMP) (cf. (40) in Section III) as well as develop DMP-preserving numerical schemes.

The research topic has attracted considerable attention from researchers since 1970’s and success has been made for elliptic diffusion problems; for example, see Refs. [12, 19–36]. For example, it is shown in Refs. [19, 21] that for isotropic diffusion problems, the requirement of all element angles of the mesh to be nonobtuse is sufficient for the linear finite element approximation to satisfy DMP. In two dimensions (2D), this nonobtuse angle condition can be replaced by a weaker, so-called Delaunay condition [34], which requires the sum of any pair of angles facing a common interior edge to be less than or equal to \( \pi \). For anisotropic diffusion problems, Drăgănescu et al. [22] show that the nonobtuse angle condition fails to guarantee DMP satisfaction for a linear finite element approximation. Various techniques have been proposed to reduce spurious oscillations, including local matrix modification [26, 29], mesh optimization [12], and mesh adaptation [28]. An anisotropic nonobtuse angle condition, which uses element angles measured in the metric specified by \( D^{-1} \) instead of angles measured in the Euclidean metric (as in the nonobtuse angle condition), is developed in Ref. [27] to guarantee DMP satisfaction for anisotropic diffusion problems. A weaker, Delaunay-type mesh condition is obtained in Ref. [23] for 2D problems. The results of Refs. [23, 27] are extended in Ref. [30] to problems containing convection and reaction terms.

On the other hand, less progress has been made for time-dependent problems; for example, see Refs. [6, 37–50]. Most of the existing research has focused on isotropic diffusion problems. For example, Fujii [44] considers the heat equation and shows that the time step size should be bounded from below and above for a linear finite element approximation to satisfy DMP when the mesh satisfies the nonobtuse angle condition. He also shows that the lower bound requirement can be removed when a lumped mass matrix is used. The study is extended in Ref. [38] to a more

*Numerical Methods for Partial Differential Equations* DOI 10.1002/num
general isotropic diffusion problem with a reaction term. Thomée and Wahlbin [48] consider
general anisotropic diffusion problems and show that a semidiscrete conventional finite element
solution does not satisfy DMP in general. Slope limiters are used in Ref. [6] to improve DMP
satisfaction for anisotropic thermal conduction in magnetized plasmas. Nonlinear finite volume
methods are developed by Le Potier [51, 52] for time-dependent problems.

The objective of this article is to investigate conditions for the finite element approximation of
IBVP (1) to satisfy DMP for a general diffusion matrix function. We are particularly interested
in lower and upper bounds on the time step size when the \( \theta \)-method and the conventional linear
finite element method are used for temporal and spatial discretization, respectively. Two types of
simplicial mesh are considered, meshes satisfying the anisotropic nonobtuse angle condition [27]
or a Delaunay-type mesh condition [23]. It is known that those meshes lead to DMP-satisfaction
linear finite element approximations to steady-state anisotropic diffusion problems. A lumped
mass matrix is also studied. The results obtained in this article can be viewed as a generalization
of Fujii’s [44] to anisotropic diffusion problems although such generalization is not trivial.

The outline of this article is as follows. In Section II, the linear finite element solution of
IBVP (1) is described. Section III is devoted to the development of DMP-satisfaction conditions.
Numerical examples are presented in Section IV to verify the theoretical findings. Finally, Section
V contains conclusions.

II. LINEAR FINITE ELEMENT FORMULATION

Consider the linear finite element solution of IBVP (1). Assume that an affine family of simplicial
triangulations \( \{T_h\} \) is given for the physical domain \( \Omega \). Define

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

Denote the linear finite element space associated with mesh \( T_h \) by \( U_h^g \). A linear finite element
solution \( u^h(t) \in U_h^g \) for \( t \in (0, T] \) to IBVP (1) is defined by

\[
\int_{\Omega} \frac{\partial u^h}{\partial t} v^h dx + \int_{\Omega} (\nabla v^h)^T \mathbb{D} \nabla u^h dx = \int_{\Omega} f v^h dx, \quad \forall v^h \in U_0^h
\]

where \( U_0^h = U_h^g \) with \( g = 0 \). This equation can be rewritten as

\[
\sum_{K \in T_h} \int_K \frac{\partial u^h}{\partial t} v^h dx + \sum_{K \in T_h} |K| (\nabla v^h)^T \mathbb{D}_K \nabla u^h dx = \sum_{K \in T_h} \int_K f v^h dx, \quad \forall v^h \in U_0^h
\]

where \( |K| \) is the volume of element \( K \) and

\[
\mathbb{D}_K = \frac{1}{|K|} \int_K \mathbb{D} dx.
\]

Equation (4) can be expressed in a matrix form. Denote the numbers of the elements, vertices,
and interior vertices of \( T_h \) by \( N_e, N_v, \) 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

\[
U_0^h = \text{span}\{\phi_1, \ldots, \phi_{N_v}\},
\]

\[
u^h = \sum_{j=1}^{N_{vi}} u_j \phi_j + \sum_{j=N_{vi}+1}^{N_v} u_j \phi_j,
\]

(5)

where \( \phi_j \) is the linear basis function associated with the \( j \)th vertex, \( a_j \). We approximate the boundary and initial conditions in (1) as

\[
u_j(t) = g_j \equiv g(a_j, t), \quad j = N_{vi} + 1, \ldots, N_v
\]

(6)

\[
u_j(0) = u_0(a_j), \quad j = 1, \ldots, N_v.
\]

(7)

Substituting (5) into (4), taking \( \nu^h = \phi_i \) (\( i = 1, \ldots, N_{vi} \)), and combining the resulting equations with (6), we obtain the linear algebraic system

\[
M \frac{du}{dt} + A u = f,
\]

(8)

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

\[
M = \begin{bmatrix}
M_{11} & M_{12} \\
0 & 0
\end{bmatrix}, \quad A = \begin{bmatrix}
A_{11} & A_{12} \\
0 & I
\end{bmatrix},
\]

(9)

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

\[
m_{ij} = \sum_{K \in T_h} \int_K \phi_j \phi_i \, dx, \quad i = 1, \ldots, N_{vi}, \quad j = 1, \ldots, N_v \]

(10)

\[
a_{ij} = \sum_{K \in T_h} |K| (\nabla \phi_i)^T \nabla \phi_j, \quad i = 1, \ldots, N_{vi}, \quad j = 1, \ldots, N_v \]

(11)

\[
f_i = \sum_{K \in T_h} \int_K f \phi_i \, dx, \quad i = 1, \ldots, N_{vi}.
\]

(12)

We use the \( \theta \)-method with a constant time step \( \Delta t \) for time integration. Let \( u^n \) and \( u^{n+1} \) be the computed solutions at the current and next time steps, respectively. Applying the \( \theta \)-method to the first \( N_{vi} \) equations, we get

\[
[M_{11} M_{12}] \frac{u^{n+1} - u^n}{\Delta t} + [A_{11} A_{12}][(1 - \theta)u^n + \theta u^{n+1}] = \tilde{f}^{n+\theta},
\]

where

\[
\tilde{f}^{n+\theta} = \begin{bmatrix}
f_1(t_n + \theta \Delta t), \ldots, f_{N_{vi}}(t_n + \theta \Delta t)
\end{bmatrix}^T.
\]

Numerical Methods for Partial Differential Equations DOI 10.1002/num
For the last $N_v - N_{vi}$ equations (corresponding to the boundary condition), we use

$$u_j^{n+1} = g(a_j, t_{n+1}), \quad j = N_{vi} + 1, \ldots, N_v. \quad (14)$$

Combining (13) and (14), we have

$$Bu^{n+1} = Cu^n + \Delta t f^{n+\theta}, \quad (15)$$

where

$$B = \begin{bmatrix} M_{11} & M_{12} \\ 0 & I \end{bmatrix} + \theta \Delta t \begin{bmatrix} A_{11} & A_{12} \\ 0 & 0 \end{bmatrix}, \quad (16)$$

$$C = \begin{bmatrix} M_{11} & M_{12} \\ 0 & 0 \end{bmatrix} - (1 - \theta) \Delta t \begin{bmatrix} A_{11} & A_{12} \\ 0 & 0 \end{bmatrix}, \quad (17)$$

$$f^{n+\theta} = \left( f_1(t_n + \theta \Delta t), \ldots, f_{N_{vi}}(t_n + \theta \Delta t), \frac{1}{\Delta t} g(a_{N_{vi}} + 1, t_{n+1}), \ldots, \frac{1}{\Delta t} g(a_{N_v}, t_{n+1}) \right)^T, \quad (18)$$

$$u^0 = u_0 = (u_0(a_1), \ldots, u_0(a_{N_v}))^T. \quad (19)$$

It is worth noting that the right-hand side vector, $f^{n+\theta}$, is formed from the values of the right-hand side function $f(x, t)$ and the boundary function $g(x, t)$. We are interested in conditions under which the scheme satisfies DMP.

**III. CONDITIONS FOR DMP SATISFACTION**

In this section, we develop the conditions (on the mesh and time step size) under which scheme (15) satisfies DMP. The main tool is a result from Ref. [33] which states that the solution of a linear algebraic system satisfies DMP when the corresponding coefficient matrix is an $M$-matrix and has nonnegative row sums. We first discuss the general dimensional case along with the anisotropic nonobtuse angle condition developed in Ref. [27] and then study the 2D case with the Delaunay-type mesh condition developed in Ref. [23].

We introduce some notation. Consider a generic element $K \in \mathcal{T}_h$ and denote its vertices by $a^K_1, a^K_2, \ldots, a^K_{d+1}$. Denote the face opposite to vertex $a^K_i$ (i.e., the face not having $a^K_i$ as its vertex) by $S^K_i$ and its unit inward (pointing to $a^K_i$) normal by $n^K_i$. The distance (or height) from vertex $a^K_i$ to face $S^K_i$ is denoted by $h^K_i$. Define $q$-vectors as

$$q^K_i = \frac{n^K_i}{h^K_i}, \quad i = 1, \ldots, d + 1. \quad (20)$$

Obviously, we have $h^K_i = 1/\|q^K_i\|$. We now consider the mapping $\mathbb{D}^{-\frac{1}{2}}_K : K \to \widetilde{K}$; see Fig. 1. The $q$-vectors and heights associated with $\widetilde{K}$ are denoted by $\widetilde{q}^K_i$, and $\widetilde{h}^K_i$. We have relations

$$\widetilde{a}^K_i = \mathbb{D}^{-\frac{1}{2}}_K a^K_i, \quad \widetilde{S}^K_i = \mathbb{D}^{-\frac{1}{2}}_K S^K_i, \quad |\widetilde{K}| = \det(\mathbb{D}_K)^{-\frac{1}{2}} |K|, \quad \widetilde{q}^K_i = \mathbb{D}^{-\frac{1}{2}}_K q^K_i, \quad \widetilde{h}^K_i = \frac{1}{\|\widetilde{q}^K_i\|}. \quad (21)$$
The dihedral angle between surfaces $\tilde{S}_i^K$ and $\tilde{S}_j^K$ ($i \neq j$) is denoted by $\tilde{\alpha}_{ij}^K$. It can be expressed as

$$
\cos(\tilde{\alpha}_{ij}^K) = -\frac{(\tilde{q}_i^K)^T \tilde{q}_j^K}{\|\tilde{q}_i^K\| \cdot \|\tilde{q}_j^K\|}, \quad i \neq j
$$

where $\|\tilde{q}_i^K\|_{D_K} = \sqrt{(\tilde{q}_i^K)^T D_K^{-1} \tilde{q}_i^K}$. Note that $\tilde{\alpha}_{ij}^K$ can be considered as a dihedral angle of $K$ measured in the metric specified by $D_K^{-1}$.

A. General Dimensional Case: $d \geq 1$

We now are ready for the development of the DMP satisfaction conditions for scheme (15) for the general dimensional case. We first have the following four lemmas.

**Lemma 3.1.** For any element $K \in T_h$ and $i, j = 1, \ldots, d + 1$,

$$
(\nabla \phi_i)^T D_K \nabla \phi_j = \begin{cases} 
\cos(\tilde{\alpha}_{ij}^K), & \text{for } i \neq j \\
-\frac{h_i^K h_j^K}{1} & \text{for } i = j
\end{cases}
$$

where $\phi_i$ and $\phi_j$ are the linear basis functions associated with the vertices $a_i^K$ and $a_j^K$, respectively. In two dimensions ($d = 2$),

$$
|K|(\nabla \phi_i)^T D_K \nabla \phi_j = -\frac{\sqrt{\det(D_K)}}{2} \cot(\tilde{\alpha}_{ij}^K), \quad i \neq j, \quad i, j = 1, 2, 3.
$$

**Proof.** see Refs. [23, 30].

**Lemma 3.2.** The stiffness matrix $A$ defined in (9) and (11) is an $M$-matrix and has nonnegative row sums if the mesh satisfies the anisotropic nonobtuse angle condition

$$
0 < \tilde{\alpha}_{ij}^K \leq \frac{\pi}{2}, \quad \forall i, j = 1, \ldots, d + 1, i \neq j, \forall K \in T_h.
$$

**Proof.** See [Ref. [27], Theorem 2.1 and its proof].

*Numerical Methods for Partial Differential Equations* DOI 10.1002/num
Lemma 3.3. Matrix $B$ defined in (16) ($0 < \theta \leq 1$) is an $M$-matrix if the mesh satisfies (25) and the time step size satisfies

$$\Delta t \geq \frac{1}{\theta(d+1)(d+2)} \max_{k \in T_h} \max_{i,j=1,\ldots,d+1} \frac{h_i^K h_j^K}{\cos(\alpha_{ij}^K)} \lambda_{\min}(D_K).$$

(26)

Proof. We first show that $M + \theta \Delta t A$ is a $Z$-matrix, that is, it has positive diagonal and nonpositive off-diagonal entries. From (9), we only need to show

$$m_{ii} + \theta \Delta t a_{ii} > 0, \quad i = 1, \ldots, N_v$$

(27)

$$m_{ij} + \theta \Delta t a_{ij} \leq 0 \quad \forall i \neq j, \quad i = 1, \ldots, N_v, \quad j = 1, \ldots, N_v.$$  

(28)

Let $\omega_i$ be the patch of the elements containing vertex $a_i$. Notice that $\nabla \phi_i = 0$ when $K \notin \omega_i$. Recall from [53] that

$$\int_K \phi_i \phi_j \, dx = \frac{|K|}{(d+1)(d+2)}, \quad \int_K \phi_i^2 \, dx = \frac{2|K|}{(d+1)(d+2)}.$$

Then (27) follows immediately from (10) and Lemma 3.2.

For (28), from (10), (11), and (29) we have

$$m_{ij} + \theta \Delta t a_{ij} = \sum_{K \in T_h} \int_K \phi_i \phi_j \, dx + \theta \Delta t \sum_{K \in T_h} |K| (\nabla \phi_i)^T D_K \nabla \phi_j$$

$$= \sum_{K \in \omega_i \cap \omega_j} \left( \int_K \phi_i \phi_j \, dx + \theta \Delta t |K| (\nabla \phi_i)^T D_K \nabla \phi_j \right)$$

$$= \sum_{K \in \omega_i \cap \omega_j} \left( \int_K \phi_i \phi_j \, dx + \theta \Delta t |K| (\nabla \phi_{iK})^T D_K \nabla \phi_{jK} \right),$$

(30)

where $i_K$ and $j_K$ denote the local indices (on element $K$) of vertices $a_i$ and $a_j$. From (29) and Lemma 3.1, we get

$$m_{ij} + \theta \Delta t a_{ij} = \sum_{K \in \omega_i \cap \omega_j} |K| \left( \frac{1}{(d+1)(d+2)} - \theta \Delta t \frac{\cos(\alpha_{iK,jK})}{h_i^K h_j^K} \right).$$

(31)

The right-hand side term is nonpositive if

$$\Delta t \geq \frac{1}{\theta(d+1)(d+2)} \max_{k \in T_h} \max_{i,j=1,\ldots,d+1} \frac{\tilde{h}_i^K \tilde{h}_j^K}{\cos(\alpha_{ij}^K)}.$$

(32)

Moreover, (21) implies

$$\tilde{h}_i^K = \frac{1}{\|q_i\|} \frac{1}{\sqrt{\|q_i^T D_K q_i\|}}.$$
Thus, we have
\[
\frac{h_K^i}{\sqrt{\lambda_{\text{max}}(D_K)}} \leq \tilde{h}_i^K \leq \frac{h_K^i}{\sqrt{\lambda_{\text{min}}(D_K)}}. \tag{33}
\]
From this, we can see that (26) implies (32). Hence, we have shown that \( B \) is a \( Z \)-matrix when (26) holds.

To show \( B \) is an \( M \)-matrix, we recall from (16) that
\[
B = \begin{bmatrix}
M_{11} + \theta \Delta t A_{11} & M_{12} + \theta \Delta t A_{12} \\
0 & I
\end{bmatrix}.
\]
The fact that \( B \) is a \( Z \)-matrix means that \( M_{11} + \theta \Delta t A_{11} \) is also a \( Z \)-matrix and \( M_{12} + \theta \Delta t A_{12} \leq 0 \). It is easy to show that \( M_{11} + \theta \Delta t A_{11} \) is positive definite, which in turn implies \( M_{11} + \theta \Delta t A_{11} \) is an \( M \)-matrix. Notice
\[
B^{-1} = \begin{bmatrix}
(M_{11} + \theta \Delta t A_{11})^{-1} & -(M_{11} + \theta \Delta t A_{11})^{-1}(M_{12} + \theta \Delta t A_{12}) \\
0 & I
\end{bmatrix}.
\]
This means \( B^{-1} \geq 0 \) and hence \( B \) is an \( M \)-matrix.

**Lemma 3.4.** Matrix \( C \) defined in (17) \((0 \leq \theta \leq 1)\) is nonnegative if the mesh satisfies (25) and the time step size satisfies
\[
\Delta t \leq \frac{2}{(1-\theta)(d+1)(d+2)} \min_{K} \min_{i=1,...,d+1} \frac{(h_K^i)^2}{\lambda_{\text{max}}(D_K)}. \tag{34}
\]

**Proof.** For off-diagonal entries \((i \neq j, i = 1, \ldots, N_v, j = 1, \ldots, N_v)\), \( m_{ij} - (1-\theta)\Delta t a_{ij} \), are nonnegative because \( a_{ij} \leq 0 \) under condition (25) (cf. Lemma 3.2) and \( m_{ij} \geq 0 \) from definition (10). To see if the diagonal entries are also nonnegative, from (10), (11), and (29) we have
\[
m_{ii} - (1-\theta)\Delta t a_{ii} = \sum_{K \in \omega_i} |K| \left( \frac{2}{(d+1)(d+2)} - \frac{(1-\theta)\Delta t}{(\tilde{h}_i^K)^2} \right). \tag{35}
\]
The right-hand side term is nonnegative if
\[
\Delta t \leq \frac{2}{(1-\theta)(d+1)(d+2)} \min_{K} \min_{i=1,...,d+1} (\tilde{h}_i^K)^2.
\]
From (33), we see that this condition holds when (34) is satisfied.

We are now in a position to prove our first main theoretical result.

**Theorem 3.1.** Scheme (15) satisfies a DMP if the mesh satisfies the anisotropic nonobtuse angle condition (25) and the time step size satisfies (26) and (34), that is,
\[
\frac{1}{\theta(d+1)(d+2)} \max_{i,j=1,...,d+1} \max_{i \neq j} \frac{h_K^i h_K^j}{\cos(\tilde{\alpha}_{ij}^K) \lambda_{\text{min}}(D_K)} \leq \Delta t \leq \frac{2}{(1-\theta)(d+1)(d+2)} \min_{K} \min_{i=1,...,d+1} \frac{(h_K^i)^2}{\lambda_{\text{max}}(D_K)}. \tag{36}
\]

*Numerical Methods for Partial Differential Equations* DOI 10.1002/num
Proof. Scheme (15) can be expressed as

$$AU = F,$$  \hspace{1cm} (37)

where

$$A = \begin{bmatrix} I & 0 \\ -C & B \\ & -C & B \\ & & \ddots & \ddots \\ & & & -C & B \end{bmatrix}, \quad U = \begin{bmatrix} u^0 \\ u^1 \\ \vdots \\ u^N \end{bmatrix}, \quad F = \begin{bmatrix} u_0 \\ \Delta t f^0 \\ \Delta t f^1 \\ \vdots \\ \Delta t f^{N-1+\theta} \end{bmatrix}, \hspace{1cm} (38)$$

and $B$ and $C$ are defined in (17). Scheme (37) satisfies a DMP if coefficient matrix $A$ is an $M$-matrix and has nonnegative row sums. From Lemmas 3.3 and 3.4, we know that $B$ is an $M$-matrix and $C \geq 0$. As a result, $A$ is a $Z$-matrix. Moreover, we can show $A^{-1} \geq 0$. Indeed, from (37) we know that $u_0 = u_0$ and thus if $u_0 \geq 0$, we have $u^0 \geq 0$. Next, from the scheme, we have $u^1 = B^{-1} \Delta t f^0 + B^{-1} C u^0$. Recall that $C \geq 0$ and $B$ is an $M$-matrix and thus $B^{-1} \geq 0$. Combining these results, we can conclude that $f^0 \geq 0$ implies $u^1 \geq 0$. Similarly, we can show $u^n \geq 0$ if $f^{n-1+\theta} \geq 0$, $n = 2, \ldots, N$. Thus, we have shown that $F \geq 0$ implies $U \geq 0$. This implies $A^{-1} \geq 0$ and $A$ is an $M$-matrix.

We notice that the sum of each of the second to the last (block) rows is

$$B - C = \begin{bmatrix} \Delta t A_{11} & \Delta t A_{12} \\ 0 & I \end{bmatrix}.$$

As $A$ has nonnegative row sums (cf. Lemma 3.2), $A$ has nonnegative row sums. Thus, we have proven that $A$ is an $M$-matrix and has nonnegative row sums.

Form Ref. [33][Theorem 1], we conclude that the solution of (37) satisfies

$$\max_{i=1, \ldots, (N+1)N_v} U_i = \max \left\{ 0, \max_{i \in S(F^+)} U_i \right\}, \hspace{1cm} (39)$$

where $S(F^+)$ is the set of the indices with $F_i > 0$. When $f(x, t) \leq 0$, from (18) we know that $F_i > 0$ holds only for those indices corresponding to the boundary points on $\partial \Omega$. Moreover, from (16), (17), and (38) we see that at the boundary points, $U_i$ is equal to either the boundary function $g$ or the initial function $u_0$. Because a piecewise linear function attains its maximum value at vertices, (39) implies that when $f(x, t) \leq 0$, the solution of (15) satisfies a DMP

$$\max_{n=0, \ldots, N} \max_{x \in \Omega} U^n(x) = \max \left\{ 0, \max_{n=1, \ldots, N} \max_{x \in \partial \Omega} U^n(x), \max_{x \in \Omega} U^0(x) \right\}, \hspace{1cm} (40)$$

where

$$U^n(x) = \sum_{j=1}^{N_v} u^n_j \phi_j(x) + \sum_{j=N_v+1}^{N_v+N} u^n_j \phi_j(x), \hspace{1cm} n = 0, \ldots, N.$$ 

Hence, we have proven that scheme (15) satisfies DMP. 

\hfill \blacksquare
Remark 3.1. Consider a special case with $\mathbb{D} = \alpha I$, where $\alpha$ is a positive constant. It is known (e.g., see Emert and Nelson [54]) that the height (or altitude), volume, and cosine of the dihedral angles of a regular $d$-dimensional simplex $K$ are given by

$$h_K = e_K \sqrt{\frac{d+1}{2d}}, \quad |K| = \frac{\sqrt{d+1}}{d!\sqrt{2^d}} e_K^d, \quad \cos(\alpha_K^{ij}) = \frac{1}{d},$$

(41)

where $e_K$ is the edge length. Thus, if the elements of $\mathcal{T}_h$ are all regular simplexes, (36) reduces to

$$\max_{K \in \mathcal{T}_h} \frac{e^2_K}{2\theta \alpha (d+2)} \leq \Delta t \leq \min_{K \in \mathcal{T}_h} \frac{e^2_K}{(1-\theta)\alpha d (d+2)}.$$

(42)

If further the mesh is uniform (and thus all mesh elements have the same volume and same edge length ($e$)), the above condition becomes

$$\frac{e^2}{2\theta \alpha (d+2)} \leq \Delta t \leq \frac{e^2}{(1-\theta)\alpha d (d+2)},$$

(43)

which is exactly the result of Theorem 20 of Ref. [38] where the maximum principle of linear finite element approximation of isotropic diffusion problems is studied. Interestingly, we can rewrite (43) in terms of the number of the elements, $N_e$. Indeed, because the mesh is uniform, the elements have a constant volume $|\Omega|/N_e$. From (41), we have

$$e = \sqrt{2} N_e^{-\frac{1}{d}} \left( \frac{|\Omega|d!}{\sqrt{d+1}} \right)^{\frac{1}{d}}.$$

Inserting this into (43), we get

$$\frac{N_e^{-\frac{1}{d}}}{\theta \alpha (d+2)} \left( \frac{|\Omega|d!}{\sqrt{d+1}} \right)^{\frac{1}{d}} \leq \Delta t \leq \frac{2N_e^{-\frac{1}{d}}}{(1-\theta)\alpha d (d+2)} \left( \frac{|\Omega|d!}{\sqrt{d+1}} \right)^{\frac{1}{d}}.$$

(44)

□

Remark 3.2. Another special case is that the mesh is uniform in the metric specified by $\mathbb{D}^{-1}$. It is known [55] that such a mesh satisfies the so-called alignment and equidistribution conditions

$$\frac{1}{d} \text{tr}((F_K^t)^T \mathbb{D}^{-1}_K F_K^t) = 1, \quad \forall K \in \mathcal{T}_h$$

(45)

$$|K| \sqrt{\text{det}(\mathbb{D}^{-1}_K)} = \frac{\sigma_h}{N_e}, \quad \forall K \in \mathcal{T}_h$$

(46)

where $\text{tr}(\cdot)$ and $\text{det}(\cdot)$ denote the trace and determinant of a matrix, $F_K$ is the Jacobian matrix of the affine mapping $F_K$ from the reference element $\hat{K}$ to element $K$, and

$$\sigma_h = \sum_{K \in \mathcal{T}_h} |K| \sqrt{\text{det}(\mathbb{D}^{-1}_K)}.$$

(47)
Geometrically, the alignment condition (45) implies that the element \( \tilde{K} \) in Fig. 1 is a regular simplex, whereas the equidistribution condition indicates that all elements have a constant volume \( \sigma_h/N_v \) in the metric \( \mathbb{D}^{-1} \).

For such a mesh, it is more suitable to replace (36) by

\[
\frac{1}{\theta(d+1)(d+2)} \max_{K \in T_h} \max_{i \neq j} \frac{\tilde{h}_i^K \tilde{h}_j^K}{\cos(\tilde{\alpha}_{ij}^K)} \leq \frac{2}{(1-\theta)(d+1)(d+2)} \min_{K \in T_h} \min_{i = 1, \ldots, d+1} (\tilde{h}_i^K)^2.
\]

(48)

Using the same procedure as in Remark 3.1 and noticing that \( \tilde{K} \) is regular, we can get

\[
\frac{N_v^{-\frac{2}{d}}}{\theta(d+2)} \left( \frac{\sigma_h d!}{\sqrt{d+1}} \right)^{\frac{2}{d}} \leq \Delta t \leq \frac{2N_v^{-\frac{2}{d}}}{(1-\theta)d(d+2)} \left( \frac{\sigma_h d!}{\sqrt{d+1}} \right)^{\frac{2}{d}}.
\]

(49)

Notice that the difference between (44) and (49) lies in that the factor, \( |\Omega|/\alpha \), has been replaced by the volume of \( \Omega \) in the metric \( \mathbb{D}^{-1} \), \( \sigma_h \).

**Remark 3.3.** It is known [27] that a mesh, generated as a uniform mesh in the metric specified by \( M_K = \theta_K \mathbb{D}_K^{-1} \) for all \( K \in T_h \), where \( \theta_K \) is an arbitrary piecewise constant, scalar function defined on \( \Omega \), satisfies the anisotropic nonobtuse angle condition (25). The reader is referred to Ref. [27] for more information on the generation of such meshes.

The lower bound requirement on \( \Delta t \) in (36) can be avoided by using a lumped mass matrix. In this case, scheme (15) is modified into

\[
\begin{bmatrix}
\tilde{M}_{11} & 0 \\
0 & I
\end{bmatrix} + \theta \Delta t \begin{bmatrix}
A_{11} & A_{12} \\
0 & 0
\end{bmatrix} \begin{bmatrix}
u_n \\
u_{n+1}
\end{bmatrix} = \begin{bmatrix}
\tilde{M}_{11} & 0 \\
0 & 0
\end{bmatrix} - (1-\theta) \Delta t \begin{bmatrix}
A_{11} & A_{12} \\
0 & 0
\end{bmatrix} \begin{bmatrix}
u_n \\
u_{n+1}
\end{bmatrix} + \Delta t f^{n+\theta},
\]

(50)

where \( \tilde{M}_{11} \) is the lumped mass matrix with diagonal entries

\[
\tilde{m}_{ii} = \sum_{j=1}^{N_v} m_{ij}, \quad i = 1, \ldots, N_v.
\]

The following theorem can be proven in a similar manner as for Theorem 3.1.

**Theorem 3.2.** Scheme (50) with a lumped mass matrix satisfies a DMP if the mesh satisfies the anisotropic nonobtuse angle condition (25) and the time step size satisfies

\[
\Delta t \leq \frac{1}{(1-\theta)(d+1)} \min_{K \in T_h} \min_{i = 1, \ldots, d+1} \frac{(h_i^K)^2}{\lambda_{\max}(\mathbb{D}_K)}.
\]

(51)

**Remark 3.4.** If the mesh is uniform in the metric specified by \( \mathbb{D}^{-1} \), the condition (51) reduces to

\[
\Delta t \leq \frac{N_v^{-\frac{2}{d}}}{(1-\theta)d} \left( \frac{\sigma_h d!}{\sqrt{d+1}} \right)^{\frac{2}{d}},
\]

(52)

where \( \sigma_h \) is defined in (47).

Numerical Methods for Partial Differential Equations DOI 10.1002/num
B. Two Dimensional Case: $d = 2$

The results in the previous subsection are valid for all dimensions. However, it is known [23] that a Delaunay-type mesh condition, which is weaker than the nonobtuse angle condition (25), is sufficient for a linear finite element approximation to satisfy DMP in 2D for steady-state problems. It is interesting to know if this is also true for time-dependent problems.

Consider an arbitrary interior edge $e_{ij}$. Denote the two vertices of the edge by $a_i$ and $a_j$ and the two elements sharing this common edge by $K$ and $K'$. Let the local indices of the vertices on $K$ be $i_K$ and $j_K$. The angle of $K$ opposite $e_{ij}$ is denoted by $\alpha_{K i_K j_K}$ (when measured in the Euclidean metric) and by $\tilde{\alpha}_{K i_K j_K}$ when measured in the metric $D_K^{-1}$. Similarly, we have $\alpha_{K' i_{K'} j_{K'}}$ and $\tilde{\alpha}_{K' i_{K'} j_{K'}}$.

**Lemma 3.5.** The stiffness matrix $A$ defined in (9) and (11) is an $M$-matrix and has nonnegative row sums if the mesh satisfies the Delaunay-type mesh condition

$$\frac{1}{2} \left[ \tilde{\alpha}_{i_K j_K} + \text{arccot} \left( \frac{\text{det}(D_K)}{\text{det}(D_{K'})} \text{cot}(\tilde{\alpha}_{i_K j_K}) \right) \right] + \tilde{\alpha}_{i_{K'} j_{K'}} + \text{arccot} \left( \frac{\text{det}(D_{K'})}{\text{det}(D_K)} \text{cot}(\tilde{\alpha}_{i_{K'} j_{K'}}) \right) \leq \pi, \quad \forall \text{ interior edges } e_{ij}. \quad (53)$$

**Proof.** See Ref. [23, Theorem 4.1].

**Lemma 3.6.** Matrix $B$ defined in (16) ($0 < \theta \leq 1$) is an $M$-matrix if the mesh satisfies (53) and the time step size satisfies

$$\Delta t \geq \frac{1}{6\theta} \max_{e_{ij}} \frac{|K| + |K'|}{\sqrt{\text{det}(D_K)} \text{cot}(\tilde{\alpha}_{i_K j_K}) + \sqrt{\text{det}(D_{K'})} \text{cot}(\tilde{\alpha}_{i_{K'} j_{K'}})}, \quad (54)$$

where the maximum is taken over all interior edges and $K$ and $K'$ are the two elements sharing the common edge $e_{ij}$.

**Proof.** Inequality (54) follows from (24), (29), and (30).

**Lemma 3.7.** Matrix $C$ defined in (17) ($0 < \theta \leq 1$) is nonnegative if the mesh satisfies (53) and the time step size satisfies

$$\Delta t \leq \frac{1}{6(1 - \theta)} \min_i \sum_{K \in \omega_i} \frac{|\omega_i|}{|K| \lambda_{\max}(D_K) (h_{i_K})^{-2}}, \quad (55)$$

where the minimum is taken over all interior vertices and $\omega_i$ is the patch of the elements containing $a_i$ as its vertex.

**Proof.** The proof is similar to that of Lemma 3.4. Indeed, Lemma 3.5 implies that the off-diagonal entries of $C$ are nonnegative under condition (53). For diagonal entries, from (35) we get

$$m_{ii} - (1 - \theta) \Delta t a_{ii} = \frac{|\omega_i|}{6} - (1 - \theta) \Delta t \sum_{K \in \omega_i} \frac{|K|}{(h_{i_K})^2}. \quad (56)$$
From (33), we can see that the right-side term of the above equation is nonnegative when (55) holds.

Using the above results, we can prove the following theorems in a similar manner as for Theorems 3.1 and 3.2.

**Theorem 3.3.** In two dimensions, scheme (15) satisfies a DMP if the mesh satisfies the Delaunay-type mesh condition (53) and the time step size satisfies (54) and (55), that is,

\[
\frac{1}{6\theta \max_{e_{ij}}} \frac{|K| + |K'|}{\sqrt{\text{det}(D_K)} \cot(\tilde{\alpha}_{K,jk}) + \sqrt{\text{det}(D_{K'})} \cot(\tilde{\alpha}_{K',ik'})} \\
\leq \Delta t \leq \frac{1}{6(1 - \theta)} \min_i \sum_{K \in \omega_i} |K| \lambda_{\max}(D_K) (h_{iK}^K)^{-2},
\]

(56)

where the maximum is taken over all interior edges, \( K \) and \( K' \) are the two elements sharing the common edge \( e_{ij} \), and the minimum is taken over all interior vertices and \( \omega_i \) is the patch of the elements containing \( a_i \) as its vertex.

**Theorem 3.4.** In 2D, scheme (50) with a lumped mass matrix satisfies a DMP if the mesh satisfies the Delaunay-type mesh condition (53) and the time step size satisfies

\[
\Delta t \leq \frac{1}{3(1 - \theta)} \min_i \sum_{K \in \omega_i} |K| \lambda_{\max}(D_K) (h_{iK}^K)^{-2}.
\]

(57)

**Remark 3.5.** Conditions (56) and (57) (for \( d = 2 \)) reduce to (49) and (52), respectively, for a uniform mesh in the metric specified by \( D^{-1} \) but are weaker than conditions (36) and (51) for general meshes.

**IV. NUMERICAL RESULTS**

In this section, we present numerical results obtained for three examples in 2D to demonstrate the significance of both mesh conditions (25, 53) and time step conditions (56) and (57) for DMP satisfaction. Three types of mesh are considered. The first is denoted by Mesh45 where the elements are isosceles right triangles with longest sides in the northeast direction. The second one is denoted by Mesh135 where the elements are isosceles right triangles with longest sides in the northwest direction. Examples of Mesh45 and Mesh135 are shown in Figs. 2(a,b). The third type of mesh, denoted by \( M_{DMP} \), is a uniform mesh in the metric \( M_{DMP} = D^{-1} \) which guarantees satisfaction of mesh condition (25) (cf. Remark 3.3).

The implicit Euler method (corresponding to \( \theta = 1 \) in (15)) is used in our computation. For this method, conditions (36), (51), (56), and (57) place no constraint on the upper bound of \( \Delta t \). For this reason, we consider only the lower bound for the time step size. The lower bound in (36) (related to the anisotropic nonobtuse angle condition) is denoted by \( \Delta t_{Ani} \) and that in (56) (related to the Delaunay-type mesh condition) by \( \Delta t_{Del} \). Unless stated otherwise, the presented results are obtained after 10 steps of time integration.
Example 4.1. The first example is in the form of IBVP (1) with
\[ f \equiv 0, \quad \Omega = [0, 1]^2 \setminus [0.4, 0.6]^2, \quad g = 0 \text{ on } \Gamma_{\text{out}}, \quad g = 4 \text{ on } \Gamma_{\text{in}}, \]
where \( \Gamma_{\text{out}} \) and \( \Gamma_{\text{in}} \) are the outer and inner boundaries of \( \Omega \), respectively; see Fig. 3(a). The initial solution \( u_0(x, y) \) is given as
\[ u_0(x, y) = \begin{cases} 
4, & \text{on } \Gamma_{\text{in}} \\
0, & \text{in } \Omega \setminus [0.2, 0.8]^2 \\
\text{increases linearly,} & \text{from } \Gamma_{\text{mid}} \text{ to } \Gamma_{\text{in}}
\end{cases} \]
where \( \Gamma_{\text{mid}} \) is the boundary of subdomain \([0.2, 0.8]^2\); see Fig. 3. The diffusion matrix is taken as
\[ \mathbb{D} = \begin{bmatrix} 50.5 & 49.5 \\ 49.5 & 50.5 \end{bmatrix}, \]
which has eigenvalues 100 and 1. The principal eigenvectors are in the northeast direction.

This example satisfies the maximum principle and the exact solution (whose analytical expression is unavailable) stays between 0 and 4. Our goal is to produce a numerical solution which also satisfies DMP and stays between 0 and 4.

We first consider Mesh45 and Mesh135. Mesh45 satisfies the anisotropic nonobtuse angle condition (25) as its maximum angle in the metric \( M = \mathbb{D}^{-1} \) is 0.47\pi. It is known [23] that (25) implies the Delaunay-type mesh condition, (53). By direct calculation, we can find that the maximum of the left-hand-side term of (53) is 0.94\pi. On the other hand, Mesh135 satisfies neither of (25) and (53), with the maximum angle in the metric \( M = \mathbb{D}^{-1} \) being 0.94\pi and the maximum of the left-hand-side term of (53) being 1.87\pi.
FIG. 3. The physical domain, boundary condition, and initial solution for Example 4.1. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

The solution contours (after 10 time steps) using Mesh45 and Mesh135 with $h = 2.5 \times 10^{-2}$ and $\Delta t = 1.5 \times 10^{-4}$ are shown in Fig. 4, where $h$ denotes the maximal height of triangular elements of the mesh and $u_{\text{min}}$ is the minimum of the numerical solution. No undershoot occurs in the numerical solution obtained with Mesh45.

FIG. 4. Solution contours obtained for Mesh45 and Mesh135 with $h = 2.5 \times 10^{-2}$ and $\Delta t = 1.5 \times 10^{-4}$ for Example 4.1. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]
The results for Mesh45 are listed in Table I. They show that for meshes with $h \leq 2.5 \times 10^{-2}$, $\Delta t_D$ is smaller than the step size $\Delta t = 1.5 \times 10^{-4}$ used in the computation. As a consequence, time condition (56) and mesh condition (53) is satisfied and Theorem 3.3 implies that the numerical solution satisfies DMP. Table I confirms that no undershoot occurs in the numerical solution or $u_{\min} = 0$. On the other hand, for $h = 5.0 \times 10^{-2}$, neither of time conditions (36) and (56) is satisfied and undershoot with $u_{\min} = -1.41 \times 10^{-7}$ is observed.

The table also records the numerical results obtained for $h = 2.5 \times 10^{-2}$ and $h = 1.25 \times 10^{-2}$ with decreasing $\Delta t$. One can see that no undershoot occurs when $\Delta t \geq \Delta t_D$. However, undershoot occurs when $\Delta t$ continues to decrease and pass $\Delta t_D$. This is consistent with Theorem 3.3.

It is pointed out that $\Delta t_D < \Delta t_A$ for all the cases listed in the table. Moreover, for some cases, we have $\Delta t_D < \Delta t < \Delta t_A$ and no undershoot occurs in the numerical solution. These indicate that time condition (56) (related to the Delaunay-type mesh condition) is weaker than (36) (related to the anisotropic nonobtuse angle condition).

Recall that Mesh135 does not satisfy mesh condition (25) nor (53). Thus, there is no guarantee that the numerical solution obtained with Mesh135 satisfies DMP. Indeed, Table II shows that undershoot occurs in all numerical solutions obtained with various sizes of Mesh135 and various $\Delta t$.

Next, we consider $M_{\text{DMP}}$ meshes which are generated as (quasi-)uniform ones in the metric specified by $M = D^{-1}$. Recall from Remark 3.3 that such meshes satisfy the anisotropic nonobtuse angle condition (25). In our computation, $M_{\text{DMP}}$ meshes are generated using (bidimensional anisotropic mesh generator) code developed by Hecht [56]. An example is shown in Fig. 5(b). Notice that the elements are aligned with the principal diffusion direction (northeast). Because the diffusion tensor $D$ is constant, the mesh is generated initially based on $M_{\text{DMP}} = D^{-1}$ and then kept for the subsequent time steps.

| $h$     | $\Delta t_A$ | $\Delta t_D$ | $\Delta t$ | $u_{\min}$ |
|---------|--------------|--------------|-------------|-------------|
| 5.0e-2  | 1.48e-3      | 2.08e-6      | 1.5e-4      | -8.99e-2    |
| 2.5e-2  | 3.70e-5      | 5.21e-7      | 1.5e-4      | -6.57e-2    |
| 1.25e-2 | 9.25e-6      | 1.30e-7      | 1.5e-4      | -1.58e-2    |
| 1.25e-2 | 9.25e-6      | 1.30e-7      | 1.0e-7      | -2.26e-2    |
| 6.25e-3 | 2.31e-6      | 3.26e-8      | 5.0e-4      | -1.59e-3    |
| 6.25e-3 | 2.31e-6      | 3.26e-8      | 1.5e-5      | -1.43e-2    |
| 6.25e-3 | 2.31e-6      | 3.26e-8      | 1.5e-6      | -2.11e-2    |
The results obtained with $M_{DMP}$ meshes are similar to those obtained with Mesh45. For example, for the $M_{DMP}$ mesh shown in Fig. 5(b), it is found numerically that $\Delta t_{\text{Ani}} = 4.30 \times 10^{-2}$ and $\Delta t_{\text{Del}} = 1.63 \times 10^{-3}$. Theorem 3.3 ensures that no undershoot occurs in the numerical solution when $\Delta t \geq \Delta t_{\text{Del}}$. It is emphasized that (53) and (56) are not necessary for DMP satisfaction and the numerical solution may be free of undershoot for some smaller values of $\Delta t$. In fact, no undershoot is observed numerically for $\Delta t \geq 10^{-4}$. An undershoot-free solution obtained with the mesh shown in Fig. 5(b) and time step size $\Delta t = 1.5 \times 10^{-4}$ is shown in Fig. 5(a). For the same mesh with $\Delta t = 1.0 \times 10^{-5}$, undershoot is observed with $u_{\text{min}} = -1.45 \times 10^{-6}$.

Finally, we consider the lumped mass method. Theorem 3.4 implies that there is no constraint placed on $\Delta t$ for the DMP satisfaction of the numerical solution with the lumped mass matrix and implicit Euler discretization. Indeed, for all Mesh45 meshes and $\Delta t$ considered in Table I, no undershoot is observed numerically for the lumped mass method. The same also holds for $M_{DMP}$ meshes. For example, for the mesh shown in Fig. 5(b), no undershoot is observed in the numerical solution for $\Delta t = 10^{-4}, 10^{-5},$ and $10^{-6}$. For Mesh135 meshes, mesh condition (25) or (53) is not satisfied and thus Theorem 3.4 does not hold. For example, for a case with a Mesh135 mesh with $h = 1.25 \times 10^{-2}$ and $\Delta t = 1.5 \times 10^{-4}$, the numerical solution violates DMP and has a minimum $u_{\text{min}} = -1.60 \times 10^{-2}$.

**Example 4.2.** The second example is the same as Example 4.1 except that the diffusion matrix is taken as a function of $x$ and $y$, i.e.,

$$
\mathbb{D} = \begin{pmatrix}
\cos \theta & -\sin \theta \\
\sin \theta & \cos \theta
\end{pmatrix}
\begin{pmatrix}
k_1 & 0 \\
0 & k_2
\end{pmatrix}
\begin{pmatrix}
\cos \theta & \sin \theta \\
-\sin \theta & \cos \theta
\end{pmatrix},
$$

(58)

where $k_1 = 100$, $k_2 = 1$, and $\theta = \theta(x, y)$ is the angle of the tangential direction at point $(x, y)$ along circles centered at $(0.5, 0.5)$. This diffusion matrix $\mathbb{D}$ also has eigenvalues 1 and 100 but has its principal eigen-direction along the tangential direction of circles centered at $(0.5, 0.5)$.
A physical example with such a diffusion matrix is the toroidal magnetic field in a Tokamak device confining fusion plasma [57]. This problem also satisfies the maximum principle and the solution stays between 0 and 4.

For this example, neither Mesh45 nor Mesh135 (cf. Fig. 2) satisfies the Delaunay-type mesh condition (53). In the metric specified by $M = D^{-1}$, the maximum of the left-hand side of the inequality is $1.87\pi$ for both Mesh45 and Mesh135. Due to the symmetry of the diffusion matrix, both Mesh45 and Mesh135 lead to almost the same results for this example except that undershoot occurs at different locations. Figure 6 shows the results obtained with these meshes for $\Delta t = 5 \times 10^{-5}$. 

Numerical Methods for Partial Differential Equations DOI 10.1002/num
Table III lists numerical results obtained with Mesh45 and $M_{\text{DMP}}$ meshes. Recall that Theorem 3.3 does not apply to Mesh45 meshes as they do not satisfy (53). As a matter of fact, numerical solutions obtained with this type of meshes with or without mass lumping violate DMP and exhibit undershoot. On the other hand, $M_{\text{DMP}}$ meshes generated with $M = D^{-1}$ satisfy the mesh condition. For the lumped mass method, no undershoot occurs in the numerical solution for all values of $\Delta t$. This is consistent with Theorem 3.4. For the standard finite element method, there is no undershoot for relatively large $\Delta t$. It is interesting to point out that for this example with variable $D$, the lower bounds $\Delta t_{\text{Ani}}$ and $\Delta t_{\text{Del}}$ are far too pessimistic. A several magnitude smaller $\Delta t$ can still lead to numerical solutions free of undershoot.

Example 4.3. This example is the same as the previous examples except that the diffusion matrix is taken as in the form (58) with

$$\theta = \frac{1}{2} \arctan \left( \cos \left( \frac{\pi x}{4} \right) \right), \quad k_1 = 100 \cos \left( \left( x^2 + y^2 \right) \frac{\pi}{6} \right), \quad k_2 = 10 \sin \left( \left( x^2 + y^2 + 1 \right) \frac{\pi}{6} \right).$$

Notice that $D$ is a function of $x$ and $y$ and both its eigenvalues and eigenvectors vary with location.

Numerical results are shown in Table IV and Fig. 7. Similar observations can be made as in the previous example. More specifically, both Mesh45 and Mesh135 does not satisfy the Delauney-type mesh condition (53) and thus there is no guarantee that the obtained numerical solution is undershoot-free. On the other hand, $M_{\text{DMP}}$ meshes generated with $M = D^{-1}$ satisfy (53). The numerical solution is guaranteed to be undershoot-free for sufficiently large $\Delta t$ for the standard linear finite element method and for all $\Delta t$ for the lumped mass method.

Table IV. Results obtained with $M_{\text{DMP}}$ meshes for Example 4.3.

| $N_e$ | $\Delta t_{\text{Ani}}$ | $\Delta t_{\text{Del}}$ | $\Delta t$ | $u_{\text{min}}$ | $u_{\text{min}}$ (lumped mass) |
|-------|-----------------|-----------------|-------------|-----------------|-----------------|
| 3180  | 1.83e-2         | 6.38e-4         | 1.0e-4      | 0               | 0               |
|       | 5.0e-5          | 0               | 0           | 0               | 0               |
|       | 1.0e-5          | 0               | 0           | 0               | 0               |
|       | 2.5e-6          | $-7.67e-5$      | 0           | 0               | 0               |
|       | 1.0e-6          | $-6.21e-3$      | 0           | 0               | 0               |
V. CONCLUSIONS

In the previous sections, we have studied the conditions under which a full discretization for IBVP (1) with a general diffusion matrix function satisfies a DMP. The discretization is realized using the $\theta$-method in time and the linear finite element method in space. The main theoretical results are given in Theorems 3.1, 3.2, 3.3, and 3.4.

Specifically, the numerical solution obtained with the full discrete scheme satisfies a DMP when the mesh satisfies the anisotropic nonobtuse angle condition (25) and the time step size satisfies condition (36). As shown in Ref. [27], a mesh satisfying (25) can be generated as a uniform
mesh in the metric specified by $\alpha \mathbb{D}^{-1}$ with $\alpha$ being a scalar function defined on $\Omega_T$. On the other hand, condition (36) essentially requires the time step size to satisfy

$$C_1 h^2 \leq \Delta t \leq \frac{C_2}{1 - \theta} h^2,$$

where $C_1$ and $C_2$ are positive constants, $h$ is the maximal element diameter, and $\theta \in (0, 1]$ is the parameter used in the $\theta$-method. Obviously, this condition is restrictive. This is especially true when the numerical scheme with $\theta \in [0.5, 1]$ is known to be unconditionally stable and no constraint is placed on $\Delta t$ for the sake of stability. Moreover, the presence of the lower bound for $\Delta t$ and the numerical results showing the violation of the maximum principle as $\Delta t \to 0$ seem to support the finding of Thomée and Wahlbin [48] that a semidiscrete standard Galerkin finite element solution violates DMP as the semidiscrete scheme can be considered as the limit of the full discrete scheme as $\Delta t \to 0$. Furthermore, Theorems 3.2 and 3.4 show that the lower bound requirement on $\Delta t$ can be removed when a lumped mass matrix is used. Finally, in 2D, the mesh and time step conditions can be replaced with weaker conditions (53) and (56), respectively. Numerical results in Section IV confirm the theoretical findings.

References

1. P. E. Farrell, M. D. Piggott, C. C. Pain, G. J. Gorman, and C. R. Wilson, Conservative interpolation between unstructured meshes via supermesh construction, Comput Methods Appl Mech Engrg 198 (2009), 2632–2642.
2. S. Günter and K. Lackner, A mixed implicit-explicit finite difference scheme for heat transport in magnetised plasmas, J Comput Phys 228 (2009), 282–293.
3. S. Günter, K. Lackner, and C. Tichmann, Finite element and higher order difference formulations for modelling heat transport in magnetised plasmas, J Comput Phys 226 (2007), 2306–2316.
4. S. Günter, Q. Yu, J. Kruger, and K. Lackner, Modelling of heat transport in magnetised plasmas using non-aligned coordinates, J Comput Phys 209 (2005), 354–370.
5. K. Nishikawa and M. Wakatani, Plasma physics, Springer-Verlag Berlin Heidelberg, New York, 2000.
6. P. Sharma and G. W. Hammett, Preserving monotonicity in anisotropic diffusion, J Comput Phys 227 (2007), 123–142.
7. T. Stix, Waves in plasmas, American Institute of Physics, New York, 1992.
8. I. Aavatsmark, T. Barkve, Bøe, Ø., and T. Mannseth, Discretization on unstructured grids for inhomogeneous, anisotropic media I. Derivation of the methods, SIAM J Sci Comput 19 (1998), 1700–1716.
9. I. Aavatsmark, T. Barkve, Bøe, Ø., and T. Mannseth, Discretization on unstructured grids for inhomogeneous, anisotropic media II. Discussion and numerical results, SIAM J Sci Comput 19 (1998), 1717–1736.
10. P. I. Crumpton, G. J. Shaw, and A. F. Ware, Discretisation and multigrid solution of elliptic equations with mixed derivative terms and strongly discontinuous coefficients, J Comput Phys 116 (1995), 343–358.
11. T. Ertekin, J. H. Abou-Kassem, and G. R. King, Basic applied reservoir simulation, SPE textbook series, Vol. 7, Richardson, Texas, 2001.
12. M. J. Mlacinik and L. J. Durlofsky, Unstructured grid optimization for improved monotonicity of discrete solutions of elliptic equations with highly anisotropic coefficients, J Comput Phys 216 (2006), 337–361.
13. T. F. Chan and J. Shen, Non-texture inpainting by curvature driven diffusions (CDD), J Vis Commun Image Rep 12 (2000), 436–449.
14. T. F. Chan, J. Shen, and L. Vese, Variational PDE models in image processing, Not AMS J 50 (2003), 14–26.
15. D. A. Karras and G. B. Mertzios, New PDE-based methods for image enhancement using SOM and Bayesian inference in various discretization schemes, Meas Sci Technol 20 (2009), 104012.
16. D. Mumford and J. Shah, Optimal approximations by piecewise smooth functions and associated variational problems, Commun Pure Appl Math 42 (1989), 577–685.
17. P. Perona and J. Malik, Scale-space and edge detection using anisotropic diffusion, IEEE Trans Pattern Anal Mach Intel 12 (1990), 629–639.
18. J. Weickert, Anisotropic diffusion in image processing, Teubner-Verlag, Stuttgart, Germany, 1998.
19. J. Brandts, S. Korotov, and M. Křížek, The discrete maximum principle for linear simplicial finite element approximations of a reaction-diffusion problem, Lin Alg Appl 429 (2008), 2344–2357.
20. P. G. Ciarlet, Discrete maximum principle for finite difference operators, Aequationes Math 4 (1970), 338–352.
21. P. G. Ciarlet and P. A. Raviart, Maximum principle and uniform convergence for the finite element method, Comput Meth Appl Mech Engrg 2 (1973), 17–31.
22. A. Drăgănescu, T. F. Dupont, and L. R. Scott, Failure of the discrete maximum principle for an elliptic finite element problem, Math Comp 74 (2004), 1–23.
23. W. Huang, 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 (2011), 319–334. (arXiv:1008.0562v1).
24. J. Karátson and S. Korotov, Discrete maximum principles for finite element solutions of nonlinear elliptic problems with mixed boundary conditions, Numer Math 99 (2005), 669–698.
25. J. Karátson, S. Korotov, and M. Křížek, On discrete maximum principles for nonlinear elliptic problems, Math Comput Sim 76 (2007), 99–108.
26. 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), 3448–3463.
27. X. P. Li and W. Huang, An anisotropic mesh adaptation method for the finite element solution of heterogeneous anisotropic diffusion problems, J Comput Phys 229 (2010), 8072–8094. (arXiv:1003.4530v2).
28. X. P. Li, D. Svyatskiy, and M. Shashkov, 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.
29. R. Liska and M. Shashkov, Enforcing the discrete maximum principle for linear finite element solutions of second-order elliptic problems, Comm Comput Phys 3 (2008), 852–877.
30. C. Lu, W. Huang, and J. Qiu., Maximum principle in linear finite element approximations of anisotropic diffusion-convection-reaction problems, 2012. submitted (arXiv:1201.3564v1).
31. Z. Sheng and G. Yuan, The finite volume scheme preserving extremum principle for diffusion equations on polygonal meshes, J Comput Phys 230 (2011), 2588–2604.
32. G. Stoyan, On a maximum principle for matrices, and on conservation of monotonicity. With applications to discretization methods, Z Angew Math Mech 62 (1982), 375–381.
33. G. Stoyan, On maximum principles for monotone matrices, Lin Alg Appl 78 (1986), 147–161.
34. G. Strang and G. J. Fix, An analysis of the finite element method, Prentice Hall, Englewood Cliffs, NJ, 1973.
35. J. Wang and R. Zhang, Maximum principle for P1-conforming finite element approximations of quasi-linear second order elliptic equations, SIAM J Numer Anal 50 (2012), 626–642.
36. C. Yuan and Z. Sheng, Monotone finite volume schemes for diffusion equations on polygonal meshes, J Comput Phys 227 (2008), 6288–6312.

Numerical Methods for Partial Differential Equations DOI 10.1002/num
37. L. Dascal, Well-posedness and maximum principle for PDE based models in image processing. PhD thesis, Tel-Aviv University, PO Box 39040, Ramat-Aviv 69978, Israel, 2006.

38. I. Faragó, Discrete maximum principle for finite element parabolic models in higher dimensions, Math Comput Simulation 80 (2010), 1601–1611.

39. I. Faragó and R. Horváth, Discrete maximum principle and adequate discretizations of linear parabolic problems, SIAM J Sci Comput 28 (2006), 2313–2336.

40. I. Faragó and R. Horváth, A review of reliable numerical models for three-dimensional linear parabolic problems, Int J Numer Meth Engng 70 (2007), 25–45.

41. I. Faragó and R. Horváth, Continuous and discrete parabolic operators and their qualitative properties, IMA J Numer Anal 29 (2009), 606–631.

42. I. Faragó, R. Horváth, and S. Korotov, Discrete maximum principle for linear parabolic problems solved on hybrid meshes, Appl Numer Math 53 (2005), 249–264.

43. I. Faragó, Karátson, and S. Korotov, Discrete maximum principles for nonlinear parabolic pde systems, Technical Report 93, Department of Mathematics, Tampere University of Technology, Korkeakoulunkatu 10, FI-33720 Tampere, FINLAND, 2009.

44. H. Fujii, Some remarks on finite element analysis of time-dependent field problems, Theory and proactice in finite element structural analysis, University of Tokyo, Tokyo, 1973, pp. 91–106.

45. I. Harari, Stability of semidiscrete formulations for parabolic problems at small time steps, Comput Methods Appl Mech Engrg 193 (2004), 1491–1516.

46. M. Lobo and A. F. Emery, The discrete maximum principle in finite-element thermal radiation analysis, Numer Heat Transfer Part B 24 (1993), 209–227.

47. V. Murti, S. Valliappan, and N. Khalili-Naghadeh, Time step constraints in finite element analysis of the poisson type equations, Comput Struct 31 (1989), 269–273.

48. V. Thomée and L. B. Wahlbin, On the existence of maximum principles in parabolic finite element equations, Math Comput 77 (2008), 11–19.

49. T. Vejchodský, S. Korotov, and A. Hannukainen, Discrete maximum principle for parabolic problems solved by prismatic finite elements, Technical Report 77, Institute of Mathematics, AS CR, Prague, 2008.

50. C. Yang and Y. Gu, Minimum time-step criteria for the galerkin finite element methods applied to one-dimensional parabolic partial differential equations, Numer Meth P D E 22 (2006), 259–273.

51. C. Le Potier, Schéma volumes finis monotone pour des opérateurs de diffusion fortement anisotropes sur des maillages de triangles non structurés, C R Math Acad Sci Paris 341 (2005), 787–792.

52. C. Le Potier, A nonlinear finite volume scheme satisfying maximum and minimum principles for diffusion operators, Int J Finite Vol 6 (2009), electronic only.

53. P. G. Ciarlet, The finite element method for elliptic problems, Amsterdam, North-Holland, 1978.

54. J. Emert and R. Nelson, Volume and surface area for polyhedra and polytopes, Math Mag 70 (1997), 365–371.

55. W. Huang and R. D. Russell, Adaptive moving mesh methods, Springer-Verlag Berlin Heidelberg, New York, 2011.

56. F. Hecht, BAMG – Bidimensional Anisotropic Mesh Generator homepage, available at: http://www.ann.jussieu.fr/~hecht/ftp/bamg/bamg-v1.01.tar.gz, 2010.

57. A. R. Sanderson, G. Chen, X. Tricoche, D. Pugmire, S. Kruger, and J. Breslau, Analysis of recurrent patterns in toroidal magnetic fields, IEEE Trans Vis Comput Graph 16 (2010), 1431–1440.