Commun. Comput. Phys.
doi: 10.4208/cicp.180914.121214a

Vol. **18**, No. 1, pp. 65-90
July 2015

Discrete Maximum Principle for the Weak Galerkin Method for Anisotropic Diffusion Problems

Weizhang Huang[1] and Yanqiu Wang[2,*]

[1] *Department of Mathematics, The University of Kansas, Lawrence, KS 66045, U.S.A.*
[2] *Department of Mathematics, Oklahoma State University, Stillwater, OK 74078, U.S.A.*

Received 18 September 2014; Accepted 12 December 2014

Abstract. A weak Galerkin discretization of the boundary value problem of a general anisotropic diffusion problem is studied for preservation of the maximum principle. It is shown that the direct application of the M-matrix theory to the stiffness matrix of the weak Galerkin discretization leads to a strong mesh condition requiring all of the mesh dihedral angles to be strictly acute (a constant-order away from 90 degrees). To avoid this difficulty, a reduced system is considered and shown to satisfy the discrete maximum principle under weaker mesh conditions. The discrete maximum principle is then established for the full weak Galerkin approximation using the relations between the degrees of freedom located on elements and edges. Sufficient mesh conditions for both piecewise constant and general anisotropic diffusion matrices are obtained. These conditions provide a guideline for practical mesh generation for preservation of the maximum principle. Numerical examples are presented.

AMS subject classifications: 65N30, 65N50

Key words: Discrete maximum principle, weak Galerkin method, anisotropic diffusion.

1 Introduction

We are concerned with the discrete maximum principle for a weak Galerkin discretization of the boundary value problem (BVP) of a two-dimensional diffusion problem,

$$\begin{cases} -\nabla\cdot(\mathcal{A}\nabla u) = f, & \text{in } \Omega, \\ u = g, & \text{on } \partial\Omega, \end{cases} \tag{1.1}$$

where $\Omega \subset \mathbb{R}^2$ is a polygonal domain, f and g are given functions, and $\mathcal{A}$ is a symmetric and uniformly positive definite diffusion matrix defined on Ω. The problem is isotropic

*Corresponding author. *Email addresses:* whuang@ku.edu (W. Huang), yanqiu.wang@okstate.edu (Y. Wang)

 ©2015 Global-Science Press

https://doi.org/10.4208/cicp.180914.121214a Published online by Cambridge University Press

when the diffusion matrix takes the form $\mathcal{A}=\alpha(\mathbf{x})I$ for some scalar function $\alpha(\mathbf{x})$ and anisotropic otherwise. In this work we are interested in the anisotropic situation. It is known (e.g., see Evans [10]) that the classical solution of (1.1) satisfies the maximum principle,

$$f\leq 0 \text{ in } \Omega \quad \Longrightarrow \quad \max_{\mathbf{x}\in\Omega\cup\partial\Omega} u(\mathbf{x})=\max_{\mathbf{x}\in\partial\Omega} u(\mathbf{x}). \tag{1.2}$$

It is theoretically and practically important to investigate if a numerical approximation to (1.1) preserves such a property. Indeed, preservation of the maximum principle has attracted considerable attention from researchers; e.g., see [4, 6, 8, 9, 14–18, 20–22, 24–26, 31–35, 39–42]. For example, it is shown by Ciarlet and Raviart [8] and Brandts et al. [4] that P1 conforming finite element (FE) solutions to isotropic diffusion problems satisfy a discrete maximum principle (DMP) if all of the mesh elements have nonobtuse dihedral angles. This nonobtuse angle condition can be replaced in two dimensions by a weaker condition (the Delaunay condition) [34] requiring the sum of any pair of angles facing a common interior edge to be less than or equal to π. For anisotropic diffusion problems, Drăgănescu et al. [9] show that the nonobtuse angle condition fails to guarantee the satisfaction of DMP for a P1 conforming FE approximation. Various techniques, including local matrix modification [17, 24], mesh optimization [26], and mesh adaptation [22], have been proposed to reduce spurious oscillations. More recently, it is shown by Li and Huang [20] that P1 conforming FE solutions to anisotropic diffusion problems can be guaranteed to satisfy DMP if the mesh satisfies an anisotropic nonobtuse angle condition where mesh dihedral angles are measured in the metric specified by $\mathcal{A}^{-1}$ instead of the Euclidean metric. The result is extended to two dimensional problems [14], problems with convection and reaction terms [25], and time dependent problems [21]. It is emphasized that while DMP has been well studied for conforming FE discretizations, it is less explored for nonconforming or mixed/mixed-hybrid FE methods. Noticeably, DMP has been proven by Gu [11] for a nonconforming FE discretization and by Hoteit et al. [13] and Vohralík and Wohlmuth [36] for mixed/mixed-hybrid FE discetizations. However, their results focus on isotropic diffusion problems. Little is known about those discretizations for anisotropic diffusion problems.

The objective of this paper is to investigate the preservation of the maximum principle by a weak Galerkin approximation of BVP (1.1) with a general anisotropic diffusion matrix $\mathcal{A}$. The weak Galerkin method, recently introduced by Wang and Ye [38], is a FE method which uses a discontinuous FE space and approximates derivatives with weakly defined ones on functions with discontinuity. It can be easy to implement for meshes containing arbitrary polygonal/polyhedral elements [27, 29, 37, 38]. The method has been successfully applied to various model problems [28, 29], and its optimal order convergence has been established for second order elliptic equations [29, 37, 38]. On the other hand, the weak Galerkin method has not been studied in the aspect of preserving the maximum principle. Such studies are useful in practice to avoid unphysical numerical solutions. They are also beneficial in theory since they provide in-depth understandings of the newly developed weak Galerkin method. It should be pointed out that such

https://doi.org/10.4208/cicp.180914.121214a Published online by Cambridge University Press

studies are not trivial. A commonly used and effective tool in the study of preservation of the maximum principle is the theory of M-matrices, matrices in the form of $sI-B$, where I is the identity matrix of some order $n>0$, s is a positive number, and B is a nonnegative matrix ($B(i,j)\geq 0$) with spectral radius less than s. In principle, the theory can be directly applied to the current situation where the weak Galerkin method defines the degrees of freedom separately on edges and inside elements. (For the current work, we consider a simplest and lowest order weak Galerkin method where solutions are approximated using functions that are piecewise constant on edges and inside elements.) Unfortunately, this direct application leads to a strong mesh condition requiring all of the mesh dihedral angles to be strictly acute ($\mathcal{O}(1)$ smaller than 90 degrees) for DMP satisfaction (cf. Remark 3.3). To avoid this difficulty, we use a two-step procedure to study DMP preservation. We first obtain a reduced system involving only the degrees of freedom on edges, and show that it satisfies DMP if the mesh is sufficiently fine and meets an $\mathcal{O}(h^2)$-acute anisotropic angle condition (which requires the angles to be only $\mathcal{O}(h^2)$ away from 90 degrees), where h is the maximal element diameter. We then show that the weak Galerkin approximation to the solution on elements also satisfies DMP. The mesh condition provides a guideline for practical generation of DMP-preserving meshes for the weak Galerkin discretization of general anisotropic diffusion problems.

An outline of the paper is given as follows. A weak Galerkin discretization for BVP (1.1) is given in Section 2. A weak gradient is defined and the properties of the discrete system are discussed in Section 3. Preservation of the maximum principle is studied in Section 4, followed by numerical examples in Section 5. Finally, conclusions are drawn in Section 6.

2 The weak Galerkin formulation

In this section we describe a simplest and lowest order weak Galerkin discretization for BVP (1.1).

We start with introducing some notation. For any given polygonal domain D, we use the standard notation for Sobolev spaces $H^s(D)$ and $H_0^s(D)$ with $s\geq 0$. The inner-product, norm, and semi-norms in $H^s(D)$ are denoted by $(\cdot,\cdot)_{s,D}$, $\|\cdot\|_{s,D}$, and $|\cdot|_{r,D}$ ($0\leq r\leq s$), respectively. When $s=0$, $H^0(D)$ coincides with $L^2(D)$, the space of square integrable functions. In this case, the subscript s is suppressed from these notation. So is the subscript D when $D=\Omega$. For $s<0$, the space $H^s(D)$ is defined as the dual of $H_0^{-s}(D)$. The above notation is extended in a straightforward manner to vector-valued and matrix-valued functions and to an edge, a domain with a lower dimension. Particularly, $\|\cdot\|_{s,e}$ and $\|\cdot\|_e$ denote the norm in $H^s(e)$ and $L^2(e)$, respectively. Functions or quantities on the boundary of Ω or boundary edges of a mesh will be denoted by $(\cdot)^\partial$.

The variational form of BVP (1.1) reads as: Given $f\in H^{-1}(\Omega)$ and $g\in H^{\frac{1}{2}}(\partial\Omega)$, find $u\in H^1(\Omega)$ such that $u=g$ on $\partial\Omega$ and

$$(\mathcal{A}\nabla u,\nabla v)=\langle f,v\rangle,\qquad \forall v\in H_0^1(\Omega), \tag{2.1}$$

https://doi.org/10.4208/cicp.180914.121214a Published online by Cambridge University Press

where $\langle\cdot,\cdot\rangle$ denotes the duality form on Ω.

To define the weak Galerkin approximation of (2.1), let $\mathcal{T}_h$ be a given triangular mesh on Ω. For each triangle $K\in\mathcal{T}_h$, denote the interior and boundary of K by K_0 and ∂K, respectively. Also, denote the diameter (i.e., the length of the longest edge) of K by h_K and let $h=\max_{K\in\mathcal{T}_h}h_K$. The boundary ∂K of K consists of three edges. Denote by $\mathcal{E}_h$ the set of all edges in $\mathcal{T}_h$. For simplicity, hereafter we use "$\lesssim$" to denote "less than or equal to up to a general constant independent of the mesh size or functions appearing in the inequality". We denote by $P_0(K_0)$ the set of constant polynomials on the interior K_0 of triangle K. Likewise, $P_0(e)$ is the set of constant polynomials on $e\in\mathcal{E}_h$. Following [38], we define a weak discrete space on mesh $\mathcal{T}_h$ by

$$V_h=\{v: v|_{K_0}\in P_0(K_0) \text{ for } K\in\mathcal{T}_h;\ v|_e\in P_0(e) \text{ for } e\in\mathcal{E}_h\}.$$

Note that V_h does not require the continuity of its functions across interior edges. A function in V_h is characterized by its values (v_0) on the interior of the elements and those (v_b) on edges. It is often convenient to represent it with two components, $v=\{v_0,v_b\}$. V_h is one of the lowest order weak Galerkin space defined on triangular meshes [38]. To cope with the boundary conditions, for a given piecewise constant function g_h defined on $\mathcal{E}_h\cap\partial\Omega$ we denote

$$V_h^{g_h}=\{v: v\in V_h \text{ and } v_b|_e=g_h|_e \text{ for } e\in\mathcal{E}_h\cap\partial\Omega\}.$$

When $g_h\equiv 0$, $V_h^{g_h}$ becomes V_h^0.

The weak Galerkin method seeks an approximation $u_h\in V_h^{g_h}$ to the solution of (2.1), where g_h is an approximation to the actual boundary data g. Notice that $V_h\not\subset H^1(\Omega)$ and the gradient operator is not defined for functions in V_h. For the moment we assume that a weak gradient, denoted by ∇_w, is defined for functions in V_h. (A definition will be given in the next section.) Then, a weak Galerkin FE approximation is defined as $u_h=\{u_0,u_b\}\in V_h^{g_h}$ such that

$$(\mathcal{A}\nabla_w u_h,\nabla_w v_h)=\langle f,v_0\rangle,\qquad \forall\, v_h=\{v_0,v_b\}\in V_h^0. \tag{2.2}$$

The well-posedness and error estimates of the weak Galerkin formulation (2.2) have been discussed in [27,38].

Equation (2.2) can be cast in a matrix form. Denote the numbers of the triangles, interior edges, and boundary edges in $\mathcal{T}_h$ by N_0, N_b, and N_b^∂, respectively. Let $\phi_{0,i}$ ($i=1,\cdots,N_0$) be the basis function in V_h associated with the i^{th} element such that its value is 1 on the triangle and 0 on other elements or all edges. Similarly, let $\phi_{b,i}$ ($i=1,\cdots,N_b$) and $\phi_{b,i}^\partial$ ($i=1,\cdots,N_b^\partial$) be the basis functions in V_h associated with i^{th} interior and boundary edges, respectively. Then $u_h=\{u_0,u_b\}\in V_h$ can be expressed as

$$u_h=\sum_{i=1}^{N_0}u_{0,i}\phi_{0,i}+\sum_{i=1}^{N_b}u_{b,i}\phi_{b,i}+\sum_{i=1}^{N_b^\partial}u_{b,i}^\partial\phi_{b,i}^\partial. \tag{2.3}$$

https://doi.org/10.4208/cicp.180914.121214a Published online by Cambridge University Press

For convenience, we define the vector representation of u_h as

$$\mathbf{u}=\begin{bmatrix}\mathbf{u}_0\\ \mathbf{u}_b\\ \mathbf{u}_b^\partial\end{bmatrix},\quad \text{with}\quad \mathbf{u}_0=\begin{bmatrix}u_{0,1}\\ u_{0,2}\\ \vdots\\ u_{0,N_0}\end{bmatrix},\qquad \mathbf{u}_b=\begin{bmatrix}u_{b,1}\\ u_{b,2}\\ \vdots\\ u_{b,N_b}\end{bmatrix},\qquad \mathbf{u}_b^\partial=\begin{bmatrix}u_{b,1}^\partial\\ u_{b,2}^\partial\\ \vdots\\ u_{b,N_b^\partial}^\partial\end{bmatrix}.$$

Inserting (2.3) into (2.2) and taking v_h to be the basis functions, we get

$$M\mathbf{u}=\mathbf{F}, \tag{2.4}$$

where

$$M=\begin{bmatrix}M_{0,0} & M_{0,b} & M_{0,b}^\partial\\ M_{b,0} & M_{b,b} & M_{b,b}^\partial\\ 0 & 0 & I\end{bmatrix},\qquad \mathbf{F}=\begin{bmatrix}\mathbf{F}_0\\ 0\\ \mathbf{g}_h\end{bmatrix},$$

$$M_{0,0}=\left[(\mathcal{A}\nabla_w\phi_{0,j},\nabla_w\phi_{0,i})\right]\in\mathbb{R}^{N_0\times N_0},\qquad M_{b,0}=\left[(\mathcal{A}\nabla_w\phi_{0,j},\nabla_w\phi_{b,i})\right]\in\mathbb{R}^{N_b\times N_0},$$

$$M_{0,b}=\left[(\mathcal{A}\nabla_w\phi_{b,j},\nabla_w\phi_{0,i})\right]\in\mathbb{R}^{N_0\times N_b},\qquad M_{b,b}=\left[(\mathcal{A}\nabla_w\phi_{b,j},\nabla_w\phi_{b,i})\right]\in\mathbb{R}^{N_b\times N_b},$$

$$M_{0,b}^\partial=\left[(\mathcal{A}\nabla_w\phi_{b,j}^\partial,\nabla_w\phi_{0,i})\right]\in\mathbb{R}^{N_0\times N_b^\partial},\qquad M_{b,b}^\partial=\left[(\mathcal{A}\nabla_w\phi_{b,j}^\partial,\nabla_w\phi_{b,i})\right]\in\mathbb{R}^{N_b\times N_b^\partial},$$

$$\mathbf{F}_0=\left[\langle f,\phi_{0,i}\rangle\right]\in\mathbb{R}^{N_0},$$

and $\mathbf{g}_h\in\mathbb{R}^{N_b^\partial}$ is the vector representation of the discrete boundary data g_h.

We are interested in the preservation of the maximum principle by the weak Galerkin approximation defined above. A commonly used and effective tool for this type of study is the theory of M-matrices. In principle, the theory can be directly applied to the system (2.4). However, as will be seen in Remark 3.3, the mesh condition ensuring all of the off-diagonal entries of $M_{b,b}$ to be nonpositive is generally stronger than that obtained with a reduced system. Such reduced system is obtained by eliminating the variable $\mathbf{u}_0$ in (2.4), i.e.,

$$\begin{bmatrix}M_{b,b}-M_{b,0}M_{0,0}^{-1}M_{0,b} & M_{b,b}^\partial-M_{b,0}M_{0,0}^{-1}M_{0,b}^\partial\\ 0 & I\end{bmatrix}\begin{bmatrix}\mathbf{u}_b\\ \mathbf{u}_b^\partial\end{bmatrix}=\begin{bmatrix}-M_{b,0}M_{0,0}^{-1}\mathbf{F}_0\\ \mathbf{g}_h\end{bmatrix}. \tag{2.5}$$

In the next section, we shall show that the stiffness matrix of (2.5) can be an M-matrix under suitable, weaker mesh conditions. Notice that the stiffness matrix involves the inverse of $M_{0,0}$. It is easy to see that $M_{0,0}$ is diagonal since the support of any basis function $\phi_{0,i}$ does not overlap with the support of other basis functions $\phi_{0,j}$ with $j\neq i$. Thus, the involvement of the inverse of $M_{0,0}$ will not complicate the analysis of the system. More properties of (2.5) are discussed in the next section.

For convenience, we rewrite (2.5) as

$$\begin{bmatrix}A & A^\partial\\ 0 & I\end{bmatrix}\begin{bmatrix}\mathbf{u}_b\\ \mathbf{u}_b^\partial\end{bmatrix}=\begin{bmatrix}-M_{b,0}M_{0,0}^{-1}\mathbf{F}_0\\ \mathbf{g}_h\end{bmatrix}, \tag{2.6}$$

https://doi.org/10.4208/cicp.180914.121214a Published online by Cambridge University Press

where

$$\bar{A}=\begin{bmatrix} A & A^{\partial} \\ 0 & I \end{bmatrix}, \qquad A=M_{b,b}-M_{b,0}M_{0,0}^{-1}M_{0,b}, \qquad A^{\partial}=M_{b,b}^{\partial}-M_{b,0}M_{0,0}^{-1}M_{0,b}^{\partial}.$$

3 Weak gradient and properties of the discrete system

In this section we present a definition of the weak gradient operator and study the properties of the discrete system (2.6).

We use a definition of the weak gradient operator proposed in [38]. For any element $K\in\mathcal{T}_h$, we denote the space of the lowest order Raviart-Thomas element [30] on K by $RT_0(K)$, i.e.,

$$RT_0(K)=(P_0(K))^2+\mathbf{x}P_0(K).$$

The degrees of freedom of $RT_0(K)$ consist of 0^{th} order moments of normal components on each edge of K. The functions in $RT_0(K)$ can be written in the form of $c(\mathbf{x}-\mathbf{x}_0)$ for some constant c and vector $\mathbf{x}_0$. Define

$$\Sigma_h=\left\{\mathbf{q}\in(L^2(\Omega))^2:\mathbf{q}|_K\in RT_0(K)\text{ for }K\in\mathcal{T}_h\right\}.$$

A discrete weak gradient [38] of $v_h=\{v_0,v_b\}\in V_h$ is defined as $\nabla_w v_h\in\Sigma_h$ such that on each $K\in\mathcal{T}_h$,

$$(\nabla_w v_h,\mathbf{q})_K=-(v_0,\nabla\cdot\mathbf{q})_K+\langle v_b,\mathbf{q}\cdot\mathbf{n}\rangle_{\partial K},\quad\forall\,\mathbf{q}\in RT_0(K), \tag{3.1}$$

where $\mathbf{n}$ is the unit outward normal on ∂K. Such a discrete weak gradient is well defined on V_h. Moreover, $\nabla_w\phi_{0,i}$, $\nabla_w\phi_{b,i}$, and $\nabla_w\phi_{b,i}^{\partial}$ can be found explicitly. To this end, we denote the centroid and area of $K\in\mathcal{T}_h$ by $\mathbf{x}_K$ and $|K|$, respectively, and the length of $e\in\mathcal{E}_h$ by $|e|$.

Lemma 3.1. *Letting K be the i^{th} triangle in $\mathcal{T}_h$, then*

$$\nabla_w\phi_{0,i}|_K=-C_K(\mathbf{x}-\mathbf{x}_K),$$

where

$$C_K=\frac{2|K|}{\|\mathbf{x}-\mathbf{x}_K\|_K^2}. \tag{3.2}$$

Proof. Taking $v_h=\phi_{0,i}$ and $\mathbf{q}=\begin{bmatrix}1\\0\end{bmatrix}$ and $\begin{bmatrix}0\\1\end{bmatrix}$ in (3.1), we have

$$\left(\nabla_w\phi_{0,i},\begin{bmatrix}1\\0\end{bmatrix}\right)_K=-\left(\phi_{0,i},\nabla\cdot\begin{bmatrix}1\\0\end{bmatrix}\right)_K=0,$$

$$\left(\nabla_w\phi_{0,i},\begin{bmatrix}0\\1\end{bmatrix}\right)_K=-\left(\phi_{0,i},\nabla\cdot\begin{bmatrix}0\\1\end{bmatrix}\right)_K=0,$$

https://doi.org/10.4208/cicp.180914.121214a Published online by Cambridge University Press

which implies $\int_K \nabla_w \phi_{0,i} d\mathbf{x} = \mathbf{0}$. Since both components of $\nabla_w \phi_{0,i}$ are linear polynomials, we get $\nabla_w \phi_{0,i} = c(\mathbf{x}-\mathbf{x}_K)$ for some constant c. To determine c, we take $\mathbf{q} = \mathbf{x}-\mathbf{x}_K$ in (3.1) and have

$$(\nabla_w \phi_{0,i}, \mathbf{x}-\mathbf{x}_K)_K = -(\phi_{0,i}, \nabla\cdot(\mathbf{x}-\mathbf{x}_K))_K = -2|K|.$$

Combining this with $\nabla_w \phi_{0,i} = c(\mathbf{x}-\mathbf{x}_K)$, we obtain $c = C_K$. □

Remark 3.1. From the definition of $\mathbf{x}_K$, one can easily see that

$$\left(\mathbf{x}-\mathbf{x}_K, \begin{bmatrix} a \\ b \end{bmatrix}\right)_K = 0, \quad \forall \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^2. \tag{3.3}$$

This can also be verified by direct calculation. Identity (3.3) will be used frequently in the following analysis.

Lemma 3.2. *Assume that the i^{th} interior edge e_i is on ∂K. Then,*

$$\nabla_w \phi_{b,i}|_K = \frac{C_K}{3}(\mathbf{x}-\mathbf{x}_K) + \frac{|e_i|}{|K|}\mathbf{n}_{i,K},$$

where $\mathbf{n}_{i,K}$ is the unit outward normal on e_i with respect to K and C_K is given in (3.2). *The formula also applies to the boundary edge e_i^∂, viz.,*

$$\nabla_w \phi_{b,i}^\partial|_K = \frac{C_K}{3}(\mathbf{x}-\mathbf{x}_K) + \frac{|e_i^\partial|}{|K|}\mathbf{n}_{i,K}^\partial.$$

Proof. We only consider $\nabla_w \phi_{b,i}|_K$ since the proof for $\nabla_w \phi_{b,i}^\partial|_K$ is exactly the same. Taking $v_h = \phi_{b,i}$ and $\mathbf{q} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ in (3.1), we have

$$\left(\nabla_w \phi_{b,i}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}\right)_K = \left\langle \phi_{b,i}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}\cdot\mathbf{n}_{i,K} \right\rangle_{e_i} = |e_i| \begin{bmatrix} 1 \\ 0 \end{bmatrix}\cdot\mathbf{n}_{i,K},$$

$$\left(\nabla_w \phi_{b,i}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}\right)_K = \left\langle \phi_{b,i}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}\cdot\mathbf{n}_{i,K} \right\rangle_{e_i} = |e_i| \begin{bmatrix} 0 \\ 1 \end{bmatrix}\cdot\mathbf{n}_{i,K},$$

which implies $\int_K \nabla_w \phi_{b,i} d\mathbf{x} = |e_i|\mathbf{n}_{i,K}$. Again, since both components of $\nabla_w \phi_{b,i}$ are linear polynomials, we get

$$\nabla_w \phi_{b,i}|_K = c(\mathbf{x}-\mathbf{x}_K) + \frac{|e_i|}{|K|}\mathbf{n}_{i,K}.$$

To determine c, we take $\mathbf{q} = \mathbf{x}-\mathbf{x}_K$ in (3.1) and get

$$\left(c(\mathbf{x}-\mathbf{x}_K) + \frac{|e_i|}{|K|}\mathbf{n}_{i,K}, \mathbf{x}-\mathbf{x}_K\right)_K = \langle \phi_{b,i}, (\mathbf{x}-\mathbf{x}_K)\cdot\mathbf{n}_{i,K}\rangle_{e_i}.$$

https://doi.org/10.4208/cicp.180914.121214a Published online by Cambridge University Press

From (3.3), the left-hand side of the above equation becomes $c\|\mathbf{x}-\mathbf{x}_K\|_K^2$. For the right-hand side, we observe that for any $\mathbf{x}\in e_i$, $(\mathbf{x}-\mathbf{x}_K)\cdot\mathbf{n}_{i,K}$ is equal to one third of the height of triangle K with e_i as the base. This implies that the right-hand side is equal to $\frac{2}{3}|K|$. Combining these results, we obtain $c=C_K/3$ and therefore the expression for $\nabla_w\phi_{b,i}|_K$. □

Remark 3.2. From the definitions of $\phi_{0,i}$ and $\phi_{b,i}$ and Lemmas 3.1 and 3.2, we see that the support of $\nabla_w\phi_{0,i}$ consists of the i^{th} element and that of $\nabla_w\phi_{b,i}$ consists of the elements sharing e_i as a common edge.

Having obtained $\nabla_w\phi_{0,i}$, $\nabla_w\phi_{b,i}$, and $\nabla_w\phi_{b,i}^{\partial}$, we now are ready to find the matrices $M_{0,0}$, $M_{0,b}$, $M_{0,b}^{\partial}$, $M_{b,0}$, $M_{b,b}$, and $M_{b,b}^{\partial}$.

Lemma 3.3. *$M_{0,0}$ is a diagonal matrix with diagonal entries*

$$M_{0,0}(i,i)=C_K^2\|\mathbf{x}-\mathbf{x}_K\|_{\mathcal{A},K}^2,$$

where K is the i^{th} triangle and

$$\|\mathbf{x}-\mathbf{x}_K\|_{\mathcal{A},K}=\big(\mathcal{A}(\mathbf{x}-\mathbf{x}_K),\mathbf{x}-\mathbf{x}_K\big)_K^{\frac{1}{2}}. \tag{3.4}$$

Proof. $M_{0,0}$ is diagonal since the support of the basis function $\phi_{0,i}$ does not overlap with the support of other basis functions $\phi_{0,j}$ with $j\neq i$. Moreover, from Lemma 3.1,

$$\begin{aligned}M_{0,0}(i,i)&=(\mathcal{A}\nabla_w\phi_{0,i},\nabla_w\phi_{0,i})_K\\&=\big(-\mathcal{A}C_K(\mathbf{x}-\mathbf{x}_K),-C_K(\mathbf{x}-\mathbf{x}_K)\big)_K\\&=C_K^2\|\mathbf{x}-\mathbf{x}_K\|_{\mathcal{A},K}^2.\end{aligned}$$

□

Lemma 3.4. *If the i^{th} interior edge e_i is an edge of the j^{th} triangle $K\in\mathcal{T}_h$, then*

$$M_{b,0}(i,j)=M_{0,b}(j,i)=-\frac{1}{3}C_K^2\|\mathbf{x}-\mathbf{x}_K\|_{\mathcal{A},K}^2-C_K\frac{|e_i|}{|K|}\big(\mathcal{A}(\mathbf{x}-\mathbf{x}_K),\mathbf{n}_{i,K}\big)_K.$$

Otherwise, $M_{b,0}(i,j)=M_{0,b}(j,i)=0$. Similarly, if the i^{th} boundary edge e_i^{∂} is an edge of the j^{th} triangle $K\in\mathcal{T}_h$, then

$$M_{0,b}^{\partial}(j,i)=-\frac{1}{3}C_K^2\|\mathbf{x}-\mathbf{x}_K\|_{\mathcal{A},K}^2-C_K\frac{|e_i^{\partial}|}{|K|}\Big(\mathcal{A}(\mathbf{x}-\mathbf{x}_K),\mathbf{n}_{i,K}^{\partial}\Big)_K.$$

Otherwise, $M_{0,b}^{\partial}(j,i)=0$.

https://doi.org/10.4208/cicp.180914.121214a Published online by Cambridge University Press

Proof. From Remark 3.2, we know $M_{b,0}(i,j)=M_{0,b}(j,i)=0$ when e_i is not an edge of the jth triangle K. On the other hand, when e_i is an edge of the jth triangle K, by Lemmas 3.1 and 3.2 we have

$$\begin{aligned}M_{b,0}(i,j)=M_{0,b}(j,i)&=(\mathcal{A}\nabla_w\phi_{0,j},\nabla_w\phi_{b,i})_K\\&=\left(-\mathcal{A}C_K(\mathbf{x}-\mathbf{x}_K),\frac{1}{3}C_K(\mathbf{x}-\mathbf{x}_K)+\frac{|e_i|}{|K|}\mathbf{n}_{i,K}\right)_K\\&=-\frac{1}{3}C_K^2\|\mathbf{x}-\mathbf{x}_K\|_{\mathcal{A},K}^2-C_K\frac{|e_i|}{|K|}(\mathcal{A}(\mathbf{x}-\mathbf{x}_K),\mathbf{n}_{i,K})_K.\end{aligned}$$

This completes the proof for $M_{b,0}(i,j)$ and $M_{0,b}(j,i)$. The proof for $M_{0,b}^{\partial}(j,i)$ is similar. □

For the calculation of $M_{b,b}$ and $M_{b,b}^{\partial}$, we need to know how many elements are sharing a given edge. We first consider the situation of two interior edges which can be the same. Denote by $\mathcal{T}_{i,j}$ the collection of triangles in $\mathcal{T}_h$ that contain both e_i and e_j as edges, i.e.,

$$\mathcal{T}_{i,j}=\{K\in\mathcal{T}_h: e_i,e_j\in\partial K\}.$$

When e_i and e_j are the same, $T_{i,j}$ contains two elements sharing the edge. If they are not the same, they can be either the edges of a triangle or two different triangles. $T_{i,j}$ contains an element in the former case and none in the latter. To summarize, $T_{i,j}$ is given by

$$\mathcal{T}_{i,j}=\begin{cases}\{K,K'\}, & \text{for } e_i=e_j \text{ (where } K \text{ and } K' \text{ are elements satisfying } e_i=e_j\in\partial K\cap\partial K'),\\ \{K\}, & \text{for } e_i\neq e_j \text{ and if there exists an element } K \text{ such that } e_i,e_j\in\partial K,\\ \emptyset, & \text{for } e_i\neq e_j \text{ and if there is no element } K \text{ such that } e_i,e_j\in\partial K.\end{cases}$$

For the situation where e_i is an interior edge and e_j^{∂} is a boundary edge, $\mathcal{T}_{i,j}^{\partial}$ contains at most one triangle in $\mathcal{T}_h$ that takes both e_i and e_j^{∂} as its edges.

Lemma 3.5. *Letting e_i and e_j be two interior edges, then*

$$\begin{aligned}M_{b,b}(i,j)=\sum_{K\in\mathcal{T}_{i,j}}\Bigg[&\frac{C_K^2}{9}\|\mathbf{x}-\mathbf{x}_K\|_{\mathcal{A},K}^2+\frac{C_K}{3|K|}(\mathcal{A}(\mathbf{x}-\mathbf{x}_K),|e_i|\mathbf{n}_{i,K}+|e_j|\mathbf{n}_{j,K})_K\\&+\frac{|e_i||e_j|}{|K|^2}(\mathcal{A}\mathbf{n}_{j,K},\mathbf{n}_{i,K})_K\Bigg].\end{aligned}$$

For the case where e_i is an interior edge and e_j^{∂} is a boundary edge,

$$\begin{aligned}M_{b,b}^{\partial}(i,j)=\sum_{K\in\mathcal{T}_{i,j}^{\partial}}\Bigg[&\frac{C_K^2}{9}\|\mathbf{x}-\mathbf{x}_K\|_{\mathcal{A},K}^2+\frac{C_K}{3|K|}(\mathcal{A}(\mathbf{x}-\mathbf{x}_K),|e_i|\mathbf{n}_{i,K}+|e_j^{\partial}|\mathbf{n}_{j,K}^{\partial})_K\\&+\frac{|e_i||e_j^{\partial}|}{|K|^2}(\mathcal{A}\mathbf{n}_{j,K}^{\partial},\mathbf{n}_{i,K})_K\Bigg].\end{aligned}$$

https://doi.org/10.4208/cicp.180914.121214a Published online by Cambridge University Press

Proof. The results follow directly from Lemma 3.2 and

$$M_{b,b}(i,j)=\sum_{K\in\mathcal{T}_{i,j}}(\mathcal{A}\nabla_w\phi_{b,j},\nabla_w\phi_{b,i})_K,$$
$$M^{\partial}_{b,b}(i,j)=\sum_{K\in\mathcal{T}^{\partial}_{i,j}}(\mathcal{A}\nabla_w\phi^{\partial}_{b,j},\nabla_w\phi_{b,i})_K.$$

□

Remark 3.3. From this lemma, we can see that the requirement of the off-diagonal entries of $M_{b,b}$ being nonpositive yields the mesh condition

$$\frac{C_K^2}{9}\|\mathbf{x}-\mathbf{x}_K\|^2_{\mathcal{A},K}+\frac{C_K}{3|K|}\left(\mathcal{A}(\mathbf{x}-\mathbf{x}_K),|e_i|\mathbf{n}_{i,K}+|e_j|\mathbf{n}_{j,K}\right)_K+\frac{|e_i||e_j|}{|K|^2}(\mathcal{A}\mathbf{n}_{j,K},\mathbf{n}_{i,K})_K\leq 0.$$

To get a feel for this condition, we consider a simple situation with $\mathcal{A}=I$ and e_i and e_j being two different edges of a triangle K. From (3.2) and (3.3), the inequality reduces to

$$\frac{4|K|^2}{9\|\mathbf{x}-\mathbf{x}_K\|^2_K}+\frac{|e_i||e_j|}{|K|}\mathbf{n}_{j,K}\cdot\mathbf{n}_{i,K}\leq 0.$$

Denote the internal angle of K formed by edges e_i and e_j by θ. From $|K|=\frac{1}{2}|e_i||e_j|\sin\theta$, the above condition becomes

$$\cot\theta\geq\frac{2|K|^2}{9\|\mathbf{x}-\mathbf{x}_K\|^2_K}. \tag{3.5}$$

Since the right-hand side has a lower bound

$$\frac{2|K|^2}{9\|\mathbf{x}-\mathbf{x}_K\|^2_K}\geq\frac{2|K|^2}{9\int_K h_K^2\,dx}=\frac{2|K|}{9h_K^2},$$

therefore, for an element with $|K|=\mathcal{O}(h_K^2)$ the mesh condition requires $\cot\theta$ to be $\mathcal{O}(1)$ away from zero, i.e., θ be $\mathcal{O}(1)$ away from $\pi/2$. This is much stronger than that to be obtained with the reduced system. As will be seen in Theorem 4.3 in the next section, the mesh condition obtained with the reduced system only requires θ to be nonobtuse for the current situation $\mathcal{A}=I$ (and for a more general situation with piecewise constant $\mathcal{A}$).

Lemma 3.6. *For any two interior edges e_i and e_j,*

$$A(i,j)=\sum_{K\in\mathcal{T}_{i,j}}\frac{|e_i||e_j|}{|K|^2}\left[(\mathcal{A}\mathbf{n}_{j,K},\mathbf{n}_{i,K})_K-\frac{(\mathcal{A}(\mathbf{x}-\mathbf{x}_K),\mathbf{n}_{i,K})_K(\mathcal{A}(\mathbf{x}-\mathbf{x}_K),\mathbf{n}_{j,K})_K}{\|\mathbf{x}-\mathbf{x}_K\|^2_{\mathcal{A},K}}\right].$$

Moreover, for any interior edge e_i and boundary edge e_j^{∂},

$$A^{\partial}(i,j)=\sum_{K\in\mathcal{T}^{\partial}_{i,j}}\frac{|e_i||e_j^{\partial}|}{|K|^2}\left[\left(\mathcal{A}\mathbf{n}^{\partial}_{j,K},\mathbf{n}_{i,K}\right)_K-\frac{(\mathcal{A}(\mathbf{x}-\mathbf{x}_K),\mathbf{n}_{i,K})_K(\mathcal{A}(\mathbf{x}-\mathbf{x}_K),\mathbf{n}^{\partial}_{j,K})_K}{\|\mathbf{x}-\mathbf{x}_K\|^2_{\mathcal{A},K}}\right].$$

https://doi.org/10.4208/cicp.180914.121214a Published online by Cambridge University Press

Proof. Denote by K_k the k^{th} triangle in $\mathcal{T}_h$. Using the sparsity pattern of $M_{b,0}$, $M_{0,0}$ and $M_{0,b}$, it is not hard to see that

$$A(i,j)=M_{b,b}(i,j)-\sum_{K_k\in\mathcal{T}_{i,j}} M_{b,0}(i,k)M_{0,0}^{-1}(k,k)M_{0,b}(k,j).$$

The rest is a straightforward calculation using Lemmas 3.4 and 3.5. The calculation of $A^{\partial}(i,j)$ is similar. □

Remark 3.4. If the diffusion coefficient $\mathcal{A}$ is piecewise constant on $\mathcal{T}_h$, then by (3.3) we have

$$A(i,j)=\sum_{K\in\mathcal{T}_{i,j}}\frac{|e_i||e_j|}{|K|^2}(\mathcal{A}\mathbf{n}_{j,K},\mathbf{n}_{i,K})_K,\quad A^{\partial}(i,j)=\sum_{K\in\mathcal{T}_{i,j}^{\partial}}\frac{|e_i||e_j^{\partial}|}{|K|^2}(\mathcal{A}\mathbf{n}_{j,K}^{\partial},\mathbf{n}_{i,K})_K.$$

In this case, it is not difficult to see that system (2.6) is exactly the same as the discrete system of the lowest order Crouzeix-Raviart element (P_1 nonconforming FE) for (2.1). Since the mixed-hybrid FE discretization is also equivalent to the P_1 nonconforming FE discretization when $\mathcal{A}$ is piecewise constant [2], we know that the weak Galerkin method is equivalent to the mixed-hybrid FE discretization in this case. This implies that for piecewise constant $\mathcal{A}$, our DMP analysis also applies to P_1 nonconforming and mixed-hybrid FE discretizations which have been studied very little in the past for anisotropic diffusion problems.

It should also be pointed out that the equivalence is valid only in the sense that the weak Galerkin solution u_b on edges is equal to the Lagrange multiplier used in the mixed-hybrid FE discretization. On the other hand, the flux $\mathcal{A}\nabla_w u_h$ and the values of u_0 in the weak Galerkin method are generally different from the dual and primal variables in the mixed-hybrid FE discretization. They are identical only when $\mathcal{A}$ is of the form cI for some constant c.

Remark 3.5. When $\mathcal{A}$ is not piecewise constant, the weak Galerkin method is generally different from the nonconforming or mixed-hybrid FE method. The difference between the weak Galerkin method and the nonconforming FE method can be seen by comparing the entries of the coefficient matrix $\bar{A}$. The difference between the weak Galerkin method and the mixed-hybrid FE method, on the other hand, can be observed from the following fact. In the weak Galerkin method, the weak gradient $\nabla_w u$ lies in the discrete Raviart-Thomas space Σ_h but the flux $\mathcal{A}\nabla_w u$ does not, whereas in the mixed-hybrid FE, the flux $\mathcal{A}\nabla u$ lies in the Raviart-Thomas space but the gradient ∇u does not.

Lemma 3.7. *Matrix A is symmetric and positive definite.*

Proof. It is easy to see that A is symmetric. Next we show that A is positive semi-definite.

https://doi.org/10.4208/cicp.180914.121214a Published online by Cambridge University Press

For any given vector $\mathbf{v}\in\mathbb{R}^{N_b}$, we denote $v_h=\sum_{i=1}^{N_b}v_i\phi_{b,i}$. Noticing $M_{0,b}=M_{b,0}^T$, we have

$$\begin{aligned}\mathbf{v}^TA\mathbf{v}&=\mathbf{v}^TM_{b,b}\mathbf{v}-(M_{0,b}\mathbf{v})^TM_{0,0}^{-1}(M_{0,b}\mathbf{v})\\&=(\mathcal{A}\nabla_w v_h,\nabla_w v_h)-\sum_{i=1}^{N_0}\frac{(\mathcal{A}\nabla_w v_h,\nabla_w\phi_{0,i})_{K_i}^2}{(\mathcal{A}\nabla_w\phi_{0,i},\nabla_w\phi_{0,i})_{K_i}}\\&=\sum_{i=1}^{N_0}\left((\mathcal{A}\nabla_w v_h,\nabla_w v_h)_{K_i}-\frac{(\mathcal{A}\nabla_w v_h,\nabla_w\phi_{0,i})_{K_i}^2}{(\mathcal{A}\nabla_w\phi_{0,i},\nabla_w\phi_{0,i})_{K_i}}\right).\end{aligned}$$

Using the Schwartz inequality on each K_i, it is not difficult to see that $\mathbf{v}^TA\mathbf{v}\geq 0$. To show A is nonsingular, we notice that matrix $\bar{A}=\begin{bmatrix}A & A^{\partial}\\0 & I\end{bmatrix}$ comes from the Schur complement of matrix M. Since M is non-singular, its Schur complement must also be non-singular. Hence, A is non-singular. □

Lemma 3.8. *All row sums of $\bar{A}$ are nonnegative.*

Proof. Let e_i be an interior edge. For each triangle $K\in\mathcal{T}_h$, denote by (x_i,y_i), $i=1,2,3$, the vertices of K and by $\tilde{e}_i$ ($i=1,2,3$) the locally indexed edge opposite to vertex (x_i,y_i). Also denote the unit outward normal vector on these three edges by $\mathbf{n}_1$, $\mathbf{n}_2$ and $\mathbf{n}_3$. Let $\mathcal{T}_i$ be the collection of two triangles sharing edge e_i. By Lemma 3.6, the sum of all entries on the i^{th} row, $1\leq i\leq N_b$, of matrix $\bar{A}$ is

$$\begin{aligned}\sum_{j=1}^{N_b+N_b^{\partial}}\bar{A}(i,j)=\sum_{K\in\mathcal{T}_i}\frac{|e_i|}{|K|^2}\Bigg(&\Big(\mathcal{A}(|\tilde{e}_1|\mathbf{n}_1+|\tilde{e}_2|\mathbf{n}_2+|\tilde{e}_3|\mathbf{n}_3),\mathbf{n}_{i,K}\Big)_K\\&-\frac{(\mathcal{A}(\mathbf{x}-\mathbf{x}_K),\mathbf{n}_{i,K})_K(\mathcal{A}(\mathbf{x}-\mathbf{x}_K),|\tilde{e}_1|\mathbf{n}_1+|\tilde{e}_2|\mathbf{n}_2+|\tilde{e}_3|\mathbf{n}_3)_K}{\|\mathbf{x}-\mathbf{x}_K\|_{\mathcal{A},K}^2}\Bigg).\end{aligned}$$

Notice that

$$|\tilde{e}_1|\mathbf{n}_1+|\tilde{e}_2|\mathbf{n}_2+|\tilde{e}_3|\mathbf{n}_3=\begin{bmatrix}y_3-y_2\\-(x_3-x_2)\end{bmatrix}+\begin{bmatrix}y_1-y_3\\-(x_1-x_3)\end{bmatrix}+\begin{bmatrix}y_2-y_1\\-(x_2-x_1)\end{bmatrix}=\mathbf{0}. \tag{3.6}$$

Combining the above results, we know that the sum of each of the first N_b rows of matrix $\bar{A}$ is 0. The rest of the row sums are just equal to 1. □

4 Discrete maximum principle

We now study the maximum principle for the weak Galerkin approximation (2.6). The weak Galerkin approximation to the solution of BVP (1.1) on edges is said to satisfy DMP if

$$f(\mathbf{x})\leq 0,\quad \forall\mathbf{x}\in\Omega\quad\Longrightarrow\quad\max_{1\leq i\leq N_b}u_{b,i}\leq\max\left\{0,\max_{1\leq i\leq N_b^{\partial}}u_{b,i}^{\partial}\right\}. \tag{4.1}$$

https://doi.org/10.4208/cicp.180914.121214a Published online by Cambridge University Press

The maximum principle has been studied extensively in the past for systems in the form (2.6). For example, Ciarlet [6] shows that the DMP

$$-M_{b,0}M_{0,0}^{-1}\mathbf{F}_0 \le 0 \quad \Longrightarrow \quad \max_{1\le i\le N_b} u_{b,i} \le \max\left\{0, \max_{1\le i\le N_b^\partial} u_{b,i}^\partial\right\} \tag{4.2}$$

holds if and only if

(a) $\bar{A}$ is monotone, i.e., $\bar{A}$ is nonsingular and $\bar{A}^{-1}\ge 0$; and

(b) $\boldsymbol{\xi}+A^{-1}A^\partial\boldsymbol{\xi}^\partial\ge 0$, where $\boldsymbol{\xi}\in\mathbb{R}^{N_b}$ and $\boldsymbol{\xi}^\partial\in\mathbb{R}^{N_b^\partial}$ are vectors consisting of all ones,

where, and hereafter, unless stated otherwise the sign "$\le$" or "$\ge$" is in the elementwise sense when used for vectors or matrices.

The following Lemma is well-known. For completeness, a brief proof is provided.

Lemma 4.1. *The above conditions (a) and (b) hold if*

(i) A is positive definite; and

(ii) All of the off-diagonal entries of $\bar{A}$ are nonpositive; and

(iii) All of the row sums of $\bar{A}$ are nonnegative.

Proof. Conditions (ii) and (iii) imply that $\bar{A}$ is a Z-matrix (defined as a matrix with nonpositive off-diagonal entries and nonnegative diagonal entries) and therefore, A is a Z-matrix too. This, together with (i), implies that A is an M-matrix and thus $A^{-1}\ge 0$. Condition (a) follows by directly examining $\bar{A}^{-1}=\begin{bmatrix} A^{-1} & -A^{-1}A^\partial \\ 0 & I\end{bmatrix}$ and using (ii). Condition (b) follows from (iii) and the fact that A is monotone. □

We should point out that there is a difference between (4.1) and (4.2). Generally speaking, $f(\mathbf{x})\le 0$ does not guarantee

$$-M_{b,0}M_{0,0}^{-1}\mathbf{F}_0 \le 0. \tag{4.3}$$

Thus, we need to include (4.3) as a part of the condition for the weak Galerkin approximation to satisfy DMP.

We now examine system (2.6) more closely. From Lemmas 3.7, 3.8, and 4.1, to verify the maximum principle we need to check the sign of the off-diagonal entries of $\bar{A}$ and the condition (4.3). The off-diagonal entries of $\bar{A}$ are given in Lemma 3.6. When $i\ne j$, we know that either $\mathcal{T}_{i,j}$ is empty, in which case $A(i,j)=0$, or $\mathcal{T}_{i,j}$ contains the only triangle $K\in\mathcal{T}_h$ that has both e_i and e_j as edges. In this case, we have

$$A(i,j)=\frac{|e_i||e_j|}{|K|^2}\left[(\mathcal{A}\mathbf{n}_{j,K},\mathbf{n}_{i,K})_K-\frac{(\mathcal{A}(\mathbf{x}-\mathbf{x}_K),\mathbf{n}_{i,K})_K(\mathcal{A}(\mathbf{x}-\mathbf{x}_K),\mathbf{n}_{j,K})_K}{\|\mathbf{x}-\mathbf{x}_K\|_{\mathcal{A},K}^2}\right]. \tag{4.4}$$

https://doi.org/10.4208/cicp.180914.121214a Published online by Cambridge University Press

Similarly, when interior edge e_i and boundary edge e_j^∂ are edges of triangle K,

$$A^\partial(i,j)=\frac{|e_i||e_j^\partial|}{|K|^2}\left[(\mathcal{A}\mathbf{n}_{j,K}^\partial,\mathbf{n}_{i,K})_K-\frac{(\mathcal{A}(\mathbf{x}-\mathbf{x}_K),\mathbf{n}_{i,K})_K(\mathcal{A}(\mathbf{x}-\mathbf{x}_K),\mathbf{n}_{j,K}^\partial)_K}{\|\mathbf{x}-\mathbf{x}_K\|_{\mathcal{A},K}^2}\right]. \tag{4.5}$$

Theorem 4.1. *If the mesh satisfies*

$$(\mathcal{A}\mathbf{n}_{i,K},\mathbf{n}_{j,K})_K\le\frac{(\mathcal{A}(\mathbf{x}-\mathbf{x}_K),\mathbf{n}_{i,K})_K(\mathcal{A}(\mathbf{x}-\mathbf{x}_K),\mathbf{n}_{j,K})_K}{\|\mathbf{x}-\mathbf{x}_K\|_{\mathcal{A},K}^2},$$
$$\forall K\in\mathcal{T}_h,\quad e_i,e_j,\in\partial K,\quad e_i\neq e_j, \tag{4.6}$$
$$\left|(\mathcal{A}(\mathbf{x}-\mathbf{x}_K),\mathbf{n}_{i,K})_K\right|\le\frac{C_K|K|}{3|e_i|}\|\mathbf{x}-\mathbf{x}_K\|_{\mathcal{A},K}^2,\qquad\forall K\in\mathcal{T}_h,\quad e_i\in\partial K, \tag{4.7}$$

then, u_b, the weak Galerkin approximation (2.6) *to the solution of BVP* (1.1) *on edges, satisfies the discrete maximum principle* (4.1).

Proof. From (4.4) and (4.5), (4.6) implies that all off-diagonal entries of matrix $\bar{A}$ are non-positive. Combining this with Lemmas 3.7 and 3.8, we know that the conditions in Lemma 4.1 are satisfied.

For the condition (4.3), from Lemma 3.4 we see that (4.7) implies that the entries of $M_{b,0}$ are all nonpositive. Moreover, from Lemma 3.3, we know that $M_{0,0}^{-1}$ is a diagonal matrix with positive diagonal entries. From the definitions of $\phi_{0,i}$ and $\mathbf{F}_0$, we have $\mathbf{F}_0\le 0$ when $f(\mathbf{x})\le 0$. Thus, (4.7) implies (4.3) and the solution of (2.6) satisfies the DMP (4.1). □

Theorem 4.2. *Under the assumptions of Theorem* 4.1, *u_0 satisfies the DMP*

$$f(\mathbf{x})\le 0,\quad\forall\mathbf{x}\in\Omega\quad\Longrightarrow\quad\max_{1\le i\le N_0}u_{0,i}\le\max\left\{0,\max_{1\le i\le N_b^\partial}u_{b,i}^\partial\right\}. \tag{4.8}$$

Therefore, the weak Galerkin approximation (2.3) *satisfies the DMP*

$$f(\mathbf{x})\le 0,\quad\forall\mathbf{x}\in\Omega\quad\Longrightarrow\quad\max_{\substack{\mathbf{x}\in\Omega\\ \mathbf{x}\text{ is not a vertex}}}u_h(\mathbf{x})\le\max\left\{0,\max_{1\le i\le N_b^\partial}u_{b,i}^\partial\right\}, \tag{4.9}$$

where the values of $u_h(\mathbf{x})$ on vertices are excluded because $u_h(\mathbf{x})$ assumes multiple values on each vertex due to the discontinuity nature of the weak Galerkin approximation.

Proof. From (2.4), we have

$$\mathbf{u}_0=M_{0,0}^{-1}\left(\mathbf{F}_0-M_{0,b}\mathbf{u}_b-M_{0,b}^\partial\mathbf{u}_b^\partial\right).$$

https://doi.org/10.4208/cicp.180914.121214a Published online by Cambridge University Press

From Lemma 3.4, (4.7) implies $M_{0,b}(i,j)\le 0$ and $M^{\partial}_{0,b}(i,j)\le 0$. Moreover, $f(\mathbf{x})\le 0$ means $F_{0,i}\le 0$. Thus, letting K be the i^{th} element, from Theorem 4.1 we have

$$\begin{aligned}
u_{0,i} &= \frac{1}{M_{0,0}(i,i)}\left(F_{0,i}-\sum_{e_j\in\partial K} M_{0,b}(i,j)u_{b,j}-\sum_{e^{\partial}_j\in\partial K} M^{\partial}_{0,b}(i,j)u^{\partial}_{b,j}\right)\\
&\le \frac{1}{M_{0,0}(i,i)}\left(\sum_{e_j\in\partial K}\left(-M_{0,b}(i,j)\right)\cdot\max_{1\le k\le N_b} u_{b,k}+\sum_{e^{\partial}_j\in\partial K}\left(-M^{\partial}_{0,b}(i,j)\right)\cdot\max_{1\le k\le N^{\partial}_b} u^{\partial}_{b,k}\right)\\
&\le \frac{1}{M_{0,0}(i,i)}\left(\sum_{e_j\in\partial K}\left(-M_{0,b}(i,j)\right)+\sum_{e^{\partial}_j\in\partial K}\left(-M^{\partial}_{0,b}(i,j)\right)\right)\max\left\{0,\max_{1\le k\le N^{\partial}_b} u^{\partial}_{b,k}\right\}.
\end{aligned}$$

From Lemmas 3.3 and 3.4 and the identity (3.6), we get

$$\begin{aligned}
u_{0,i} &\le \frac{\max\{0,\max_{1\le k\le N^{\partial}_b} u^{\partial}_{b,k}\}}{C_K\|\mathbf{x}-\mathbf{x}_K\|^2_{\mathcal{A},K}}\sum_{e_j\in\partial K}\left(\frac{1}{3}C_K\|\mathbf{x}-\mathbf{x}_K\|^2_{\mathcal{A},K}+\frac{|e_j|}{|K|}\left(\mathcal{A}(\mathbf{x}-\mathbf{x}_K),\mathbf{n}_{j,K}\right)_K\right)\\
&= \frac{\max\{0,\max_{1\le k\le N^{\partial}_b} u^{\partial}_{b,k}\}}{C_K\|\mathbf{x}-\mathbf{x}_K\|^2_{\mathcal{A},K}}\left(C_K\|\mathbf{x}-\mathbf{x}_K\|^2_{\mathcal{A},K}+\frac{1}{|K|}(\mathcal{A}(\mathbf{x}-\mathbf{x}_K),\sum_{e_j\in\partial K}|e_j|\mathbf{n}_{j,K})_K\right)\\
&= \frac{\max\{0,\max_{1\le k\le N^{\partial}_b} u^{\partial}_{b,k}\}}{C_K\|\mathbf{x}-\mathbf{x}_K\|^2_{\mathcal{A},K}}\left(C_K\|\mathbf{x}-\mathbf{x}_K\|^2_{\mathcal{A},K}+0\right)=\max\left\{0,\max_{1\le k\le N^{\partial}_b} u^{\partial}_{b,k}\right\},
\end{aligned}$$

which implies (4.8). Combining this with Theorem 4.1 gives (4.9). □

Remark 4.1. In actual computation, the L^2 inner-products on a triangle K involved in the weak Galerkin approximation (2.6) are typically calculated using quadrature rules. Since most of these quadrature rules still define an inner-product on polynomial functions, the above analysis as well as those to be given below in Section 4.1 and Section 4.2 can be extended to the situation with numerical integration. In this case, we need to replace the L^2 inner-products in the analysis by the discrete L^2 inner-product associated with the quadrature rule and to require that the discrete inner-product satisfy condition (3.3), which is true as long as the quadrature is exact for linear polynomials.

Next we look into the conditions in Theorem 4.1 in more detail. Let $\mathcal{A}_K$ be the average of $\mathcal{A}$ over K, i.e.,

$$\mathcal{A}_K=\frac{1}{|K|}\int_K \mathcal{A}d\mathbf{x}.$$

Then we can rewrite the left-hand side of (4.6) as

$$(\mathcal{A}\mathbf{n}_{i,K},\mathbf{n}_{j,K})_K=|K|\,\mathbf{n}^T_{j,K}\mathcal{A}_K\mathbf{n}_{i,K}.$$

https://doi.org/10.4208/cicp.180914.121214a Published online by Cambridge University Press

Denote the unit directions (with the vertices of K being ordered counterclockwisely) along edges e_i and e_j by $\mathbf{e}_i$ and $\mathbf{e}_j$, respectively. By direct calculation one has

$$\mathbf{n}_{j,K}^T \mathcal{A}_K \mathbf{n}_{i,K} = \det(\mathcal{A}_K)\, \mathbf{e}_j^T \mathcal{A}_K^{-1} \mathbf{e}_i.$$

Denote by $\alpha_{i,j,\mathcal{A}_K^{-1}}$ the angle (in K) formed by e_i and e_j and measured in the metric specified by $\mathcal{A}_K^{-1}$. By definition, we have

$$\cos\left(\alpha_{i,j,\mathcal{A}_K^{-1}}\right) = -\frac{\mathbf{e}_j^T \mathcal{A}_K^{-1} \mathbf{e}_i}{\sqrt{\mathbf{e}_i^T \mathcal{A}_K^{-1} \mathbf{e}_i} \cdot \sqrt{\mathbf{e}_j^T \mathcal{A}_K^{-1} \mathbf{e}_j}}.$$

Combining the above results, we get

$$(\mathcal{A}\mathbf{n}_{i,K}, \mathbf{n}_{j,K})_K = -|K| \det(\mathcal{A}_K) \cos\left(\alpha_{i,j,\mathcal{A}_K^{-1}}\right) \sqrt{\mathbf{e}_i^T \mathcal{A}_K^{-1} \mathbf{e}_i} \cdot \sqrt{\mathbf{e}_j^T \mathcal{A}_K^{-1} \mathbf{e}_j}. \tag{4.10}$$

From this, we can see that the statements that $(\mathcal{A}\mathbf{n}_{i,K}, \mathbf{n}_{j,K})_K \le 0$ and the angle $\alpha_{i,j,\mathcal{A}_K^{-1}}$ is nonobtuse are equivalent.

It is noted that the conditions (4.6) and (4.7) can be simplified significantly when $\mathcal{A}$ is piecewise constant on $\mathcal{T}_h$. For this reason, we study this situation first in the following and discuss the general situation afterward.

4.1 The case with piecewise constant $\mathcal{A}$

For this case, from (3.3) the conditions (4.6) and (4.7) reduce to

$$(\mathcal{A}\mathbf{n}_{i,K}, \mathbf{n}_{j,K})_K \le \frac{(\mathcal{A}(\mathbf{x}-\mathbf{x}_K), \mathbf{n}_{i,K})_K (\mathcal{A}(\mathbf{x}-\mathbf{x}_K), \mathbf{n}_{j,K})_K}{\|\mathbf{x}-\mathbf{x}_K\|_{\mathcal{A},K}^2} = 0, \qquad \forall K \in \mathcal{T}_h, \quad e_i, e_j \in \partial K, \quad e_i \ne e_j,$$

$$\left|(\mathcal{A}(\mathbf{x}-\mathbf{x}_K), \mathbf{n}_{i,K})_K\right| = 0 \le \frac{C_K |K|}{3|e_i|} \|\mathbf{x}-\mathbf{x}_K\|_{\mathcal{A},K}^2, \quad \forall K \in \mathcal{T}_h, \quad e_i \in \partial K.$$

Thus, (4.7) is satisfied automatically. Moreover, from (4.10) one can see that (4.6) is true if all the angles of the mesh are nonobtuse when measured in the metric specified by $\mathcal{A}_K^{-1}$. Combining this with Theorem 4.2, we have the following theorem.

Theorem 4.3. *If $\mathcal{A}$ is piecewise constant on $\mathcal{T}_h$ and all of the mesh angles are nonobtuse when measured in the metric specified by $\mathcal{A}_K^{-1}$, the weak Galerkin approximation defined in* (2.2) *and* (2.3) *satisfies the DMP* (4.9).

Remark 4.2. The mesh condition in Theorem 4.3 is referred to as the anisotropic nonobtuse angle condition by Li and Huang [20], a generalization of the well-known nonobtuse angle condition [4, 8] to the case with a general anisotropic diffusion matrix. They show

https://doi.org/10.4208/cicp.180914.121214a Published online by Cambridge University Press

that the P1 conforming FE approximation to BVP (1.1) satisfies a DMP when the mesh condition holds. Like the isotropic diffusion case [19,34,40], it is also shown in [14] that the condition can be replaced by a weaker, Delaunay-type mesh condition in two dimensions. Unfortunately, this may not be true for the weak Galerkin approximation. This is because in the P1 conforming FE approximation, basis functions are associated with vertices and the support of basis functions associated with any pair of neighboring vertices can overlap over two triangles. It is this two-triangle overlap that leads to a weaker condition in two dimensions. On the other hand, the system (2.6) involves basis functions associated with edges and the support of basis functions based on any pair of neighboring edges overlaps over at most a triangle, which unlikely leads to a weaker mesh condition.

4.2 The case with a general anisotropic matrix $\mathcal{A}$

The general case is considered as a perturbation of the piecewise constant case. Define

$$\lambda_{\min,K}(\mathcal{A}) = \min_{\mathbf{x}\in K} \lambda_{\min}(\mathcal{A}(\mathbf{x})),$$

where $\lambda_{\min}(\mathcal{A}(\mathbf{x}))$ denotes the minimal eigenvalue of $\mathcal{A}(\mathbf{x})$. We assume that $\mathcal{A}$ is Lipschitz continuous on each element, i.e., for any $K \in \mathcal{T}_h$, there exists a constant $L_{\mathcal{A},K}$ such that

$$|\mathcal{A}(\mathbf{x}) - \mathcal{A}(\mathbf{y})| \le L_{\mathcal{A},K}|\mathbf{x}-\mathbf{y}|, \quad \forall \mathbf{x},\mathbf{y} \in K.$$

Then, by the mean value theorem we have

$$|\mathcal{A}(\mathbf{x}) - \mathcal{A}_K| \le L_{\mathcal{A},K} h_K, \qquad \forall \mathbf{x} \in K.$$

Theorem 4.4. *Assume that $\mathcal{A}$ is Lipschitz continuous on each element of $\mathcal{T}_h$. If the mesh satisfies*

$$\frac{L_{\mathcal{A},K}^2 h_K^2}{\lambda_{\min,K}^2(\mathcal{A})} \le \cos\left(\alpha_{i,j,\mathcal{A}_K^{-1}}\right), \quad \forall e_i, e_j \in \partial K, \quad e_i \ne e_j, \quad \forall K \in \mathcal{T}_h, \tag{4.11}$$

$$\frac{h_K^3}{|K|} \le \frac{2\lambda_{\min,K}(\mathcal{A})}{3L_{\mathcal{A},K}}, \quad \forall K \in \mathcal{T}_h, \tag{4.12}$$

then the weak Galerkin approximation defined in (2.2) *and* (2.3) *satisfies the DMP* (4.9).

Proof. We first consider the condition (4.6). Notice that $\det(\mathcal{A}_K) = \lambda_{\max}(\mathcal{A}_K)\lambda_{\min}(\mathcal{A}_K)$ and

$$\mathbf{e}_i^T \mathcal{A}_K^{-1} \mathbf{e}_i \ge \frac{\mathbf{e}_i^T \mathbf{e}_i}{\lambda_{\max}(\mathcal{A}_K)} = \frac{1}{\lambda_{\max}(\mathcal{A}_K)}, \qquad \mathbf{e}_j^T \mathcal{A}_K^{-1} \mathbf{e}_j \ge \frac{1}{\lambda_{\max}(\mathcal{A}_K)}.$$

https://doi.org/10.4208/cicp.180914.121214a Published online by Cambridge University Press

Assuming that $\alpha_{i,j,\mathcal{A}_K^{-1}}$ is nonobtuse, from (4.10) we have

$$\begin{aligned}(\mathcal{A}\mathbf{n}_{i,K},\mathbf{n}_{j,K})_K &= -|K|\lambda_{\max}(\mathcal{A}_K)\lambda_{\min}(\mathcal{A}_K)\cos\left(\alpha_{i,j,\mathcal{A}_K^{-1}}\right)\sqrt{\mathbf{e}_i^T\mathcal{A}_K^{-1}\mathbf{e}_i}\cdot\sqrt{\mathbf{e}_j^T\mathcal{A}_K^{-1}\mathbf{e}_j}\\ &\le -\frac{|K|\lambda_{\max}(\mathcal{A}_K)\lambda_{\min}(\mathcal{A}_K)\cos\left(\alpha_{i,j,\mathcal{A}_K^{-1}}\right)}{\lambda_{\max}(\mathcal{A}_K)}\\ &\le -|K|\lambda_{\min,K}(\mathcal{A})\cos\left(\alpha_{i,j,\mathcal{A}_K^{-1}}\right).\end{aligned} \tag{4.13}$$

For the right-hand side of (4.6), we have

$$\begin{aligned}&\frac{(\mathcal{A}(\mathbf{x}-\mathbf{x}_K),\mathbf{n}_{i,K})_K(\mathcal{A}(\mathbf{x}-\mathbf{x}_K),\mathbf{n}_{j,K})_K}{\|\mathbf{x}-\mathbf{x}_K\|_{\mathcal{A},K}^2}\\ =&\frac{\left((\mathcal{A}-\mathcal{A}_K)(\mathbf{x}-\mathbf{x}_K),\mathbf{n}_{i,K}\right)_K\left((\mathcal{A}-\mathcal{A}_K)(\mathbf{x}-\mathbf{x}_K),\mathbf{n}_{j,K}\right)_K}{\|\mathbf{x}-\mathbf{x}_K\|_{\mathcal{A},K}^2}\\ \ge&-\frac{\left|\left((\mathcal{A}-\mathcal{A}_K)(\mathbf{x}-\mathbf{x}_K),\mathbf{n}_{i,K}\right)_K\left((\mathcal{A}-\mathcal{A}_K)(\mathbf{x}-\mathbf{x}_K),\mathbf{n}_{j,K}\right)_K\right|}{\|\mathbf{x}-\mathbf{x}_K\|_{\mathcal{A},K}^2}\\ \ge&-\frac{\left(\int_K|(\mathcal{A}-\mathcal{A}_K)(\mathbf{x}-\mathbf{x}_K)|d\mathbf{x}\right)^2}{\|\mathbf{x}-\mathbf{x}_K\|_{\mathcal{A},K}^2}\\ \ge&-\frac{L_{\mathcal{A},K}^2h_K^2\left(\int_K|\mathbf{x}-\mathbf{x}_K|d\mathbf{x}\right)^2}{\lambda_{\min,K}(\mathcal{A})\|\mathbf{x}-\mathbf{x}_K\|_K^2}\\ \ge&-\frac{L_{\mathcal{A},K}^2|K|h_K^2}{\lambda_{\min,K}(\mathcal{A})}.\end{aligned}$$

From this and (4.13), we know that (4.6) is true when (4.11) holds.

We now consider the condition (4.7). For the left-hand side, we have

$$\begin{aligned}\left|\left(\mathcal{A}(\mathbf{x}-\mathbf{x}_K),\mathbf{n}_{i,K}\right)_K\right| &= \left|\left((\mathcal{A}-\mathcal{A}_K)(\mathbf{x}-\mathbf{x}_K),\mathbf{n}_{i,K}\right)_K\right|\\ &\le L_{\mathcal{A},K}h_K\int_K|\mathbf{x}-\mathbf{x}_K|d\mathbf{x}\\ &\le L_{\mathcal{A},K}h_K^2|K|.\end{aligned} \tag{4.14}$$

For the right-hand side, we get

$$\frac{C_K|K|}{3|e_i|}\|\mathbf{x}-\mathbf{x}_K\|_{\mathcal{A},K}^2\ge\frac{C_K|K|}{3|e_i|}\lambda_{\min,K}(\mathcal{A})\|\mathbf{x}-\mathbf{x}_K\|_K^2=\frac{2\lambda_{\min,K}(\mathcal{A})|K|^2}{3|e_i|}\ge\frac{2\lambda_{\min,K}(\mathcal{A})|K|^2}{3h_K}.$$

From this and (4.14), we know that (4.7) is true when (4.12) holds.

The conclusion is then drawn from Theorem 4.2. □

https://doi.org/10.4208/cicp.180914.121214a Published online by Cambridge University Press

Remark 4.3. When $\mathcal{A}$ is piecewise constant on $\mathcal{T}_h$, we will have $L_{\mathcal{A},K}=0$ for all $K\in\mathcal{T}_h$. It is easy to see that Theorem 4.4 reduces to Theorem 4.3 in this case.

Remark 4.4. The mesh condition (4.11) requires that the mesh be $\mathcal{O}(h^2)$-acute, i.e., all of the mesh angles, measured in the metric specified by $\mathcal{A}_K^{-1}$, are $\mathcal{O}(h^2)$ away from being the right angle. On the other hand, the mesh condition (4.12) is less restrictive, which can be satisfied as long as the mesh is sufficiently fine and the elements are not very skew.

5 Numerical results

In this section we present some numerical results to illustrate the theoretical analysis in the previous sections.

Example 5.1. The first test problem is in the form (1.1) with $\Omega=(0,16)\times(0,16)$,

$$\mathcal{A}=\begin{bmatrix}500.5 & 499.5\\ 499.5 & 500.5\end{bmatrix},\quad f=0,\quad \text{and}\quad g=\begin{cases}1, & \text{for } 0\le x\le 14, y=16,\\ 8-0.5x, & \text{for } 14<x<16, y=16,\\ 1, & \text{for } x=0, 2\le y\le 16,\\ 0.5y, & \text{for } x=0, 0<y<2,\\ 0, & \text{otherwise.}\end{cases}$$

This example has been studied in [14, 20]. Notice that the diffusion coefficient matrix is constant on Ω. We solve this problem using the weak Galerkin method on three types of mesh as shown in Fig. 1. Among them, `mesh45` and `mesh90` satisfy the mesh conditions in Theorem 4.3 whereas `mesh135` does not. The maximum and minimum values of both u_b and u_0 are reported in Table 1. These results confirm the theoretical predictions in Theorem 4.3: both u_b and u_0 obtained with `mesh45` and `mesh90` remain within the range between 0 and 1 but those obtained with `mesh135` have undershoots and overshoots. Contour plots, drawn using the average values at vertices, are shown in Fig. 2. They are consistent with the above observation.

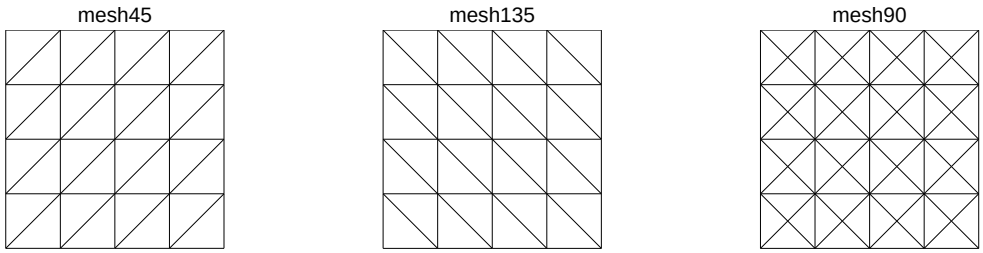


Figure 1: Three types of mesh are used for Example 5.1.

https://doi.org/10.4208/cicp.180914.121214a Published online by Cambridge University Press

Table 1: Maximum and minimum values of the numerical solutions obtained with different meshes for Example 5.1.

	mesh45				mesh135				mesh90			
Size	Max.		Min.		Max.		Min.		Max.		Min.	
	u_b	u_0	u_b	u_0	u_b	u_0	u_b	u_0	u_b	u_0	u_b	u_0
8×8	1	1	0	0	1.038	1.019	$-5.14E-2$	$-2.57E-2$	1	1	0	0
16×16	1	1	0	0	1.041	1.026	$-5.01E-2$	$-3.20E-2$	1	1	0	0
32×32	1	1	0	0	1.035	1.028	$-4.05E-2$	$-3.36E-2$	1	1	0	0
64×64	1	1	0	0	1.028	1.027	$-3.19E-2$	$-3.09E-2$	1	1	0	0

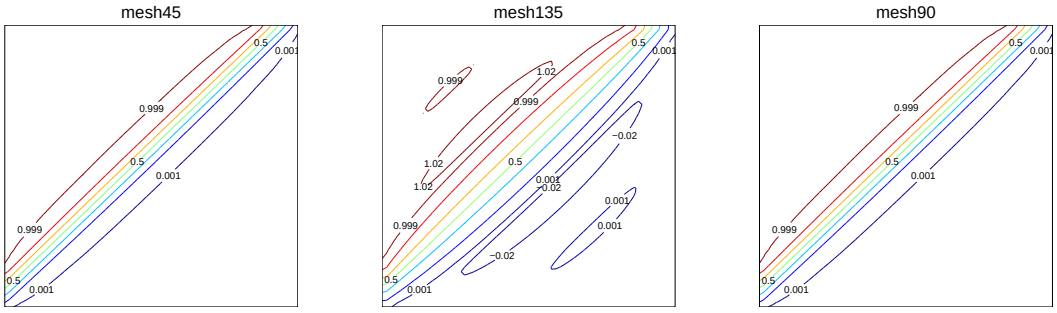


Figure 2: Contours of the numerical solutions obtained with three types of mesh (with size 64×64) for Example 5.1.

Example 5.2. To test the discrete maximum principle for non-constant diffusion coefficients, we consider an example in the form (1.1) with $\Omega=(0,1)\times(0,1)$. Denote by (r,θ) the polar coordinates with the pole centered at $(x,y)=(-0.1,0.5)$. The diffusion matrix is defined as

$$\mathcal{A}=\begin{bmatrix}\cos(\theta) & \sin(\theta)\\ -\sin(\theta) & \cos(\theta)\end{bmatrix}\begin{bmatrix}k_1 & 0\\ 0 & k_2\end{bmatrix}\begin{bmatrix}\cos(\theta) & -\sin(\theta)\\ \sin(\theta) & \cos(\theta)\end{bmatrix},$$

where $k_1=1$, $k_2=1+\gamma e^{-200(r-0.5)^2}$, and γ be a positive parameter. Notice that k_2 is a Gaussian distribution in r and peaks at $r=0.5$ with a maximum value $\gamma+1$. The diffusion matrix becomes more anisotropic around $r=0.5$ for larger γ. We choose $f=0$ and $g=\sin(x+0.5)\pi$. The maximum principle implies that the exact solution of the BVP stays between -1 and 1.

We now test this problem on all three meshes in Fig. 1. Because of symmetry, `mesh45` and `mesh135` give almost identical results up to a reflection across $y=0.5$, thus we examine here only the maximum and minimum values on `mesh45` and `mesh90`. In Tables 2–5, the maximum and minimum values of u_0 and u_b, computed on `mesh45` and `mesh90` using various values of γ, are reported. Both u_b and u_0 have overshoots on some meshes. Moreover, we have marked the triangles and edges, on which u_h has an overshoot, in Figs. 3 and 4, for $\gamma=99$ and various mesh sizes. Behavior for other values of γ and mesh sizes are similar and hence omitted.

https://doi.org/10.4208/cicp.180914.121214a Published online by Cambridge University Press

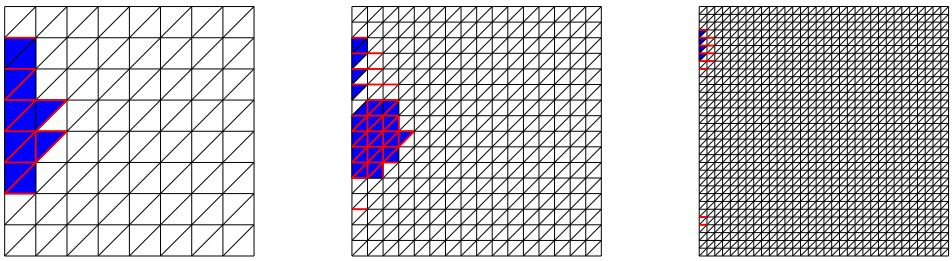


Figure 3: Triangles and edges on which overshoots are observed, using `mesh45` with sizes 8×8, 16×16 and 32×32 for Example 5.2, with $\gamma = 99$.

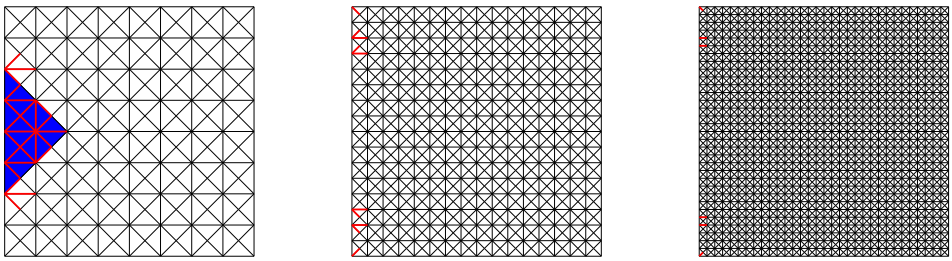


Figure 4: Triangles and edges on which overshoots are observed, using `mesh90` with sizes 8×8, 16×16 and 32×32 for Example 5.2, with $\gamma = 99$.

It is interesting to point out that both `mesh45` and `mesh90` do not satisfy the mesh condition (4.11) for any value of γ or any mesh size. To see this, we first notice that the diffusion matrix $\mathcal{A}$ has negative off-diagonal entries, $(k_2-k_1)\sin\theta\cos\theta$, at all points below the line $y = 0.5$ and so does $\mathcal{A}_K$ at all triangles below this line. From (4.10), one can see that $\cos(\alpha_{i,j,\mathcal{A}_K^{-1}})$ are negative when e_i and e_j are horizontal and vertical edges of those triangles. Thus, `mesh45` does not satisfy (4.11). For `mesh90`, we consider the four triangles within an arbitrary square and denote the unit normal of the diagonal lines by $\tilde{\mathbf{n}}_1$ and $\tilde{\mathbf{n}}_2$. We assume that h is sufficiently small so that $\mathcal{A}_K$ is almost the same on these triangles. Then, the left-hand side of (4.10) takes value $\tilde{\mathbf{n}}_1^T\mathcal{A}_K\tilde{\mathbf{n}}_2$ on two of those triangles and $-\tilde{\mathbf{n}}_1^T\mathcal{A}_K\tilde{\mathbf{n}}_2$ on the other. This means that $\cos(\alpha_{i,j,\mathcal{A}_K^{-1}})$ takes negative values on two of the triangles and therefore `mesh90` violates (4.11).

The above analysis explains why the maximum principle is violated for most cases shown in Tables 2–5. On the other hand, the tables also show that the magnitudes of the undershoots and overshoots decrease as $h\to 0$, which is consistent with the fact that the weak Galerkin approximation is convergent [27, 38]. Moreover, one can see from the tables that the maximum principle is satisfied for some cases even when the mesh condition (4.11) is violated. This does not contradict the theoretical analysis since (4.11) is only a sufficient condition.

Another observation from Table 2–5 is that increasing the value of γ worsens the vio-

https://doi.org/10.4208/cicp.180914.121214a Published online by Cambridge University Press

Table 2: Maximum and minimum values of u_b obtained with different γ for Example 5.2 on `mesh45`.

	$\gamma=20$		$\gamma=40$		$\gamma=60$		$\gamma=99$	
Size	Max.	Min.	Max.	Min.	Max.	Min.	Max.	Min.
8×8	1.02	-1	1.038	-1	1.045	-1	1.051	-1
16×16	1.002	-1	1.012	-1	1.016	-1	1.019	-1
32×32	1	-1	1.002	-1	1.004	-1	1.006	-1
64×64	1	-1	1	-1	1.001	-1	1.002	-1

Table 3: Maximum and minimum values of u_0 obtained with different γ for Example 5.2 on `mesh45`.

	$\gamma=20$		$\gamma=40$		$\gamma=60$		$\gamma=99$	
Size	Max.	Min.	Max.	Min.	Max.	Min.	Max.	Min.
8×8	.992	$-.971$	1.004	$-.971$	1.01	$-.971$	1.015	$-.971$
16×16	.996	$-.991$	.998	$-.991$	1.001	$-.991$	1.005	$-.991$
32×32	.998	$-.997$	.999	$-.997$	.999	$-.997$	1.001	$-.997$
64×64	.999	$-.999$	.999	$-.999$	.999	$-.999$	1.000	$-.999$

Table 4: Maximum and minimum values of u_b obtained with different γ for Example 5.2 on `mesh90`.

	$\gamma=20$		$\gamma=40$		$\gamma=60$		$\gamma=99$	
Size	Max.	Min.	Max.	Min.	Max.	Min.	Max.	Min.
8×8	1.0098	-1	1.006	-1	1.002	-1	1.010	-1
16×16	1.003	-1	1.003	-1	1.002	-1	1.003	-1
32×32	1.0004	-1	1.0005	-1	1.0004	-1	1.0004	-1
64×64	1	-1	1	-1	1	-1	1	-1

Table 5: Maximum and minimum values of u_0 obtained with different γ for Example 5.2 on `mesh90`.

	$\gamma=20$		$\gamma=40$		$\gamma=60$		$\gamma=99$	
Size	Max.	Min.	Max.	Min.	Max.	Min.	Max.	Min.
8×8	.995	$-.981$	.997	$-.981$	.999	$-.981$	1.007	$-.981$
16×16	.997	$-.994$	.998	$-.994$	.999	$-.993$	.999	$-.993$
32×32	.999	$-.998$	.999	$-.998$	.999	$-.998$	.999	$-.998$
64×64	.999	$-.999$	.999	$-.999$	.999	$-.999$	.999	$-.999$

lation of the maximum principle. This is because the problem becomes more anisotropic when γ gets larger.

Next, we solve the Example 5.2 on meshes generated by the Delaunay-type triangulator BAMG (Bidimensional Anisotropic Mesh Generator, developed by Hecht [12]). BAMG is designed to generate triangular meshes with a given metric tensor. Meshes

https://doi.org/10.4208/cicp.180914.121214a Published online by Cambridge University Press

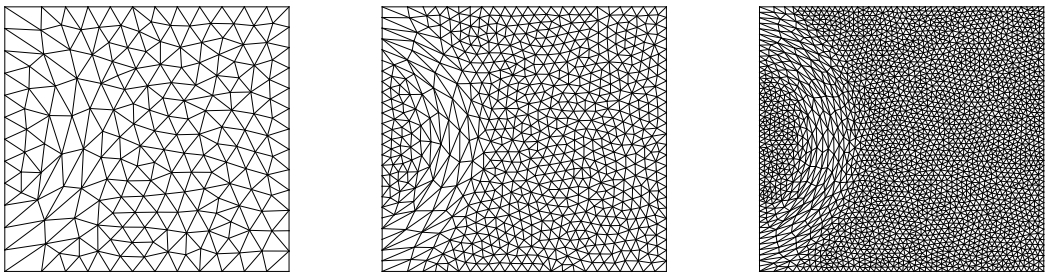


Figure 5: Meshes generated by BAMG with metric $\mathcal{A}^{-1}$ and $\gamma=20$ for Example 5.2.

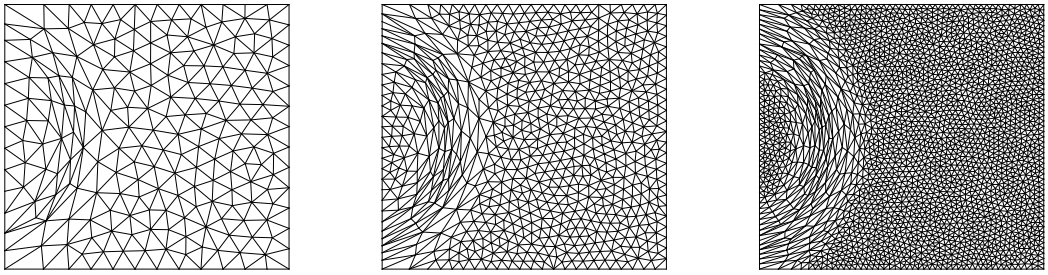


Figure 6: Meshes generated by BAMG with metric $\mathcal{A}^{-1}$ and $\gamma=99$ for Example 5.2.

generated by BAMG with $\mathcal{A}^{-1}$ as the metric tensor are shown in Figs. 5 and 6 for different values of γ. These meshes match well with the diffusion coefficient $\mathcal{A}$, which becomes more anisotropic around $r=0.5$ and remains nearly isotropic away from $r=0.5$. By comparing meshes with different values of γ, we can see that the triangles become more skewed around $r=0.5$ as γ becomes larger. Numerical experiments show that weak Galerkin solutions on these meshes satisfy the discrete maximum principle (the results are not shown to save space): all entries of u_b as well as u_0 lie between -1 and 1, which agrees with the theoretical prediction given in Theorem 4.4.

6 Conclusions

In the previous sections we have studied the discrete maximum principle for a simplest and lowest order weak Galerkin discretization of the anisotropic diffusion BVP (1.1). The main results are stated in Theorems 4.3 and 4.4.

Theorem 4.3 states that the weak Galerkin approximation to BVP (1.1) satisfies a discrete maximum principle if the diffusion matrix $\mathcal{A}$ is piecewise constant on the mesh and for any $K\in\mathcal{T}_h$, all of the angles of K are nonobtuse when measured in the metric specified by $\mathcal{A}_K^{-1}$, where $\mathcal{A}_K$ is the average of $\mathcal{A}$ over K. For the general anisotropic diffusion situation (cf. Theorem 4.4), the mesh is required to be sufficiently fine, not very skewed (cf. (4.12)), and $\mathcal{O}(h^2)$-acute (cf. (4.11)) when measured in the metric specified

https://doi.org/10.4208/cicp.180914.121214a Published online by Cambridge University Press

by $\mathcal{A}_K^{-1}$. These conditions are comparable to the mesh conditions for P1 conforming finite elements in three and higher dimensions but stronger in two dimensions where a Delaunay-type condition is sufficient to guarantee a P1 conforming FE approximation to satisfy a discrete maximum principle.

Finally, it is worth pointing out that although the analysis has been carried out in this work in two dimensions, it applies to three and higher dimensions without major modifications.

Acknowledgments

This work was supported in part by the NSF under Grant DMS-1115118.

References

[1] R. A. Adams. *Sobolev Spaces*. Academic Press, New York, 1975.
[2] D. Arnold and F. Brezzi. Mixed and nonconforming finite element methods: implementation, postprocessing and error estimates. *RAIRO Modél. Math. Anal. Numér.*, 19:7–32, 1985.
[3] A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. Society for Industrial and Applied Mathematics, Philadelphia, 1994.
[4] 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.*, 429:2344–2357, 2008.
[5] E. Burman and A. Ern. Discrete maximum principle for Galerkin approximations of the Laplace operator on arbitrary meshes. *C. R. Acad. Sci. Paris*, Ser.I 338:641–646, 2004.
[6] P. G. Ciarlet. Discrete maximum principle for finite difference operators. *Aequationes Math.*, 4:338–352, 1970.
[7] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems*. North-Holland, Amsterdam, 1978.
[8] P. G. Ciarlet and P.-A. Raviart. Maximum principle and uniform convergence for the finite element method. *Comput. Meth. Appl. Mech. Engrg.*, 2:17–31, 1973.
[9] A. Drăgănescu, T. F. Dupont, and L. R. Scott. Failure of the discrete maximum principle for an elliptic finite element problem. *Math. Comp.*, 74:1–23, 2004.
[10] L. C. Evans. *Partial Differential Equations*. American Mathematical Society, Providence, Rhode Island, 1998. Graduate Studies in Mathematics, Volume 19.
[11] J. Gu. *Domain Decomposition Methods for Nonconforming Finite Element Discretizations*. Nova Science Publishers, Inc., New York, 1999.
[12] F. Hecht. BAMG – Bidimensional Anisotropic Mesh Generator homepage. `http://www.ann.jussieu.fr/~hecht/ftp/bamg/`, 1997.
[13] H. Hoteit, R. Mosé, B. Philippe, Ph. Ackerer and J. Erhel. The maximum principle violations of the mixed-hybrid finite-element method applied to diffusion equations. *Int. J. Numer. Meth. Engng.*, 55:1373–1390, 2002.
[14] 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.*, 4:319–334, 2011. (arXiv:1008.0562).

https://doi.org/10.4208/cicp.180914.121214a Published online by Cambridge University Press

[15] J. Karátson and S. Korotov. Discrete maximum principles for finite element solutions of nonlinear elliptic problems with mixed boundary conditions. *Numer. Math.*, 99:669–698, 2005.
[16] J. Karátson, S. Korotov, and M. Křížek. On discrete maximum principles for nonlinear elliptic problems. *Math. Comput. Sim.*, 76:99–108, 2007.
[17] 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.*, 228:3448–3463, 2009.
[18] M. Křížek and Q. Lin. On diagonal dominance of stiffness matrices in 3D. *East-West J. Numer. Math.*, 3:59–69, 1995.
[19] F. W. Letniowski. Three-dimensional Delaunay triangulations for finite element approximations to a second-order diffusion operator. *SIAM J. Sci. Stat. Comput.*, 13:765–770, 1992.
[20] X. P. Li and W. Huang. An anisotropic mesh adaptation method for the finite element solution of heterogeneous anisotropic diffusion problems. *J. Comput. Phys.*, 229:8072–8094, 2010. (arXiv:1003.4530).
[21] X. P. Li and W. Huang. Maximum principle for the finite element solution of time dependent anisotropic diffusion problems. *Numer Meth. P. D. E.*, 29:1963–1985, 2013. (arXiv:1209.5657).
[22] 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.
[23] K. Lipnikov, M. Shashkov, D. Svyatskiy, and Y. Vassilevski. Monotone finite volume schemes for diffusion equations on unstructured triangular and shape-regular polygonal meshes. *J. Comput. Phys.*, 227:492–512, 2007.
[24] R. Liska and M. Shashkov. Enforcing the discrete maximum principle for linear finite element solutions of second-order elliptic problems. *Comm. Comput. Phys.*, 3:852–877, 2008.
[25] C. Lu, W. Huang, and J. Qiu. Maximum principle in linear finite element approximations of anisotropic diffusion-convection-reaction problems. *Numer. Math.*, 127:515–537, 2014. (arXiv:1201.3564).
[26] M. J. Mlacnik and L. J. Durlofsky. Unstructured grid optimization for improved monotonicity of discrete solutions of elliptic equations with highly anisotropic coefficients. *J. Comput. Phys.*, 216:337–361, 2006.
[27] L. Mu, J. Wang, Y. Wang, and X. Ye. A weak Galerkin mixed finite element method for biharmonic equations. In O.P. Iliev et.al., editors, *Numerical Solution of Partial Differential Equations: Theory, Algorithms, and Their Applications*, volume 45 of *Springer Proceedings in Mathematics & Statistics*, New York, 2013. Springer-Verlag. (arXiv:1210.3818).
[28] L. Mu, J. Wang, Y. Wang, and X. Ye. A computational study of the weak Galerkin method for second order elliptic equations. *Numer. Alg.*, 63:753–777, 2013. (arXiv:1111.0618).
[29] L. Mu, J. Wang, and X. Ye. Weak Galerkin finite element methods on polytopal meshes. *Int. J. Numer. Anal. Model.*, 12:31–53, 2015. (arXiv:1204.3655).
[30] P. Raviart and J. Thomas. A mixed finite element method for second order elliptic problems. In I. Galligani and E. Magenes, editors, *Mathematical Aspects of the Finite Element Method*, volume 606 of *Lectures Notes in Mathematics*, New York, 1977. Springer-Verlag.
[31] Z. Sheng and G. Yuan. The finite volume scheme preserving extremum principle for diffusion equations on polygonal meshes. *J. Comput. Phys.*, 230:2588–2604, 2011.
[32] G. Stoyan. On a maximum principle for matrices, and on conservation of monotonicity. With applications to discretization methods. *Z. Angew. Math. Mech.*, 62:375–381, 1982.
[33] G. Stoyan. On maximum principles for monotone matrices. *Lin. Alg. Appl.*, 78:147–161, 1986.

https://doi.org/10.4208/cicp.180914.121214a Published online by Cambridge University Press

[34] G. Strang and G. J. Fix. *An Analysis of the Finite Element Method*. Prentice Hall, Englewood Cliffs, NJ, 1973.
[35] R. S. Varga. On a discrete maximum principle. *SIAM J. Numer. Anal.*, 3:355–359, 1966.
[36] M. Vohralík and B.I. Wohlmuth. Mixed finite element methods: Implementation with one unknown per element, local flux expressions, positivity, polygonal meshes, and relations to other methods. *Math. Models Methods Appl. Sci.*, 23:803–838, 2013.
[37] J. Wang and X. Ye. A weak Galerkin mixed finite element method for second-order elliptic problems. *Math. Comp.*, 83:2101–2126, 2014. (arXiv:1202.3655).
[38] J. Wang and X. Ye. A weak Galerkin finite element method for second-order elliptic problems. *J. Comput. Appl. Math.*, 241:103–115, 2013. (arXiv:1104.2897).
[39] J. Wang and R. Zhang. Maximum principle for P1-conforming finite element approximations of quasi-linear second order elliptic equations. *SIAM J. Numer. Anal.*, 50:626–642, 2012. (arXiv:1105.1466).
[40] J. Xu and L. Zikatanov. A monotone finite element scheme for convection-diffusion equations. *Math. Comput.*, 69:1429–1446, 1999.
[41] G. Yuan and Z. Sheng. Monotone finite volume schemes for diffusion equations on polygonal meshes. *J. Comput. Phys.*, 227:6288–6312, 2008.
[42] Y. Zhang, X. Zhang, and C.-W. Shu. Maximum-principle-satisfying second order discontinuous Galerkin schemes for convection-diffusion equations on triangular meshes. *J. Comput. Phys.*, 234:295–316, 2013.

https://doi.org/10.4208/cicp.180914.121214a Published online by Cambridge University Press