Commun. Comput. Phys.
doi: 10.4208/cicp.280315.140815a

Vol. **19**, No. 2, pp. 473-495
February 2016

Monotone Finite Difference Schemes for Anisotropic Diffusion Problems via Nonnegative Directional Splittings

Cuong Ngo[1] and Weizhang Huang[1,*]

[1] *Department of Mathematics, University of Kansas, Lawrence, KS 66045, USA.*

Received 28 March 2015; Accepted (in revised version) 14 August 2015

Abstract. Nonnegative directional splittings of anisotropic diffusion operators in the divergence form are investigated. Conditions are established for nonnegative directional splittings to hold in a neighborhood of an arbitrary interior point. The result is used to construct monotone finite difference schemes for the boundary value problem of anisotropic diffusion operators. It is shown that such a monotone scheme can be constructed if the underlying diffusion matrix is continuous on the closure of the physical domain and symmetric and uniformly positive definite on the domain, the mesh spacing is sufficiently small, and the size of finite difference stencil is sufficiently large. An upper bound for the stencil size is obtained, which is determined completely by the diffusion matrix. Loosely speaking, the more anisotropic the diffusion matrix is, the larger stencil is required. An exception is the situation with a strictly diagonally dominant diffusion matrix where a three-by-three stencil is sufficient for the construction of a monotone finite difference scheme. Numerical examples are presented to illustrate the theoretical findings.

AMS subject classifications: 65N06, 65M06

Key words: Anisotropic diffusion, anisotropic coefficient, discrete maximum principle, monotone scheme, finite difference, nonnegative directional splitting.

1 Introduction

We consider the finite difference solution of the boundary value problem

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

where $\Omega=(0,1)\times(0,1)$, f and g are given functions, and the diffusion matrix,

$$\mathbb{D}=\begin{bmatrix} a(x,y) & b(x,y) \\ b(x,y) & c(x,y) \end{bmatrix}, \tag{1.2}$$

*Corresponding author. *Email addresses:* cngo@ku.edu (C. Ngo), whuang@ku.edu (W. Huang)

 ©2016 Global-Science Press

https://doi.org/10.4208/cicp.280315.140815a Published online by Cambridge University Press

is assumed to be continuous on $\overline{\Omega}\equiv\Omega\cup\partial\Omega$ and symmetric and uniformly positive definite on Ω. We are interested in the case where $\mathbb{D}$ depends on spatial location (heterogeneous diffusion) and has unequal eigenvalues at least on some portion of Ω (anisotropic diffusion). For this case, (1.1) is often called a heterogeneous anisotropic diffusion problem. Anisotropic diffusion arises from various areas of science and engineering, including plasma physics [9], petroleum reservoir simulation [7,20], and image processing [12,31].

When a conventional method such as a finite element, a finite difference, or a finite volume method, is applied to this problem, spurious oscillations can occur in the numerical solution. This difficulty can be overcome by numerical schemes satisfying a discrete maximum principle (DMP). The development and studies of DMP-preserving schemes have received considerable attention in the past particularly in the context of finite element and finite volume methods, e.g., see [3–6, 9–11, 13–19, 24–30, 32, 33]. On the other hand, relatively less work has been done in the context of finite difference (FD) methods. The major effort has been made to study monotone (or called positive-type or nonnegative-type) schemes, which belong to a special class of DMP-preserving schemes with the coefficient matrix (or the Jacobian matrix in the nonlinear case) of the corresponding algebraic system being an M-matrix (e.g., see [30] for the definition of M-matrices). For example, Motzkin and Wasow [21] prove that a monotone FD scheme exists for any linear second-order elliptic problem when the mesh is sufficiently fine. Greenspan and Jain [8], using nonnegative directional splittings (see the definition below), propose a way to construct such schemes for elliptic operators in the nondivergence form

$$L[u]\equiv a(x,y)u_{xx}+2b(x,y)u_{xy}+c(x,y)u_{yy} \quad \text{with} \quad b(x,y)^2<a(x,y)c(x,y). \tag{1.3}$$

The results are extended to elliptic problems in the divergence form (1.1) by Weickert [31]. Consistent and stable monotone FD schemes are shown to be convergent (to the solution) for linear second-order elliptic problems by Bramble et al. [2] and (to the viscosity solution) for nonlinear second-order degenerate elliptic or parabolic partial differential equations by Barles and Souganidis [1]. Oberman [23] studies degenerate elliptic schemes (a special type of monotone scheme) for a general class of nonlinear degenerate elliptic problems.

The objective of this work is to study the construction of monotone FD schemes for problems in the form of (1.1) using nonnegative directional splittings. Weickert's results are improved in two aspects. We first present a condition under which nonnegative directional splittings hold for a neighborhood of an arbitrary point. As we will see below, it is necessary to consider nonnegative directional splittings in a neighborhood for the construction of monotone FD schemes. We then extend the result to the situation where the coefficient $b(x,y)$ can change sign over the domain. To be more specific, we recall that Weickert [31] considers the directional splitting

$$\nabla\cdot(\mathbb{D}\nabla u)(x_0,y_0)=\partial_x(\gamma_0\partial_x u)(x_0,y_0)+\partial_\beta(\gamma_1\partial_\beta u)(x_0,y_0)+\partial_y(\gamma_2\partial_y u)(x_0,y_0), \tag{1.4}$$

https://doi.org/10.4208/cicp.280315.140815a Published online by Cambridge University Press

where (x_0,y_0) is a given point in Ω, $\partial_\beta=(\sin(\beta),\cos(\beta))^T\cdot\nabla$, $\gamma_i=\gamma_i(x,y)$ $(i=0,1,2)$ are functions, and β is a constant (angle) (but can vary with (x_0,y_0)). Weickert [31, Page 90] shows that the condition

$$a(x_0,y_0)-b(x_0,y_0)\cot(\beta)\geq 0 \quad \text{and} \quad c(x_0,y_0)-b(x_0,y_0)\tan(\beta)\geq 0 \tag{1.5}$$

is sufficient to guarantee that γ_i $(i=0,1,2)$ are nonnegative at (x_0,y_0) (nonnegative directional splitting). Unfortunately, such a point-wise result is insufficient for the construction of monotone FD schemes for the divergence form (1.4) since an FD discretization has to use the information of the coefficients γ_i in a neighborhood of any mesh point. To avoid this difficulty, we develop a condition under which nonnegative directional splittings hold in a neighborhood of a given point. Moreover, we study how the condition can be extended to the situation where $b(x,y)$ changes sign over the domain. An upper bound for the stencil size is obtained, which is completely determined by the diffusion matrix. Loosely speaking, the more anisotropic the diffusion matrix is, the larger stencil is required. An exception is the situation with a strictly diagonally dominant diffusion matrix where a three-by-three stencil is sufficient for the construction of a monotone finite difference scheme (cf. Theorem 3.2).

An outline of this paper is as follows. Nonnegative directional splittings for (1.1) are studied in Section 2. Construction of monotone FD schemes using the results in Section 2 is discussed in Section 3. Numerical examples are presented in Section 4, followed by the conclusions in Section 5.

2 Nonnegative directional splitting

In this section we consider nonnegative directional splittings for problems in the form of (1.1). Recall that our goal is to use the splittings to construct monotone FD schemes. Thus we are interested in the situations where such a splitting holds for a neighborhood of a given point. We first consider the case where $b(x,y)$ does not change sign and then extend the result to the sign changing situation. Finally, we establish a condition under which the nonnegative directional splitting holds uniformly over the domain for any neighborhood of a fixed size. The result is needed in constructing monotone FD schemes on a uniform mesh.

Lemma 2.1. *Consider an arbitrary point (x_0,y_0) in Ω and a neighborhood Ω_0 of (x_0,y_0). We assume that $b(x,y)$ does not change sign on Ω_0. If there exists a constant β satisfying*

$$\begin{cases} \sup\limits_{\Omega_0}\dfrac{b(x,y)}{a(x,y)}\leq\tan\beta\leq\inf\limits_{\Omega_0}\dfrac{c(x,y)}{b(x,y)}, & \text{for } b\geq 0 \text{ and } b\not\equiv 0 \text{ on } \Omega_0,\\ \sup\limits_{\Omega_0}\dfrac{c(x,y)}{b(x,y)}\leq\tan\beta\leq\inf\limits_{\Omega_0}\dfrac{b(x,y)}{a(x,y)}, & \text{for } b\leq 0 \text{ and } b\not\equiv 0 \text{ on } \Omega_0,\\ \beta \text{ is any finite value with } \sin(\beta)\cos(\beta)\neq 0, & \text{for } b\equiv 0 \text{ on } \Omega_0, \end{cases} \tag{2.1}$$

https://doi.org/10.4208/cicp.280315.140815a Published online by Cambridge University Press

then we have the nonnegative directional splitting

$$\nabla\cdot(\mathbb{D}\nabla u)(x_0,y_0)=\partial_x(\gamma_0\partial_x u)(x_0,y_0)+\partial_\beta(\gamma_1\partial_\beta u)(x_0,y_0)+\partial_y(\gamma_2\partial_y u)(x_0,y_0), \tag{2.2}$$

where the coefficients γ_0, γ_1, and γ_2, all nonnegative on Ω_0, are given by

$$\begin{cases}\gamma_0(x,y)=a(x,y)-b(x,y)\cot\beta\geq 0,\\ \gamma_1(x,y)=\dfrac{b(x,y)}{\cos\beta\sin\beta}\geq 0,\\ \gamma_2(x,y)=c(x,y)-b(x,y)\tan\beta\geq 0.\end{cases} \tag{2.3}$$

Proof. For the situation with $b(x,y)\equiv 0$ on Ω_0, it is obvious that (2.2) and (2.3) hold with any β satisfying $\sin(\beta)\cos(\beta)\neq 0$. For other situations, condition (2.1) implies that $0<\beta<\frac{\pi}{2}$ or $-\frac{\pi}{2}<\beta<0$. Since $\mathbb{D}(x,y)$ is symmetric and uniformly positive definite on $\Omega_0\subset\Omega$, we have $a(x,y)>0$ and $c(x,y)>0$ for each $(x,y)\in\Omega_0$. For each $(x,y)\in\Omega_0$, we denote the eigenvalues of $\mathbb{D}$ by λ_1 and λ_2 (without loss of generality we assume $\lambda_1\geq\lambda_2$) and the angle formed by the principal eigenvector and the x-axis by ψ. Then, the diffusion matrix at (x,y) has the eigen-decomposition as

$$\begin{aligned}\mathbb{D}&=\begin{bmatrix}a & b\\ b & c\end{bmatrix}=\lambda_1\begin{bmatrix}\cos(\psi)\\ \sin(\psi)\end{bmatrix}\begin{bmatrix}\cos(\psi) & \sin(\psi)\end{bmatrix}+\lambda_2\begin{bmatrix}-\sin(\psi)\\ \cos(\psi)\end{bmatrix}\begin{bmatrix}-\sin(\psi) & \cos(\psi)\end{bmatrix}\\ &=\begin{bmatrix}\lambda_1\cos^2(\psi)+\lambda_2\sin^2(\psi) & (\lambda_1-\lambda_2)\cos(\psi)\sin(\psi)\\ (\lambda_1-\lambda_2)\cos(\psi)\sin(\psi) & \lambda_1\sin^2(\psi)+\lambda_2\cos^2(\psi)\end{bmatrix}.\end{aligned} \tag{2.4}$$

In the following, we assume $b\geq 0$ on Ω_0. (The analysis for the case with $b\leq 0$ is similar.) In this case, $\psi\in[0,\frac{\pi}{2}]$. We now show that $\nabla\cdot(\mathbb{D}\nabla u)$ can be written into the form of (2.2) with nonnegative γ_0, γ_1, and γ_2. By direct calculation, we have

$$\begin{aligned}\partial_\beta(\gamma_1\partial_\beta u)&=\nabla\cdot\left(\gamma_1\begin{bmatrix}\cos^2\beta & \cos\beta\sin\beta\\ \cos\beta\sin\beta & \sin^2\beta\end{bmatrix}\nabla u\right),\\ \partial_x(\gamma_0\partial_x u)&=\nabla\cdot\left(\gamma_0\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}\nabla u\right),\\ \partial_y(\gamma_2\partial_y u)&=\nabla\cdot\left(\gamma_2\begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix}\nabla u\right).\end{aligned}$$

Then, (2.2) is equivalent to

$$\nabla\cdot(\mathbb{D}\nabla u)=\nabla\cdot\left(\begin{bmatrix}\gamma_0+\gamma_1\cos^2\beta & \gamma_1\cos\beta\sin\beta\\ \gamma_1\cos\beta\sin\beta & \gamma_1\sin^2\beta+\gamma_2\end{bmatrix}\nabla u\right),$$

which requires

$$\begin{bmatrix}a & b\\ b & c\end{bmatrix}=\begin{bmatrix}\gamma_0+\gamma_1\cos^2\beta & \gamma_1\cos\beta\sin\beta\\ \gamma_1\cos\beta\sin\beta & \gamma_1\sin^2\beta+\gamma_2\end{bmatrix}.$$

https://doi.org/10.4208/cicp.280315.140815a Published online by Cambridge University Press

This can be expressed into

$$\begin{bmatrix} 1 & \cos^2\beta & 0 \\ 0 & \cos\beta\sin\beta & 0 \\ 0 & \sin^2\beta & 1 \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{bmatrix} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}.$$

Since $0<\beta<\frac{\pi}{2}$, we can solve the above equation and get (2.3). The condition (2.1) implies that $\gamma_0 \geq 0$ and $\gamma_2 \geq 0$ on Ω_0. From (2.4), we have

$$b=(\lambda_1-\lambda_2)\cos\psi\sin\psi.$$

Thus,

$$\gamma_1=\frac{b}{\sin\beta\cos\beta}=\frac{(\lambda_1-\lambda_2)\cos\psi\sin\psi}{\cos\beta\sin\beta}\geq 0,$$

since $\psi\in[0,\frac{\pi}{2}]$ and $\beta\in(0,\frac{\pi}{2})$. □

The above lemma shows that for the case where b does not change sign, (1.1) has a nonnegative directional splitting on Ω_0 if (2.1) is satisfied. In the following lemma, we consider the general case where b changes sign on Ω_0. Hereafter, we use the superscripts " $^+$" and " $^-$" to indicate the regions associated with $b(x,y)>0$ and $b(x,y)<0$, respectively. For example, we denote

$$\Omega_0^+=\{(x,y)\in\Omega_0 : b(x,y)>0\}, \qquad \Omega_0^-=\{(x,y)\in\Omega_0 : b(x,y)<0\}.$$

Remark 2.1. The set Ω_0^+ or Ω_0^- may be empty. When this happens, many operations associated with such an empty in the following analysis may not make sense. For this reason, we assume that the formulas with an empty set will be ignored. This remark also applies to other empty sets.

Lemma 2.2. *Consider an arbitrary point (x_0,y_0) in Ω and a neighborhood Ω_0 of (x_0,y_0). If there exist constants β_1 and β_2 satisfying*

$$\begin{cases} \sup\limits_{\Omega_0^+}\dfrac{b(x,y)}{a(x,y)}<\tan\beta_1<\inf\limits_{\Omega_0^+}\dfrac{c(x,y)}{b(x,y)}, \\ \sup\limits_{\Omega_0^-}\dfrac{c(x,y)}{b(x,y)}<\tan\beta_2<\inf\limits_{\Omega_0^-}\dfrac{b(x,y)}{a(x,y)}, \end{cases} \tag{2.5}$$

then we have the nonnegative directional splitting

$$\begin{aligned} \nabla\cdot(\mathbb{D}\nabla u)(x_0,y_0)=&\partial_x(\gamma_0\partial_x u)(x_0,y_0)+\partial_{\beta_1}(\gamma_1^+\partial_{\beta_1}u)(x_0,y_0) \\ &+\partial_{\beta_2}(\gamma_1^-\partial_{\beta_2}u)(x_0,y_0)+\partial_y(\gamma_2\partial_y u)(x_0,y_0), \end{aligned} \tag{2.6}$$

https://doi.org/10.4208/cicp.280315.140815a Published online by Cambridge University Press

where $\gamma_0(x,y)$, $\gamma_1^+(x,y)$, $\gamma_1^-(x,y)$, *and* $\gamma_2(x,y)$, *all nonnegative on* Ω_0, *are given by*

$$\gamma_0(x,y)=\begin{cases}a(x,y)-b(x,y)\cot\beta_1, & \text{for } b(x,y)\geq 0,\\ a(x,y)-b(x,y)\cot\beta_2, & \text{for } b(x,y)<0,\end{cases} \tag{2.7}$$

$$\gamma_1^+(x,y)=\begin{cases}\dfrac{b(x,y)}{\cos\beta_1\sin\beta_1}, & \text{for } b(x,y)\geq 0,\\ 0, & \text{for } b(x,y)<0,\end{cases} \tag{2.8}$$

$$\gamma_1^-(x,y)=\begin{cases}0, & \text{for } b(x,y)\geq 0,\\ \dfrac{b(x,y)}{\cos\beta_2\sin\beta_2}, & \text{for } b(x,y)<0,\end{cases} \tag{2.9}$$

$$\gamma_2(x,y)=\begin{cases}c(x,y)-b(x,y)\tan\beta_1, & \text{for } b(x,y)\geq 0,\\ c(x,y)-b(x,y)\tan\beta_2, & \text{for } b(x,y)<0.\end{cases} \tag{2.10}$$

Proof. For any given constants $\epsilon>0$ and $M>0$ we define a decompose of the diffusion matrix as

$$\mathbb{D}(x,y)=\mathbb{D}_1(x,y)+\mathbb{D}_2(x,y),$$

where

$$\mathbb{D}_1=\begin{bmatrix}a_1 & b_1\\ b_1 & c_1\end{bmatrix}\equiv\begin{cases}\begin{bmatrix}a-\epsilon & b+\epsilon/M\\ b+\epsilon/M & c-\epsilon\end{bmatrix}, & \text{for } b(x,y)\geq 0,\\ \begin{bmatrix}\epsilon & 0\\ 0 & \epsilon\end{bmatrix}, & \text{for } b(x,y)<0,\end{cases} \tag{2.11}$$

$$\mathbb{D}_2=\begin{bmatrix}a_2 & b_2\\ b_2 & c_2\end{bmatrix}\equiv\begin{cases}\begin{bmatrix}\epsilon & -\epsilon/M\\ -\epsilon/M & \epsilon\end{bmatrix}, & \text{for } b(x,y)\geq 0,\\ \begin{bmatrix}a-\epsilon & b\\ b & c-\epsilon\end{bmatrix}, & \text{for } b(x,y)<0.\end{cases} \tag{2.12}$$

Then we have

$$\nabla\cdot(\mathbb{D}\nabla u)=\nabla\cdot(\mathbb{D}_1\nabla u)+\nabla\cdot(\mathbb{D}_2\nabla u). \tag{2.13}$$

Since

$$b_1\geq 0 \quad\text{and}\quad b_2<0, \quad \forall(x,y)\in\Omega_0,$$

we can apply Lemma 2.1 to each term on the right-hand side of (2.13). To this end, we notice that (2.5) implies, for sufficiently small ϵ and large M, that

$$\sup_{\Omega_0^+}\frac{b}{a}<\sup_{\Omega_0^+}\frac{b+\epsilon/M}{a-\epsilon}<\tan\beta_1<\inf_{\Omega_0^+}\frac{c-\epsilon}{b+\epsilon/M}<\inf_{\Omega_0^+}\frac{c}{b}, \tag{2.14}$$

$$\sup_{\Omega_0^-}\frac{c}{b}<\sup_{\Omega_0^-}\frac{c-\epsilon}{b}<\tan\beta_2<\inf_{\Omega_0^-}\frac{b}{a-\epsilon}<\inf_{\Omega_0^-}\frac{b}{a}. \tag{2.15}$$

https://doi.org/10.4208/cicp.280315.140815a Published online by Cambridge University Press

From (2.11), we have

$$\sup_{\Omega_0^-}\frac{b_1}{a_1}=0\leq\sup_{\Omega_0^+}\frac{b_1}{a_1}=\sup_{\Omega_0^+}\frac{b+\epsilon/M}{a-\epsilon},\qquad \inf_{\Omega_0^-}\frac{c_1}{b_1}=+\infty>\inf_{\Omega_0^+}\frac{c_1}{b_1}=\inf_{\Omega_0^+}\frac{c-\epsilon}{b+\epsilon/M}.$$

Combining this with (2.14), we get

$$\sup_{\Omega_0}\frac{b_1}{a_1}=\sup_{\Omega_0^+}\frac{b_1}{a_1}<\tan\beta_1<\inf_{\Omega_0}\frac{c_1}{b_1}=\inf_{\Omega_0^+}\frac{c_1}{b_1}.$$

Thus, the condition of Lemma 2.1 is satisfied, which implies that

$$\nabla\cdot(\mathbb{D}_1\nabla u)=\partial_x(\gamma_{1,0}\partial_x u)+\partial_{\beta_1}(\gamma_{1,1}\partial_{\beta_1}u)+\partial_y(\gamma_{1,2}\partial_y u),\tag{2.16}$$

where $\gamma_{1,0}$, $\gamma_{1,1}$, and $\gamma_{1,2}$ are nonnegative and given by

$$\gamma_{1,0}=\begin{cases}(a-\epsilon)-(b+\epsilon/M)\cot(\beta_1), & \text{for } b\geq 0,\\ \epsilon, & \text{for } b<0,\end{cases}$$

$$\gamma_{1,1}=\begin{cases}\dfrac{b+\epsilon/M}{\cos\beta_1\sin\beta_1}, & \text{for } b\geq 0,\\ 0, & \text{for } b<0,\end{cases}$$

$$\gamma_{1,2}=\begin{cases}(c-\epsilon)-(b+\epsilon/M)\tan\beta_1, & \text{for } b\geq 0,\\ \epsilon, & \text{for } b<0.\end{cases}$$

Similarly, for sufficiently large M, from (2.12) we have

$$\sup_{\Omega_0^+}\frac{c_2}{b_2}=-M\leq\sup_{\Omega_0^-}\frac{c_2}{b_2}=\sup_{\Omega_0^-}\frac{c-\epsilon}{b},\qquad \inf_{\Omega_0^+}\frac{b_2}{a_2}=-\frac{1}{M}\geq\inf_{\Omega_0^-}\frac{b_2}{a_2}=\inf_{\Omega_0^-}\frac{b}{a-\epsilon}.$$

Combining this with (2.15) gives

$$\sup_{\Omega_0}\frac{c_2}{b_2}=\sup_{\Omega_0^-}\frac{c_2}{b_2}<\tan\beta_2<\inf_{\Omega_0}\frac{b_2}{a_2}=\inf_{\Omega_0^-}\frac{b_2}{a_2}.$$

Then from Lemma 2.1 we have

$$\nabla\cdot(\mathbb{D}_2\nabla u)=\partial_x(\gamma_{2,0}\partial_x u)+\partial_{\beta_2}(\gamma_{2,1}\partial_{\beta_2}u)+\partial_y(\gamma_{2,2}\partial_y u),\tag{2.17}$$

https://doi.org/10.4208/cicp.280315.140815a Published online by Cambridge University Press

where $\gamma_{2,0}$, $\gamma_{2,1}$, and $\gamma_{2,2}$, all nonnegative, are given by

$$\gamma_{2,0}=\begin{cases}\epsilon+\epsilon/M\cot\beta_2, & \text{for } b\geq 0,\\ (a-\epsilon)-b\cot\beta_2, & \text{for } b<0,\end{cases}$$

$$\gamma_{2,1}=\begin{cases}\dfrac{-\epsilon/M}{\cos\beta_2\sin\beta_2}, & \text{for } b\geq 0,\\ \dfrac{b}{\cos\beta_2\sin\beta_2}, & \text{for } b<0,\end{cases}$$

$$\gamma_{2,2}=\begin{cases}\epsilon+\epsilon/M\tan\beta_2, & \text{for } b\geq 0,\\ (c-\epsilon)-b\tan\beta_2, & \text{for } b<0.\end{cases}$$

By combining (2.13), (2.16), and (2.17), we get

$$\nabla\cdot(\mathbb{D}\nabla u)=\partial_x(\gamma_0\partial_x u)+\partial_{\beta_1}(\gamma_1^+\partial_{\beta_1}u)+\partial_{\beta_2}(\gamma_1^-\partial_{\beta_2}u)+\partial_y(\gamma_2\partial_y u),$$

where γ_0, γ_1^+, γ_1^-, and γ_2, all nonnegative, are given by

$$\gamma_0=\gamma_{1,0}+\gamma_{2,0}=\begin{cases}a-(b+\epsilon/M)\cot\beta_1+\epsilon/M\cot\beta_2, & \text{for } b\geq 0,\\ a-b\cot\beta_2, & \text{for } b<0,\end{cases}$$

$$\gamma_2=\gamma_{1,2}+\gamma_{2,2}=\begin{cases}c-(b+\epsilon/M)\tan\beta_1+\epsilon/M\tan\beta_2, & \text{for } b\geq 0,\\ c-b\tan\beta_2, & \text{for } b<0,\end{cases}$$

$$\gamma_1^+=\gamma_{1,1}=\begin{cases}\dfrac{b+\epsilon/M}{\cos\beta_1\sin\beta_1}, & \text{for } b\geq 0,\\ 0, & \text{for } b<0,\end{cases}$$

$$\gamma_1^-=\gamma_{2,1}=\begin{cases}\dfrac{-\epsilon/M}{\cos\beta_2\sin\beta_2}, & \text{for } b\geq 0,\\ \dfrac{b}{\cos\beta_2\sin\beta_2}, & \text{for } b<0.\end{cases}$$

Since the above result holds for any sufficiently small ϵ, we take the limit as $\epsilon\to 0$ and obtain the conclusion of the lemma. □

In the next lemma, we show that (2.5) can be satisfied uniformly over Ω under a reasonable assumption on the diffusion matrix $\mathbb{D}$. For a positive number R, we denote

$$B_R(x_0,y_0)=\{(x,y)\in\Omega:|(x,y)-(x_0,y_0)|<R\}.$$

Lemma 2.3. *Assume that $\Omega\subset\mathbb{R}^2$ is a bounded domain and $\mathbb{D}$ is continuous on $\overline{\Omega}\equiv\Omega\cup\partial\Omega$ and symmetric and uniformly positive definite on Ω. Then, there exists $R>0$ such that for any*

https://doi.org/10.4208/cicp.280315.140815a Published online by Cambridge University Press

$(x_0,y_0)\in\Omega$,

$$\sup_{B_R^+(x_0,y_0)}\frac{b(x,y)}{a(x,y)}+\frac{\overline{\alpha}}{3\alpha}\leq\inf_{B_R^+(x_0,y_0)}\frac{c(x,y)}{b(x,y)}, \tag{2.18}$$

$$\sup_{B_R^-(x_0,y_0)}\frac{c(x,y)}{b(x,y)}+\frac{\overline{\alpha}}{3\alpha}\leq\inf_{B_R^-(x_0,y_0)}\frac{b(x,y)}{a(x,y)}, \tag{2.19}$$

$$\sup_{B_R^+(x_0,y_0)}\frac{b(x,y)}{c(x,y)}+\frac{\overline{\alpha}}{3\alpha}\leq\inf_{B_R^+(x_0,y_0)}\frac{a(x,y)}{b(x,y)}, \tag{2.20}$$

$$\sup_{B_R^-(x_0,y_0)}\frac{a(x,y)}{b(x,y)}+\frac{\overline{\alpha}}{3\alpha}\leq\inf_{B_R^-(x_0,y_0)}\frac{b(x,y)}{c(x,y)}, \tag{2.21}$$

where

$$\overline{\alpha}=\inf_{\Omega}\left(a(x,y)c(x,y)-b(x,y)^2\right), \tag{2.22}$$

$$\alpha=\max\left(\sup_{\Omega}a(x,y)(|b(x,y)|+1),\sup_{\Omega}c(x,y)(|b(x,y)|+1)\right). \tag{2.23}$$

Proof. The inequalities (2.20) and (2.21) can be obtained from (2.18) and (2.19) by the symmetry between a and c. To prove (2.18) and (2.19), we first notice that the uniform positive definiteness of $\mathbb{D}$ implies that $\overline{\alpha}>0$ and

$$a(x,y)c(x,y)-b(x,y)^2\geq\overline{\alpha},\quad\forall(x,y)\in\Omega.$$

Dividing both sides by ab, we have

$$\frac{c(x,y)}{b(x,y)}\geq\frac{b(x,y)}{a(x,y)}+\frac{\overline{\alpha}}{a(x,y)b(x,y)},\quad\text{for}\quad b(x,y)>0$$
$$\frac{c(x,y)}{b(x,y)}\leq\frac{b(x,y)}{a(x,y)}+\frac{\overline{\alpha}}{a(x,y)b(x,y)},\quad\text{for}\quad b(x,y)<0.$$

Denoting

$$F(x,y)=\frac{c(x,y)}{b(x,y)},\quad G=\frac{b(x,y)}{a(x,y)}, \tag{2.24}$$

from the definition of α (2.22) we have

$$F(x,y)\geq G(x,y)+\frac{\overline{\alpha}}{\alpha},\quad\text{for}\quad b(x,y)>0$$
$$F(x,y)\leq G(x,y)-\frac{\overline{\alpha}}{\alpha},\quad\text{for}\quad b(x,y)<0.$$

Define

$$M=\sup_{\Omega}|G(x,y)|+\frac{\overline{\alpha}}{\alpha}. \tag{2.25}$$

https://doi.org/10.4208/cicp.280315.140815a Published online by Cambridge University Press

To prove (2.18), we introduce a cut-off function of F as

$$F^+(x,y)=\begin{cases}F(x,y), & \text{for } b(x,y)>0 \text{ and } F(x,y)<M,\\ M, & \text{for } b(x,y)>0 \text{ and } F(x,y)\geq M,\\ M, & \text{for } b(x,y)\leq 0.\end{cases} \tag{2.26}$$

This function is continuous on $\overline{\Omega}$ and has the following properties,

$$F^+(x,y)\leq F(x,y), \quad \text{for } b(x,y)>0, \tag{2.27}$$

$$F^+(x,y)\geq G(x,y)+\frac{\overline{\alpha}}{\alpha}, \quad \text{for } (x,y)\in\Omega. \tag{2.28}$$

Since Ω is bounded, F^+ and G are uniformly continuous on $\overline{\Omega}$. Thus, there exists a constant $R_1>0$ (which depends on F^+, G, $\overline{\alpha}$, and α and therefore the diffusion matrix $\mathbb{D}$ but not on (x_0,y_0)) such that

$$F^+(x,y)\geq F^+(x_0,y_0)-\frac{\overline{\alpha}}{3\alpha}, \quad G(x,y)\leq G(x_0,y_0)+\frac{\overline{\alpha}}{3\alpha}, \qquad \forall (x,y)\in B_{R_1}(x_0,y_0) \tag{2.29}$$

which in turn implies that

$$\inf_{B_{R_1}(x_0,y_0)} F^+(x,y)\geq F^+(x_0,y_0)-\frac{\overline{\alpha}}{3\alpha}, \quad \sup_{B_{R_1}(x_0,y_0)} G(x,y)\leq G(x_0,y_0)+\frac{\overline{\alpha}}{3\alpha}.$$

Combining these results with (2.28), we have

$$\sup_{B_{R_1}(x_0,y_0)} G(x,y)\leq G(x_0,y_0)+\frac{\overline{\alpha}}{3\alpha}\leq F^+(x_0,y_0)-\frac{2\overline{\alpha}}{3\alpha}\leq \inf_{B_{R_1}(x_0,y_0)} F^+(x,y)-\frac{\overline{\alpha}}{3\alpha}$$

or

$$\sup_{B_{R_1}(x_0,y_0)} G(x,y)\leq \inf_{B_{R_1}(x_0,y_0)} F^+(x,y)-\frac{\overline{\alpha}}{3\alpha}.$$

From this and (2.27), we have

$$\begin{aligned}\sup_{B^+_{R_1}(x_0,y_0)} G(x,y)\leq \sup_{B_{R_1}(x_0,y_0)} G(x,y)&\leq \inf_{B_{R_1}(x_0,y_0)} F^+(x,y)-\frac{\overline{\alpha}}{3\alpha}\\ &\leq \inf_{B^+_{R_1}(x_0,y_0)} F^+(x,y)-\frac{\overline{\alpha}}{3\alpha}\leq \inf_{B^+_{R_1}(x_0,y_0)} F(x,y)-\frac{\overline{\alpha}}{3\alpha},\end{aligned}$$

which gives (2.18) (with $R=R_1$).

Similarly, we define

$$F^-(x,y)=\begin{cases}F(x,y), & \text{for } b(x,y)<0 \text{ and } F(x,y)>-M,\\ -M, & \text{for } b(x,y)<0 \text{ and } F(x,y)\leq -M,\\ -M, & \text{for } b(x,y)\geq 0.\end{cases} \tag{2.30}$$

https://doi.org/10.4208/cicp.280315.140815a Published online by Cambridge University Press

It is continuous on $\overline{\Omega}$ and satisfies

$$F^-(x,y) \geq F(x,y), \quad \text{for} \quad b(x,y) < 0, \tag{2.31}$$

$$F^-(x,y) \leq G(x,y) - \frac{\overline{\alpha}}{\alpha}, \quad \text{for} \quad (x,y) \in \Omega. \tag{2.32}$$

From the uniform continuity of F^- and G on $\overline{\Omega}$, there exists a constant $R_2 > 0$ (which depends only on the diffusion matrix $\mathbb{D}$) such that

$$F^-(x,y) \leq F^-(x_0,y_0) + \frac{\overline{\alpha}}{3\alpha}, \quad G(x,y) \geq G(x_0,y_0) - \frac{\overline{\alpha}}{3\alpha}, \quad \forall (x,y) \in B_{R_2}(x_0,y_0). \tag{2.33}$$

Thus,

$$\sup_{B_{R_2}(x_0,y_0)} F^-(x,y) \leq F^-(x_0,y_0) + \frac{\overline{\alpha}}{3\alpha} \leq G(x_0,y_0) - \frac{2\overline{\alpha}}{3\alpha} \leq \inf_{B_{R_2}(x_0,y_0)} G(x,y) - \frac{\overline{\alpha}}{3\alpha}$$

or

$$\sup_{B_{R_2}(x_0,y_0)} F^-(x,y) \leq \inf_{B_{R_2}(x_0,y_0)} G(x,y) - \frac{\overline{\alpha}}{3\alpha}.$$

Then, the inequality (2.19) (with $R = R_2$) follows from this and (2.31).

Using the symmetry between a and c we can show from (2.18) and (2.19) that (2.20) and (2.21) hold with some positive numbers R_3 and R_4, respectively. Taking

$$R = \min(R_1, R_2, R_3, R_4),$$

we have proven that (2.18) – (2.21) hold for the same R. □

Remark 2.2. The constants $\overline{\alpha}$ and α defined in (2.22) and (2.23) and the functions F, F^+, F^-, and G defined in (2.24), (2.25), (2.26), and (2.30) are determined completely by the diffusion matrix $\mathbb{D}$. So is the constant R. If F^+, F^-, and G are Lipschitz continuous with constants L_{F^+}, L_{F^-}, and L_G, respectively, then we can choose R as

$$R = \frac{\overline{\alpha}}{3\alpha \max(L_{F^+}, L_{F^-}, L_G)}. \tag{2.34}$$

Theorem 2.1. *Assume that $\Omega \subset \mathbb{R}^2$ is a bounded domain and the diffusion matrix $\mathbb{D}$ is continuous on $\overline{\Omega}$ and symmetric and uniformly positive definite on Ω. Then, there exists a constant $R > 0$ (which depends only on $\mathbb{D}$) such that for any $(x_0,y_0) \in \Omega$, the intervals*

$$\left(\sup_{B_R^+(x_0,y_0)} \frac{b(x,y)}{a(x,y)}, \quad \inf_{B_R^+(x_0,y_0)} \frac{c(x,y)}{b(x,y)} \right), \qquad \left(\sup_{B_R^-(x_0,y_0)} \frac{c(x,y)}{b(x,y)}, \quad \inf_{B_R^-(x_0,y_0)} \frac{b(x,y)}{a(x,y)} \right)$$

are not empty when $B_R^+(x_0,y_0)$ and $B_R^-(x_0,y_0)$ are not empty, respectively. Moreover, for any constants β_1 and β_2 with $\tan\beta_1$ and $\tan\beta_2$ in the intervals, the nonnegative directional splitting (2.6) holds with $\Omega_0 = B_R(x_0,y_0)$ and nonnegative coefficients given in (2.7)-(2.10).

https://doi.org/10.4208/cicp.280315.140815a Published online by Cambridge University Press

Proof. The theorem is obtained by combining Lemma 2.2 and Lemma 2.3. □

Remark 2.3. The result in the above theorem is different from Weickert's result [31] in the sense that the coefficients in the directional splitting (2.6) are nonnegative in a neighborhood of (x_0,y_0) whereas those of (1.4) are nonnegative only at (x_0,y_0). This is crucial to the construction of monotone FD schemes since any FD scheme for the divergence form (1.1) uses the values of the coefficients on a neighborhood or a stencil of a mesh point for the discretization at the point.

Remark 2.4. As mentioned in Remark 2.2 the constant R depends on the diffusion matrix $\mathbb{D}$ and is independent of the location. This implies that a uniform mesh can be used to construct a monotone FD scheme as long as the mesh spacing is sufficiently small and the maximal size of the FD stencils (used for all mesh points) is bounded (see Section 3).

3 Construction of a monotone FD discretization

In this section we develop a monotone FD discretization for (1.1) based on the nonnegative directional splitting (2.6). We consider a uniform mesh $(x_j,y_k)=(jh,kh)$, $j,k=0,\cdots,N$, where N is a positive integer and $h=1/N$.

Let (x_j,y_k) be an interior mesh point. Denote by $\omega_{j,k}$ a FD stencil of size $(2m_{j,k}+1)\times(2m_{j,k}+1)$ centered at (x_j,y_k). We assume that

$$m_{j,k}\le\left[\frac{3\alpha}{\overline{\alpha}}\right]+1, \tag{3.1}$$

where $\overline{\alpha}$ and α are defined in (2.22) and (2.23) and $[x]$ denotes the nearest integer smaller than or equal to x. We will show that the right-hand side of (3.1) is an upper bound of the stencil size to guarantee a monotone FD discretization for (1.1). We also assume that the mesh spacing h is sufficiently small such that the stencil $\omega_{j,k}\subset B_R(x_j,y_k)$, with R being the radius determined in Lemma 2.3 and Theorem 2.1. Mathematically, this is equivalent to

$$\sqrt{2}h\max_{j,k}m_{j,k}\le R. \tag{3.2}$$

Using (2.34) and (3.1), we obtain a sufficient condition for (3.2) as

$$\sqrt{2}h\left(\left[\frac{3\alpha}{\overline{\alpha}}\right]+1\right)\le\frac{\overline{\alpha}}{3\alpha\max(L_{F^+},L_{F^-},L_G)}, \tag{3.3}$$

which can be achieved by choosing h sufficiently small.

We first consider the case where $\omega_{j,k}\subset\overline{\Omega}$. Under the above assumptions, Theorem 2.1 implies that the nonnegative directional splitting (2.6), with the coefficients being non-

https://doi.org/10.4208/cicp.280315.140815a Published online by Cambridge University Press

negative on $\omega_{j,k}$, holds for any constants β_1 and β_2 satisfying

$$A_{j,k} \equiv \sup_{\omega_{j,k}^+} \frac{b(x,y)}{a(x,y)} < \tan\beta_1 < B_{j,k} \equiv \inf_{\omega_{j,k}^+} \frac{c(x,y)}{b(x,y)}, \tag{3.4}$$

$$C_{j,k} \equiv \sup_{\omega_{j,k}^-} \frac{c(x,y)}{b(x,y)} < \tan\beta_2 < D_{j,k} \equiv \inf_{\omega_{j,k}^-} \frac{b(x,y)}{a(x,y)}. \tag{3.5}$$

It is not difficult to see that

$$A_{j,k}^{-1} = \inf_{\omega_{j,k}^+} \frac{a(x,y)}{b(x,y)}, \quad B_{j,k}^{-1} = \sup_{\omega_{j,k}^+} \frac{b(x,y)}{c(x,y)}, \quad C_{j,k}^{-1} = \inf_{\omega_{j,k}^-} \frac{b(x,y)}{c(x,y)}, \quad D_{j,k}^{-1} = \sup_{\omega_{j,k}^-} \frac{a(x,y)}{b(x,y)}.$$

From Lemma 2.3, we have

$$B_{j,k} - A_{j,k} \geq \frac{\overline{\alpha}}{3\alpha}, \quad A_{j,k}^{-1} - B_{j,k}^{-1} \geq \frac{\overline{\alpha}}{3\alpha}, \tag{3.6}$$

$$D_{j,k} - C_{j,k} \geq \frac{\overline{\alpha}}{3\alpha}, \quad C_{j,k}^{-1} - D_{j,k}^{-1} \geq \frac{\overline{\alpha}}{3\alpha}, \tag{3.7}$$

where $\overline{\alpha}$ and α are defined in (2.22) and (2.23).

The central FD discretization of the first and fourth terms of (2.6) is straightforward. It can be done based on points $(x_{j-1},y_k), (x_j,y_k), (x_{j+1},y_k)$ and $(x_j,y_{k-1}), (x_j,y_k), (x_j,y_{k+1})$, respectively. (These points are on a 3×3 stencil centered at (x_j,y_k).) On the other hand, the FD discretization of the second and third terms are more complicated. Take the second term as an example. The main challenge of the discretization is to find β_1 with $\tan\beta_1 \in (A_{j,k},B_{j,k})$ such that the straight line having the slope $\tan\beta_1$ and passing the point (x_j,y_k) intersects at least two more mesh points on $\omega_{j,k}$ other than (x_j,y_k). Once such a β_1 has been found, the three mesh points are then used to discretize the term (which is the second order directional derivative along $\tan\beta_1$) using central finite differences.

To find β_1, we follow [31] to define $4m_{j,k}$ principal directions associated with the outer mesh nodes of $\omega_{j,k}$ (cf. Fig. 1),

$$\hat{\beta}_i = \begin{cases} \arctan\left(\frac{i}{m_{j,k}}\right), & \text{for } -m_{j,k} \leq i \leq m_{j,k}, \\ \arctan\left(\frac{m_{j,k}}{2m_{j,k}-i}\right), & \text{for } m_{j,k}+1 \leq i \leq 2m_{j,k}, \\ \arctan\left(\frac{m_{j,k}}{-2m_{j,k}-i}\right), & \text{for } -2m_{j,k}+1 \leq i \leq -m_{j,k}-1. \end{cases} \tag{3.8}$$

When $m_{j,k}$ is sufficiently large, we can choose β_1 to be one of $\hat{\beta}_i$, $i=1,\cdots,2m_{j,k}$ that satisfies (3.4). Indeed, there are three possibilities $A_{j,k}$ and $B_{j,k}$ are located: $A_{j,k}<1<B_{j,k}$, $A_{j,k}<B_{j,k}\leq 1$, or $1\leq A_{j,k}<B_{j,k}$. For the case $A_{j,k}<1<B_{j,k}$, we can choose $\beta_1=\hat{\beta}_{m_{j,k}}=\pi/4$ and $m_{j,k}=1$. It is easy to see that (3.4) is satisfied and the three mesh points on the diagonal

https://doi.org/10.4208/cicp.280315.140815a Published online by Cambridge University Press

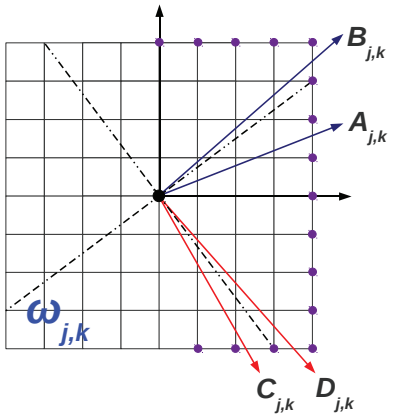


Figure 1: Sketch of a FD stencil and principal directions.

line running from the southwest corner to the northeast corner can be used to discretize the directional derivative.

For the case $A_{j,k}<B_{j,k}\leq 1$, we need to consider the principal directions $\hat{\beta}_i$, $i=1,\cdots,m_{j,k}-1$. Thus, we require that at least there is an i $(1\leq i<m_{j,k}-1)$ such that

$$A_{j,k}<\frac{i}{m_{j,k}}<B_{j,k}.$$

(In this case, β_1 is chosen as $\hat{\beta}_i$.) Equivalently, $m_{j,k}$ should be sufficiently large such that $(m_{j,k}A_{j,k},m_{j,k}B_{j,k})$ contains at least an integer. A sufficient condition for this is that the length of the interval is bigger than one, i.e.,

$$m_{j,k}>\frac{1}{B_{j,k}-A_{j,k}}. \tag{3.9}$$

For the case $1\leq A_{j,k}<B_{j,k}$, we need to consider the principal directions $\hat{\beta}_i$, $i=m_{j,k}+1,\cdots,2m_{j,k}-1$. We require

$$A_{j,k}<\frac{m_{j,k}}{2m_{j,k}-i}<B_{j,k}$$

or

$$m_{j,k}B_{j,k}^{-1}<2m_{j,k}-i<m_{j,k}A_{j,k}^{-1}$$

for some $m_{j,k}<i<2m_{j,k}$. This is mathematically equivalent to the requirement that $(m_{j,k}B_{j,k}^{-1},m_{j,k}A_{j,k}^{-1})$ contains at least an integer. A sufficient condition for this is

$$m_{j,k}>\frac{1}{A_{j,k}^{-1}-B_{j,k}^{-1}}.$$

https://doi.org/10.4208/cicp.280315.140815a Published online by Cambridge University Press

The above analysis shows that we should choose $m_{j,k}$ such that

$$\begin{cases} m_{j,k}=1, & \text{for } A_{j,k}<1<B_{j,k}, \\ (m_{j,k}A_{j,k}, m_{j,k}B_{j,k}) \text{ contains at least an integer}, & \text{for } A_{j,k}<B_{j,k}\le 1, \\ (m_{j,k}B_{j,k}^{-1}, m_{j,k}A_{j,k}^{-1}) \text{ contains at least an integer}, & \text{for } 1\le A_{j,k}<B_{j,k}. \end{cases} \tag{3.10}$$

A sufficient condition for this is

$$\begin{cases} m_{j,k}\ge 1, & \text{for } A_{j,k}<1<B_{j,k}, \\ m_{j,k}>\frac{1}{B_{j,k}-A_{j,k}}, & \text{for } A_{j,k}<B_{j,k}\le 1, \\ m_{j,k}>\frac{1}{A_{j,k}^{-1}-B_{j,k}^{-1}}, & \text{for } 1\le A_{j,k}<B_{j,k}. \end{cases} \tag{3.11}$$

From (3.6), one can see that the above condition can be satisfied when $m_{j,k}$ is taken as $m_{j,k}=\left[\frac{3\alpha}{\bar{\alpha}}\right]+1$. Since this is only a sufficient condition, so we have (3.1).

Similarly, the condition that we can choose β_2 as one of the principal directions $\hat{\beta}_i$, $i=-2m_{j,k},\cdots,-1$ satisfying (3.5) is

$$\begin{cases} m_{j,k}=1, & \text{for } C_{j,k}<-1<D_{j,k}, \\ (m_{j,k}C_{j,k}, m_{j,k}D_{j,k}) \text{ contains at least an integer}, & \text{for } -1\le C_{j,k}<D_{j,k}, \\ (m_{j,k}D_{j,k}^{-1}, m_{j,k}C_{j,k}^{-1}) \text{ contains at least an integer}, & \text{for } C_{j,k}<D_{j,k}\le -1. \end{cases} \tag{3.12}$$

A sufficient condition for this is

$$\begin{cases} m_{j,k}\ge 1, & \text{for } C_{j,k}<-1<D_{j,k}, \\ m_{j,k}>\frac{1}{D_{j,k}-C_{j,k}}, & \text{for } -1\le C_{j,k}<D_{j,k}, \\ m_{j,k}>\frac{1}{C_{j,k}^{-1}-D_{j,k}^{-1}}, & \text{for } C_{j,k}<D_{j,k}\le -1. \end{cases} \tag{3.13}$$

From (3.7), one can see that the above condition can be satisfied when $m_{j,k}$ is taken as the right-hand side of (3.1).

We now consider the situation when the stencil $\omega_{j,k}$ is not in $\overline{\Omega}$. This can happen when (x_j,y_k) is close to the domain boundary and $m_{j,k}>1$. In this case, we can determine $m_{j,k}$, β_1, and β_2 as for the case where $\omega_{j,k}$ is in $\overline{\Omega}$. But, one of the three mesh points found along direction $\tan\beta_1$ or $\tan\beta_2$ lays outside of the domain; see Fig. 2. A treatment for this difficulty is to replace the mesh point by the intersection of the straight line (with slope $\tan\beta_1$ or $\tan\beta_2$) and the domain boundary and then discretize the corresponding directional derivative using central finite differences based on this boundary point and other two mesh points.

It is worth pointing out that each of the terms on the right-hand side of (2.6) has been discretized using central three-point finite differences along the corresponding direction. Since the coefficients of the differential operator are nonnegative, it is not difficult to see

https://doi.org/10.4208/cicp.280315.140815a Published online by Cambridge University Press

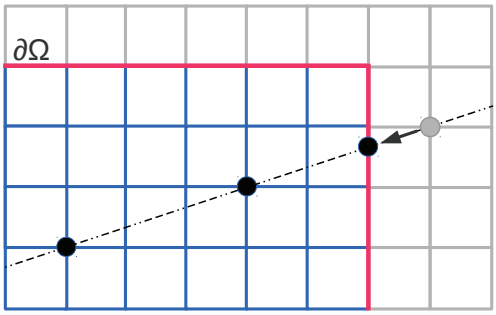


Figure 2: FD stencil near the domain boundary.

that the coefficient matrix of the FD equations is a Z-matrix (with off-diagonal entries being nonpositive) and has nonnegative diagonal entries. Moreover, it is easy to verify that the matrix is diagonally dominant and irreducible. Thus, the coefficient matrix is an M-matrix and the FD scheme is monotone.

For easy reference, we summarize the above analysis in the following theorem.

Theorem 3.1. *Assume that the diffusion matrix $\mathbb{D}$ is continuous on $\overline{\Omega}$ and symmetric and uniformly positive definite on Ω and a uniform mesh with spacing h is given for Ω. If h is sufficiently small, a monotone FD scheme exists for problem (1.1) and can be constructed based on central (directional) finite differences.*

Remark 3.1. Generally speaking, the upper bound for $m_{j,k}$ defined in (3.1) is very conservative. In practice, we can use stencils of variable size for different mesh points. The stencil size $m_{j,k}$ can be found using (3.10) and (3.12). The constant R can be calculated using (2.34). The mesh spacing can be verified using

$$\sqrt{2}h\max_{j,k} m_{j,k} \le R. \tag{3.14}$$

To conclude this section, we would like to see what condition is if only the three-by-three ($m_{j,k}=1$) stencil is used for all interior mesh points. In this case, we need to use the mesh points on the diagonal lines of $\omega_{j,k}$ to discretize the second and third terms on the right-hand side of (2.6). This means that $A_{j,k}<1<B_{j,k}$ ($\tan\beta_1=1$) and $C_{j,k}<-1<D_{j,k}$ ($\tan\beta_2=-1$). From (3.4) and (3.5), this implies

$$a(x,y)>|b(x,y)|, \quad c(x,y)>|b(x,y)|, \quad \forall (x,y)\in\Omega \tag{3.15}$$

or, in words, $\mathbb{D}$ is strictly diagonally dominant for any point in Ω. Thus, we have the following theorem.

Theorem 3.2. *Assume that the diffusion matrix $\mathbb{D}$ is continuous on $\overline{\Omega}$ and symmetric and uniformly positive definite on Ω and a uniform mesh with spacing h is given thereon. If (3.15) is satisfied and h is sufficiently small, then a monotone FD scheme can be constructed for (1.1) based on a three-by-three stencil and central finite differences.*

https://doi.org/10.4208/cicp.280315.140815a Published online by Cambridge University Press

It is pointed out that the condition (3.15) has been obtained by Weickert [31] and Mrázek and Navara [22].

4 Numerical results

In this section we present some numerical results for four examples to verify the theoretical findings in the previous sections. We first note that it is useful to have an estimate on the values of parameters R (the uniform radius defined in Lemma 2.3) and $m_{j,k}$ (the size of FD stencil) in the actual computation. The former can be estimated using (2.34) where $\bar{\alpha}$, α, L_{F^+}, L_{F^-}, and L_G can be approximated numerically. For the latter, (3.1) is generally too conservative to use in practical computation. Instead, we use (3.10) and (3.12) to estimate $m_{j,k}$, with $A_{j,k}$, $B_{j,k}$, $C_{j,k}$, $D_{j,k}$ being computed on $B_R(x_j,y_k)$. Notice that we use $B_R(x_j,y_k)$ instead of $\omega_{j,k}$ as defined in (3.4) and (3.5). This is because the former is larger than the latter and the former is independent of $m_{j,k}$ which is to be sought. Once $m_{j,k}$ has been determined, we can find two principal directions (cf. (3.8)) to satisfy (3.4) and (3.5) and finally construct the FD discretization of (1.1).

Example 4.1. This example is in the form of (1.1) with

$$\mathbb{D}=\begin{bmatrix} 9 & 4\sin(2\pi xy) \\ 4\sin(2\pi xy) & 3 \end{bmatrix}. \tag{4.1}$$

We seek to verify that the developed monotone scheme satisfy DMP and choose

$$f(x,y)=0, \quad g(x,y)=\cos(\pi xy)+y.$$

This example does not satisfy (3.15). So we do not expect that a 3×3 stencil will work for this example. A direct numerical calculation shows that

$$\overline{\alpha}=11, \qquad \alpha=45.$$

Thus the upper bound in (3.1) is $[2\alpha/\overline{\alpha}]+1=13$, which requires a stencil of size 27×27 for the monotone FD discretization. On the other hand, a numerical calculation based on (3.10) and (3.12) shows that a stencil of size 5×5 is sufficient to guarantee a monotone FD discretization.

The numerical results obtained with a stencil of size 5×5 for this example are shown in Table 1. They confirm that the computed solution satisfies the maximum principle and stays between the minimum and maximum values of the solution on the domain boundary.

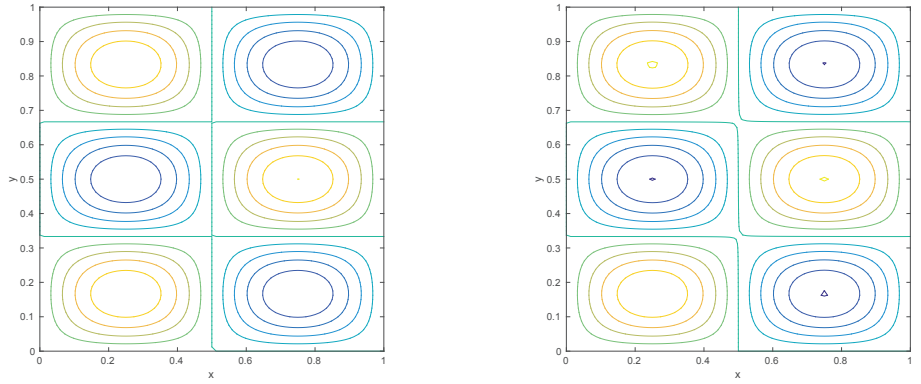
Example 4.2. In this example, we seek to verify the accuracy of the monotone scheme. We use the same differential operator as in Example 4.1, i.e., in the form of (1.1) with (4.1), but choose f and g such that the exact solution of the BVP is given by $u(x,y)=\sin(2\pi x)\sin(3\pi y)$.

https://doi.org/10.4208/cicp.280315.140815a Published online by Cambridge University Press

Table 1: Example 4.1. Extrema of the numerical solutions in the interior and on the boundary of the physical domain.

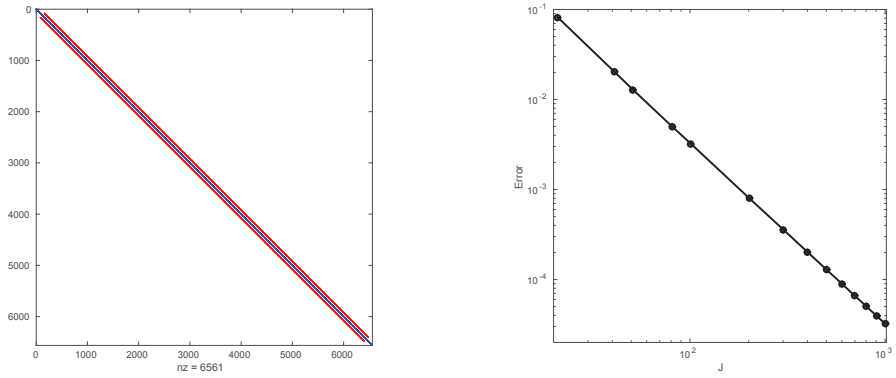
$J=K$	Boundary Min	Interior Min	Boundary Max	Interior Max
21	-5.105652e-2	1.040961e-2	2.000000	1.912261
51	-5.105652e-2	-2.444841e-2	2.000000	1.972163
101	-5.105652e-2	-3.753179e-2	2.000000	1.987642
501	-5.109830e-2	-4.837255e-2	2.000000	1.997862
1001	-5.110190e-2	-4.973654e-2	2.000000	1.998961

The numerical results obtained for this example are shown in Figs. 3 and 4. The contours of the exact and computed solutions are shown in Fig. 3. The sign pattern in Fig. 4(a) shows that the coefficient matrix has nonpositive off-diagonal entries and nonnegative



(a) Exact solution (b) Numerical solution with $J=81$

Figure 3: Example 4.2. Contours of the exact and computed solutions.



(a) Sign pattern for $J=81$ (b) Convergence history

Figure 4: Example 4.2. Sign pattern of the coefficient matrix and the convergence history of the maximum norm of the error of the FD solution.

https://doi.org/10.4208/cicp.280315.140815a Published online by Cambridge University Press

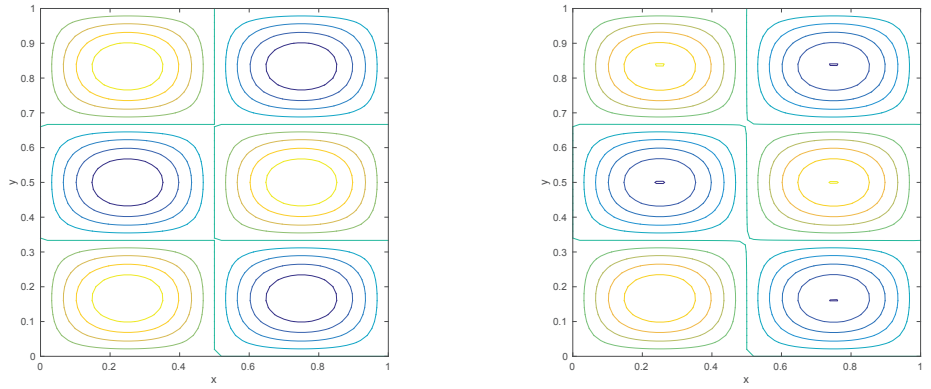
diagonal entries, as stated in Theorem 3.1. Since the scheme is monotone, from the result of Barles and Souganidis [1] we may expect the computed solution to converge in second order, which indeed can be seen from the convergence history shown in Fig. 4(b).

Example 4.3. For this example, the diffusion matrix is given by

$$\mathbb{D} = \begin{bmatrix} 1.1 & \sin(2\pi xy) \\ \sin(2\pi xy) & 1.1 \end{bmatrix}. \tag{4.2}$$

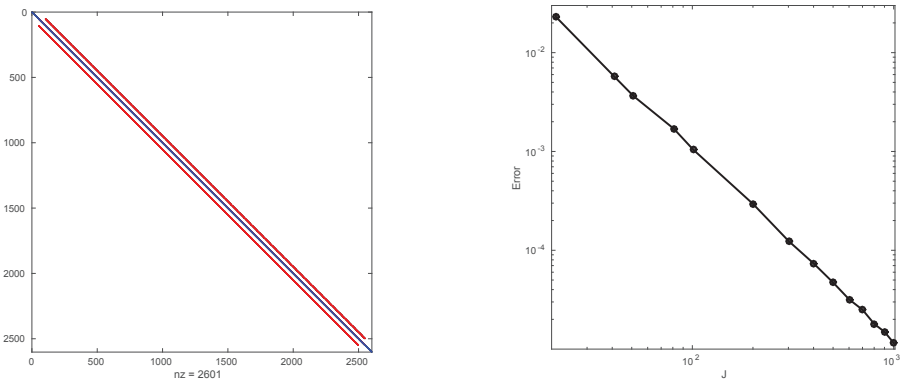
The functions f and g are chosen such that the exact solution of the BVP reads as $u(x,y) = \sin(2\pi x)\sin(3\pi y)$. This example satisfies (3.15) and by Theorem 3.2, we can use a 3×3 stencil for constructing a monotone scheme.

The numerical results for this example are shown in Figs. 5 and 6. One can see that the



(a) Exact solution (b) Numerical solution with $J=51$

Figure 5: Example 4.3. Contours of the exact and computed solutions.



(a) Sign pattern for $J=51$ (b) Convergence history

Figure 6: Example 4.3. Sign pattern of the coefficient matrix and the convergence history for the maximum norm of the error of the FD solution.

https://doi.org/10.4208/cicp.280315.140815a Published online by Cambridge University Press

coefficient matrix has nonpositive off-diagonal entries and nonnegative diagonal entries and the computed solution converges to the exact solution at a rate of second order.

Example 4.4. In this final example, the diffusion matrix is given by

$$\mathbb{D}=\begin{bmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{bmatrix}\begin{bmatrix}k & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{bmatrix}, \tag{4.3}$$

where $\theta=\pi\sin(x)\cos(y)$ and k is a parameter. The bigger k is, the more anisotropic the diffusion matrix is. The more anisotropic $\mathbb{D}$ is, the larger stencil is required for the construction of a monotone FD scheme. We consider two cases with $k=10$ and $k=100$. For both of these cases, we again choose f and g in (1.1) such that the exact solution is $u(x,y)=\sin(2\pi x)\sin(3\pi y)$.

The numerical calculation suggests that a 7×7 stencil be needed to construct a monotone scheme for $k=10$ and a stencil of 53×53 for the case $k=100$. The second order convergence can be seen in Fig. 7 for both cases although the error is almost two magnitude larger for the much more anisotropic case with $k=100$.

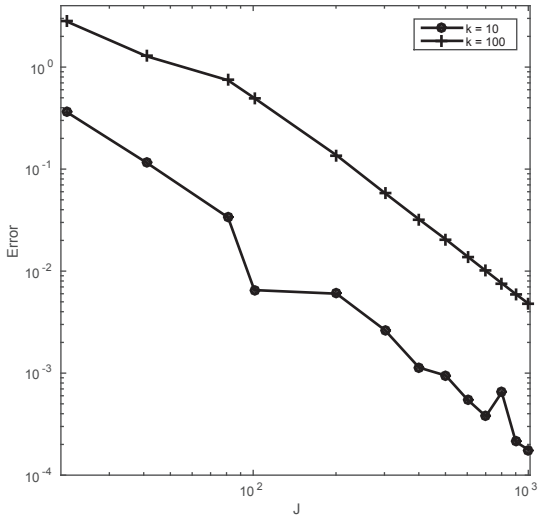


Figure 7: Example 4.4. The convergence history for the maximum norm of the error of the FD solution.

5 Conclusions

In the previous sections we have studied the construction of monotone finite difference schemes for the divergence form (1.1) using nonnegative directional splittings. The main results are stated in Theorems 2.1, 3.1, and 3.2. Theorem 2.1 states that the diffusion operator in (1.1) has a nonnegative directional splitting at any arbitrary interior point (cf. (2.6)) when the diffusion matrix $\mathbb{D}$ is continuous on $\overline{\Omega}$ and symmetric and uniformly

https://doi.org/10.4208/cicp.280315.140815a Published online by Cambridge University Press

positive definite in Ω. This splitting holds in a neighborhood of the point in the sense that the corresponding coefficients are nonnegative in the neighborhood. Moreover, the size of the neighborhood depends only on $\mathbb{D}$. This result is an important improvement over that in [31] where the coefficients of the splitting are nonnegative only at a point. Another improvement is that the function $b(x,y)$ in (1.2) is allowed to change sign on Ω. These improvements are crucial to the construction of monotone finite difference schemes for (1.1) since any finite difference discretization for the divergence form (1.1) requires information of the coefficients in a neighborhood of a mesh point.

The construction of monotone finite difference schemes using the nonnegative directional splitting (2.6) has been discussed in Section 3. A key to the construction is to define principal directions for the underlying finite difference stencil [31]. We have shown that some of these principal directions can be used as the splitting directions when the mesh spacing is sufficiently small and the size of finite difference stencil is sufficiently large. Note that the stencil size has a uniform bound determined by $\mathbb{D}$; see (3.1). This bound can become large when $\mathbb{D}$ is strongly anisotropic. In that case, not only a large stencil but a very fine mesh have to be used. In such a situation, the constructed finite difference scheme may be impractical and a more sophisticated way to construct monotone finite difference schemes may be needed, which is a research topic worth more investigations in the future. On the other hand, Theorem 3.2 shows that a 3×3 stencil can be used to construct monotone finite difference schemes when $\mathbb{D}$ is strictly diagonally dominant. Numerical results in Section 4 are consistent with the theoretical findings.

Acknowledgments

This work was supported in part by the National Science Foundation through grant DMS-1115118.

References

[1] G. Barles and P. Souganidis. Convergence of approximation schemes for fully nonlinear second order equations. *Asymptotic Analysis*, 4(3):271–283, 1991.

[2] J. H. Bramble, B. Hubbard, and V. Thomée. Convergence estimates for essentially positive type discrete dirichlet problems. *Math. Comp.*, 23:695–709, 1969.

[3] P. Ciarlet. Discrete maximum principle for finite difference operators. *Aequationes Math.*, 4:338–352, 1970.

[4] P. Ciarlet and P.-A. Raviart. Maximum principle and uniform convergence for the finite element method. *Comput. Meth. Appl. Mech. Eng.*, 2:17–31, 1973.

[5] 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.*, 116:343–358, 1995.

[6] 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.

https://doi.org/10.4208/cicp.280315.140815a Published online by Cambridge University Press

[7] T. Ertekin, J. H. Abou-Kassem, and G. R. King. *Basic Applied Reservoir Simulation*. SPE textbook series, Vol. 7, Richardson, Texas, 2001.
[8] D. Greenspan and P. C. Jain. On non-negative difference analogues of elliptic differential equations. *J. Franklin Inst.*, 279:360–365, 1965.
[9] S. Gűnter and K. Lackner. A mixed implicit-explicit finite difference scheme for heat transport in magnetised plasmas. *J. Comput. Phys.*, 228:282–293, 2009.
[10] S. Günter, K. Lackner, and C. Tichmann. Finite element and higher order difference formulations for modelling heat transport in magnetised plasmas. *J. Comput. Phys.*, 226:2306–2316, 2007.
[11] S. Günter, Q. Yu, J. Kruger, and K. Lackner. Modelling of heat transport in magnetised plasmas using non-aligned coodinates. *J. Comput. Phys.*, 209:354–370, 2005.
[12] 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.*, 20:104012, 2009.
[13] D. Kuzmin, J. 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.
[14] 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).
[15] 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).
[16] X. P. Li, D. Svyatskiy, and M. Shashkov. Mesh adaptation and discrete maximum principle for 2d anisotropic diffusion problems. Technical report, LANL, 2007. Final Report of the Summer Student Program.
[17] 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.
[18] 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.
[19] M. J. Mlacnik and L. J. Durlofsky. Unstructured grid optimizatioin for improved monotonicity of discrete solutions of elliptic equations with highly anisotropic cofficients. *J. Comput. Phys.*, 216:337–361, 2006.
[20] 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.
[21] T. S. Motzkin and W. Wasow. On the approximation of linear elliptic differential equations by difference equations with positive coefficients. *J. Math. Physics*, 31:253–259, 1953.
[22] P. Mrázek and M. Navara. Consistent positive directional splitting of anisotropic diffusion. In B. Likar, editor, *Proc. of Computer Vision Winter Workshop*, pages 37–48, Bled, Slovenia, February 2001.
[23] A. Oberman. Convergent difference schemes for degenerate elliptic and parabolic equations: Hamilton-jacobi equations and free boundary problems. *SIAM J. Numer. Anal*, 44(2):879–895, 2006.
[24] C. L. 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*, 341:787–792, 2005.

https://doi.org/10.4208/cicp.280315.140815a Published online by Cambridge University Press

[25] C. L. Potier. A nonlinear finite volume scheme satisfying maximum and minimum principles for diffusion operators. *Int. J. Finite Vol.*, 6:20, 2009.
[26] C. L. Potier. Un schéma linéaire vérifiant le principe du maximum pour des opérateurs de diffusion très anisotropes sur des maillages déformés. *C. R. Math. Acad. Sci. Paris*, 347:105–110, 2009.
[27] P. Sharma and G. W. Hammett. Preserving monotonicity in anisotropic diffusion. *J. Comput. Phys.*, 227:123–142, 2007.
[28] G. Stoyan. On a maximum principle for matrices, and on conservation of monotonicity, with applications to discretization methods. *ZAMM*, 62:375–381, 1982.
[29] G. Stoyan. On maximum principles for monotone matrices. *Lin. Alg. Appl.*, 78:147–161, 1986.
[30] R. Varga. *Matrix Iterative Analysis*. Prentice-Hall, New Jersey, 1962.
[31] J. Weickert. *Anisotropic Diffusion in Image Processing*. B.G Teubner Stuttgart, 1998.
[32] G. Yuan and Z. Sheng. Monotone finite volume schemes for diffusion equations on polygonal meshes. *J. Comput. Phys.*, 227:6288–6312, 2008.
[33] 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.280315.140815a Published online by Cambridge University Press