A Monotone Finite Volume Scheme with Second Order Accuracy for Convection-Diffusion Equations on Deformed Meshes

Bin Lan¹, Zhiqiang Sheng² and Guangwei Yuan²,*

¹ The Graduate School of China Academy of Engineering Physics, P.O. Box 2101, Beijing, China, 100088.
² Laboratory of Computational Physics, Institute of Applied Physics and Computational Mathematics, P.O. Box 8009, Beijing, China, 100088.

Communicated by Kun Xu

Received 6 June 2017; Accepted (in revised version) 27 November 2017

Abstract. In this paper, we present a new monotone finite volume scheme for the steady state convection-diffusion equation. The discretization of diffusive flux [33] is utilised and a new corrected upwind scheme with second order accuracy for the discretization of convective flux is proposed based on some available informations of diffusive flux. The scheme is locally conservative and monotone on deformed meshes, and has only cell-centered unknowns. Numerical results are presented to show that the scheme obtains second-order accuracy for the solution and first-order accuracy for the flux.

AMS subject classifications: 35J15, 35J25, 65M08
Key words: Convection-diffusion equation, nonlinear, monotone, deformed meshes.

1 Introduction

It is widely recognized that the discrete maximum principle [11, 30] plays an important role in proving the existence and uniqueness of discrete solution, enforcing numerical stability, and deriving convergence (a priori error estimates) for a sequence of approximate solutions [23]. However, it is not very easy to construct a scheme satisfying the discrete maximum principle on distorted meshes for convection-diffusion equation, especially in the case that the magnitude of convection velocity is much larger than the diffusive coefficient. In this paper, we consider monotonicity (i.e. positivity-preserving) that it can only guarantee nonnegative bound of the numerical solution.

*Corresponding author. Email addresses: yuan_guangwei@iapcm.ac.cn (G. W. Yuan), lanb_nx@126.com (B. Lan), sheng_zhiqiang@iapcm.ac.cn (Z. Q. Sheng)
Many works have been proposed to improve the monotonicity of diffusion scheme, e.g., some restrictive conditions on meshes and diffusive coefficients are given in [12, 21, 24], and some preprocessing or postprocessing methods are proposed in [2, 8, 36]. Recently, some nonlinear schemes without these restrictive conditions have been constructed to guarantee monotonicity in [10, 18, 19, 26, 27, 29, 31–33, 38].

Compared with diffusive term, some distinctive features have to be considered for the discretization of convective term such as upwind characteristic. As far as we know, gradient reconstruction [5, 6, 13, 15, 20, 22, 35, 37] is one of the most popular method, where the convective flux can be approximated via the upwind approach [3] and controlled by different slope limiting techniques [6, 7, 9, 17]. In [5], a MUSCL-like cell-centered finite volume method is proposed to approximate the solution of steady advection-diffusion equations. The method is based on a least square reconstruction of the vertex values. Moreover, the slope limiter, which is required to prevent the formation and growth of spurious numerical oscillations, is designed to guarantee the existence of the discrete solution of the nonlinear scheme. In [20], a second order accurate monotone finite volume method of the steady-state advection-diffusion equation is presented, which uses a second-order upwind method with a specially designed minimal nonlinear correction for the discretization of advective flux. In [35], a monotone finite volume scheme is also presented, in which the approximation of the advective flux is based on the second-order upwind method with slope limiter. In [40], the integral mean of the cell is used to reconstruct the gradient and the slope technique [6] is used to control the numerical oscillations, however, the accuracy of these methods is lost in some cases of the diffusive coefficient being discontinuous.

In this paper, we focus on the discretization of convective flux. We know that the standard upwind scheme has only first-order accuracy. Hence, in order to improve the accuracy, a corrected method with second order accuracy is proposed. The main feature of the method is that our scheme achieves second order accuracy with a more simpler construction procedure than those in [20, 35, 40]. Moreover, it is efficient for some large deformed meshes, such as Kershaw mesh. It is also efficient for the problem with discontinuous, anisotropic and heterogeneous full tensor coefficients on deformed meshes.

The article is organized as follows. We describe the problem and introduce some notations in Section 2. The discretization of diffusive flux is given in Section 3 and the discretization of convective flux is given in Section 4. In Section 5, we show that our scheme is monotone, and give the detailed algorithm. In Section 6, some numerical results are presented to illustrate the features of our scheme. At last, some conclusions are given in Section 7.

2 The problem

Consider the stationary convection-diffusion problem for unknown function $u = u(x)$:
\(-\nabla \cdot (\kappa \nabla u - vu) = f, \text{ in } \Omega, (2.1)\)
\[ u = g, \text{ on } \partial \Omega, (2.2) \]

where \(\Omega\) is a bounded polygonal domain in \(\mathbb{R}^2\) with boundary \(\partial \Omega\), \(v = v(x)\) is a velocity vector field and \(\kappa = \kappa(x)\) is a known diffusive coefficient. Assume that the functions \(v(x), f(x)\) and \(g(x)\) satisfy the following constraints:
\[
\nabla \cdot v \geq 0, \quad v \in C^1(\overline{\Omega})^2, \quad (2.3)
\]
\[
f \in L^2(\Omega), \quad (2.4)
\]
\[
g \in H^{1/2}(\partial \Omega) \cap C(\partial \Omega), \quad (2.5)
\]
and there are two positive constants \(\lambda_1\) and \(\lambda_2\) such that
\[
\lambda_1|\xi|^2 \leq \kappa(x)\xi \cdot \xi \leq \lambda_2|\xi|^2, \quad \forall \xi \in \mathbb{R}^2.
\]

We use a mesh on \(\Omega\) made up of arbitrary convex polygon cells. The set of all cells, edges and nodes are denoted by \(\mathcal{T}, \mathcal{E} \text{ and } \mathcal{N}\), respectively.

We denote the cell by \(K, L, \ldots\), and the cell center is also denoted by \(K, L, \ldots\). In addition, the common edge of two cells \(K\) and \(L\) is denoted by \(\sigma\), i.e., \(\sigma = K|L \in \mathcal{E}\). The cell-edge \(\sigma\) is also denoted by \(AB\), and the midpoint of \(\sigma\) is denoted by \(M\). Let \(\bar{n}_{K,\sigma}\) (or \(\bar{n}_{L,\sigma}\)) be the unit outer normal vector on the edge \(\sigma\) of cell \(K\) (or \(L\)). \(\kappa^T\) be the transpose of matrix \(\kappa\) and \(\mathcal{E}_K\) be the set of all edges of cell \(K\). Denote \(\mathcal{E}_{\text{int}} = \mathcal{E} \cap \Omega, \mathcal{E}_{\text{ext}} = \mathcal{E} \cap \partial \Omega\). Denote \(h = (\text{sup}_{K \in \mathcal{T}} \text{diam}(K))^{1/2}\), where \(\text{diam}(K)\) is the diameter of cell \(K\).

Integrate (2.1) on the cell \(K\), to obtain
\[
\sum_{\sigma \in \mathcal{E}_K} (\mathcal{F}_{K,\sigma} + \mathcal{G}_{K,\sigma}) = \int_K f(x) dx, \quad (2.6)
\]
where
\[
\mathcal{F}_{K,\sigma} = -\int_{\sigma} \nabla u(x) \cdot \kappa^T(x) \bar{n}_{K,\sigma} dl, \quad (2.7)
\]
and
\[
\mathcal{G}_{K,\sigma} = \int_{\sigma} vu(x) \cdot \bar{n}_{K,\sigma} dl. \quad (2.8)
\]

3 The diffusive flux

We use the method in [33] to discretize the diffusive flux, that is Eq. (2.7). In the method, the two nonnegative parameters are introduced to define a nonlinear two-point flux, in which one point is the cell-center and the other is the midpoint of cell-edge. At last, the continuity of normal flux on the cell-edge is used to give the final diffusive flux. Now we give a brief review of [33].
A ray originating at the point $K$ along the direction $\kappa^T \tilde{n}_{K,\sigma}$ must intersect one segment connecting two neighboring midpoints of edge of cell $K$, where the two midpoints are denoted by $P_1$ and $P_2$ and the cross point is denoted by $O_1$. Similarly, a ray originating at the midpoint $M$ along the direction $-\kappa^T \tilde{n}_{K,\sigma}$ must intersect certain segment $KP_4$, where $P_4$ must be one vertex of $\sigma$ and the cross point is denoted by $O_2$. In addition, let $\tilde{t}_{KP_1}, \tilde{t}_{KP_2}, \tilde{t}_{MK}$ and $\tilde{t}_{MP_4}$ be some unit tangential vectors along their corresponding directions, respectively. $\theta_i$ ($i = 1, \cdots, 4$) are some angles (see Fig. 1).

Then, the following relations are established:

\[
\begin{align*}
\frac{\kappa^T \tilde{n}_{K,\sigma}}{|\kappa^T \tilde{n}_{K,\sigma}|} &= \frac{\sin \theta_2}{\sin(\theta_1 + \theta_2)} \tilde{t}_{KP_1} + \frac{\sin \theta_1}{\sin(\theta_1 + \theta_2)} \tilde{t}_{KP_2}, \\
-\frac{\kappa^T \tilde{n}_{K,\sigma}}{|\kappa^T \tilde{n}_{K,\sigma}|} &= \frac{\sin \theta_4}{\sin(\theta_3 + \theta_4)} \tilde{t}_{MK} + \frac{\sin \theta_3}{\sin(\theta_3 + \theta_4)} \tilde{t}_{MP_4}.
\end{align*}
\]

Substituting Eq. (3.1) into Eq. (2.7), and neglecting the high order terms, we obtain the following approximation of diffusive flux

\[
\begin{align*}
\tilde{F}_1 &= -|\kappa^T \tilde{n}_{K,\sigma}| |\sigma| \left( \frac{\sin \theta_2}{\sin(\theta_1 + \theta_2)} \frac{u_{K_1} - u_K}{|KP_1|} + \frac{\sin \theta_1}{\sin(\theta_1 + \theta_2)} \frac{u_{K_2} - u_K}{|KP_2|} \right) \\
&= a_1 (u_K - u_{P_1}) + a_2 (u_K - u_{P_2}),
\end{align*}
\]

where \( a_1 = \frac{|\kappa^T \tilde{n}_{K,\sigma}| |\sigma|}{|KP_1|} \frac{\sin \theta_2}{\sin(\theta_1 + \theta_2)} \) and \( a_2 = \frac{|\kappa^T \tilde{n}_{K,\sigma}| |\sigma|}{|KP_2|} \frac{\sin \theta_1}{\sin(\theta_1 + \theta_2)} \). Similarly, substituting Eq. (3.2) into Eq. (2.7), we have another approximate flux

\[
\begin{align*}
\tilde{F}_2 &= |\kappa^T \tilde{n}_{K,\sigma}| |\sigma| \left( \frac{\sin \theta_4}{\sin(\theta_3 + \theta_4)} \frac{u_K - u_M}{|MK|} + \frac{\sin \theta_3}{\sin(\theta_3 + \theta_4)} \frac{u_{P_4} - u_M}{|MP_4|} \right) \\
&= a_3 (u_K - u_M) + a_4 (u_{P_4} - u_M),
\end{align*}
\]
where \( a_3 = \frac{|\mathbf{e}_x \times \mathbf{e}_y| |\sigma|}{M \sin \theta_1 + \sin \theta_2} \) and \( a_4 = \frac{|\mathbf{e}_x \times \mathbf{e}_y| |\sigma|}{M \sin \theta_1 + \sin \theta_2} \).

We define the discrete normal flux on edge \( \sigma \) by using Eqs. (3.3) and (3.4) as follows:

\[
F_{K,\sigma} = \mu_1 \hat{F}_1 + \mu_2 \hat{F}_2
\]

\[
= \mu_1 a_1 (u_K - u_P) + \mu_1 a_2 (u_K - u_P) + \mu_2 a_3 (u_K - u_M) + \mu_2 a_4 (u_P - u_M),
\]

(3.5)

where \( \mu_1 \) and \( \mu_2 \) are two coefficients satisfying convex combination \( \mu_1 + \mu_2 = 1 \), which will be determined later.

In order to assure \( \mu_1 \) and \( \mu_2 \) are positive, two additional parameters \( \omega_1 \) and \( \omega_2 \) are introduced later. There exists three cases according to the different positions of \( P_1 \) and \( P_2 \).

When \( u_P \neq u_M \) and \( u_P \neq u_M, \) the normal flux (3.5) can be rewritten as

\[
F_{K,\sigma} = (\mu_1 (a_1 + a_2) + \mu_2 a_3) u_K - \mu_2 (a_3 + a_4) u_M - \mu_1 (a_1 u_P + a_2 u_P) + \mu_2 a_4 u_P
\]

\[
+ \mu_1 (a_1 (u_P + \omega_1 u_K) + a_2 (u_P + \omega_1 u_K)) + \mu_2 a_4 (u_P + \omega_2 u_M).
\]

(3.6)

In order to obtain the two-point flux approximation, the last two terms of the above expression should be vanished, hence, \( \mu_1 \) and \( \mu_2 \) are taken as follows:

\[
\mu_1 = \frac{a_4 (u_P + \omega_2 u_M)}{a_1 (u_P + \omega_1 u_K) + a_2 (u_P + \omega_1 u_K) + a_4 (u_P + \omega_2 u_M)},
\]

(3.7)

\[
\mu_2 = \frac{a_4 (u_P + \omega_1 u_K) + a_2 (u_P + \omega_1 u_K) + a_4 (u_P + \omega_2 u_M)}{a_1 (u_P + \omega_1 u_K) + a_2 (u_P + \omega_1 u_K) + a_4 (u_P + \omega_2 u_M)}.
\]

(3.8)

Hence, (3.6) can be expressed as follows

\[
F_{K,\sigma} = A_{K,1} u_K - A_{K,2} u_M,
\]

(3.9)

where \( A_{K,1} = \mu_1 (a_1 + a_2) (1 + \omega_1) + \mu_2 a_3 \) and \( A_{K,2} = \mu_2 (a_3 + a_4 (1 + \omega_2)) \). In order to assure \( \mu_1 \) and \( \mu_2 \) are positive, we can choose \( \omega_1 \) and \( \omega_2 \) such that \( a_1 (u_P + \omega_1 u_K) + a_2 (u_P + \omega_1 u_K) \geq 0 \) and \( a_4 (u_P + \omega_2 u_M) \geq 0 \). If \( a_1 (u_P + \omega_1 u_K) + a_2 (u_P + \omega_1 u_K) = a_4 (u_P + \omega_2 u_M) = 0 \), we let \( \mu_1 = \mu_2 = \frac{1}{2} \).

When \( u_P = u_M \) or \( u_P = u_M \), (3.5) can be rewritten as the similar form (3.9) by using the above method.

Similar to the expression (3.9), a discrete flux on edge \( \sigma \) of cell \( L \) can be obtained

\[
F_{L,\sigma} = A_{L,1} u_L - A_{L,2} u_M.
\]

(3.10)

Using the continuity of normal flux \( F_{K,\sigma} + F_{L,\sigma} = 0 \), we have

\[
u_M = \frac{A_{K,1} u_K + A_{L,1} u_L}{A_{K,2} + A_{L,2}}.
\]

(3.11)

Substitute the above expression into Eq. (3.9) to obtain the monotone nonlinear two-point diffusive flux on \( \sigma = K|L \)

\[
F_{K,\sigma} = A_{K,\sigma} u_K - A_{L,\sigma} u_L,
\]

(3.12)

where \( A_{K,\sigma} = \frac{A_{K,1} A_{L,2}}{A_{K,2} + A_{L,2}} \) and \( A_{L,\sigma} = \frac{A_{K,2} A_{L,1}}{A_{K,2} + A_{L,2}} \).

By (3.11), we can know \( u_M \) is positive as long as \( u_K \) and \( u_L \) are positive.
4 The convective flux

In this section, we focus on the discretization of convective flux. \( \forall K|L \in \mathcal{E}_{int} \).

By (2.8), we have

\[
G_{K,\sigma}(L) = \bar{u}_M \int_{\sigma} v \cdot \vec{n}_{K,\sigma} dl + O(h^2)
\]

\[
= \bar{u}_M (v^+_K - v^-_K) + O(h^2)
\]

\[
= v^+_K R_{K,\sigma}(M) - v^-_K R_{L,\sigma}(M) + O(h^2),
\]

(4.1)

where \( \bar{u}_M \) is an approximate value of \( u(M) \) and

\[
v^+_K = \frac{|v^+_K| + v^+_K}{2}, \quad v^-_K = \frac{|v^-_K| - v^-_K}{2}, \quad v_K = \int_{\sigma} v \cdot \vec{n}_{K,\sigma} dl.
\]

In Eq. (4.1), \( \bar{u}_M \) can not be substituted by \( u_M \) which is mentioned in the previous section directly. According to the definition in [33], \( u_M \) comes from the weighted average of \( u_K \) and \( u_L \) and does not satisfy the property of upwind. In order to preserve the property of upwind, we need to define a new value \( \bar{u}_M \). For example, \( R_{K,\sigma}(M) \) is a value of \( \bar{u}_M \) at the midpoint \( M \) of cell-edge \( \sigma \) that only relies on the cell \( K \).

Thus, we can obtain the approximate expression of upwind formula [16]

\[
G_{K,\sigma}(M) \approx v^+_K R_{K,\sigma}(M) - v^-_K R_{L,\sigma}(M).
\]

(4.2)

Usually, the first-order accuracy approximation is as follows:

\[
R_{K,\sigma}(M) = u_K + O(h) \approx u_K.
\]

(4.3)

In the following, we propose a method to obtain an approximation with second-order accuracy.

A local stencil is given in Fig. 2. For the cell \( K \), the ray originated at the midpoint \( M \) of cell-edge \( \sigma \) along the direction \( \overrightarrow{MK} \) must intersect another cell-edge \( \sigma' \) of cell \( K \), and we let \( C \) and \( D \) (two cell-vertices of \( K \)) be two endpoints of \( \sigma' \), and \( M' \) be the cross point.

In order to improve the accuracy, the approximate value at \( \sigma \) can be given by Taylor series expansion at the cell-edge \( \sigma \) to the cell-center \( K \) as follows:

\[
R_{K,\sigma}(M) = u_K + \nabla u|_{K} \cdot \overrightarrow{KM} + O(h^2).
\]

(4.4)

Noticing that

\[
\nabla u|_{K} \cdot \overrightarrow{KM} = \nabla u|_{K} \cdot \vec{t}_{KM} |\overrightarrow{KM}|
\]

\[
= \nabla u|_{K} \cdot \frac{M' \overrightarrow{KM}}{|M' \overrightarrow{KM}|} |\overrightarrow{KM}|
\]

\[
= \frac{u_M - u'_M}{|M' \overrightarrow{KM}|} |\overrightarrow{KM}| + O(h^2),
\]

(4.5)
where $\vec{t}_{KM}$ is the unit tangential vectors on the line $KM$, we have

$$\mathcal{R}_{K,\sigma}(M) = u_K + \frac{u_M - u_{M'}}{|M'M|} |\vec{KM}| + \mathcal{O}(h^2)$$

$$\approx u_K + \frac{u_M - u_{M'}}{|M'M|} |\vec{KM}|. \quad (4.6)$$

Similarly, the approximate value at $\sigma$ can be given by Taylor’s series expansion at the cell-edge $\sigma$ to the cell-center $L$ as follows:

$$\mathcal{R}_{L,\sigma}(M) = u_L + \nabla u|_{L \cdot L'M} + \mathcal{O}(h^2),$$

and

$$\nabla u|_{L \cdot L'M} = \frac{u_M - u_{M''}}{|M''M|} |\vec{LM}| + \mathcal{O}(h^2). \quad (4.8)$$

Then, we can obtain

$$\mathcal{R}_{L,\sigma}(M) = u_L + \frac{u_M - u_{M''}}{|M''M|} |\vec{LM}| + \mathcal{O}(h^2)$$

$$\approx u_L + \frac{u_M - u_{M''}}{|M''M|} |\vec{LM}|. \quad (4.9)$$

According to the definition of the convective flux in Eqs. (4.1) and (4.2), the form of its discretization can be given (the following nonlinear structure can be found in [20, 35, 40]). When $\sigma \in E_{int}$, we define

$$G_{K,\sigma} = B_{K,\sigma} u_K - B_{L,\sigma} u_L,$$
where \( B_{K,\sigma} = v_{K,\sigma}^+ R_{K,\sigma}(M) / (u_K + d_0) \), \( B_{L,\sigma} = v_{L,\sigma}^+ R_{L,\sigma}(M) / (u_L + d_0) \), and \( d_0 \) is a small positive number to be determined later.

When \( \sigma \in \mathcal{E}_{\text{ext}} \), we define
\[
G_{K,\sigma} = B_{K,\sigma} u_K - b_{K,\sigma},
\]
(4.11)

where \( B_{K,\sigma} = v_{K,\sigma}^+ R_{K,\sigma}(M) / (u_K + d_0) \), \( b_{K,\sigma} = v_{K,\sigma}^- g_M \).

Remark 4.1. From (4.6) and (4.9), \( B_{K,\sigma} \) and \( B_{L,\sigma} \) maybe negative. In order to guarantee the monotonicity of the scheme, we will correct them in the iteration step. The detailed algorithm will be introduced in Section 5.

Remark 4.2. For the point \( M' \) on the cell-edge \( \sigma' \), the following expression can be used to compute its value
\[
u_{M'} = \frac{|M'D| u_C + |MC| u_D}{|CD|}.
\]

Remark 4.3. For quadrilateral meshes, if the cell-center is the simple average of four cell-vertices, the point \( M' \) is exactly the midpoint of cell-edge \( \sigma' \).

Remark 4.4. From the definition of Eqs. (4.10)-(4.11), the method can avoid the case that the denominators are very close to zero or equal to zero without reducing the precision. It is different from [20, 35] and [40].

Remark 4.5. In [35], a linear reconstruction function of concentration is given and a constrained minimization problem is considered to avoid nonphysical oscillation. In [40], the similar linear reconstruction of approximate solution is utilized and the slope limiter factor is used to avoid nonphysical solutions. The main difference between the scheme in this paper and these in [35] and [40] is that the discretizations of convective flux are different. Although our new scheme has a similar nonlinear structure, a more simpler construction procedure for discrete gradient is proposed. The numerical results show that the new method is suitable for large deformed meshes in the Section 6.

Analysis. The expression of convective flux in Eqs. (4.10)-(4.11) has second order accuracy when the constant \( d_0 \) satisfies some restrictive conditions. Noticing Eqs. (4.1) and (4.10), we assume \( v_{K,\sigma}^+ > 0 \), then Eq. (4.1) can be written as the following form
\[
G_{K,\sigma} = \frac{v_{K,\sigma}^+ R_{K,\sigma}(M)}{u_K + d_0} (u_K + d_0) + \mathcal{O}(h^2)
\]
\[
= B_{K,\sigma} u_K + d_0 \frac{v_{K,\sigma}^+ R_{K,\sigma}(M)}{u_K + d_0} + \mathcal{O}(h^2).
\]

In order to obtain second order accuracy, the second term on the right hand needs to satisfy
\[
d_0 \frac{v_{K,\sigma}^+ R_{K,\sigma}(M)}{u_K + d_0} \leq Ch^2,
\]
where $C > 0$ is a constant independent of $h$. Hence, there is

$$\frac{v^+_{K,\sigma} R_{K,\sigma}(M)}{Ch^2} \leq \frac{u_K}{d_0}.$$ 

If $v^+_{K,\sigma} R_{K,\sigma}(M) \leq Ch^2$, then $d_0 > 0$ satisfies the requirement, else

$$d_0 \leq \frac{u_K Ch^2}{v^+_{K,\sigma} R_{K,\sigma}(M) - Ch^2}.$$ 

5 The monotone finite volume scheme

5.1 The scheme

According to the definition of discrete flux, the finite volume scheme can be written as follows:

$$\sum_{\sigma \in E_K} (F_{K,\sigma} + G_{K,\sigma}) = f_K m(K), \quad \forall K \in T,$$

$$u_m = g_m, \quad \forall M \in \partial \Omega,$$ 

where $F_{K,\sigma}$ is defined in (3.12), $G_{K,\sigma}$ is defined in (4.10) and (4.11), $g_m = g(M)$ and $f_K = f(K)$.

5.2 The nonlinear system

From the scheme (5.1)-(5.2), a nonlinear algebraic system can be formed. Let $U$ be the discrete unknown vector and $A(U)$ be the coefficient matrix of this system, the matrix $A(U)$ will consist of $2 \times 2$ matrix (for the inner edges)

$$A_{\sigma}(U) = \begin{pmatrix} M_{K,\sigma}(U) & -M_{L,\sigma}(U) \\ -M_{K,\sigma}(U) & M_{L,\sigma}(U) \end{pmatrix},$$

and $1 \times 1$ matrix (edges on the boundary) $A_{\sigma}(U) = M_{K,\sigma}(U)$, where

$$M_{K,\sigma}(U) = A_{K,\sigma}(U) + B_{K,\sigma}(U),$$

$$M_{L,\sigma}(U) = A_{L,\sigma}(U) + B_{L,\sigma}(U).$$

Then the global discrete nonlinear system reads as:

$$A(U)U = F.$$ 

where

$$A(U) = \sum_{\sigma \in E} N_{\sigma} A_{\sigma}(U) N_{\sigma}^T,$$ 

$N_{\sigma}$ is assembling matrix whose elements are 0 or 1.

The matrix $A(U)$ is nonsymmetric and has some characters as follows:
1. All diagonal entries of matrix $A(U)$ are positive;
2. All off-diagonal entries of $A(U)$ are non-positive;
3. Each column sum in $A(U)$ is non-negative and there at least exists a column with a positive sum.

So $A(U)$ is weak diagonal dominance in column. We use the Picard nonlinear iteration method to solve the nonlinear system (5.3): choose a small value $\varepsilon_{\text{non}} > 0$ and initial vector $U^0 > 0$, and repeat for nonlinear iteration index $s = 1, 2, \cdots$.

1. Solve $A(U^{s-1})U^s = F$,
2. Stop if $||A(U^s)U^s - F||_2 \leq \varepsilon_{\text{non}}||A(U^0)U^0 - F||_2$.

The linear system with nonsymmetric coefficient matrix $A(U^{s-1})$ is solved by the Bi-Conjugate Gradient Stabilized (BiCGSTab) method, and the linear iterations are terminated when relative norm of the initial residual becomes smaller than $\varepsilon_{\text{lin}}$.

5.3 The algorithm

In this subsection, we describe the detailed algorithm.

1. Initial $U^0 > 0$, $\varepsilon_{\text{non}}$ and $\varepsilon_{\text{lin}}$.
2. When $s = 0$,
   (a) compute $A_{K,\sigma}^{(0)}$, $B_{K,\sigma}^{(0)}$,
   (b) correct $B_{K,\sigma}^{(0)}$: if $A_{K,\sigma}^{(0)} + B_{K,\sigma}^{(0)} < 0$, correct it as Remark 4.5;
   (c) compute initial residual $||A(U^{(0)})U^{(0)} - F||_2$.
3. When $s = 1, 2, \cdots$,
   (a) solve $A(U^{(s-1)})U^{(s)} = F$;
   (b) compute $u^{(s)}_P$, $u^{(s)}_M$, $\forall P \in N$, $M \in \sigma$;
   (c) compute $u^{(s)}_{M'}$, $\forall M' \in \sigma$;
   (d) compute $A_{K,\sigma}^{(s)}$, $B_{K,\sigma}^{(s)}$;
   (e) correct $B_{K,\sigma}^{(s)}$: if $A_{K,\sigma}^{(s)} + B_{K,\sigma}^{(s)} < 0$, correct it as Remark 4.5;
   (f) compute residual $||A(U^{(s)})U^{(s)} - F||_2$;
   (g) whether $||A(U^{(s)})U^{(s)} - F||_2 \leq \varepsilon_{\text{non}}||A(U^{(0)})U^{(0)} - F||_2$, if true, then go to 4, otherwise, go to 3.
4. Stop.
Remark 5.1. In order to preserve the monotonicity of the scheme, $A_{K,\sigma}^{(s)}$ and $B_{K,\sigma}^{(s)}$ must be nonnegative. According to the construction of diffusive flux in Section 3, $A_{K,\sigma}^{(s)}$ is always nonnegative. However, $B_{K,\sigma}^{(s)}$ maybe negative. For this case, if $A_{K,\sigma}^{(s)} + B_{K,\sigma}^{(s)} < 0$, we can take the following nonnegative value
\[
B_{K,\sigma}^{(s)} = \frac{u_{M}^{(s)} v_{K,\sigma}^{+}}{u_{K}^{(s)}},
\]
if $R_{K,\sigma}(M) < 0$, it can be substituted by the value $u_{M}^{(s)}$ of previous iteration, which preserves the property of upwind.

5.4 The monotonicity

From the construction of scheme, we know that all diagonal entries of the coefficient matrix $A(U^{(s)})$ are positive. Next, in order to show that the scheme is monotone, we introduce the following lemma [38].

Lemma 5.1 ([4]). For an irreducible matrix $A = (a_{ij})_{n \times n}$ satisfying $a_{ii} > 0 (1 \leq i \leq n)$ and $a_{ij} \leq 0 (1 \leq i, j \leq n, i \neq j)$, if $A$ is weak diagonal dominance in rows, that is
\[
\sum_{j=1}^{n} a_{ij} \geq 0, \quad i = 1, 2, \cdots, n,
\]
with strict inequality for at least one of Eq. (5.5), then the matrix $A$ is an M-matrix [1,11].

Now we state that our scheme is monotone.

For the scheme presented for the convection-diffusion equation, we can obtain the following theorem [38]:

Theorem 5.1. Let $F \geq 0$, $U^{0} \geq 0$, and linear systems in Picard iterations are solved exactly. Then all iterates $U^{k}$ are non-negative vectors $U^{k} \geq 0$.

The detailed proof of monotonicity is given in [38].

6 Numerical experiments

In order to demonstrate the accuracy and robustness of the scheme, we test some problems on Random meshes and Kershaw meshes [14]. We give discrete $L_2$-norms: $\epsilon_{u}^{\ell} = [\sum_{K \in T} (u_{K} - u(K))^2 m(K)]^{1/2}$ and $\epsilon_{f}^{\ell} = [\sum_{\sigma \in \mathcal{E}} (F_{K,\sigma} - F_{K,\sigma})^2 |\sigma|]^{1/2}$ to evaluate approximate
errors, and take $\epsilon_{\text{non}} = 1.0 \times 10^{-6}$ and $\epsilon_{\text{lin}} = 1.0 \times 10^{-10}$. The convergent order can be obtained by the following formula:

$$\text{Order} = \frac{\log(\text{Error}(N_1) / \text{Error}(N_2))}{\log(N_2 / N_1)},$$

where $N_1$ and $N_2$ represent different number of cells, $\text{Error}(N_1)$ and $\text{Error}(N_2)$ are the corresponding errors.

The random meshes over the physical domain are formed by random disturbance on the inner vertices of rectangular meshes. For each inner vertex $(x, y)$, we have

$$x := x + \sigma \xi_x h, \quad y := y + \sigma \xi_y h,$$

where $\xi_x$ and $\xi_y$ is random number between $-0.5$ and $0.5$, $\sigma \in [0, 1]$ is a parameter of disturbance. In this paper, we take $\sigma = 0.6$ for all random meshes.

### 6.1 The problem with continuous coefficient

Consider the problem (2.1)-(2.2) with Dirichlet boundary condition on $\Omega = [0,1] \times [0,1]$. Take $v = (2, 3)^T$, and the exact solution is as follows

$$u(x, y) = e^{-\frac{(x-1)^2 + (y-1)^2}{2\eta^2}},$$

where $\eta = 0.1$.

Let’s go back to the Eqs. (4.10)-(4.11) and Remark 4.4 in Section 4, our scheme is different from [20, 35] and [40], because our method can avoid the case that the denominators are very close to zero or equal to zero without reducing the precision. Here, the exact solution is very close to zero nearby $x = 0$ and $y = 0$. In order to show the accuracy and efficiency of our scheme, we test it on two kinds of quadrilateral meshes (see Fig. 3). And

Figure 3: Random quadrilateral and Kershaw meshes.
Table 1: Accuracy for the problem with boundary layers ($\kappa = 1.0$).

<table>
<thead>
<tr>
<th>cells</th>
<th>Random meshes</th>
<th>Kershaw meshes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\epsilon^u_2$</td>
<td>Order</td>
</tr>
<tr>
<td>12 x 12</td>
<td>5.43E-3</td>
<td>-</td>
</tr>
<tr>
<td>24 x 24</td>
<td>1.30E-3</td>
<td>2.06</td>
</tr>
<tr>
<td>48 x 48</td>
<td>4.44E-4</td>
<td>1.55</td>
</tr>
<tr>
<td>96 x 96</td>
<td>1.04E-4</td>
<td>2.09</td>
</tr>
<tr>
<td>192 x 192</td>
<td>2.66E-5</td>
<td>1.97</td>
</tr>
</tbody>
</table>

Table 2: Accuracy for the problem with boundary layers ($\kappa = 10^{-2}$).

<table>
<thead>
<tr>
<th>cells</th>
<th>Random meshes</th>
<th>Kershaw meshes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\epsilon^u_2$</td>
<td>Order</td>
</tr>
<tr>
<td>12 x 12</td>
<td>7.64E-3</td>
<td>-</td>
</tr>
<tr>
<td>24 x 24</td>
<td>1.57E-3</td>
<td>2.28</td>
</tr>
<tr>
<td>48 x 48</td>
<td>4.42E-4</td>
<td>1.83</td>
</tr>
<tr>
<td>96 x 96</td>
<td>9.63E-5</td>
<td>2.20</td>
</tr>
<tr>
<td>192 x 192</td>
<td>2.45E-5</td>
<td>1.98</td>
</tr>
</tbody>
</table>

Table 3: Accuracy for the problem with boundary layers ($\kappa = 10^{-4}$).

<table>
<thead>
<tr>
<th>cells</th>
<th>Random meshes</th>
<th>Kershaw meshes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\epsilon^u_2$</td>
<td>Order</td>
</tr>
<tr>
<td>12 x 12</td>
<td>9.86E-3</td>
<td>-</td>
</tr>
<tr>
<td>24 x 24</td>
<td>2.32E-3</td>
<td>1.89</td>
</tr>
<tr>
<td>48 x 48</td>
<td>7.41E-4</td>
<td>1.84</td>
</tr>
<tr>
<td>96 x 96</td>
<td>1.59E-4</td>
<td>2.22</td>
</tr>
<tr>
<td>192 x 192</td>
<td>3.58E-5</td>
<td>2.15</td>
</tr>
</tbody>
</table>

the numerical results with different diffusive coefficients ($\kappa = 1, 10^{-2}, 10^{-4}, 10^{-6}$) are given in Tables 1, 2, 3 and 4, respectively.

From these tables, we find that our scheme obtains good accuracy, especially for the convection-dominated cases. When $\kappa = 10^{-4}$ and $10^{-6}$, the convergent order for the solution and the flux is about second order on the two kinds of meshes, as number of cells is increased. When $\kappa = 1.0$, the convergent order of solution and flux is higher than first order on Kershaw meshes, as number of cells is increased.

In addition, although the part of solutions are very close to zero ($u_{\min}(x, y) = 3.82E-43$ on $48 \times 48$ meshes), i.e., the values of $u_K, u_L \to 0$ on denominators, the computed results show that the definitions of Eq. (4.10) and Eq. (4.11) are efficient.
Table 4: Accuracy for the problem with boundary layers ($\kappa = 10^{-6}$).

<table>
<thead>
<tr>
<th>cells</th>
<th>Random meshes</th>
<th>Kershaw meshes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\varepsilon_2^1$</td>
<td>Order</td>
</tr>
<tr>
<td>$12 \times 12$</td>
<td>9.89E-3</td>
<td>-</td>
</tr>
<tr>
<td>$24 \times 24$</td>
<td>2.64E-3</td>
<td>1.91</td>
</tr>
<tr>
<td>$48 \times 48$</td>
<td>7.46E-4</td>
<td>1.82</td>
</tr>
<tr>
<td>$96 \times 96$</td>
<td>1.61E-4</td>
<td>2.21</td>
</tr>
<tr>
<td>$192 \times 192$</td>
<td>3.62E-5</td>
<td>2.15</td>
</tr>
</tbody>
</table>

Table 5: The minimal value for the different cases on the Kershaw meshes.

<table>
<thead>
<tr>
<th>cells</th>
<th>$\kappa$</th>
<th>no correction</th>
<th>correction</th>
</tr>
</thead>
<tbody>
<tr>
<td>$12 \times 12$</td>
<td>$10^{-4}$</td>
<td>-1.76E-4</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>$10^{-6}$</td>
<td>-1.64E-4</td>
<td>0</td>
</tr>
<tr>
<td>$24 \times 24$</td>
<td>$10^{-4}$</td>
<td>-1.79E-7</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>$10^{-6}$</td>
<td>-1.18E-7</td>
<td>0</td>
</tr>
</tbody>
</table>

On the other hand, for the discretization of the convective term, it is possible to give negative values in nonlinear iteration and then the final solution maybe negative. In Table 5, we present the minimal value for the cases without correction and with correction on Kershaw quadrilateral meshes, respectively. From this table, we find that there exists negative solution for the case without correction, and has no negative solution for the case with correction.

6.2 The problem with discontinuous coefficient

Consider the problem (2.1)-(2.2) with Dirichlet boundary condition on $\Omega = [0,1] \times [0,1]$, and take $v = (1,1)^T$. The exact solution is as follows:

$$u(x,y) = \begin{cases} 
\sin \frac{\pi}{2}x + \sin \frac{\pi}{2}y, & x < \frac{1}{2}, \\
\sqrt{2} \csc \left(\frac{x - \frac{1}{2}}{4}\right) + \sin \frac{\pi}{2}y + \frac{\pi}{2}, & x \geq \frac{1}{2},
\end{cases}$$

and the diffusive coefficient is as follows:

$$\kappa = \begin{cases} 
c \times k_0, & x < \frac{1}{2}, \\
k_0, & x \geq \frac{1}{2},
\end{cases}$$

where $c = 40$.

We test the problem on two kinds of quadrilateral meshes: Kershaw meshes in Fig. 3 and the random quadrilateral meshes in Fig. 4. The numerical results for different diffusive coefficients $k_0 = 1, 10^{-2}, 10^{-4}$ and $10^{-6}$ are given in Tables 6, 7, 8 and 9, respectively.
Figure 4: Random quadrilateral meshes.

Table 6: Accuracy for the problem with discontinuous coefficient ($k_0 = 1.0$).

<table>
<thead>
<tr>
<th>cell</th>
<th>Random meshes</th>
<th>Kershaw meshes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\epsilon_u^2$</td>
<td>Order</td>
</tr>
<tr>
<td>12 $\times$ 12</td>
<td>9.19E-4</td>
<td>-</td>
</tr>
<tr>
<td>24 $\times$ 24</td>
<td>2.39E-4</td>
<td>1.94</td>
</tr>
<tr>
<td>48 $\times$ 48</td>
<td>6.48E-5</td>
<td>1.88</td>
</tr>
<tr>
<td>96 $\times$ 96</td>
<td>1.65E-5</td>
<td>1.97</td>
</tr>
<tr>
<td>192 $\times$ 192</td>
<td>4.19E-6</td>
<td>1.98</td>
</tr>
</tbody>
</table>

Table 7: Accuracy for the problem with discontinuous coefficient ($k_0 = 10^{-2}$).

<table>
<thead>
<tr>
<th>cell</th>
<th>Random meshes</th>
<th>Kershaw meshes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\epsilon_u^2$</td>
<td>Order</td>
</tr>
<tr>
<td>12 $\times$ 12</td>
<td>6.52E-4</td>
<td>-</td>
</tr>
<tr>
<td>24 $\times$ 24</td>
<td>1.90E-4</td>
<td>1.78</td>
</tr>
<tr>
<td>48 $\times$ 48</td>
<td>5.06E-5</td>
<td>1.91</td>
</tr>
<tr>
<td>96 $\times$ 96</td>
<td>1.28E-5</td>
<td>1.98</td>
</tr>
<tr>
<td>192 $\times$ 192</td>
<td>3.24E-6</td>
<td>1.98</td>
</tr>
</tbody>
</table>

When $k_0 = 10^{-4}$ and $10^{-6}$, we also find that our scheme obtains second-order accuracy for the solution and flux. When $k_0 = 1.0$ and $10^{-2}$, we obtain second-order accuracy for the solution, and first-order accuracy for the flux, as number of cells is increased. These numerical results show that our scheme can obtains good accuracy.

Then, in order to show the efficiency of the present scheme (Method 2), the comparison of accuracy between Method 1 in [40] and Method 2 is given in Table 10. It is obvious that our method is more efficient.
### Table 8: Accuracy for the problem with discontinuous coefficient ($k_0 = 10^{-4}$).

<table>
<thead>
<tr>
<th>cell</th>
<th>Random meshes</th>
<th>Kershaw meshes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\varepsilon^2$</td>
<td>Order</td>
</tr>
<tr>
<td>12×12</td>
<td>1.00E-3</td>
<td>-</td>
</tr>
<tr>
<td>24×24</td>
<td>2.26E-4</td>
<td>2.15</td>
</tr>
<tr>
<td>48×48</td>
<td>5.62E-5</td>
<td>2.01</td>
</tr>
<tr>
<td>96×96</td>
<td>1.35E-5</td>
<td>2.06</td>
</tr>
<tr>
<td>192×192</td>
<td>3.29E-6</td>
<td>2.04</td>
</tr>
</tbody>
</table>

### Table 9: Accuracy for the problem with discontinuous coefficient ($k_0 = 10^{-6}$).

<table>
<thead>
<tr>
<th>cell</th>
<th>Random meshes</th>
<th>Kershaw meshes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\varepsilon^2$</td>
<td>Order</td>
</tr>
<tr>
<td>12×12</td>
<td>1.05E-3</td>
<td>-</td>
</tr>
<tr>
<td>24×24</td>
<td>2.41E-4</td>
<td>2.12</td>
</tr>
<tr>
<td>48×48</td>
<td>6.26E-5</td>
<td>1.95</td>
</tr>
<tr>
<td>96×96</td>
<td>1.54E-5</td>
<td>2.02</td>
</tr>
<tr>
<td>192×192</td>
<td>3.86E-6</td>
<td>2.00</td>
</tr>
</tbody>
</table>

### Table 10: Comparison of accuracy with discontinuous coefficient ($k_0 = 10^{-5}$).

<table>
<thead>
<tr>
<th>cell</th>
<th>Method 1</th>
<th>Method 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Uniform meshes</td>
<td>Random meshes</td>
</tr>
<tr>
<td></td>
<td>$\varepsilon^2$</td>
<td>Order</td>
</tr>
<tr>
<td>16×16</td>
<td>1.10E-1</td>
<td>-</td>
</tr>
<tr>
<td>32×32</td>
<td>3.98E-2</td>
<td>1.47</td>
</tr>
<tr>
<td>64×64</td>
<td>1.42E-2</td>
<td>1.49</td>
</tr>
<tr>
<td>128×128</td>
<td>5.02E-3</td>
<td>1.50</td>
</tr>
</tbody>
</table>

### 6.3 The problem with anisotropic diffusion tensor

Consider the problem (2.1)-(2.2) with Dirichlet boundary condition on $\Omega = [0,1] \times [0,1]$, and take $v = (-1,1)^T$. The exact solution is

$$ u(x,y) = \cos\left(\frac{\pi x}{2}\right)e^y, $$

and the diffusive coefficient is

$$ \kappa = \begin{pmatrix} 10k_0 & 0 \\ 0 & 0.1k_0 \end{pmatrix}. $$
We test the problem on two kinds of triangular meshes (see Fig. 5). The numerical results with different parameters \(k_0 = 1, 10^{-2}, 10^{-4}, 10^{-6}\) are given in Tables 11, 12, 13 and 14. Similarly, when \(k_0 = 10^{-4}, 10^{-6}\), we also obtain second order accuracy for the solution and flux. When \(k_0 = 1\), the convergence order of solution is higher than 1.5, and the convergent order of flux is higher than 1 on two kinds of triangular meshes, as number of cells is increased. These results show that our scheme is effective for triangular meshes.
Table 13: Accuracy for the problem with anisotropic coefficient \((k_0 = 10^{-4})\).

<table>
<thead>
<tr>
<th>cells</th>
<th>Random triangular meshes</th>
<th>Kershaw triangular meshes</th>
</tr>
</thead>
<tbody>
<tr>
<td>24 × 24</td>
<td>2.95E-3</td>
<td>5.31E-3</td>
</tr>
<tr>
<td>48 × 48</td>
<td>7.31E-4</td>
<td>1.31E-3</td>
</tr>
<tr>
<td>96 × 96</td>
<td>1.78E-4</td>
<td>3.21E-4</td>
</tr>
<tr>
<td>192 × 192</td>
<td>4.31E-5</td>
<td>7.60E-5</td>
</tr>
<tr>
<td>384 × 384</td>
<td>1.11E-5</td>
<td>2.09E-5</td>
</tr>
</tbody>
</table>

Table 14: Accuracy for the problem with anisotropic coefficient \((k_0 = 10^{-6})\).

<table>
<thead>
<tr>
<th>cells</th>
<th>Random triangular meshes</th>
<th>Kershaw triangular meshes</th>
</tr>
</thead>
<tbody>
<tr>
<td>24 × 24</td>
<td>2.98E-3</td>
<td>5.36E-3</td>
</tr>
<tr>
<td>48 × 48</td>
<td>7.41E-4</td>
<td>1.34E-3</td>
</tr>
<tr>
<td>96 × 96</td>
<td>1.84E-4</td>
<td>3.37E-4</td>
</tr>
<tr>
<td>192 × 192</td>
<td>4.56E-5</td>
<td>8.33E-5</td>
</tr>
<tr>
<td>384 × 384</td>
<td>1.14E-5</td>
<td>2.08E-5</td>
</tr>
</tbody>
</table>

6.4 The positivity

Consider the problem (2.1)-(2.2) with homogeneous Dirichlet boundary conditions on the unit square \(\Omega = [0,1]^2\). Take \(v = (2x, 3y)^T\) and diffusion coefficient as follows:

\[
\kappa = \begin{pmatrix}
    y^2 + \varepsilon x^2 & -(1 - \varepsilon)xy \\
    -(1 - \varepsilon)xy & \varepsilon y^2 + x^2
\end{pmatrix}, \quad \varepsilon = 5 \times 10^{-3}.
\]

The source term is

\[
f = \begin{cases}
    1, & \text{if } (x, y) \in [3/8, 5/8]^2, \\
    0, & \text{others}.
\end{cases}
\]

The exact solution \(u(x, y)\) is unknown, but the maximum principle states that it is non-negative. This problem is tested on random quadrilateral meshes with 128 × 128 cells, and the corresponding distribution of these meshes is similarly shown in Fig. 6. We give the numerical solutions in Fig. 7. The minimum value is 0 and the maximum value is 4.64E − 2, which show that our scheme preserves the positivity of the solution and does not produce any non-physical oscillations.
Figure 6: Random quadrilateral meshes.

Figure 7: The numerical solutions on the random quadrilateral meshes, $u_{\text{min}} = 0, u_{\text{max}} = 4.64E - 2$.

7 Conclusions

A nonlinear monotone finite volume scheme for the convection-diffusion equation is presented on convex polygonal meshes. The discretization of diffusive term is based on the method in [33]. For the discretization of convective term, we propose a corrected upwind method. The main idea is that we use the information of the discretization for diffusive flux in the discretization of convective term, and introduce the Taylor series expansion to achieve the second order accuracy, and some auxiliary unknowns are used to compute the gradient. All information in the discretization of convective term comes from the up-stream cell in order to keep monotonicity.

Our scheme includes only cell-centered unknowns, and rigorously treats material discontinuities. Numerical experiments show that our scheme obtains good accuracy for the both of diffusion-dominated and convection-dominated problems.
Acknowledgments

This work is supported by Science Challenge Project (No. TZ2016002), the National Natural Science Foundation of China (11571047, 11571048, 11772165, 11361045), NSAF (U1630249) and the Foundation of CEAP (2015B0202042).

References

Fluids, 70 (2012), 1188-1205.


