Journal of Computational Mathematics
Vol.xx, No.x, 2025, 1–27.

http://www.global-sci.org/jcm
doi:10.4208/jcm.2502-m2024-0081

A FINITE VOLUME SCHEME PRESERVING STRONG EXTREMUM PRINCIPLE FOR THREE-DIMENSIONAL DIFFUSION EQUATIONS AND ITS ANDERSON ACCELERATION*

Fei Zhao
College of Science, North China University of Technology, Beijing 100144, China
Email: zhaofei2020@ncut.edu.cn
Zhiqiang Sheng
Laboratory of Computational Physics, Institute of Applied Physics and Computational Mathematics, P.O. Box 8009, Beijing 100088, China;
HEDPS, Center for Applied Physics and Technology, and College of Engineering, Peking University, Beijing 100871, China
Email: sheng_zhiqiang@iapcm.ac.cn
Guangwei Yuan[1)]
Laboratory of Computational Physics, Institute of Applied Physics and Computational Mathematics, P.O. Box 8009, Beijing 100088, China
Email: yuan_guangwei@iapcm.ac.cn

Abstract

In this paper, we introduce a nonlinear finite volume scheme preserving discrete strong extremum principle (DSEP) for diffusion equations on tetrahedral meshes. In the construction of our nonlinear scheme, the key is to reformulate a discrete normal flux with local extremum principle structure, which is based on a modification of a second order linear scheme. In the construction of existing cell-centered finite volume schemes that maintain the discrete maximum principle, it is required to represent auxiliary unknowns as convex combinations of primary unknowns, which results in strong constraints on the smoothness of the mesh and diffusion coefficient. By contrast, our new scheme avoids this kind of constraints. Moreover, we will prove that there holds the DSEP for any solution of our scheme and there exists at least one solution preserving DSEP for our scheme. Furthermore, a modified Picard iteration with the Anderson acceleration (mP-AA) for solving the nonlinear scheme is proposed, and the nonlinear convergence of the modified Picard iteration is also proved. Finally, numerical examples are presented to show that the new scheme preserves DSEP and obtains second order accuracy, as well as the mP-AA method is effective.

Mathematics subject classification: 65M08, 65M12, 65B99.
Key words: Diffusion equation, Extremum principle, Tetrahedral meshes, Anderson acceleration, Existence.

1. Introduction

The strong extremum principle is an important property of diffusion equation (see [13]). In the content of heat conduction problems, this property is closely related with the second law

* Received April 14, 2024 / Revised version received September 12, 2024 / Accepted February 17, 2025 / Published online April 10, 2025 /
[1)] Corresponding author

of thermodynamics, which specifies the direction of heat flow. If a discrete scheme satisfies the discrete strong extremum principle, then it ensures that the direction of the discrete heat flow is correct, and thereby avoids non-physical numerical oscillations appearing in the numerical solution. Therefore, for the numerical solution of diffusion problems, we hope to design a discrete scheme that has good numerical approximation properties, while maintaining the main physical characteristics, e.g., conservativeness, extremum principles. Some cell-centered finite volume schemes preserving DSEP have been constructed by designing nonlinear discrete normal flux with local extremum principle (LEP) structure (see, e.g., [3, 6–8, 12, 22, 23] and the references therein).

Another approach of designing finite volume schemes preserving DSEP is based on certain nonlinear correction of linear second order schemes ([5,24,31,33]). Cancès *et al.* [5] introduced a nonlinear correction method for existing linear schemes so that some important properties can be satisfied. However, numerical results show that the order of convergence of those corrected schemes degrades to first order, even though the original scheme is second order accurate. In order to avoid negative numerical solution caused by the sink term for the difference method of the two-dimensional (2D) Poisson equation, in [20] a non-linearization technique was devised by rewriting the sink term of the Poisson equation as a nonlinear formulation. In [4], this nonlinear technique is also applied to the numerical discretization of a class of stiff ordinary differential equations, and obtained a high-order positivity-preserving scheme. The nonlinear correction method is also applied to get a scheme satisfying the discrete maximum principle unconditionally for 2D diffusion problems on arbitrary polygonal meshes in [24, 31], while the existence of solution for both of nonlinear schemes has not been addressed. Moreover, two algebraic approaches to designing diffusion schemes preserving DSEP are proposed by introducing two different limiters to the so-called nine-point scheme (often named as diamond scheme) in [33] and [34] respectively.

In order to accurately distinguish discontinuities, it is necessary to introduce auxiliary unknowns on each cell to construct discrete gradients, and these auxiliary quantities should be represented by primary unknowns to reduce computational complexity. In order to ensure that the obtained discrete scheme satisfies DSEP, it is often necessary for this representation to be a convex combination in existing construction methods. However, obtaining the convex combination representation is very difficult, especially in the case of numerically solving multi-material problems on three-dimensional (3D) general polyhedral meshes. Moreover, this has also led to more stringent restrictions on the regularities of the mesh-cell geometry and diffusion coefficients, resulting in poor robustness of the scheme and its inability to be well suited for numerical simulations of Lagrangian radiation hydrodynamics coupling problems.

The commonly used method of solving nonlinear schemes is Picard iteration due to its advantages of easy implementation. However, the convergence rate of Picard iteration is only first order and the calculation is time-consuming especially for 3D problems. So certain accelerated iteration convergence method should be applied to improve computational efficiency. The Anderson acceleration (AA) method is an algorithm that accelerates the convergence of fixed-point iterations, including Picard method, which was first proposed by Anderson [2] to solve a class of nonlinear integral equations. In [10] two classes of multisecant methods for nonlinear acceleration were discussed: One class is the Broyden-like, while AA method is a special case among them, and another class is the nonlinear Eirola-Nevanlinna-type. The convergence of AA method is analyzed in [9, 25]. Lipnikov *et al.* [17] apply the AA method to the iterative solution of a positivity-preserving finite volume scheme for convection-diffusion equation,

and realize the fast solution of the nonlinear scheme. Up to now, AA method has been used successfully in many fields to accelerate the nonlinear iteration, such as electronic-structure computations [32], neutron transport systems [30], variably saturated flow modeling [18], thermal coupled fluid-solid problems [11], three-temperature energy equations [1], incompressible Newtonian fluid [21], and so on.

In this paper, we develop a cell-centered finite volume scheme, which preserves DSEP and second order accuracy, for diffusion problems with strong anisotropic and discontinuous coefficients on general tetrahedral meshes. The design procedure of the new scheme is as follows. First, a discrete normal flux on one-side of cell-face is constructed based on a nonlinear correction of a linear scheme in [14] to obtain a one-sided flux with LEP structure, then, the harmonic averaging technique is used to obtain the final conservative flux on each cell-face. Compared with these DSEP preserving schemes in [3, 6–8, 12, 22, 23], the first novelty of our scheme is that it is unnecessary to impose severe restrictions on regularities of diffusion coefficients and mesh-cell geometries, since we have successfully avoided the requirement of introducing convex combinations that represent auxiliary unknowns as primary unknowns. We need only an interpolation method with second order accuracy for auxiliary unknowns. As a result, our scheme is appropriate for solving general diffusion coefficient problems on arbitrary tetrahedral meshes. Compared with that in [5], our scheme is also a kind of nonlinear correction of a linear scheme with second order accuracy, and its advantage lies in not only maintaining DSEP, but also keeping second order accuracy. Moreover, another novelty appears in the proof of existence of a solution for our nonlinear scheme, i.e., a new map is introduced and proved to be continuous in the fixed-point argument, which is different from that used in [8,34]. It is interesting that the standard Picard iteration cannot be used directly for solving our scheme, thus a new nonlinear iteration method is designed based on the new map, which is a modification of the standard Picard iteration. The third novelty is that we prove the convergence of the modified Picard iteration for solving the nonlinear scheme. And the fourth novelty of this article is the fast solution of the 3D DSEP scheme by adopting the AA method, i.e., a modified Picard iteration with the Anderson acceleration for solving our nonlinear scheme is proposed. In numerical experiments, several test examples are presented to compare the computational efficiency of the modified Picard iteration and mP-AA method, and the numerical results show that the mP-AA method is more efficient and stable.

The remainder of this paper is organized as follows. In Section 2, the problem and notations are given. In Section 3, the nonlinear finite volume scheme is established and the DSEP is given. The existence of numerical solution is proved in Section 4. The modified Picard iteration and mP-AA method are represented in Section 5. Then in Section 6 some numerical tests are presented to show that the scheme not only satisfies DSEP, but also obtains second order accuracy, and the efficiency of mP-AA method is also demonstrated. Section 7 closes this paper with conclusions.

2. Problem and Notations

In this paper, we consider a cell-centered finite volume scheme for the following diffusion equation with unknown $\breve{u}(\boldsymbol{x})$:

$$-\nabla\cdot(\boldsymbol{\kappa}\nabla\breve{u}) = f \quad \text{in } \Omega, \tag{2.1}$$

$$\breve{u} = g \quad \text{on } \partial\Omega, \tag{2.2}$$

where Ω is a bounded polyhedral domain in $\mathbb{R}^3$ with boundary $\partial\Omega$, and $f(\boldsymbol{x})$ is a right-hand term, $g(\boldsymbol{x})$ is a given boundary value function. $\boldsymbol{\kappa} = (\kappa_{ij}(\boldsymbol{x}))$ is a diffusion tensor, which is V-elliptic, i.e., there exist constants $0 < \lambda \leq \Lambda$ such that

$$\lambda|\boldsymbol{\xi}|^2 \leq \kappa_{ij}(\boldsymbol{x})\xi_i\xi_j \leq \Lambda|\boldsymbol{\xi}|^2$$

for a.e. $\boldsymbol{x} \in \Omega$ and all $\boldsymbol{\xi} \in \mathbb{R}^3$.

We use a mesh on Ω made up of tetrahedron cells. Denote the cells, as well as their cell-centers, by $K, L, \ldots$, the vertices by $A, B, \ldots$, the cell face by S. If the cell face S is the common face of cells K and L with vertices A, B, C, we denote $S = K|L = ABC$, see Fig. 2.1.

Let the set of all cells, vertices and faces be $\mathcal{J}, \mathcal{N}, \mathcal{E}$, respectively. And let $\mathcal{J}_{in}$ be the set of cell center which is in Ω, $\mathcal{J}_{out}$ be the set of face center which is on $\partial\Omega$. In addition, let $\mathcal{E}_K$ be the set of all faces of cell K. Denote $\mathcal{E}_{int} = \mathcal{E} \cap \Omega$ and $\mathcal{E}_{ext} = \mathcal{E} \cap \partial\Omega$. Denote $h = \sup_{K\in\mathcal{J}}\{\text{diam } K\}$, where diam$K$ is the diameter of cell K. Let $\vec{n}_{KS}$ (respectively $\vec{n}_{LS}$) be the unit outer normal vector on the face S of cell K (respectively L), and there holds $\vec{n}_{KS} = -\vec{n}_{LS}$ for $S = K|L$.

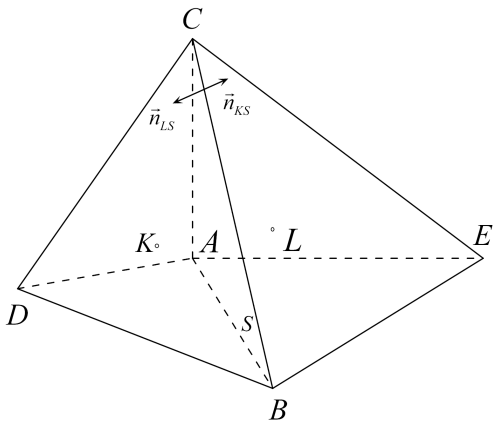


Fig. 2.1. Notations related to two adjacent tetrahedron cells.

3. Construction of the Scheme

In this section, we will introduce the construction of the new scheme preserving DSEP for stationary diffusion equations (2.1) and (2.2). Firstly, starting from a linear scheme with second order accuracy in [14], the linear discrete normal flux is nonlinearly reconstructed to obtain a single-sided flux with LEP structure. Then, the conservative flux is obtained by nonlinear harmonic averaging of the two single-sided fluxes, and finite volume scheme preserving DSEP is obtained.

3.1. The linear flux

Integrate (2.1) over cell K and use the divergence theorem to obtain

$$\sum_{S\in\mathcal{E}_K} \mathcal{F}_{K,S} = \int_K f(\boldsymbol{x})\mathrm{d}x, \tag{3.1}$$

where

$$\mathcal{F}_{K,S} = -\int_S \nabla \breve{u} \cdot \boldsymbol{\kappa}^\top \vec{n}_{KS} \mathrm{d}s. \tag{3.2}$$

Suppose the ray originated at the point K (respectively L) along the direction $\boldsymbol{\kappa}^\top(K)\vec{n}_{KS}$ (respectively $\boldsymbol{\kappa}^\top(L)\vec{n}_{LS}$) intersects the plan where the common face $S = ABC$ lies, and the intersection is K' (respectively L'), as shown in Fig. 3.1. Followed the construction of the scheme in [14], the discrete normal flux is given by

$$\begin{aligned} F_{K,S} = & \frac{\left|\boldsymbol{\kappa}^\top(K)\vec{n}_{KS}\right| \left|\boldsymbol{\kappa}^\top(L)\vec{n}_{LS}\right| |S|}{\left|\boldsymbol{\kappa}^\top(K)\vec{n}_{KS}\right| \left|\overrightarrow{LL'}\right| + \left|\boldsymbol{\kappa}^\top(L)\vec{n}_{LS}\right| \left|\overrightarrow{KK'}\right|} \\ & \times \left[u_K - u_L - \frac{(u_A\vec{n}_{CB} + u_B\vec{n}_{AC} + u_C\vec{n}_{BA}) \cdot \overrightarrow{K'L'}}{2|S|}\right], \end{aligned} \tag{3.3}$$

where $\vec{n}_{CB} = \vec{n} \times \overrightarrow{CB}, \vec{n}_{AC} = \vec{n} \times \overrightarrow{AC}, \vec{n}_{BA} = \vec{n} \times \overrightarrow{BA}$, and $\vec{n}$ is the unit normal vector to the face ΔABC.

Note that the scheme in [14] does not satisfy DSEP, which can refer to Section 6.1 of this paper for details.

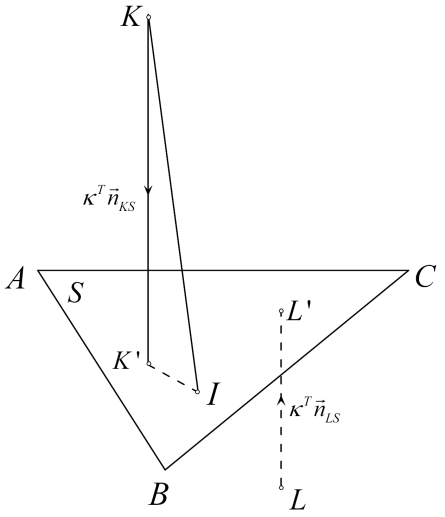


Fig. 3.1. The discretization stencil of flux.

3.2. The reconstructed flux with LEP structure

Let $\mathcal{N}(K)$ denote the set of cells around cell K, i.e., $\mathcal{N}(K) = \{L \in \mathcal{J} | \bar{K} \cap \bar{L} \neq \emptyset\}$. And choose $K_{\min} \in \mathcal{N}(K)$ such that $u_{K_{\min}} = \min_{L\in\mathcal{N}(K)} u_L$. It should be noted that the local minimum value may be achieved at several cells. At this time, if $u_K = \min_{L\in\mathcal{N}(K)} u_L$, then we set $K_{\min} = K$. Similarly, choose $K_{\max} \in \mathcal{N}(K)$ such that $u_{K_{\max}} = \max_{L\in\mathcal{N}(K)} u_L$, especially, if $u_K = \max_{L\in\mathcal{N}(K)} u_L$, then we set $K_{\max} = K$. Now by modifying (3.3), we can obtain the single-sided flux with LEP structure as follows:

$$\begin{aligned} F_{K,S} &= \tau_S(u_K - u_L + D_S) \\ &= \tau_S\big(u_K - u_L + D_S^+ - D_S^-\big) \end{aligned}$$

$$
\begin{aligned}
&= \tau_S \Bigg[u_K - u_L + \frac{D_S^+}{u_K - u_{K_{\min}} + s_0}(u_K - u_{K_{\min}} + s_0) \\
&\qquad - \frac{D_S^-}{u_K - u_{K_{\max}} - s_0}(u_K - u_{K_{\max}} - s_0) \Bigg] \\
&\approx \tau_S \Bigg[u_K - u_L + \frac{D_S^+}{u_K - u_{K_{\min}} + s_0}(u_K - u_{K_{\min}}) \\
&\qquad + \frac{D_S^-}{u_{K_{\max}} - u_K + s_0}(u_K - u_{K_{\max}}) \Bigg] \\
&= \tau_S(u_K - u_L) + \frac{\tau_S D_S^+}{u_K - u_{K_{\min}} + s_0}(u_K - u_{K_{\min}}) \\
&\quad + \frac{\tau_S D_S^-}{u_{K_{\max}} - u_K + s_0}(u_K - u_{K_{\max}}),
\end{aligned} \tag{3.4}
$$

where s_0 is a small positive constant to be determined later, and

$$
\begin{aligned}
\tau_S &= \frac{\left|\boldsymbol{\kappa}^\top(K)\vec{n}_{KS}\right| \left|\boldsymbol{\kappa}^\top(L)\vec{n}_{LS}\right| |S|}{\left|\boldsymbol{\kappa}^\top(K)\vec{n}_{KS}\right| \left|\overrightarrow{LL'}\right| + \left|\boldsymbol{\kappa}^\top(L)\vec{n}_{LS}\right| \left|\overrightarrow{KK'}\right|} > 0, \\
D_S &= -\frac{(u_A \cdot \vec{n}_{CB} + u_B \cdot \vec{n}_{AC} + u_C \cdot \vec{n}_{BA}) \cdot \overrightarrow{K'L'}}{2|S|}, \\
D_S^+ &= \frac{|D_S| + D_S}{2}, \quad D_S^- = \frac{|D_S| - D_S}{2}.
\end{aligned}
$$

It is obvious that $\tau_S = \mathcal{O}(h), D_S = \mathcal{O}(1)$.

Remark 3.1. When the constant s_0 satisfies certain conditions, the normal flux has second order accuracy. Suppose $D_S > 0$, then $D_S^+ > 0, D_S^- = 0$; moreover, if $u_{K_{\min}} < u_K < u_{K_{\max}}$, then the normal flux above can be written as

$$
F_{K,S} = \tau_S(u_K - u_L) + \frac{\tau_S D_S^+}{u_K - u_{K_{\min}} + s_0}(u_K - u_{K_{\min}}) + \frac{\tau_S D_S^+}{u_K - u_{K_{\min}} + s_0} s_0.
$$

Notice that $\tau_S D_S^+ = \mathcal{O}(h)$. In order to obtain second order accuracy, the last term at the right hand of the formula above needs to satisfy

$$
\frac{\tau_S D_S^+}{u_K - u_{K_{min}} + s_0} s_0 \le Ch^3,
$$

that is to say $s_0 = \mathcal{O}(h^2)$, where C is a positive constant. In numerical experiments, we take $s_0 = 10^{-5}$.

Notice that $F_{L,S} = \tau_S(u_L - u_K - D_S)$, then we similarly redefine it as

$$
F_{L,S} = \tau_S(u_L - u_K) + \frac{\tau_S D_S^+}{u_{L_{\max}} - u_L + s_0}(u_L - u_{L_{\max}}) + \frac{\tau_S D_S^-}{u_L - u_{L_{\min}} + s_0}(u_L - u_{L_{\min}}).
$$

Let $\boldsymbol{u} = (u_K)_{K \in \mathcal{J}_{in}}$ be the discrete unknown vector. So for $S = K|L \in \mathcal{E}_{int}$, there are

$$
F_{K,S} = \tau_S(u_K - u_L) + f_{K,S}(\boldsymbol{u}), \tag{3.5}
$$

$$
F_{L,S} = \tau_S(u_L - u_K) + f_{L,S}(\boldsymbol{u}), \tag{3.6}
$$

where

$$f_{K,S}(\boldsymbol{u}) = \frac{\tau_S D_S^+}{u_K - u_{K_{\min}} + s_0}(u_K - u_{K_{\min}}) + \frac{\tau_S D_S^-}{u_{K_{\max}} - u_K + s_0}(u_K - u_{K_{\max}}),$$
$$f_{L,S}(\boldsymbol{u}) = \frac{\tau_S D_S^+}{u_{L_{\max}} - u_L + s_0}(u_L - u_{L_{\max}}) + \frac{\tau_S D_S^-}{u_L - u_{L_{\min}} + s_0}(u_L - u_{L_{\min}}).$$

Denote

$$K_m = \begin{cases} K_{\min}, & D_S \geq 0, \\ K_{\max}, & D_S < 0, \end{cases} \quad L_m = \begin{cases} L_{\max}, & D_S \geq 0, \\ L_{\min}, & D_S < 0. \end{cases}$$

So we can write $f_{K,S}(\boldsymbol{u})$ and $f_{L,S}(\boldsymbol{u})$ as

$$f_{K,S}(\boldsymbol{u}) = g_{K,S}(\boldsymbol{u})(u_K - u_{K_m}),$$
$$f_{L,S}(\boldsymbol{u}) = g_{L,S}(\boldsymbol{u})(u_L - u_{L_m}),$$

where

$$g_{K,S}(\boldsymbol{u}) = \begin{cases} \dfrac{\tau_S D_S^+}{u_K - u_{K_m} + s_0}, & D_S \geq 0, \\ \dfrac{\tau_S D_S^-}{u_{K_m} - u_K + s_0}, & D_S < 0, \end{cases} \quad g_{L,S}(\boldsymbol{u}) = \begin{cases} \dfrac{\tau_S D_S^+}{u_{L_m} - u_L + s_0}, & D_S \geq 0, \\ \dfrac{\tau_S D_S^-}{u_L - u_{L_m} + s_0}, & D_S < 0. \end{cases}$$

Notice that regardless of whether $D_S \geq 0$ or $D_S < 0$, there holds $g_{K,S}(\boldsymbol{u}) \geq 0$.

For $S \in \mathcal{E}_{ext}$, we assume that the ray originated at the point K along the direction $\boldsymbol{\kappa}^\top(K)\vec{n}_{KS}$ must intersect the face $\tilde{S} \in \mathcal{E}_{ext}$ without crossing the discontinuity, and the intersection is $K' \in \partial\Omega$. Then we have

$$F_{K,S} = -\frac{\left|\boldsymbol{\kappa}^\top(K)\vec{n}_{KS}\right||S|}{\left|\overrightarrow{KK'}\right|}(u_{K'} - u_K),$$

where $u_{K'} = g(K')$.

3.3. The conservative flux

It can be seen from the previous subsection that the single-sided flux satisfies the LEP structure, but has no local conservation. So in this subsection, we use a nonlinear harmonic average of $f_{K,S}(\boldsymbol{u})$ and $f_{L,S}(\boldsymbol{u})$ to obtain the conservative flux with LEP structure.

Now we use the nonlinear harmonic averaging technique [16, 22] to obtain the conservative fluxes. If $|f_{K,S}(\boldsymbol{u})| + |f_{L,S}(\boldsymbol{u})| \neq 0$, there are

$$F_{K,S} = \tau_S(u_K - u_L) + \frac{|f_{L,S}(\boldsymbol{u})|}{|f_{K,S}(\boldsymbol{u})| + |f_{L,S}(\boldsymbol{u})|}f_{K,S}(\boldsymbol{u}) - \frac{|f_{K,S}(\boldsymbol{u})|}{|f_{K,S}(\boldsymbol{u})| + |f_{L,S}(\boldsymbol{u})|}f_{L,S}(\boldsymbol{u}),$$
$$F_{L,S} = \tau_S(u_L - u_K) + \frac{|f_{K,S}(\boldsymbol{u})|}{|f_{K,S}(\boldsymbol{u})| + |f_{L,S}(\boldsymbol{u})|}f_{L,S}(\boldsymbol{u}) - \frac{|f_{L,S}(\boldsymbol{u})|}{|f_{K,S}(\boldsymbol{u})| + |f_{L,S}(\boldsymbol{u})|}f_{K,S}(\boldsymbol{u}).$$

Moreover, if $|f_{K,S}(\boldsymbol{u})|+|f_{L,S}(\boldsymbol{u})| = 0$, we take $F_{K,S} = \tau_S(u_K - u_L)$, and define $F_{L,S}$ analogously. It is easy to see $F_{K,S} + F_{L,S} = 0$.

Notice that $g_{K,S}(\boldsymbol{u}) \geq 0, g_{L,S}(\boldsymbol{u}) \geq 0$. In the case of $D_S \geq 0$, there are $u_{K_m} = u_{K_{\min}}$ and $u_{L_m} = u_{L_{\max}}$. Therefore, we have $f_{K,S}(\boldsymbol{u})f_{L,S}(\boldsymbol{u}) \leq 0$. In the case of $D_S < 0$, there are $u_{K_m} = u_{K_{\max}}$ and $u_{L_m} = u_{L_{\min}}$. We also have $f_{K,S}(\boldsymbol{u})f_{L,S}(\boldsymbol{u}) \leq 0$. So there is

$$|f_{L,S}(\boldsymbol{u})|f_{K,S}(\boldsymbol{u}) = -|f_{K,S}(\boldsymbol{u})|f_{L,S}(\boldsymbol{u}),$$

as a result we can obtain

$$F_{K,S} = \tau_S(u_K - u_L) + h_{K,S}(\boldsymbol{u})(u_K - u_{K_m}), \tag{3.7}$$

$$F_{L,S} = \tau_S(u_L - u_K) + h_{L,S}(\boldsymbol{u})(u_L - u_{L_m}), \tag{3.8}$$

where

$$h_{K,S}(\boldsymbol{u}) = \begin{cases} \dfrac{2|f_{L,S}(\boldsymbol{u})|}{|f_{K,S}(\boldsymbol{u})| + |f_{L,S}(\boldsymbol{u})|} g_{K,S}(\boldsymbol{u}), & \text{if} \quad f_{K,S}(\boldsymbol{u}) f_{L,S}(\boldsymbol{u}) < 0, \\ 0, & \text{if} \quad f_{K,S}(\boldsymbol{u}) f_{L,S}(\boldsymbol{u}) \geq 0, \end{cases}$$

$$h_{L,S}(\boldsymbol{u}) = \begin{cases} \dfrac{2|f_{K,S}(\boldsymbol{u})|}{|f_{K,S}(\boldsymbol{u})| + |f_{L,S}(\boldsymbol{u})|} g_{L,S}(\boldsymbol{u}), & \text{if} \quad f_{K,S}(\boldsymbol{u}) f_{L,S}(\boldsymbol{u}) < 0, \\ 0, & \text{if} \quad f_{K,S}(\boldsymbol{u}) f_{L,S}(\boldsymbol{u}) \geq 0. \end{cases}$$

3.4. The approximation of auxiliary unknowns

In existing literatures (such as [3, 6, 8, 12, 22, 23]), in order to satisfy DSEP, auxiliary unknowns need to be represented as a convex combination of primary unknowns around with at least second order accuracy. To the best knowledge of authors, it is difficult to achieve such a convex combination unless strict restrictive conditions are imposed on regularities of diffusion coefficients and mesh-cell geometries. In the discretization of our diffusion flux in (3.4), the cell-centered unknowns are the primary unknowns, moreover, we also use the vertex unknowns as auxiliary ones. Being different from those in [3, 6, 8, 12, 22, 23], the vertex unknowns appear only in the nonlinear coefficients of diffusion flux expression defined in (3.4). Thus, we need only to express vertex unknowns as certain combination of neighboring cell-centered unknowns with second order accuracy, and the requirement for convex combination is avoided. Here we adopt the interpolation method in [28] of calculating auxiliary unknowns, and any other interpolation method also can be applied provided that it is a second order accurate one.

Remark 3.2. With the method in [28], we can get the explicit expression of vertex unknowns, which saves computational time compared with the methods in [15, 19]. Though the vertex unknowns can not be always represented as certain convex combinations of neighboring cell-centered unknowns by using the existing interpolation method, it does not affect the property preserving DSEP of our scheme, which is one of the advantages of our scheme.

3.5. The finite volume scheme preserving DSEP

With the definition of $F_{K,S}$ in (3.7), the finite volume scheme is constructed as follows:

$$\sum_{S \in \mathcal{E}_K} F_{K,S} = f_K m(K), \quad \forall\, K \in \mathcal{J}_{in}, \tag{3.9}$$

$$u_M = g(M), \qquad \forall\, M \in \mathcal{J}_{out}, \tag{3.10}$$

where $f_K = f(K), m(K)$ is the volume of cell K, and $F_{K,S}$ is obtained in Sections 3.2 and 3.3, when $S = K|L \in \mathcal{E}_{int}$, we have

$$F_{K,S} = \tau_S(u_K - u_L) + h_{K,S}(\boldsymbol{u})(u_K - u_{K_m}). \tag{3.11}$$

When $S \in \mathcal{E}_{ext}$, we have

$$F_{K,S} = \lambda_{K,S}(u_K - u_{K'}), \tag{3.12}$$

and

$$\lambda_{K,S} = \frac{\left|\boldsymbol{\kappa}^{\top}(K)\vec{n}_{KS}\right||S|}{\left|\overrightarrow{KK'}\right|} > 0, \quad u_{K'} = g(K').$$

Since the coefficients in (3.11) satisfy that $\tau_S > 0$ and $h_{K,S}(\boldsymbol{u}) \geq 0$, it is easy to see that the second law of thermodynamics holds for the solution of (3.9)-(3.10), i.e., the heats flow from high temperature to low one. Now we can show that the following DSEP holds, which implies that non-physical numerical oscillation could not appear in the solution of (3.9)-(3.10).

Theorem 3.1. *(i) If $f \geq 0$, then the nonlinear scheme* (3.9)-(3.10) *satisfies the discrete strong minimum principle. That is, for any discrete solution $\{u_K\}_{K\in\mathcal{J}}$ of* (3.9)-(3.10)*, if there exists $K_0 \in \mathcal{J}_{in}$ such that $u_{K_0} = \min_{K\in\mathcal{J}} u_K$, then $u_K = u_{K_0}$ for any $K \in \mathcal{J}$.*

(ii) If $f \leq 0$, then the nonlinear scheme (3.9)-(3.10) *satisfies the discrete strong maximum principle. That is, for any discrete solution $\{u_K\}_{K\in\mathcal{J}}$ of* (3.9)-(3.10)*, if there exists $K_0 \in \mathcal{J}_{in}$ such that $u_{K_0} = \max_{K\in\mathcal{J}} u_K$, then $u_K = u_{K_0}$ for any $K \in \mathcal{J}$.*

Proof. First, we consider the case (i) $f \geq 0$. Assume there exists $K_0 \in \mathcal{J}_{in}$ such that $u_{K_0} = \min_{K\in\mathcal{J}} u_K$, then there are $u_{K_0} - u_K \leq 0$ for all $K \in \mathcal{J}$. Hence, there is

$$\sum_{S\in\mathcal{E}_{K_0}\cap\mathcal{E}_{int}} [\tau_S(u_{K_0} - u_L) + h_{K_0,S}(\boldsymbol{u})(u_{K_0} - u_{K_{0m}})] + \sum_{S\in\mathcal{E}_{K_0}\cap\mathcal{E}_{ext}} \lambda_{K_0,S}(u_{K_0} - u_{K_0'}) \leq 0,$$

i.e., $\sum_{S\in\mathcal{E}_{K_0}} F_{K_0,S} \leq 0$. However, there holds $f_{K_0}m(K_0) \geq 0$. Noticing (3.9) and $\lambda_{K_0,S} > 0$, $h_{K_0,S}(\boldsymbol{u}) \geq 0$, we get a contradiction unless $u_K = u_{K_0}$ for any $K \in \mathcal{J}$.

Then, the case (ii) $f \leq 0$ can be proved as the case (i). Hence, we finish the proof. □

4. Existence

In order to prove the existence of a solution for the scheme, we need some lemmas [27].

Definition 4.1. *An $n \times n$ matrix $\boldsymbol{A}$ is reducible if there exists an $n \times n$ permutation matrix $\boldsymbol{P}$ such that*

$$\boldsymbol{P}^T\boldsymbol{A}\boldsymbol{P} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{0} & \boldsymbol{A}_{22} \end{pmatrix},$$

where $\boldsymbol{A}_{11}, \boldsymbol{A}_{12}, \boldsymbol{A}_{22}$ are respectively $r \times r$, $r \times (n-r)$ and $(n-r) \times (n-r)$ sub-matrixes with $1 \leq r \leq n$. If no such permutation matrix exists, then $\boldsymbol{A}$ is irreducible.

Lemma 4.1. *To any matrix $\boldsymbol{A} = (a_{ij})_{n\times n}$, we associate the graph of nodes $1, 2, \ldots, n$ and of directed edges connecting x_i to x_j if $a_{ij} \neq 0$. Then $\boldsymbol{A}$ is irreducible if and only if for any pair $i \neq j$ there exists a chain of edges that allows to go from x_i to x_j,*

$$a_{i,k_1} \neq 0 \ \rightarrow \ a_{k_1,k_2} \neq 0 \ \rightarrow \ \cdots \ \rightarrow \ a_{k_m,j} \neq 0.$$

Definition 4.2. *A matrix $\boldsymbol{A} = (a_{ij})_{n\times n}$ is weak diagonal dominance in rows, if it satisfies*

$$|a_{ii}| \geq \sum_{j=1,\, j\neq i}^{n} |a_{ij}|, \quad i = 1, 2, \ldots, n$$

with strict inequality at least one of the above inequalities. Moreover, if the above inequality is strict for all i, we say that $\boldsymbol{A}$ is strictly diagonally dominant.

Lemma 4.2. *If $\boldsymbol{A} = (a_{ij})_{n\times n}$ is an irreducible weakly diagonally dominant matrix, then it is invertible.*

We now give the existence of a solution for the nonlinear scheme.

Theorem 4.1. *If $f = 0$, then the nonlinear scheme* (3.9)-(3.10) *has a solution.*

Proof. Let $\tilde{g} = \min_{M\in\mathcal{J}_{out}} g(M), \hat{g} = \max_{M\in\mathcal{J}_{out}} g(M)$. Define a closed convex set in $\mathbb{R}^J$ (J denotes the number of cells in $\mathcal{J}_{in}$) as

$$\mathcal{C} = \big\{\boldsymbol{u} = (u_K)_{K\in\mathcal{J}_{in}} \in \mathbb{R}^J \mid \tilde{g} \le u_K \le \hat{g},\ \forall\, K \in \mathcal{J}_{in}\big\}.$$

And define the map $\mathcal{T} : \mathcal{C} \to \mathbb{R}^J$ as follows: For any $\bar{\boldsymbol{u}} \in \mathcal{C}$, $\boldsymbol{u} = \mathcal{T}\bar{\boldsymbol{u}}$ is the solution of the following linear scheme:

$$\sum_{S\in\mathcal{E}_K} F_{K,S}(\boldsymbol{u}, \bar{\boldsymbol{u}})|S| = 0, \quad \forall\, K \in \mathcal{J}_{in},$$

where for $S = K|L \in \mathcal{E}_{int}$, define

$$F_{K,S}(\boldsymbol{u}, \bar{\boldsymbol{u}}) = \tau_S(u_K - u_L) + h_{K,S}(\bar{\boldsymbol{u}})(u_K - \bar{u}_{Km}),$$

for $S \in \mathcal{E}_{ext}, F_{K,S}(\boldsymbol{u}, \bar{\boldsymbol{u}})$ is defined as the same as (3.12).

Since the map $\mathcal{T}$ may be discontinuous due to the discontinuity of $h_{K,S}(\bar{\boldsymbol{u}})$ with respect to $\bar{\boldsymbol{u}}$, we consider the regularized map $\mathcal{T}_\varepsilon$ as follows: For any $\varepsilon > 0, \mathcal{T}_\varepsilon : \mathcal{C} \to \mathbb{R}^J$, where for any $\bar{\boldsymbol{u}} \in \mathcal{C}$, $\boldsymbol{u} = \mathcal{T}_\varepsilon\bar{\boldsymbol{u}}$ is defined to be the solution of the linear scheme

$$\sum_{S\in\mathcal{E}_K} F^\varepsilon_{K,S}(\boldsymbol{u}, \bar{\boldsymbol{u}})|S| = 0, \quad \forall\, K \in \mathcal{J}_{in}, \tag{4.1}$$

where for $S = K|L \in \mathcal{E}_{int}$, define

$$F^\varepsilon_{K,S}(\boldsymbol{u}, \bar{\boldsymbol{u}}) = \tau_S(u_K - u_L) + h^\varepsilon_{K,S}(\bar{\boldsymbol{u}})(u_K - \bar{u}_{Km}),$$

and

$$h^\varepsilon_{K,S}(\bar{\boldsymbol{u}}) = \begin{cases} \dfrac{2|f_{L,S}(\bar{\boldsymbol{u}})|}{|f_{K,S}(\bar{\boldsymbol{u}})| + |f_{L,S}(\bar{\boldsymbol{u}})| + \varepsilon} g_{K,S}(\bar{\boldsymbol{u}}), & \text{if} \quad f_{K,S}(\bar{\boldsymbol{u}}) f_{L,S}(\bar{\boldsymbol{u}}) < 0, \\ 0, & \text{if} \quad f_{K,S}(\bar{\boldsymbol{u}}) f_{L,S}(\bar{\boldsymbol{u}}) \ge 0. \end{cases}$$

For $S \in \mathcal{E}_{ext}, F^\varepsilon_{K,S}(\boldsymbol{u}, \bar{\boldsymbol{u}})$ is defined as the same as $F_{K,S}(\boldsymbol{u}, \bar{\boldsymbol{u}})$.

We rewrite the linear scheme (4.1) in the form of algebraic equations as

$$\boldsymbol{M}_\varepsilon(\bar{\boldsymbol{u}})\boldsymbol{u} = \boldsymbol{H}_\varepsilon(\bar{\boldsymbol{u}}), \tag{4.2}$$

where the coefficient matrix $\boldsymbol{M}_\varepsilon$ is of $J \times J$ and satisfies

$$[\boldsymbol{M}_\varepsilon(\bar{\boldsymbol{u}})]_{KL} = \begin{cases} \displaystyle\sum_{S\in\mathcal{E}_K} \big[\tau_S + h^\varepsilon_{K,S}(\bar{\boldsymbol{u}})\big], & K = L, \\ -\tau_S, & K \ne L, \quad \mathcal{E}_K \bigcap \mathcal{E}_L = S, \\ 0, & K \ne L, \quad \mathcal{E}_K \bigcap \mathcal{E}_L = \emptyset, \end{cases}$$

and the right-hand term is given as

$$\boldsymbol{H}_\varepsilon(\bar{\boldsymbol{u}}) = \left[\sum_{S\in\mathcal{E}_K} h^\varepsilon_{K,S}(\bar{\boldsymbol{u}})\bar{u}_{Km}\right]_{K\in\mathcal{J}_{in}}.$$

From Definition 4.1 and Lemma 4.1, it can be seen that $\boldsymbol{M}_\varepsilon$ is an irreducible matrix. From Definition 4.2, it can be seen that $\boldsymbol{M}_\varepsilon$ is a weakly diagonally dominant matrix. Therefore, Lemma 4.2 indicates that $\boldsymbol{M}_\varepsilon$ is invertible. So

$$\boldsymbol{u} = \mathcal{T}_\varepsilon \bar{\boldsymbol{u}} = \boldsymbol{M}_\varepsilon(\bar{\boldsymbol{u}})^{-1}\boldsymbol{H}_\varepsilon(\bar{\boldsymbol{u}}).$$

For each $\bar{\boldsymbol{u}} \in \mathcal{C}$, there is only one solution $\boldsymbol{u}$ for (4.1), So $\mathcal{T}_\varepsilon$ is well defined.

Then, we prove that $\mathcal{T}_\varepsilon$ maps $\mathcal{C}$ into itself by using the proof by contradiction. There is no harm in assuming that $\boldsymbol{u}$ takes the maximum component at a certain point $K \in \mathcal{J}_{in}$, and $u_K = B > \hat{g}$. And from (4.1), we can obtain

$$u_K = \frac{\sum_{S\in\mathcal{E}_K\cap\mathcal{E}_{int}}\left[\tau_S u_L + h^\varepsilon_{K,S}(\bar{\boldsymbol{u}})\bar{u}_{Km}\right] + \sum_{S\in\mathcal{E}_K\cap\mathcal{E}_{ext}}\lambda_{K,S}u_{K'}}{\sum_{S\in\mathcal{E}_K\cap\mathcal{E}_{int}}\left[\tau_S + h^\varepsilon_{K,S}(\bar{\boldsymbol{u}})\right] + \sum_{S\in\mathcal{E}_K\cap\mathcal{E}_{ext}}\lambda_{K,S}}, \tag{4.3}$$

where $\tau_S > 0, h^\varepsilon_{K,S}(\bar{\boldsymbol{u}}) \geq 0, \lambda_{K,S} > 0$. Due to $\bar{u}_{Km} \leq \hat{g}$, there are

$$\sum_{S\in\mathcal{E}_K\cap\mathcal{E}_{int}} h^\varepsilon_{K,S}(\bar{\boldsymbol{u}})\bar{u}_{Km} \leq \hat{g}\sum_{S\in\mathcal{E}_K\cap\mathcal{E}_{int}} h^\varepsilon_{K,S}(\bar{\boldsymbol{u}}) < B\sum_{S\in\mathcal{E}_K\cap\mathcal{E}_{int}} h^\varepsilon_{K,S}(\bar{\boldsymbol{u}}),$$
$$\sum_{S\in\mathcal{E}_K\cap\mathcal{E}_{ext}} \lambda_{K,S}u_{K'} \leq \hat{g}\sum_{S\in\mathcal{E}_K\cap\mathcal{E}_{ext}} \lambda_{K,S} < B\sum_{S\in\mathcal{E}_K\cap\mathcal{E}_{ext}} \lambda_{K,S}.$$

Because B is the maximum component of $\boldsymbol{u}, u_L \leq B$ (for any $L \in \mathcal{J}_{in}$). Thus, from (4.3), it can be obtained that $u_K < B$, which contradicts the hypothesis $u_K = B$. Similarly, it can be proven that $u_K \geq \tilde{g}$ (for any $K \in \mathcal{J}_{in}$). Therefore, $\mathcal{T}_\varepsilon$ maps $\mathcal{C}$ into itself.

Let $ML_J(\mathbb{R})$ be the general linear group over $\mathbb{R}$, which is the group of $J \times J$ invertible matrices of real number. As $\bar{\boldsymbol{u}}_n \to \bar{\boldsymbol{u}}_0$ $(n \to \infty)$ in $\mathbb{R}^J$, we know that $\boldsymbol{M}_\varepsilon(\bar{\boldsymbol{u}}_n) \to \boldsymbol{M}_\varepsilon(\bar{\boldsymbol{u}}_0)$ in $ML_J(\mathbb{R})$. As the inverse operator is continuous on $ML_J(\mathbb{R})$, we get $\boldsymbol{M}_\varepsilon(\bar{\boldsymbol{u}}_n)^{-1} \to \boldsymbol{M}_\varepsilon(\bar{\boldsymbol{u}}_0)^{-1}$ in $ML_J(\mathbb{R})$. Similarly, we get $\boldsymbol{H}_\varepsilon(\bar{\boldsymbol{u}}_n) \to \boldsymbol{H}_\varepsilon(\bar{\boldsymbol{u}}_0)$. It follows that $\mathcal{T}_\varepsilon(\bar{\boldsymbol{u}}_n) \to \mathcal{T}_\varepsilon(\bar{\boldsymbol{u}}_0)$ in $\mathbb{R}^J$, so the map $\mathcal{T}_\varepsilon$ is continuous.

By the Brouwer's fixed point theorem (see [13]), there exists a fixed point $\boldsymbol{u}_\varepsilon$ in $\mathcal{C}$, that is $\boldsymbol{u}_\varepsilon = \mathcal{T}_\varepsilon\boldsymbol{u}_\varepsilon$. So, $\boldsymbol{u}_\varepsilon$ satisfies

$$\sum_{S\in\mathcal{E}_K} F^\varepsilon_{K,S}(\boldsymbol{u}_\varepsilon, \boldsymbol{u}_\varepsilon)|S| = 0, \quad \forall\, K \in \mathcal{J}_{in}.$$

From the Weierstrass theorem, there is a subsequence $\boldsymbol{u}_{\varepsilon_i} \to \tilde{\boldsymbol{u}} \in \mathcal{C}$ $(\varepsilon_i \to 0)$. For this subsequence, when $\varepsilon_i \to 0$, we have

$$f_{K,S}(\boldsymbol{u}_{\varepsilon_i}) \;\to\; f_{K,S}(\tilde{\boldsymbol{u}}), \quad f_{L,S}(\boldsymbol{u}_{\varepsilon_i}) \;\to\; f_{L,S}(\tilde{\boldsymbol{u}}), \quad h^{\varepsilon_i}_{K,S}(\boldsymbol{u}_{\varepsilon_i}) \;\to\; h_{K,S}(\tilde{\boldsymbol{u}}).$$

Then

$$F^{\varepsilon_i}_{K,S}(\boldsymbol{u}_{\varepsilon_i}, \boldsymbol{u}_{\varepsilon_i}) \;\to\; F_{K,S}(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{u}}) = F_{K,S}(\tilde{\boldsymbol{u}}),$$
$$F^{\varepsilon_i}_{L,S}(\boldsymbol{u}_{\varepsilon_i}, \boldsymbol{u}_{\varepsilon_i}) \;\to\; F_{L,S}(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{u}}) = F_{L,S}(\tilde{\boldsymbol{u}}), \quad \varepsilon_i \;\to\; 0.$$

Therefore, $\tilde{\boldsymbol{u}}$ is the solution of (3.9)-(3.10). This concludes the proof. □

5. Discrete System

5.1. Picard iteration

The popular method for solving nonlinear DSEP scheme (3.9)-(3.10) is the following Picard iteration:

$$\begin{cases} \sum\limits_{S\in\mathcal{E}_K\cap\mathcal{E}_{int}} \Big[\tau_S\big(u_K^{v+1}-u_L^{v+1}\big)+h_{K,S}(\boldsymbol{u}^v)\big(u_K^{v+1}-u_{Km}^{v+1}\big)\Big] \\ \quad + \sum\limits_{S\in\mathcal{E}_K\cap\mathcal{E}_{ext}} \lambda_{K,S}\big(u_K^{v+1}-g(K')\big)=f_Km(K), \quad \forall\, K\in\mathcal{J}_{in}, \\ u_M=g(M), \qquad\qquad \forall\, M\in\mathcal{J}_{out}, \end{cases}$$

where v is the nonlinear iteration index. Since we can not prove that the iteration map $\boldsymbol{u}^v=\mathcal{T}\boldsymbol{u}^{v-1}$ is a well-defined continuous map, we turn to adopt the map given at the beginning of proof of Theorem 4.1, and obtain the modified Picard (m-Picard) iteration as follows:

$$\begin{cases} \sum\limits_{S\in\mathcal{E}_K\cap\mathcal{E}_{int}} \Big[\tau_S(u_K^{v+1}-u_L^{v+1})+h_{K,S}(\boldsymbol{u}^v)(u_K^{v+1}-u_{Km}^{v})\Big] \\ \quad + \sum\limits_{S\in\mathcal{E}_K\cap\mathcal{E}_{ext}} \lambda_{K,S}(u_K^{v+1}-g(K'))=f_Km(K), \quad \forall\, K\in\mathcal{J}_{in}, & (5.1a) \\ u_M=g(M), \qquad\qquad \forall\, M\in\mathcal{J}_{out}. & (5.1b) \end{cases}$$

Now we analyse the convergence of the modified Picard iteration for the DSEP scheme. We will make use of the assumptions (H1)-(H3), as well as the stability and convergence conclusions in [14, Theorems 3.1 and 3.2]. And we will need the following assumptions and lemma to prove Theorem 5.3.

Assumption 5.1. There exist $v>0$ and $\epsilon>0$ such that

$$\max\left(\max_{K\in\mathcal{J}_{in}}\left|\frac{u_K^{v+1}-u_K^v}{u_K^v-u_{Km}^v+s_0}\right|,\ \max_{A\in\mathcal{N}}\left|\frac{u_A^{v+1}-u_A^v}{u_K^v-u_{Km}^v+s_0}\right|\right)\le\epsilon.$$

Remark 5.1. If nonlinear iteration converges and constant solutions are excluded, the following equation holds when v is sufficiently large:

$$\big|u_K^{v+1}-u_K^v\big|=\big|\epsilon\big(u_K^v-u_{Km}^v+s_0\big)\big|\le\widetilde{C}\epsilon h.$$

Assumption 5.2. For $K,L\in\mathcal{J}_{in}, v>0$, at least one of $u_K^v-u_{Km}^v=\mathcal{O}(h)$ and $u_L^v-u_{Lm}^v=\mathcal{O}(h)$ is established, and $u_K^v-u_{Km}^v-(u_L^v-u_{Lm}^v)$ is at least $\mathcal{O}(h)$.

Lemma 5.1. *Let Assumptions* 5.1 *and* 5.2 *are satisfied. Then the DSEP scheme writes*

$$\begin{cases} \sum\limits_{S\in\mathcal{E}_K\cap\mathcal{E}_{int}} \left[\tau_S\big(u_K^{v+1}-u_L^{v+1}\big)-\frac{\tau_S}{2|S|}\big(u_A^{v+1}\eta_A+u_B^{v+1}\eta_B+u_C^{v+1}\eta_C\big)\right] \\ \quad + \sum\limits_{S\in\mathcal{E}_K\cap\mathcal{E}_{ext}} \lambda_{K,S}\big(u_K^{v+1}-g(K')\big)=f_Km(K)+\rho_K^v, \quad \forall\, K\in\mathcal{J}_{in}, & (5.2a) \\ u_M=g(M), \qquad\qquad \forall\, M\in\mathcal{J}_{out}, & (5.2b) \end{cases}$$

with

$$|\rho_K^v|\le C_1\epsilon h+C_2h^2,$$

where C_1 and C_2 are constants independent of h and ϵ,

$$\eta_A=\vec{n}_{CB}\cdot\overrightarrow{K'L'},\quad \eta_B=\vec{n}_{AC}\cdot\overrightarrow{K'L'},\quad \eta_C=\vec{n}_{BA}\cdot\overrightarrow{K'L'}.$$

Proof. Suppose $D_S \geq 0$, then $D_S^+ = D_S, D_S^- = 0$ and $K_m = K_{\min}, L_m = L_{\max}$. Subtracting the Eq. (5.1a) from the Eq. (5.2a), we can yield

$$\begin{aligned}
\rho_K^v &= \sum_{S\in\mathcal{E}_K\cap\mathcal{E}_{int}} \Big[-\frac{\tau_S}{2|S|}\big(u_A^{v+1}\eta_A + u_B^{v+1}\eta_B + u_C^{v+1}\eta_C\big) - h_{K,S}(\boldsymbol{u}^v)\big(u_K^{v+1} - u_{Km}^v\big)\Big] \\
&= \sum_{S\in\mathcal{E}_K\cap\mathcal{E}_{int}} \Bigg[\frac{2\big|f_{L,S}(\boldsymbol{u}^v)\big|}{\big|f_{K,S}(\boldsymbol{u}^v)\big| + \big|f_{L,S}(\boldsymbol{u}^v)\big|}\frac{\tau_S}{2|S|}\big(u_A^v\eta_A + u_B^v\eta_B + u_C^v\eta_C\big)\frac{u_K^{v+1} - u_{K_m}^v}{u_K^v - u_{K_m}^v + s_0} \\
&\qquad - \frac{\tau_S}{2|S|}\big(u_A^{v+1}\eta_A + u_B^{v+1}\eta_B + u_C^{v+1}\eta_C\big)\Bigg] \\
&= \sum_{S\in\mathcal{E}_K\cap\mathcal{E}_{int}} \Bigg[\frac{2\big|f_{L,S}(\boldsymbol{u}^v)\big|}{\big|f_{K,S}(\boldsymbol{u}^v)\big| + \big|f_{L,S}(\boldsymbol{u}^v)\big|}\frac{\tau_S}{2|S|}\big(u_A^v\eta_A + u_B^v\eta_B + u_C^v\eta_C\big)\frac{u_K^{v+1} - u_{K_m}^v}{u_K^v - u_{K_m}^v + s_0} \\
&\qquad - \frac{\tau_S}{2|S|}\big(u_A^v\eta_A + u_B^v\eta_B + u_C^v\eta_C\big)\frac{u_K^{v+1} - u_{K_m}^v}{u_K^v - u_{K_m}^v + s_0} \\
&\qquad + \frac{\tau_S}{2|S|}\big(u_A^v\eta_A + u_B^v\eta_B + u_C^v\eta_C\big)\frac{u_K^{v+1} - u_{K_m}^v}{u_K^v - u_{K_m}^v + s_0} \\
&\qquad - \frac{\tau_S}{2|S|}\big(u_A^{v+1}\eta_A + u_B^{v+1}\eta_B + u_C^{v+1}\eta_C\big)\Bigg] \\
&= \sum_{S\in\mathcal{E}_K\cap\mathcal{E}_{int}} [X_{K,S} + Y_{K,S}],
\end{aligned}$$

where

$$\begin{aligned}
X_{K,S} &= \frac{2\big|f_{L,S}(\boldsymbol{u}^v)\big|}{\big|f_{K,S}(\boldsymbol{u}^v)\big| + \big|f_{L,S}(\boldsymbol{u}^v)\big|}\frac{\tau_S}{2|S|}\big(u_A^v\eta_A + u_B^v\eta_B + u_C^v\eta_C\big)\frac{u_K^{v+1} - u_{K_m}^v}{u_K^v - u_{K_m}^v + s_0} \\
&\quad - \frac{\tau_S}{2|S|}\big(u_A^v\eta_A + u_B^v\eta_B + u_C^v\eta_C\big)\frac{u_K^{v+1} - u_{K_m}^v}{u_K^v - u_{K_m}^v + s_0}, \\
Y_{K,S} &= \frac{\tau_S}{2|S|}\big(u_A^v\eta_A + u_B^v\eta_B + u_C^v\eta_C\big)\frac{u_K^{v+1} - u_{K_m}^v}{u_K^v - u_{K_m}^v + s_0} \\
&\quad - \frac{\tau_S}{2|S|}\big(u_A^{v+1}\eta_A + u_B^{v+1}\eta_B + u_C^{v+1}\eta_C\big).
\end{aligned}$$

Furthermore,

$$\begin{aligned}
X_{K,S} &= \left(\frac{2\big|f_{L,S}(\boldsymbol{u}^v)\big|}{\big|f_{K,S}(\boldsymbol{u}^v)\big| + \big|f_{L,S}(\boldsymbol{u}^v)\big|} - 1\right)\frac{\tau_S}{2|S|}\big(u_A^v\eta_A + u_B^v\eta_B + u_C^v\eta_C\big)\frac{u_K^{v+1} - u_{K_m}^v}{u_K^v - u_{K_m}^v + s_0} \\
&= \frac{\big|f_{L,S}(\boldsymbol{u}^v)\big| - \big|f_{K,S}(\boldsymbol{u}^v)\big|}{\big|f_{K,S}(\boldsymbol{u}^v)\big| + \big|f_{L,S}(\boldsymbol{u}^v)\big|}\frac{\tau_S}{2|S|}\big(u_A^v\eta_A + u_B^v\eta_B + u_C^v\eta_C\big)\frac{u_K^{v+1} - u_{K_m}^v}{u_K^v - u_{K_m}^v + s_0},
\end{aligned}$$

where

$$\begin{aligned}
&\big|f_{L,S}(\boldsymbol{u}^v)\big| - \big|f_{K,S}(\boldsymbol{u}^v)\big| \\
&= \left|\frac{\tau_S D_S^v}{u_{Lm}^v - u_L^v + s_0}\big(u_L^v - u_{Lm}^v\big)\right| - \left|\frac{\tau_S D_S^v}{u_K^v - u_{Km}^v + s_0}\big(u_K^v - u_{Km}^v\big)\right|
\end{aligned}$$

$$
\begin{aligned}
&= \tau_S D_S^v \left(\frac{u_{Lm}^v - u_L^v}{u_{Lm}^v - u_L^v + s_0} - \frac{u_K^v - u_{Km}^v}{u_K^v - u_{Km}^v + s_0} \right) \\
&= \tau_S D_S^v s_0 \frac{u_{Lm}^v - u_L^v - (u_K^v - u_{Km}^v)}{\big(u_{Lm}^v - u_L^v + s_0\big)\big(u_K^v - u_{Km}^v + s_0\big)}, \\
&\quad \big|f_{K,S}(\boldsymbol{u}^v)\big| + \big|f_{L,S}(\boldsymbol{u}^v)\big| \\
&= \left| \frac{\tau_S D_S^v}{u_K^v - u_{Km}^v + s_0} \big(u_K^v - u_{Km}^v\big) \right| + \left| \frac{\tau_S D_S^v}{u_{Lm}^v - u_L^v + s_0} \big(u_L^v - u_{Lm}^v\big) \right| \\
&= \tau_S D_S^v \left(\frac{u_K^v - u_{Km}^v}{u_K^v - u_{Km}^v + s_0} + \frac{u_{Lm}^v - u_L^v}{u_{Lm}^v - u_L^v + s_0} \right) \\
&= \tau_S D_S^v \left(2 - \frac{s_0}{u_K^v - u_{Km}^v + s_0} - \frac{s_0}{u_{Lm}^v - u_L^v + s_0} \right).
\end{aligned}
$$

From the Assumption 5.2 and $s_0 = \mathcal{O}(h^2)$, we can obtain

$$
s_0 \frac{u_{Lm}^v - u_L^v - \big(u_K^v - u_{Km}^v\big)}{\big(u_{Lm}^v - u_L^v + s_0\big)\big(u_K^v - u_{Km}^v + s_0\big)} = \mathcal{O}(h).
$$

So

$$
\frac{\big|f_{L,S}(\boldsymbol{u}^v)\big| - \big|f_{K,S}(\boldsymbol{u}^v)\big|}{\big|f_{K,S}(\boldsymbol{u}^v)\big| + \big|f_{L,S}(\boldsymbol{u}^v)\big|} = \mathcal{O}(h).
$$

And due to $\tau_S = \mathcal{O}(h), |S| = \mathcal{O}(h^2)$, and $\eta_A = \mathcal{O}(h^2), \eta_B = \mathcal{O}(h^2), \eta_C = \mathcal{O}(h^2)$, we have

$$
\begin{aligned}
&\frac{\tau_S}{2|S|} \big(u_A^v \eta_A + u_B^v \eta_B + u_C^v \eta_C\big) = \mathcal{O}(h), \\
&\frac{u_K^{v+1} - u_{K_m}^v}{u_K^v - u_{K_m}^v + s_0} = \frac{u_K^{v+1} - u_K^v + u_K^v - u_{K_m}^v + s_0 - s_0}{u_K^v - u_{K_m}^v + s_0} = 1 + \frac{u_K^{v+1} - u_K^v - s_0}{u_K^v - u_{K_m}^v + s_0} = \mathcal{O}(1).
\end{aligned}
$$

As a result,

$$
|X_{K,S}| \le C_1' h^2,
$$

where C_1' is a constant independent of h and ϵ.

While

$$
\begin{aligned}
Y_{K,S} &= \frac{\tau_S}{2|S|} \Bigg[\big(u_A^v \eta_A + u_B^v \eta_B + u_C^v \eta_C\big) \frac{u_K^{v+1} - u_K^v + u_K^v - u_{K_m}^v + s_0 - s_0}{u_K^v - u_{K_m}^v + s_0} \\
&\qquad - \big(u_A^{v+1} \eta_A + u_B^{v+1} \eta_B + u_C^{v+1} \eta_C\big) \Bigg] \\
&= \frac{\tau_S}{2|S|} \Bigg[(u_A^v \eta_A + u_B^v \eta_B + u_C^v \eta_C) \frac{u_K^{v+1} - u_K^v}{u_K^v - u_{K_m}^v + s_0} - \big(u_A^v \eta_A + u_B^v \eta_B + u_C^v \eta_C\big) \frac{s_0}{u_K^v - u_{K_m}^v + s_0} \\
&\qquad + \big(u_A^v \eta_A + u_B^v \eta_B + u_C^v \eta_C\big) - \big(u_A^{v+1} \eta_A + u_B^{v+1} \eta_B + u_C^{v+1} \eta_C\big) \Bigg] \\
&= \frac{\tau_S}{2|S|} \Bigg[\big(u_A^v \eta_A + u_B^v \eta_B + u_C^v \eta_C\big) \frac{u_K^{v+1} - u_K^v}{u_K^v - u_{K_m}^v + s_0} - \big(u_A^v \eta_A + u_B^v \eta_B + u_C^v \eta_C\big) \frac{s_0}{u_K^v - u_{K_m}^v + s_0} \\
&\qquad - \left(\frac{u_A^{v+1} - u_A^v}{u_K^v - u_{Km}^v + s_0} \eta_A + \frac{u_B^{v+1} - u_B^v}{u_K^v - u_{Km}^v + s_0} \eta_B + \frac{u_C^{v+1} - u_C^v}{u_K^v - u_{Km}^v + s_0} \eta_C \right) \\
&\qquad \times \big(u_K^v - u_{Km}^v + s_0\big) \Bigg].
\end{aligned}
$$

Denote

$$\epsilon_K = \frac{u_K^{v+1} - u_K^v}{u_K^v - u_{K_m}^v + s_0}, \quad \epsilon_A = \frac{u_A^{v+1} - u_A^v}{u_K^v - u_{Km}^v + s_0},$$
$$\epsilon_B = \frac{u_B^{v+1} - u_B^v}{u_K^v - u_{Km}^v + s_0}, \quad \epsilon_C = \frac{u_C^{v+1} - u_C^v}{u_K^v - u_{Km}^v + s_0},$$

then from the Assumption 5.1, we have $|\epsilon_K| \le \epsilon, |\epsilon_A| \le \epsilon, |\epsilon_B| \le \epsilon, |\epsilon_C| \le \epsilon$. And due to $\tau_S = \mathcal{O}(h), |S| = \mathcal{O}(h^2), \eta_A = \mathcal{O}(h^2), \eta_B = \mathcal{O}(h^2), \eta_C = \mathcal{O}(h^2), s_0 = \mathcal{O}(h^2)$, as well as the Assumption 5.2, we can obtain

$$|Y_{K,S}| \le C_2'\epsilon h + C_3' h^2,$$

where C_2' and C_3' are constants independent of h and ϵ. Then

$$\begin{aligned}\left|\rho_K^v\right| &= \left| \sum_{K\in\mathcal{E}_K\cap\mathcal{E}_{int}} [X_{K,S} + Y_{K,S}] \right| \le \sum_{K\in\mathcal{E}_K\cap\mathcal{E}_{int}} (|X_{K,S}| + |Y_{K,S}|) \\ &\le n_K\left(C_1' h^2 + C_2'\epsilon h + C_3' h^2\right) \le C_1\epsilon h + C_2 h^2,\end{aligned}$$

where $C_1 = n_K C_2', C_2 = n_K(C_1' + C_3')$, and n_K represents the number of faces of cell K, especially for tetrahedral meshes, $n_K = 4$. This concludes the proof. □

Remark 5.2. For the case of $f_{K,S}(\boldsymbol{u}^v) f_{L,S}(\boldsymbol{u}^v) \ge 0$, since $f_{K,S}(\boldsymbol{u}^v) f_{L,S}(\boldsymbol{u}^v) \le 0$, it is actually $f_{K,S}(\boldsymbol{u}^v) f_{L,S}(\boldsymbol{u}^v) = 0$. Here, we have $h_{K,S}(\boldsymbol{u}^v) = 0$, and then the normal flux of cell K to face S in the nonlinear DSEP scheme is obtained as $\tau_S(u_K^{v+1} - u_L^{v+1})$. Moreover, for this case, at least one of $f_{K,S}(\boldsymbol{u}^v) = 0$ and $f_{L,S}(\boldsymbol{u}^v) = 0$ is established. It would be well if

$$f_{K,S}(\boldsymbol{u}^v) = \frac{\tau_S D_S^+}{u_K^v - u_{K_m}^v + s_0}\left(u_K^v - u_{K_m}^v\right) = 0$$

under the assumption $D_S \ge 0$. Therefore, at least one of the following two cases occurred:

(i) $D_S = 0$;

(ii) $K = K_m$, i.e., the local minimum (or maximum) value is achieved at K, then we set $D_S = 0$, which represents the approximation of the tangential gradient part. The reason is a consequence of a well-known fact, if u is a continuous differentiable function in a domain Ω, and u achieves local minimum (or maximum) value at the point $K \in \Omega$, then the gradient of u vanishes at the point K.

Regardless of the above case occurred, the normal flux of cell K to face S in the linear scheme [14] is $\tau_S(u_K^{v+1} - u_L^{v+1})$. So, in the above proof, we focused on the case of $f_{K,S}(\boldsymbol{u}^v) f_{L,S}(\boldsymbol{u}^v) < 0$, i.e., $h_{K,S}(\boldsymbol{u}^v) \ne 0$.

Before giving Theorem 5.3, recall that

$\check{\boldsymbol{u}} = (\check{u}_K)_{K\in\mathcal{J}_{in}}$ is the exact solution of (2.1)-(2.2);

$\hat{\boldsymbol{u}} = (\hat{u}_K)_{K\in\mathcal{J}_{in}}$ is the numerical solution of the linear scheme in [14];

$\boldsymbol{u} = (u_K)_{K\in\mathcal{J}_{in}}$ is the numerical solution of the DSEP scheme of (3.9)-(3.10);

$\boldsymbol{u}^v = (u_K^v)_{K\in\mathcal{J}_{in}}$ is the v-th iterate solution associated with the DSEP scheme, that is the solution of (5.1).

And to maintain the completeness of this paper, we present the following conclusions from [14].

Assumption 5.3 ([14, Assumption (H1)]). $\check{u}(\boldsymbol{x})\in C^2(K), \boldsymbol{\kappa}(\boldsymbol{x})\in C^1(K)$ and $f(\boldsymbol{x})\in C^1(K)$ for all $K \in \mathcal{J}$; and $g(x) \equiv 0$ on $\partial\Omega$.

Suppose the numbers of the cells around vertex A, B, C are N_A, N_B, N_C, respectively, then

$$u_A = \sum_{i=1}^{N_A} \omega_{A_i} u_{K_{A_i}}, \quad u_B = \sum_{i=1}^{N_B} \omega_{B_i} u_{K_{B_i}}, \quad u_C = \sum_{i=1}^{N_C} \omega_{C_i} u_{K_{C_i}}.$$

Denote $\mathcal{E}_S$ be the set of all the cell faces neighboring to cell face S. That is, the cell faces in $\mathcal{E}_S$ has at least a common vertex with S. Let $N_{S'}$ be the number of cell neighboring to cell face S', where $S' \in \mathcal{E}_S$.

Assumption 5.4 ([14, Assumption (H2)]). For each cell face S,

$$\tau_S D_S^2 \omega_{S'}^2 \le \frac{1-\varepsilon_0}{N_{S'}^2} \tau_{S'},$$

where ε_0 is a given small constant, and $\omega_{S'}$ is defined such that

$$\begin{aligned} &\eta_A \left(\sum_{i=1}^{N_A} \omega_{A_i} u_{K_{A_i}} - \sum_{i=1}^{N_C} \omega_{C_i} u_{K_{C_i}} \right) + \eta_B \left(\sum_{i=1}^{N_B} \omega_{B_i} u_{K_{B_i}} - \sum_{i=1}^{N_C} \omega_{C_i} u_{K_{C_i}} \right) \\ = &\sum_{S' \in \mathcal{E}_S} \omega_{S'} (u_{S',1} - u_{S',2}), \end{aligned}$$

and $u_{S',1}$ and $u_{S',2}$ are two cell-centered unknowns which are defined in the center of two cells sharing the cell face S'.

Assumption 5.5 ([14, Assumption (H3)]). For a mesh $\mathcal{J}$ on Ω, there is a constant $C > 0$ such that the following discrete Poincare inequality holds for any discrete function $\{u_K\}_{K\in\mathcal{J}}$ satisfying $u_K = 0$ for all $K \in \mathcal{J}_{out}$,

$$\sum_{K\in\mathcal{J}} |u_K|^2 m(K) \le C \sum_{S\in\mathcal{E}_{int}} \tau_S (u_L - u_K)^2.$$

Theorem 5.1 ([14, Theorem 3.1]). *Let Assumptions* 5.3 *and* 5.4 *are satisfied. Then there exists a constant C, independent of h such that*

$$\sum_{S\in\mathcal{E}} \tau_S (u_L - u_K)^2 \le C \sum_{K\in\mathcal{J}} |f_K|^2 m(k).$$

Theorem 5.2 ([14, Theorem 3.2]). *Let Assumptions* 5.3-5.5 *are satisfied. Then there exists a constant C, independent of h such that*

$$\left(\sum_{S\in\mathcal{E}} \tau_S (e_L - e_K)^2 \right)^{\frac{1}{2}} \le Ch,$$

where $e_K = \check{u}(K) - \hat{u}_K$.

Theorem 5.3. *Let that Assumptions* 5.1-5.5 *are satisfied. Then, there exist constants C_3 and C_4 independent of h and ϵ such that*

$$\|\check{\boldsymbol{u}} - \boldsymbol{u}^{v+1}\|_2 \le C_3 \epsilon h + C_4 h.$$

Proof. The linear scheme in [14] can be written as (see [14, Eq. (4.1)])

$$\boldsymbol{A}\hat{\boldsymbol{u}} = \boldsymbol{F}.$$

While system (5.2) writes

$$\boldsymbol{A}\boldsymbol{u}^{v+1} = \boldsymbol{F} + \boldsymbol{F}_{\varepsilon}^{v}$$

with $\boldsymbol{F}_{\varepsilon}^{v} = (\rho_K^v)_{K\in\mathcal{J}_{in}}$.

Thanks to Theorem 5.1, there exists a constant C_4' such that

$$\|\hat{\boldsymbol{u}} - \boldsymbol{u}^{v+1}\|_2 \le C_4' \left\|\boldsymbol{F}_{\varepsilon}^{v}\right\|_2.$$

From Lemma 5.1, there exists constants C_5', C_6' such that

$$\left\|\boldsymbol{F}_{\varepsilon}^{v}\right\|_2 \le C_5'\epsilon h + C_6' h^2.$$

Reviewing Theorem 5.2, we have

$$\|\check{\boldsymbol{u}} - \hat{\boldsymbol{u}}\|_2 \le C_7' h. \tag{5.3}$$

Then applying the triangle inequality, as well as (5.3), we obtain

$$\|\check{\boldsymbol{u}} - \boldsymbol{u}^{v+1}\|_2 \le \|\check{\boldsymbol{u}} - \hat{\boldsymbol{u}}\|_2 + \|\hat{\boldsymbol{u}} - \boldsymbol{u}^{v+1}\|_2 \le C_3\epsilon h + C_4 h,$$

where $C_3 = C_4'C_5', C_4 = C_4'C_6' + C_7'$, which concludes the proof. □

For convenience, rewrite (5.1) in the form of an algebraic system as follows:

$$\boldsymbol{N}(\boldsymbol{u}^{v-1})\boldsymbol{u}^{v} = \boldsymbol{G}(\boldsymbol{u}^{v-1}), \tag{5.4}$$

where $\boldsymbol{N}(\boldsymbol{u}^{v-1})$ is the coefficient matrix of this system and $\boldsymbol{G}(\boldsymbol{u}^{v-1})$ is a discrete vector related to the right-hand term, boundary conditions and local extremum (given by the previous iterative step). So we introduce the following modified Picard iteration named as Algorithm 5.1.

The linearized system (5.5) is solved by BICGSTAB method [26]. The iterations are terminated when the relative norm of the residual becomes smaller than ε_{lin}. Herein, we take $\varepsilon_{non} = 10^{-5}$ and $\varepsilon_{lin} = 10^{-12}$.

Algorithm 5.1: Modified Picard Iteration.

1 Choose a small positive value ε_{non} and $\boldsymbol{u}^0$.
2 **for** $v = 1, 2, \ldots$ **do**
3 Solve the linear system
$$\boldsymbol{N}(\boldsymbol{u}^{v-1})\boldsymbol{u}^{v} = \boldsymbol{G}(\boldsymbol{u}^{v-1}). \tag{5.5}$$
4 Use formula
$$\|\boldsymbol{N}(\boldsymbol{u}^{v})\boldsymbol{u}^{v} - \boldsymbol{G}(\boldsymbol{u}^{v})\| \le \varepsilon_{non}\|\boldsymbol{N}(\boldsymbol{u}^{0})\boldsymbol{u}^{0} - \boldsymbol{G}(\boldsymbol{u}^{0})\| \tag{5.6}$$
as the convergence criterion.
5 **end**

5.2. Anderson acceleration method

Empirically, the convergence of Picard iteration is slow when the distortion of meshes is severe, or the number of cells is very large, especially for high-dimensional problems. So AA method is considered. The AA method of Picard iteration is demonstrated [17]. Now we consider the modified Picard iteration with AA method named as Algorithm 5.2.

In Algorithm 5.2, the positive parameter ε_{non} and initial value $\boldsymbol{u}^0$ are given. The Anderson depth m is a given positive integer, which represents that the most recent m iterations are involved in the acceleration. Note that when $m = 1$, the mP-AA method is same as m-Picard method. Let $\boldsymbol{u}^v (v = 0, 1, \ldots)$ be a given sequence. $\tilde{\boldsymbol{u}}^v$ is defined as the solution of $\boldsymbol{N}(\boldsymbol{u}^v)\tilde{\boldsymbol{u}}^v = \boldsymbol{G}(\boldsymbol{u}^v)$ for $v \geq 3$. The solution of problem (5.7) is transformed to a saddle-point problem, see [17] in detail.

Algorithm 5.2: Modified Picard Iteration with AA method.

Choose a small positive value ε_{non} and $\boldsymbol{u}^0 \geq \boldsymbol{0}$.
Compute (5.5) and set $\tilde{\boldsymbol{u}}^v = \boldsymbol{u}^v, \delta\boldsymbol{u}^v = \tilde{\boldsymbol{u}}^v - \boldsymbol{u}^{v-1}$, $v = 1, 2$.
for $v = 2, 3, \ldots$ **do**
  $\bar{m} = \min\{m, v\}$.
  Determine coefficients $\alpha_1, \alpha_2, \ldots, \alpha_{\bar{m}}$ by solving the constrained minimization problem

$$\min_{\sum_{i=1}^{\bar{m}} \alpha_i = 1} \left\| \sum_{i=1}^{\bar{m}} \alpha_i \delta \boldsymbol{u}^{v-\bar{m}+i} \right\|. \tag{5.7}$$

  Set new iterative values

$$\boldsymbol{u}^{v+1} = \sum_{i=1}^{\bar{m}} \alpha_i \tilde{\boldsymbol{u}}^{v-\bar{m}+i}.$$

  Use formula (5.6) as the convergence criterion.
  Solve the linear system

$$\boldsymbol{N}(\boldsymbol{u}^{v+1})\tilde{\boldsymbol{u}}^{v+1} = \boldsymbol{G}(\boldsymbol{u}^{v+1}),$$

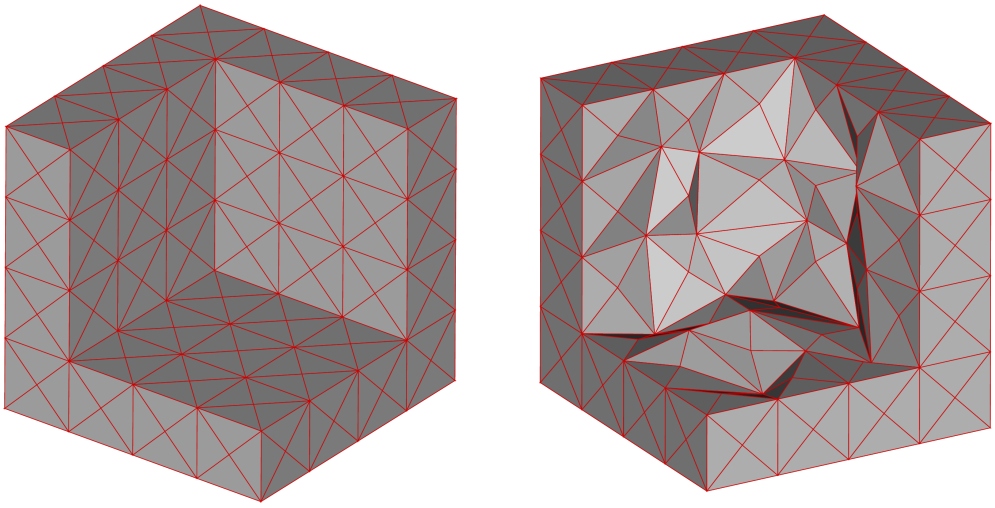
  and set $\delta\boldsymbol{u}^{v+1} = \tilde{\boldsymbol{u}}^{v+1} - \boldsymbol{u}^{v+1}$.
end

6. Numerical Examples

We present several numerical examples to demonstrate that computational performance of our scheme. The convergent order is tested on two types of meshes (see Fig. 6.1). Here we use discrete L^2-norm to evaluate approximation errors between exact solution and numerical solution

$$e_u = \left[\sum_{K \in \mathcal{J}} \big(\breve{u}(K) - u_K\big)^2 m(K) \right]^{\frac{1}{2}}.$$

Firstly, we take a uniform cubic partition on $\Omega = (0, 1)^3$ with a mesh size h, and divide each cube into 24 uniform tetrahedrons, which is called as Mesh A, or 24 random tetrahedrons with



(a) Mesh A (b) Mesh B

Fig. 6.1. The meshes with size $h = 1/4$.

randomly distorted position of mesh vertices

$$X = x + \eta_x h, \quad Y = y + \eta_y h, \quad Z = z + \eta_z h,$$

where η_x, η_y and η_z are random variables with values between -0.3 and 0.3, which is called as Mesh B.

6.1. The discrete strong extremum principle

In this subsection, we test the DSEP of the new scheme and the scheme proposed in [14], for the convenience of narration, we denote the scheme here by 3DE, and the scheme in [14] by 3DD.

Denote

$$u_{\min} = \min_{K\in\mathcal{J}_{in}\cup\mathcal{J}_{out}} u_K, \quad u^{in}_{\min} = \min_{K\in\mathcal{J}_{in}} u_K,$$
$$u_{\max} = \max_{K\in\mathcal{J}_{in}\cup\mathcal{J}_{out}} u_K, \quad u^{in}_{\max} = \max_{K\in\mathcal{J}_{in}} u_K.$$

In other words, $u^{in}_{\min}$ denotes the minimum value of internal cells, while $u_{\min}$ denotes the minimum value of internal and boundary cells. And N_c denotes the number of cells.

Now, we consider the problem (2.1)-(2.2) on $(0,1)^3$. In this problem, the diffusion coefficient is

$$\boldsymbol{\kappa} = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} k_1 & 0 & 0 \\ 0 & k_2 & 0 \\ 0 & 0 & k_3 \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

where $k_1 = 1, k_2 = 0.1, k_3 = 10$, and $\theta = \pi/6$. The Dirichlet boundary condition is

$$g = 1 + \sin\left(\frac{\pi}{2}x\right)\sin\left(\frac{\pi}{2}y\right)\sin\left(\frac{\pi}{2}z\right),$$

Table 6.1: The maximum and minimum values of the problem with zero right-hand term on Mesh A.

N_c		24×4^3	24×8^3	24×16^3	24×24^3
3DD	$u_{\min}$	9.995796e-001	9.998732e-001	9.999593e-001	9.999817e-001
	$u^{in}_{\min}$	9.995796e-001	9.998732e-001	9.999593e-001	9.999817e-001
	$u_{\max}$	1.978685e+000	1.994652e+000	1.998662e+000	1.999405e+000
	$u^{in}_{\max}$	1.966667e+000	1.992286e+000	1.998223e+000	1.999243e+000
3DE	$u_{\min}$	1.000000e+000	1.000000e+000	1.000000e+000	1.000000e+000
	$u^{in}_{\min}$	1.000057e+000	1.000035e+000	1.000013e+000	1.000007e+000
	$u_{\max}$	1.978685e+000	1.994652e+000	1.998662e+000	1.999405e+000
	$u^{in}_{\max}$	1.964976e+000	1.990512e+000	1.997566e+000	1.998902e+000

Table 6.2: The maximum and minimum values of the problem with zero right-hand term on Mesh B ($\eta_x, \eta_y, \eta_z = 0.3$).

N_c		24×4^3	24×8^3	24×16^3	24×24^3
3DD	$u_{\min}$	9.993771e-001	9.996796e-001	9.999487e-001	9.999731e-001
	$u^{in}_{\min}$	9.993771e-001	9.996796e-001	9.999487e-001	9.999731e-001
	$u_{\max}$	1.978685e+000	1.994652e+000	1.998662e+000	1.999405e+000
	$u^{in}_{\max}$	1.965566e+000	1.992250e+000	1.998120e+000	1.999215e+000
3DE	$u_{\min}$	1.000000e+000	1.000000e+000	1.000000e+000	1.000000e+000
	$u^{in}_{\min}$	1.000054e+000	1.000028e+000	1.000013e+000	1.000007e+000
	$u_{\max}$	1.978685e+000	1.994652e+000	1.998662e+000	1.999405e+000
	$u^{in}_{\max}$	1.965081e+000	1.990293e+000	1.997570e+000	1.998874e+000

the right-hand term is $f = 0$, and the exact solution is unknown. According to the extremum principle, the maximum and minimum values of this problem are all reached on the boundary.

Tables 6.1 and 6.2 give the maximum and minimum values of 3DD and 3DE schemes on two different meshes, from which we can see that the maximum and minimum values of 3DE scheme are reached on the boundary, while the minimum values of 3DD scheme are obtained in the internal cells. These results verify that 3DD scheme cannot keep DSEP, while our scheme satisfies DSEP on tetrahedral meshes.

6.2. The Anderson acceleration method

In this subsection, we will take the following examples with Dirichlet boundary condition in the unit cube $\Omega = (0,1)^3$ to demonstrate the acceleration efficiency of the mP-AA algorithm.

Example 6.1 (The Problem with Scalar Coefficient).

$$\begin{aligned}
&\boldsymbol{\kappa} = 1 + x + y + z,\\
&u = \cos(\pi x)\cos(\pi y)\cos(\pi z),\\
&f = 3\pi^2(1 + x + y + z)\cos(\pi x)\cos(\pi y)\cos(\pi z) + \pi\sin(\pi x)\cos(\pi y)\cos(\pi z)\\
&\quad + \pi\cos(\pi x)\sin(\pi y)\cos(\pi z) + \pi\cos(\pi x)\cos(\pi y)\sin(\pi z).
\end{aligned}$$

Example 6.2 (The Problem with Discontinuous Coefficient).

$$\boldsymbol{\kappa} = \begin{cases} 5, & x < \dfrac{1}{2}, \\ 1, & x \geq \dfrac{1}{2}, \end{cases}$$

$$u = \begin{cases} (x^2+10)(y-y^2)(z-z^2), & x < \dfrac{1}{2}, \\ (5x^2+9)(y-y^2)(z-z^2), & x \geq \dfrac{1}{2}, \end{cases}$$

$$f = \begin{cases} -10(y-y^2)(z-z^2) + 10(x^2+10)(z-z^2) + 10(x^2+10)(y-y^2), & x < \dfrac{1}{2}, \\ -10(y-y^2)(z-z^2) + 2(5x^2+9)(z-z^2) + 2(5x^2+9)(y-y^2), & x \geq \dfrac{1}{2}. \end{cases}$$

Example 6.3 (The Problem with Anisotropic Diffusion Coefficient).

$$\boldsymbol{\kappa} = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} k_1 & 0 & 0 \\ 0 & k_2 & 0 \\ 0 & 0 & k_3(1+x+y+z) \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

where $k_1 = 0.1, k_2 = 1.0, k_3 = 10.0$, and $\theta = \pi/6$.

$$\begin{aligned} u &= \sin(\pi x) + \sin(\pi y) + \sin(\pi z) + 1.0, \\ f &= \pi^2\left(k_1\cos^2\theta + k_2\sin^2\theta\right)\sin(\pi x) + \pi^2\left(k_1\sin^2\theta + k_2\cos^2\theta\right)\sin(\pi y) \\ &\quad + k_3(1+x+y+z)\pi^2\sin(\pi z) - k_3\pi\cos(\pi z). \end{aligned}$$

Example 6.4 (The Problem with Strong Anisotropic Diffusion Coefficient).

$$\begin{aligned} \boldsymbol{\kappa} &= \begin{pmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 0.001 \end{pmatrix}, \\ u &= e^{xy} + z^2, \\ f &= -e^{xy}(x^2+y^2) - 0.002. \end{aligned}$$

In numerical experiments, it_P denotes the number of nonlinear iterations by using m-Picard iteration, while it_{AA} represents the number of nonlinear iterations by using mP-AA method.

6.2.1. Comparison under different m

We compare the iterative numbers under different Anderson depth m for mP-AA method. First, we test Example 6.1 on Mesh B ($\eta_x, \eta_y, \eta_z = 0.3$) with $N_c = 24 \times 24^3$. The comparison results are shown in Fig. 6.2.

Then we test Example 6.2 on Mesh B ($\eta_x, \eta_y, \eta_z = 0.3$) with $N_c = 24 \times 16^3$. The comparison results are shown in Fig. 6.3.

In both examples, when $m = 1$, the mP-AA method is same as m-Picard method. From Fig. 6.2, we can see that the mP-AA method reduces the nonlinear iterative numbers of m-Picard method. For this example, $m = 4$ is a reasonable choice. Fig. 6.3 represents that $m = 4$ is the best choice for Example 6.2, and when $m = 4$, the nonlinear iterative numbers of mP-AA method is almost 1/8 of m-Picard iteration.

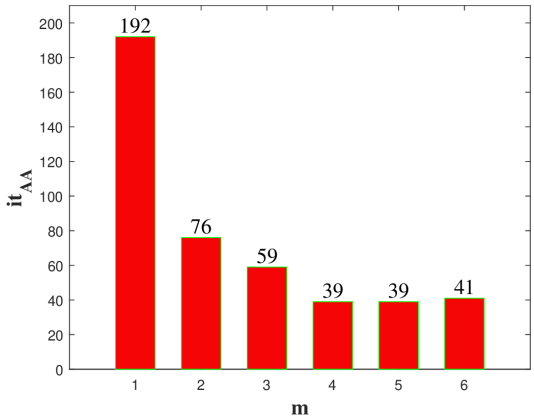


Fig. 6.2. The nonlinear iterative numbers under different m on mesh B ($\eta_x, \eta_y, \eta_z = 0.3$) with $N_c = 24 \times 24^3$ of Example 6.1.

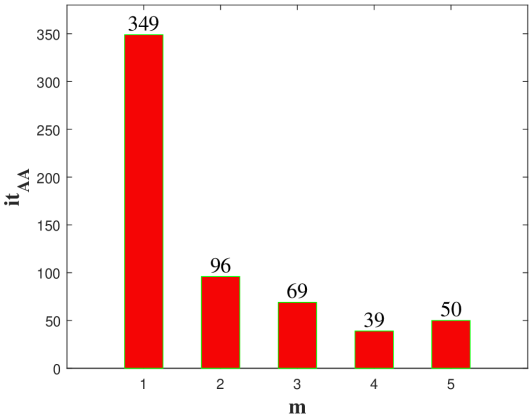


Fig. 6.3. The nonlinear iterative numbers under different m on mesh B ($\eta_x, \eta_y, \eta_z = 0.3$) with $N_c = 24 \times 16^3$ of Example 6.2.

6.2.2. Comparison under different distortion

Next we compare the iterative numbers of m-Picard iteration and mP-AA method on meshes with different distortion. We test Examples 6.2 and 6.3 on Mesh B, respectively. The distortion of Mesh B is measured by random variables η_x, η_y, η_z with values between -0.3 and 0.3. When the absolute value of the random variables get larger, the meshes become more distorted.

For Example 6.2, we take the Anderson depth $m = 4$ with $N_c = 24 \times 16^3$ and vary η_x, η_y, η_z from -0.3 to 0.3. The iterative numbers of both two iteration methods are shown in Fig. 6.4. While Fig. 6.5 represents the results of $m = 4$ with $N_c = 24 \times 8^3$ for Example 6.3.

It can be seen from Figs 6.4 and 6.5 that the nonlinear iterative numbers of mP-AA method are much less than those of m-Picard iteration, and the acceleration effect is significant.

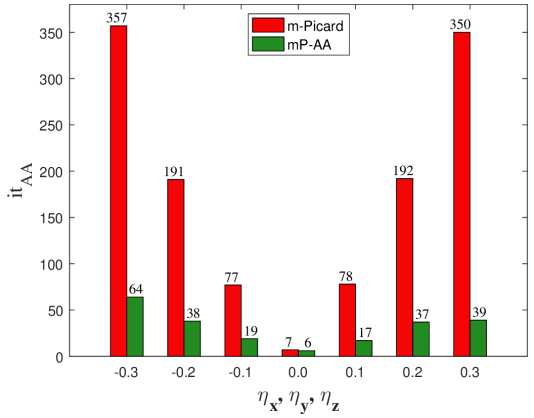


Fig. 6.4. The nonlinear iterative numbers under different distortion of Mesh B with $N_c = 24 \times 16^3$ of Example 6.2.

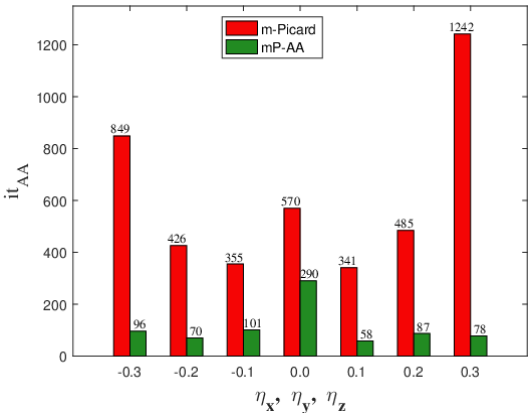


Fig. 6.5. The nonlinear iterative numbers under different distortion of Mesh B with $N_c = 24 \times 8^3$ of Example 6.3.

6.2.3. Comparison under different cell numbers

Finally, we compare the iterative numbers under different cell numbers. Here, we test Example 6.4 with $m = 6$ on Mesh A and $m = 8$ on Mesh B ($\eta_x, \eta_y, \eta_z = 0.3$) respectively. We vary the number of cells from 24×4^3 to 24×16^3. The results are shown in Table 6.3.

From Table 6.3, we can see that the increase rate of the iterative numbers of m-Picard iteration is much faster than that of mP-AA method with N_c increasing. And as N_c increases, the acceleration effect of mP-AA method becomes more obvious, especially when $N_c = 24 \times 16^3$, the nonlinear iterative numbers of mP-AA method is 1/8 of that of m-Picard iteration. These results show that the mP-AA method is more efficient than m-Picard iteration.

Table 6.3: The nonlinear iterative numbers under different cell numbers of Example 6.4.

N_c		24×4^3	24×8^3	24×12^3	24×16^3
Mesh A	it_P	40	126	251	407
	$it_{AA}(m = 6)$	14	28	39	51
Mesh B	it_P	54	182	401	704
$(\eta_x, \eta_y, \eta_z = 0.3)$	$it_{AA}(m = 8)$	17	30	56	86

6.3. The accuracy of the scheme

In this subsection, we test the accuracy of the scheme by using the examples in Section 6.2, and compare it with the accuracy of the 3DD scheme in [14].

6.3.1. Example 6.1

Table 6.4 gives L^2-norm of the error between exact solution and numerical solution for 3DE and 3DD schemes on two meshes. From this table, we can see that the two schemes obtain almost second order accuracy for the solution.

Table 6.4: Numerical results of Example 6.1 on two different meshes.

N_c	Mesh A				Mesh B ($\eta_x, \eta_y, \eta_z = 0.3$)			
	3DD		3DE		3DD		3DE	
	e_u	rate	e_u	rate	e_u	rate	e_u	rate
24×4^3	5.6508e-3	–	5.6505e-3	–	6.7496e-3	–	6.6193e-3	–
24×8^3	1.4209e-3	1.99	1.4208e-3	1.99	1.8732e-3	1.85	2.4580e-3	1.43
24×16^3	3.5601e-4	2.00	3.5595e-4	2.00	5.0591e-4	1.89	7.7103e-4	1.67
24×24^3	1.5830e-4	2.00	1.5826e-4	2.00	2.3282e-4	1.91	3.7329e-4	1.79

6.3.2. Example 6.2

Table 6.5 gives the L^2-norm of solutions of Example 6.2 for 3DE and 3DD schemes on different meshes. From this table, we can see that the two schemes can obtain almost the same accuracy, i.e., second order accuracy, on Mesh A. It also shows that compared with 3DD scheme, the errors of solutions for 3DE scheme slightly increase on Mesh B, but they are still within the acceptable range.

Table 6.5: Numerical results of Example 6.2 on two different meshes.

N_c	Mesh A				Mesh B ($\eta_x, \eta_y, \eta_z = 0.3$)			
	3DD		3DE		3DD		3DE	
	e_u	rate	e_u	rate	e_u	rate	e_u	rate
24×4^3	4.3981e-3	–	4.3979e-3	–	5.9364e-3	–	8.6226e-3	–
24×8^3	1.1045e-3	1.99	1.1044e-3	1.99	1.5424e-3	1.94	2.6009e-3	1.73
24×16^3	2.7631e-4	2.00	2.7626e-4	2.00	4.0494e-4	1.93	6.2597e-4	2.05
24×24^3	1.2280e-4	2.00	1.2277e-4	2.00	1.8286e-4	1.96	2.9674e-4	1.84

6.3.3. Example 6.3

The numerical results on different meshes of Example 6.3 are shown in Table 6.6, which shows that both of the schemes can obtain almost second order accuracy for solution. From this table, one can see that the solution errors of 3DE scheme are slightly higher than that of 3DD scheme, while the convergence order of 3DE scheme is slightly higher than that of 3DD scheme.

Table 6.6: Numerical results of Example 6.3 on two different meshes.

N_c	Mesh A				Mesh B ($\eta_x, \eta_y, \eta_z = 0.3$)			
	3DD		3DE		3DD		3DE	
	e_u	rate	e_u	rate	e_u	rate	e_u	rate
24×4^3	1.7602e-2	–	2.5553e-2	–	2.0027e-2	–	4.0347e-2	–
24×8^3	6.6721e-3	1.40	7.1068e-3	1.85	6.4572e-3	1.63	1.2722e-2	1.67
24×16^3	2.1549e-3	1.63	1.8539e-3	1.94	2.0490e-3	1.66	3.3889e-3	1.91
24×24^3	1.0272e-3	1.83	8.4148e-4	1.95	1.0032e-3	1.76	1.6054e-3	1.84

6.3.4. Example 6.4

Table 6.7 represents the numerical results of Example 6.4 calculated by using 3DD and 3DE schemes. In 3DD scheme, the parts related to vertex values in the normal flow (3.3) are placed in the source term for source iteration in order to avoid computational complexity, as detailed in [14]. From Table 6.7, it can be seen that when $N_c = 24 \times 16^3$, the 3DD scheme does not converge on Mesh B. However, by using 3DE scheme, it can be achieved almost second order accuracy for solutions.

Table 6.7: Numerical results of Example 6.4 on two different meshes.

N_c	Mesh A				Mesh B ($\eta_x, \eta_y, \eta_z = 0.3$)			
	3DD		3DE		3DD		3DE	
	e_u	rate	e_u	rate	e_u	rate	e_u	rate
24×4^3	3.7210e-3	–	3.7206e-3	–	4.9325e-3	–	4.9308e-3	–
24×8^3	1.1195e-3	1.73	1.1202e-3	1.73	1.6219e-3	1.60	1.6202e-3	1.61
24×16^3	3.0685e-4	1.87	3.0848e-4	1.86	NaN	–	4.7624e-4	1.77
24×24^3	1.4059e-4	1.92	1.4150e-4	1.92	2.1748e-4	–	2.1858e-4	1.92

7. Conclusion

In this paper, we develop a cell-centered finite volume scheme satisfying DSEP for diffusion problems on arbitrary tetrahedral meshes. In the construction of the scheme, the discrete flux of the linear scheme with second order accuracy in [14] is modified to obtain the nonlinear flux with LEP structure. So there is no need to require the vertex unknowns to be a convex combination of the cell-centered unknowns. Numerical examples show that compared with the scheme in [14], our new scheme can preserve DSEP and also obtain second order accuracy. The bandwidth of the coefficient matrix of the algebraic equation system formed by the new scheme becomes narrower. Besides, we prove our scheme has at least one solution preserving DSEP. Moreover, for solving the nonlinear scheme, a new nonlinear iteration method is designed, which is a modification of the standard Picard iteration. Also we prove the convergence of the new Picard iteration for the DSEP scheme. Furthermore, the modified Picard iteration with AA method is proposed to improve the efficiency of solving nonlinear scheme. Finally, by comparing the modified Picard iteration and mP-AA method, the numerical results show that the mP-AA method can not only accelerate convergence but also be more stable.

We use the tetrahedral meshes in this paper, but the construction idea of this scheme can be generalized to general polyhedral meshes without essential difficulty. In particular, combined with the effective face technique described in [29], the approach of designing the scheme satisfying DSEP can be generalized to general polyhedral meshes with non-planar cell-faces.

Acknowledgements. The authors are very grateful to the anonymous referees for their helpful suggestions to enhance the paper.

This work is partially supported by the National Natural Science Foundation of China (Grant Nos. 12401505, 12471340, 12071045), by the R&D Program of Beijing Municipal Education Commission (Grant No. KM202210009005), by the Youth Fund of Key Laboratory of Beijing Institute of Applied Physics and Computational Mathematics, Unveiling and Leading Project of NCUT (Grant No. 2023YZZKY19).

References

[1] H. An, X. Jia, and H.F. Walkerc, Anderson acceleration and application to the three-temperature energy equations, *J. Comput. Phys.*, **347** (2017), 1–19.

[2] D.G. Anderson, Iterative procedures for nonlinear integral equations, *J. ACM*, **12** (1965), 547–560.

[3] E. Bertolazzi and G. Manzini, A second-order maximum principle preserving finite volume method for steady convection-diffusion problems, *SIAM J. Numer. Anal.*, **43** (2005), 2172–2199.

[4] H. Burchard, E. Deleersnijder, and A. Meister, A high-order conservative Patankar-type discretisation for stiff systems of production-destruction equations, *Appl. Numer. Math.*, **47** (2003), 1–30.

[5] C. Cancès, M. Cathala, and C. Le Potier, Monotone corrections for generic cell-centered finite volume approximations of anisotropic diffusion equations, *Numer. Math.*, **125** (2013), 387–417.

[6] L. Chang, Z. Sheng, and G. Yuan, An improvement of the two-point flux approximation scheme on polygonal meshes, *J. Comput. Phys.*, **392** (2019), 187–204.

[7] J. Droniou, Finite volume schemes for diffusion equations: Introduction to and review of modern methods, *Math. Models Methods Appl. Sci.*, **24** (2014), 1575–1619.

[8] J. Droniou and C. Le Potier, Construction and convergence study of schemes preserving the elliptic local maximum principle, *SIAM J. Numer. Anal.*, **49** (2011), 459–490.

[9] C. Evans, S.N. Pollock, L.G. Rebholz, and M. Xiao, A proof that Anderson acceleration improves the convergence rate in linearly converging fixed-point methods (but not in those converging quadratically), *SIAM J. Numer. Anal.*, **58** (2020), 788–810.
[10] H.R. Fang and Y. Saad, Two classes of multisecant methods for nonlinear acceleration, *Numer. Linear Algebra Appl.*, **16** (2009), 197–221.
[11] V. Ganine, N.J. Hills, and B.L. Lapworth, Nonlinear acceleration of coupled fluid-structure transient thermal problems by Anderson mixing, *Int. J. Numer. Methods Fluids*, **71** (2013), 939–959.
[12] Z. Gao and J. Wu, A small stencil and extremum-preserving scheme for anisotropic diffusion problems on arbitrary 2D and 3D meshes, *J. Comput. Phys.*, **250** (2013), 308–331.
[13] D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equtaions of Second Order*, Springer, 2001.
[14] X. Lai, Z. Sheng, and G. Yuan, A finite volume scheme for three-dimensional diffusion equations, *Commun. Comput. Phys.*, **18** (2015), 650–672.
[15] X. Lai, Z. Sheng, and G. Yuan, Monotone finite volume scheme for three dimensional diffusion equation on tetrahedral meshes, *Commun. Comput. Phys.*, **21** (2017), 162–181.
[16] C. Le Potier, A nonlinear finite volume scheme satisfying maximum and minimum principles for diffusion operators, *Int. J. Finite Vol.*, **6** (2009), 1–20.
[17] K. Lipnikov, D. Svyatskiy, and Y. Vassilevski, Anderson acceleration for nonlinear finite volume scheme for advection-diffusion problems, *SIAM J. Sci. Comput.*, **35** (2013), A1120–A1136.
[18] P.A. Lott, H.F. Walker, C.S. Woodward, and U.M. Yang, An accelerated Picard method for nonlinear systems related to variably saturated flow, *Adv. Water Resour.*, **38** (2012), 92–101.
[19] M. Pal and M.G. Edwards, A family of multi-point flux approximation schemes for general element types in two and three dimensions with convergence performance, *Int. J. Numer. Methods Fluids*, **69** (2012), 1797–2091.
[20] S.V. Patankar, *Numerical Heat Transfer and Fluid Flow*, McGraw-Hill, 1980.
[21] Z. Sheng, M. Thiriet, and F. Hecht, An efficient numerical method for the equations of steady and unsteady flows of homogeneous incompressible Newtonian fluid, *J. Comput. Phys.*, **230** (2011), 551–571.
[22] Z. Sheng and G. Yuan, The finite volume scheme preserving extremum principle for diffusion equations on polygonal meshes, *J. Comput. Phys.*, **230** (2011), 2588–2604.
[23] Z. Sheng and G. Yuan, Analysis of the nonlinear scheme preserving the maximum principle for the anisotropic diffusion equation on distorted meshes, *Sci. China Math.*, **65** (2022), 2379–2396.
[24] Z. Sheng and G. Yuan, A nonlinear scheme preserving maximum principle for heterogeneous anisotropic diffusion equation, *J. Comput. Appl. Math.*, **436** (2024), 115438.
[25] A. Toth and C.T. Kelley, Convergence analysis for Anderson acceleration, *SIAM J. Numer. Anal.*, **53** (2013), 805–819.
[26] H.A. Van Der Vorst, Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems, *SIAM J. Sci. Stat. Comput.*, **13** (1992), 631–644.
[27] R.S. Varga, *Matrix Iterative Analysis*, Pretice Hall, 1962.
[28] S. Wang, X. Hang, and G. Yuan, A positivity-preserving finite volume scheme for diffusion equations on polyhedral meshes, *Math. Numer. Sin.*, **37** (2015), 247–263.
[29] S. Wang, X. Hang, and G. Yuan, A pyramid scheme for three-dimensional diffusion equations on polyhedral meshes, *J. Comput. Phys.*, **350** (2017), 590–606.
[30] J. Willert, W.T. Taitano, and D. Knoll, Leveraging Anderson acceleration for improved convergence of iterative solutions to transport systems, *J. Comput. Phys.*, **273** (2014), 278–286.
[31] J. Xu, F. Zhao, Z. Sheng, and G. Yuan, A nonlinear finite volume scheme preserving maximum principle for diffusion equations, Commun. Comput. Phys., **29** (2021), 747–766.
[32] C. Yang, J. Meza, B. Lee, and L.W. Wang, KSSOLV-a MATLAB toolbox for solving the Kohn-Sham equations, *ACM Trans. Math. Software*, **36** (2010), 10.
[33] G. Yuan, Diffusion schemes satisfying extremum principle on non-orthogonal meshes, *Math. Nu-*

mer. Sin., **43** (2021), 1–16. (in Chinese)

[34] G. Yuan and Y. Yu, Existence of solution of a finite volume scheme preserving maximum principle for diffusion equations: Existence of solution of a finite volume scheme, *Numer. Methods Partial Differential Equations*, **34** (2017), 80–96.