Bound-preserving discontinuous Galerkin method for compressible miscible displacement in porous media *

Hui Guo† Yang Yang‡

Abstract: In this paper, we develop bound-preserving discontinuous Galerkin (DG) methods for the coupled system of compressible miscible displacement problems. We consider the problem with two components and the (volumetric) concentration of the ith component of the fluid mixture, c_i, should be between 0 and 1. However, c_i does not satisfy the maximum principle. Therefore, the numerical techniques introduced in (X. Zhang and C.-W. Shu, Journal of Computational Physics, 229 (2010), 3091-3120) cannot be applied directly. The main idea is to apply the positivity-preserving techniques to both c_1 and c_2, respectively and enforce $c_1 + c_2 = 1$ simultaneously to obtain physically relevant approximations. By doing so, we have to treat the time derivative of the pressure dp/dt as a source in the concentration equation. Moreover, $c_i's$ are not the conservative variables, as a result, the classical bound-preserving limiter in (X. Zhang and C.-W. Shu, Journal of Computational Physics, 229 (2010), 3091-3120) cannot be applied. Therefore, another limiter will be introduced. Numerical experiments will be given to demonstrate the accuracy in L^∞-norm and good performance of the numerical technique.

Keywords: Compressible miscible displacement problem; Discontinuous Galerkin method; Bound-preserving

1 Introduction

Numerical modeling of miscible displacement in porous media is important and interesting in oil recovery and environmental pollution problem. The miscible displacement problem is described by a coupled system of nonlinear partial differential equations. The need for accurate solutions to the coupled equations challenges numerical analysts to design new methods. In this paper, we study the compressible miscible displacement problem in porous media on computational domain $\Omega = [0, 2\pi] \times [0, 2\pi]$

$$\begin{aligned} &d(c)p_t + \nabla \cdot \mathbf{u} = d(c)p_t - \nabla \cdot (\tfrac{\kappa(x,y)}{\mu(c)} \nabla p) = q, \quad (x, y) \in \Omega, \quad 0 < t \leq T, \\ &\phi c_t + b(c)p_t + \mathbf{u} \cdot \nabla c - \nabla \cdot (\mathbf{D}\nabla c) = (\tilde{c} - c)q, \quad (x, y) \in \Omega, \quad 0 < t \leq T, \end{aligned} \tag{1.1}$$

as well as its one-dimensional version. In (1.1) the dependent variables p, $\mathbf{u}$ and c are the pressure in the fluid mixture, the Darcy velocity of the mixture (volume flowing across a unit across-section

*Supported by National Natural Science Foundation of China (11571367 and 11601536) and Michigan Technological University Research Excellence Fund Scholarship and Creativity Grant 1605052

†College of Science, China University of Petroleum, Qingdao 266580, China. E-mail: sdugh@163.com

‡Michigan Technological University, Houghton, MI 49931, USA. E-mail:yyang7@mtu.edu

per unit time), and the concentration of interested species measured in amount of species per unit volume of the fluid mixture, respectively. ϕ and κ are the porosity and the permeability of the rock, respectively. μ is the concentration-dependent viscosity. q is the external volumetric flow rate, and $\tilde{c}$ is the concentration of the fluid in the external flow. $\tilde{c}$ must be specified at points where injection ($q > 0$) takes place, and is assumed to be equal to c at production points ($q < 0$). The diffusion coefficient $\mathbf{D}$ arises from two aspects: molecular diffusion, which is rather small for field-scale problems, and dispersion, which is velocity-dependent, in the petroleum engineering literature. Its form is

$$\mathbf{D} = \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix} = \phi(x,y)(d_{mol}\mathbf{I} + d_{long}|\mathbf{u}|\mathbf{E} + d_{tran}|\mathbf{u}|\mathbf{E}^{\perp}), \tag{1.2}$$

where $\mathbf{E}$, a 2×2 matrix, represents the orthogonal projection along the velocity vector and is given by

$$\mathbf{E} = (e_{ij}(\mathbf{u})) = \left(\frac{u_i u_j}{|\mathbf{u}|^2}\right), \quad \mathbf{u} = (u_1, u_2),$$

and $\mathbf{E}^{\perp} = \mathbf{I} - \mathbf{E}$ is the orthogonal complement. The diffusion coefficient d_{long} measures the dispersion in the direction of the flow and d_{tran} shows that transverse to the flow. To ensure the stability of the scheme, in almost all of the previous works $\mathbf{D}$ is assumed to be strictly positive definite. In this paper, we assume $\mathbf{D}$ to be positive semidefinite. Therefore, we have $D_{11} \geq 0$, $D_{22} \geq 0$ and $D_{12} = D_{21}$. In the numerical experiments, we also choose $\mathbf{D} = \mathbf{0}$ to challenge the algorithm. Moreover, the pressure is uniquely determined up to a constant, thus we assume $\int_\Omega p dxdy = 0$ at $t = 0$. However, this assumption is not essential. In this paper, we consider a two component displacement only. The coefficients can be stated as follows:

$$\begin{aligned} &c = c_1 = 1 - c_2, && a(c) = a(x,y,c) = \tfrac{\mu(c)}{\kappa}, \\ &b(c) = b(x,y,c) = \phi c_1\{z_1 - \sum_{j=1}^{2} z_j c_j\}, && d(c) = d(x,y,c) = \phi \sum_{j=1}^{2} z_j c_j, \end{aligned}$$

where c_i and z_i are the concentration and the compressibility factor of the ith component of the fluid mixture, respectively. In this problem, the Neumann boundary conditions are given as

$$\mathbf{u} \cdot \mathbf{n} = 0, \quad (\mathbf{D}\nabla c - c\mathbf{u}) \cdot \mathbf{n} = 0, \tag{1.3}$$

where $\mathbf{n}$ is the outer normal of the boundary $\partial\Omega$. Moreover, the initial solutions are given as

$$c(x,y,0) = c_0(x,y), \quad p(x,y,0) = p_0(x,y), \quad (x,y) \in \Omega.$$

The miscible displacement problem in porous media was first introduced in [9, 10], in which the mixed finite element methods were presented. Later, the compressible problem was studied in [11] and the optimal order estimates in L^2 and almost optimal order estimates in L^∞ were given in [5]. Subsequently, many new methods were introduced to solve the compressible miscible displacements, such as finite difference method [43, 44, 45], characteristics collocation method [21], splitting positive definite mixed element method [35] and H^1-Galerkin mixed method [3]. Besides the above, in [30], an accurate and efficient simulator is developed for problems with wells. Later, the authors introduced

an Eulerian-Lagrangian localized adjoint method (ELLAM) to solve the transport partial differential equation for concentration, while a mixed finite element method (MFEM) to solve the pressure equation [29]. Recently, discontinuous Galerkin (DG) methods became more popular in solving compressible miscible displacement problem in porous media [6, 7, 36, 38, 15, 37]. Some special numerical techniques were introduced to control jumps of numerical approximations as well as the nonlinearality of the convection term. Moreover, the superconvergence results based on the DG methods were also proved in [39, 40]. Besides the above, there were also significant works discussing the discontinuous Galerkin methods for incompressible miscible displacements, see e.g. [1, 16, 19, 23, 26, 27, 31] and for general porous media flow, see e.g. [2, 13, 12, 28] and the references therein. However, to the best knowledge of the authors, no previous works focused on the bound-preserving techniques. In many numerical simulations, the approximations of c can be placed out of the interval $[0, 1]$. Especially for problems with large gradients, the value of $d(c)$ might be negative, leading to ill-posedness of the problem, and the numerical approximations will blow up. We will use numerical experiments to demonstrate this point in Section 4. In this paper, we will discuss the bound-preserving technique for compressible problems. However, the idea can be extended to incompressible flows with some minor changes.

The DG method gained even greater popularity for good stability, high order accuracy, and flexibility on h-p adaptivity and on complex geometry. Physically bound-preserving high-order numerical methods for conservation laws have been actively studied in the last few years. Most cases, the DG schemes were coupled with suitable trouble cell indicators to avoid the order deterioration in regions where the solutions are smooth while maintaining physical solutions near shocks, e.g. [8, 18]. In 2010, the genuinely maximum-principle-satisfying high order DG and finite volume schemes were constructed in [46] by Zhang and Shu. Subsequently, this technique has been successfully extended to compressible Euler equations without or with source terms [47, 48], shallow water equations [32], and hyperbolic equations with δ-singularities [41]. The basic idea is to take the test function to be 1 in each cell to obtain an equation of the numerical cell average of the target variable, say r, and prove the cell average, $\bar{r}$, is within the desired bounds. Then we can apply the bound-preserving limiter to the numerical approximation and construct a new one

$$\tilde{r} = \bar{r} + \theta(r - \bar{r}), \quad \theta \in [0, 1]. \tag{1.4}$$

If the problem has only one lower bound zero, the technique is also called positivity-preserving technique. Thanks to the limiter, the whole algorithm were proved to be L^1 stable [42, 22] for some complicated systems. Moreover, the technique does not rely on the trouble cell detector and the limiter keeps the high-order accuracy in regions with smooth solutions for scalar equations [46]. For convection-diffusion equations, in [49], the authors applied the same idea to construct genuinely second-order maximum-principle-satisfying DG methods on unstructured meshes. Subsequently, the ideas have been extended to semilinear parabolic equations with blow-up solutions [17] and chemotaxis models [20] by using the LDG methods. Recently, the flux limiter [34, 33] and third-order maximum-preserving direct DG method [4] were also introduced. In this paper, we will extend the ideas in [46, 49] and

construct second-order bound-preserving DG methods. However, there are significant differences from previous techniques. First of all, most of the problems in [46, 49] satisfy maximum principles while the concentration c in (1.1) does not. Therefore, we will split the whole algorithm into two parts. We first rewrite the system into a conservative form and treat p_t as another source in the equation of concentration to obtain the positivity of c by the positivity-preserving technique [47, 48]. Then we choose a consistent flux pair (see Definition 2.1) in the concentration and pressure equations to obtain the positivity of $1 - c$. More precisely, in our analysis, instead of solving c and p, we rewrite (1.1) into a system of c and $c_2 = 1 - c$ and enforce $c + c_2 = 1$ by choosing a consistent flux pair. Secondly, to apply the positivity-preserving technique, we need to numerically approximate the conservative variable $r = \phi c$ instead of c. By doing so, the upper bound of r is not a constant and the limiter (1.4) may fail to work, since such a θ may not exist. Therefore, we will introduce another limiter to obtain physically relevant numerical approximations. Before we finish the introduction, we would like to point out the main difference between the proposed scheme and the TVD-like schemes. For TVD schemes, the numerical approximations are only first-order accurate at the local extrema. In this paper, our scheme is second-order accurate in L^∞-norm. In Section 4, we will construct analytical solutions to demonstrate the accuracy. To the best knowledge of the authors, this is the first analytical solution available.

The paper is organized as follows. In Sections 2 and 3, we present the bound-preserving technique for problems in one and two space dimensions. Analytical solutions will be constructed and some numerical results will be given to demonstrate the good performance of the bound-preserving DG method in Section 4. We will end in Section 5 with concluding remarks and remarks for future works.

2 Numerical algorithm in one space dimension

In this section, we proceed to construct the bound-preserving DG scheme to solve the one-dimensional compressible miscible displacement problem in porous media.

2.1 The DG scheme

We consider the one-dimensional version of (1.1) on the spatial domain $\Omega = [0, 2\pi]$ and solve

$$\begin{aligned} &d(c)p_t + u_x = d(c)p_t - (\tfrac{k(x)}{\mu(c)}p_x)_x = q, && x \in \Omega, \quad 0 < t \le T, \\ &\phi c_t + b(c)p_t + uc_x - (D(u)c_x)_x = (\tilde{c} - c)q, && x \in \Omega, \quad 0 < t \le T, \end{aligned}$$

with Neuamnn boundary conditions

$$u = 0, \quad D(u)c_x - cu = 0 \quad \text{on} \quad \partial\Omega. \tag{2.1}$$

We first rewrite the system into a conservative form

$$d(c)p_t + u_x = q, \tag{2.2}$$

$$a(c)u = -p_x, \tag{2.3}$$

$$(\phi c)_t + (uc)_x - (D(u)c_x)_x = \tilde{c}q - \phi c z_1 p_t, \tag{2.4}$$

where $a(c) = \frac{\mu(c)}{\kappa}$ and z_1 is the compressibility factor of first component of the fluid mixture. To construct the DG scheme, we divide the computational domain Ω into N cells

$$0 = x_{\frac{1}{2}} < x_{\frac{3}{2}} < \cdots < x_{N+\frac{1}{2}} = 2\pi,$$

and denote

$$I_j = \left(x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}\right), \quad j = 1, \cdots, N$$

as the cells. For simplicity, we consider uniform meshes in this paper, and denote by Δx the size of each cell. However, this assumption is not essential. Following [49], we consider second-order DG scheme only, and define

$$V_h = \left\{v \in L^2(\Omega) : v|_{I_j} \in \mathcal{P}^1(I_j), j = 1, \cdots, N\right\}$$

as the finite element space, where $\mathcal{P}^1(I_j)$ denotes the space of linear polynomials in I_j. Moreover, for any $v \in V_h$, we use $v^-_{j+\frac{1}{2}} = v(x^-_{j+\frac{1}{2}})$ to denote the left limit of v at $x_{j+\frac{1}{2}}$. Likewise for v^+. Then the jump and average of v at $x_{j+\frac{1}{2}}$ are given as $[v]_{j+\frac{1}{2}} = v^+_{j+\frac{1}{2}} - v^-_{j-\frac{1}{2}}$ and $\{v\}_{j+\frac{1}{2}} = \frac{1}{2}(v^+_{j+\frac{1}{2}} + v^-_{j-\frac{1}{2}})$, respectively. For simplicity, we define $v^-_{\frac{1}{2}} = v^+_{N+\frac{1}{2}} = 0$. Before constructing the bound-preserving DG scheme, we would like to demonstrate the following main points that are quite different from most of the previous works.

1. To approximate $r = \phi c$ directly instead of c. Due to the existence of ϕ in the time derivative in (2.4), it is not easy to extract the cell averages of c in DG methods (we cannot simply take the test function to be 1). Therefore, we introduce $r = \phi c$, and approximate r instead of c.

2. Treat p_t as a source in (2.4). We will first solve p_t in (2.2) and then use the positivity-preserving technique introduced in [47, 48] to construct positive numerical approximations of r.

3. Choose a consistent flux pair for (2.2) and (2.4). In Section 2.2, we will suitably choose the numerical fluxes and prove $\bar{r} < \bar{\phi}$, where $\bar{r}$ and $\bar{\phi}$ are the cell averages of r and ϕ, respectively.

4. The classical bound-preserving limiters (1.4) cannot be applied. Notice that, in the previous point, ϕ, the upper bound of r, may not be a constant. It is impossible to apply the bound-preserving limiter (1.4) to obtain desired new approximations between 0 and ϕ at the cell interfaces. For example, take $\phi(x) = 2 - x^2$ on [-1,1] and $\bar{r} = 1.1$. We cannot find any θ such that $0 \le \tilde{r}(x) \le \phi(x)$ at $x = \pm 1$.

5. In this paper, we would like to approximate ϕ by a piecewise linear approximation Φ, such that in cell I_j

   $$\Phi(x_{j\pm\frac{1}{2}}) = \phi(x_{j\pm\frac{1}{2}}),$$

   and we denote this interpolation operator to be $\mathcal{I}_1$. Notice that, Φ is a second-order approximation of ϕ. Since we use second-order DG method, the usage of Φ in the scheme will not kill the accuracy. Moreover, in this paper, we assume ϕ to be continuous, then Φ is also a continuous function and we denote $\Phi_{j+\frac{1}{2}} = \Phi(x_{j+\frac{1}{2}})$.

6. Introduce a new limiter. A new limiter will be given to keep the numerical cell average and modify the numerical approximations such that $0 \le r \le \Phi$ at the cell interfaces. Since r and Φ are both linear functions in each cell, then with the new limiter we have $0 \le r \le \Phi$ for all points in the computational domain, which further yields $c = \frac{r}{\Phi} \in [0, 1]$.

Now, we are ready to construct the DG scheme. For simplicity, if not otherwise stated, we use p, u, c, r as the numerical approximations from now on. then the DG scheme for (2.2)-(2.4) is the following: find $p, u, r \in V_h$ such that for any $\xi, \eta, \zeta \in V_h$

$$(\tilde{d}(r)p_t, \xi) = (u, \xi_x) + \sum_{j=1}^{N-1} \hat{u}[\xi]_{j+\frac{1}{2}} + (q, \xi)_j, \tag{2.5}$$

$$(a(c)u, \eta) = (p, \eta_x) + \sum_{j=0}^{N} \hat{p}[\eta]_{j+\frac{1}{2}}, \tag{2.6}$$

$$\begin{aligned}(r_t, \zeta) = &(uc - D(u)c_x, \zeta_x) + \sum_{j=1}^{N-1} \widehat{uc}[\zeta]_{j+\frac{1}{2}} - \sum_{j=1}^{N-1} \{D(u)c_x\}[\zeta]_{j+\frac{1}{2}} \\ &- \sum_{j=1}^{N-1} \{D(u)\zeta_x\}[c]_{j+\frac{1}{2}} - \sum_{j=1}^{N-1} \frac{\tilde{\alpha}}{h}[c][\zeta]_{j+\frac{1}{2}} + (\tilde{c}q - rz_1p_t, \zeta),\end{aligned} \tag{2.7}$$

where $c = \mathcal{I}_1\left\{\frac{r}{\Phi}\right\}$, $\tilde{d}(r) = z_1 r + z_2(\Phi(x) - r)$ and $(u, v) = \int_\Omega uvdx$. Here we take c as the piecewise linear approximation approximation of $\frac{r}{\Phi}$ such that in each cell I_j, we have

$$c^+_{j-\frac{1}{2}} = \frac{r^+_{j-\frac{1}{2}}}{\Phi_{j-\frac{1}{2}}}, \quad c^-_{j+\frac{1}{2}} = \frac{r^-_{j+\frac{1}{2}}}{\Phi_{j+\frac{1}{2}}}.$$

Remark 2.1. *In* (2.5)-(2.7)*, the usage of* $\tilde{d}(r)$ *instead of* $d(c)$ *is required by the positivity-preserving technique. In the proof of Theorem 2.2, we will subtract* (2.5) *from* (2.7) *to obtain a new source similar to the one in* (2.7)*. If we use* $d(c)$ *in* (2.5)*, then the source in* (2.7) *should be changed to* $c\Phi z_1 p_t$*.*

In (2.5)-(2.7), $\hat{u}$, $\hat{p}$ and $\widehat{uc}$ are the numerical fluxes. We use alternating fluxes for the diffusion term

$$\hat{u}_{j+\frac{1}{2}} = u^+_{j+\frac{1}{2}}, \quad \hat{p}_{j+\frac{1}{2}} = p^-_{j+\frac{1}{2}}, \quad j = 1, \cdots, N-1, \tag{2.8}$$

and at the boundary, we choose

$$\hat{p}_{\frac{1}{2}} = p^+_{\frac{1}{2}}, \quad \hat{p}_{N+\frac{1}{2}} = p^-_{N+\frac{1}{2}}.$$

For the convection term, we use the following flux

$$\widehat{uc} = u^+c^+ - \alpha[c], \tag{2.9}$$

Before we complete this subsection, we would like to introduce the following definition that will be used in the bound-preserving limiter.

Definition 2.1. *We say two fluxes* $\widehat{uc}$*,* $\hat{u}$ *are consistent if* $\widehat{uc} = \hat{u}$ *by taking* $c = 1$ *in* Ω*.*

The numerical flux pair $(\widehat{uc}, \hat{u})$ given in (2.8) and (2.9) are consistent, and this is required by the bound-preserving technique.

Remark 2.2. *There are plenty of flux pairs can be used following the procedures introduced in the next subsection. The proofs are basically the same with some minor changes, so we only list some of them below without more details.*

- $\hat{u} = u^-$, $\hat{p} = p^+$, $\widehat{uc} = u^- c^- - \alpha[c]$.
- $\hat{u} = \frac{1}{2}(u^+ + u^-)$, $\hat{p} = \frac{1}{2}(p^+ + p^-)$, $\widehat{uc} = \frac{1}{2}(u^+ c^+ + u^- c^-) - \alpha[c]$.

2.2 Bound-preserving technique

In this subsection, we consider Euler forward time discretization and apply the bound-preserving technique to construct physically relevant numerical approximations.

For simplicity, we use o_j for the numerical approximation o in I_j, and the cell average is $\bar{o}_j$ for $o = u, p, c, r$. Moreover, we use o^n to represent the solution o at time level n. Take $\zeta = 1$ in I_j in (2.7) for $j = 2, \cdots, N-1$ to obtain the equation satisfied by $\bar{r}$,

$$\bar{r}_j^{n+1} = H_j^c(r, u, c) + H_j^d(r, u, c) + H_j^s(r, \tilde{c}, q, z_1 p_t), \tag{2.10}$$

where

$$\begin{aligned}
H_j^c(r, u, c) &= \frac{1}{3}\bar{r}_j^n + \lambda\left(\widehat{uc}_{j-\frac{1}{2}} - \widehat{uc}_{j+\frac{1}{2}}\right), \\
H_j^d(r, u, c) &= \frac{1}{3}\bar{r}_j^n - \lambda\left(\{D(u)c_x\}_{j-\frac{1}{2}} + \frac{\tilde{\alpha}}{\Delta x}[c]_{j-\frac{1}{2}} - \{D(u)c_x\}_{j+\frac{1}{2}} - \frac{\tilde{\alpha}}{\Delta x}[c]_{j+\frac{1}{2}}\right), \\
H_j^s(r, \tilde{c}, q, z_1 p_t) &= \frac{1}{3}\bar{r}_j^n + \Delta t\overline{\tilde{c}q - r z_1 p_t},
\end{aligned}$$

with $\lambda = \frac{\Delta t}{\Delta x}$ being the ratio of the time and space mesh sizes, $\tilde{\alpha}$ being another parameter that will be chosen by the positivity-preserving technique, and $\overline{\tilde{c}q - r z_1 p_t}$ being the cell average of $\tilde{c}q - r z_1 p_t$. We will prove that if Δt is sufficiently small, then H_j^c, H_j^d and H_j^s are all positive, and the results are given in the following three lemmas. For simplicity of presentation, if the denominator in a fraction is zero, then the value of the fraction is defined as ∞. Let us consider H_j^c first.

Lemma 2.1. *Suppose $r^n > 0$ ($c^n > 0$), if α and λ satisfy*

$$\alpha > \max_{1 \le j \le N-1}\left\{u^+_{j+\frac{1}{2}}, 0\right\}, \quad \lambda \le \min_{2 \le j \le N-1}\left\{\frac{\Phi_m^0}{6\alpha}, \frac{\Phi_{j-\frac{1}{2}}}{6\left(\alpha - u^+_{j-\frac{1}{2}}\right)}\right\}, \tag{2.11}$$

where

$$\Phi_m^0 = \min_{2 \le j \le N-1} \Phi_{j+\frac{1}{2}},$$

then $H_j^c(r, u, c) > 0$.

Proof. It is easy to check

$$\begin{aligned} H_j^c(r,u,c) &= \frac{1}{3}\bar{r}_j^n + \lambda\left(\widehat{uc}_{j-\frac{1}{2}} - \widehat{uc}_{j+\frac{1}{2}}\right) \\ &= \frac{1}{6}\left(c^+\Phi_{j-\frac{1}{2}} + c^-\Phi_{j+\frac{1}{2}}\right) + \lambda\left(u^+c^+_{j-\frac{1}{2}} - \alpha[c]_{j-\frac{1}{2}} - u^+c^+_{j+\frac{1}{2}} + \alpha[c]_{j+\frac{1}{2}}\right) \\ &= \alpha\lambda c^-_{j-\frac{1}{2}} + \left(\frac{1}{6}\Phi_{j-\frac{1}{2}} + \lambda u^+_{j-\frac{1}{2}} - \alpha\lambda\right)c^+_{j-\frac{1}{2}} + \left(\frac{1}{6}\Phi_{j+\frac{1}{2}} - \alpha\lambda\right)c^-_{j+\frac{1}{2}} + \lambda\left(\alpha - u^+_{j+\frac{1}{2}}\right)c^+_{j+\frac{1}{2}}. \end{aligned}$$

Therefore, $H_j^c(r,u,c) > 0$, if α and λ satisfy condition (2.11). □

Now, we proceed to analyze H_j^d. Following the same analysis in [49] with some minor changes, we can show the following lemma.

Lemma 2.2. *Suppose $r^n > 0$ ($c^n > 0$), then $H_j^d(r,u,c) > 0$ under the condition*

$$\tilde{\alpha} > \frac{1}{2}D_M^0, \quad \Lambda = \frac{\Delta t}{\Delta x^2} \le \min_{2\le j\le N-1}\left\{\frac{\Phi_{j-\frac{1}{2}}}{3D^-_{j+\frac{1}{2}} + 6\tilde{\alpha}}, \frac{\Phi_{j+\frac{1}{2}}}{6\tilde{\alpha} + 3D^+_{j-\frac{1}{2}}}\right\}, \tag{2.12}$$

where

$$D_M^0 = \max_{2\le j\le N-1}\left\{D^-_{j-\frac{1}{2}}, D^+_{j+\frac{1}{2}}\right\}.$$

Proof. Notice that c is a linear function in each cell, then

$$c_x{}^-_{j-\frac{1}{2}} = \frac{c^-_{j-\frac{1}{2}} - c^+_{j-\frac{3}{2}}}{\Delta x}, \quad c_x{}^+_{j-\frac{1}{2}} = \frac{c^-_{j+\frac{1}{2}} - c^+_{j-\frac{1}{2}}}{\Delta x} = c_x{}^-_{j+\frac{1}{2}}, \quad c_x{}^+_{j+\frac{1}{2}} = \frac{c^-_{j+\frac{3}{2}} - c^+_{j+\frac{1}{2}}}{\Delta x}.$$

Therefore,

$$\begin{aligned} H_j^d(r,u,s) &= \frac{1}{3}\bar{r}_j^n - \lambda\left(\{D(u)c_x\}_{j-\frac{1}{2}} + \frac{\tilde{\alpha}}{\Delta x}[c]_{j-\frac{1}{2}} - \{D(u)c_x\}_{j+\frac{1}{2}} - \frac{\tilde{\alpha}}{\Delta x}[c]_{j+\frac{1}{2}}\right), \\ &= \frac{1}{6}\Phi_{j-\frac{1}{2}}c^+_{j-\frac{1}{2}} - \frac{1}{2}\Lambda D^-_{j-\frac{1}{2}}\left(c^-_{j-\frac{1}{2}} - c^+_{j-\frac{3}{2}}\right) - \frac{1}{2}\Lambda D^+_{j-\frac{1}{2}}\left(c^-_{j+\frac{1}{2}} - c^+_{j-\frac{1}{2}}\right) - \Lambda\tilde{\alpha}\left(c^+_{j-\frac{1}{2}} - c^-_{j-\frac{1}{2}}\right) \\ &\quad + \frac{1}{6}\Phi_{j+\frac{1}{2}}c^-_{j+\frac{1}{2}} + \frac{1}{2}\Lambda D^-_{j+\frac{1}{2}}\left(c^-_{j+\frac{1}{2}} - c^+_{j-\frac{1}{2}}\right) + \frac{1}{2}\Lambda D^+_{j+\frac{1}{2}}\left(c^-_{j+\frac{3}{2}} - c^+_{j+\frac{1}{2}}\right) + \Lambda\tilde{\alpha}\left(c^+_{j+\frac{1}{2}} - c^-_{j+\frac{1}{2}}\right) \\ &= \frac{1}{2}\Lambda D^-_{j-\frac{1}{2}}c^+_{j-\frac{3}{2}} + \Lambda\left(\tilde{\alpha} - \frac{1}{2}D^-_{j-\frac{1}{2}}\right)c^-_{j-\frac{1}{2}} + \left(\frac{1}{6}\Phi_{j-\frac{1}{2}} + \frac{1}{2}\Lambda D^+_{j-\frac{1}{2}} - \frac{1}{2}\Lambda D^-_{j+\frac{1}{2}} - \Lambda\tilde{\alpha}\right)c^+_{j-\frac{1}{2}} \\ &\quad + \left(\frac{1}{6}\Phi_{j+\frac{1}{2}} - \frac{1}{2}\Lambda D^+_{j-\frac{1}{2}} + \frac{1}{2}\Lambda D^-_{j+\frac{1}{2}} - \Lambda\tilde{\alpha}\right)c^-_{j+\frac{1}{2}} + \Lambda\left(\tilde{\alpha} - \frac{1}{2}D^+_{j+\frac{1}{2}}\right)c^+_{j+\frac{1}{2}} + \frac{1}{2}\Lambda D^+_{j+\frac{1}{2}}c^-_{j+\frac{3}{2}}, \end{aligned}$$

where $D^\pm_{j-\frac{1}{2}} = D(u^\pm_{j-\frac{1}{2}}) \ge 0$. Therefore, we have $H_j^d(r,u,s) > 0$ if (2.12) is satisfied. □

Next, we proceed to study H_j^s. We use Gaussian quadrature with two points to approximate the cell average of the source. The quadrature points on I_j are denoted as x_j^i, $i = 1, 2$. Also, we denote w_i as the corresponding weights on the interval $[-\frac{1}{2}, \frac{1}{2}]$. Then we can state the result.

Lemma 2.3. *Suppose $r^n > 0$ ($c^n > 0$), then $H_j^s(r, \tilde{c}, q, z_1 p_t) > 0$ under the conditions*

$$\Delta t \le \frac{1}{6 z_1 p_M}, \quad \Delta \le \min_{1\le j\le N} \frac{\min\{\Phi_{j-\frac{1}{2}}, \Phi_{j+\frac{1}{2}}\}}{6\max\left\{-q(x_j^1), -q(x_j^2), 0\right\}}. \tag{2.13}$$

where

$$p_M = \max_{i,j} \left(p_t(x_j^i), 0\right). \tag{2.14}$$

Proof. The cell average of the source is approximated by the Gaussian quadrature, then

$$H_j^s(r, u, c) = \frac{1}{3}\bar{r}_j^n + \Delta t\overline{\tilde{c}q - r z_1 p_t} = \sum_{i=1}^{2} w_i (L_i^1 + L_i^2),$$

where

$$L_i^1 = \left(\frac{1}{6} - \Delta t z_1 p_t(x_j^i)\right) r_j(x_j^i), \quad L_i^2 = \frac{1}{6} r_j(x_j^i) + \Delta t \tilde{c} q(x_j^i).$$

Clearly, we have $L_i^1 \ge 0$ if (2.13) is satisfied. Moreover, $L_i^2 > 0$ if $q(x_j^i) \ge 0$. We only need to consider the case with $q(x_j^i) < 0$. Without loss of generality, we assume $q(x_j^1) < 0$, then $\tilde{c}(x_j^1) = c(x_j^1)$. Notice that r and c are both linear functions in I_j then it is easy to check

$$r(x_j^1) = \mu_1 r^+_{j-\frac{1}{2}} + \mu_2 r^-_{j+\frac{1}{2}}, \quad c(x_j^1) = \mu_1 c^+_{j-\frac{1}{2}} + \mu_2 c^-_{j+\frac{1}{2}},$$

with $\mu_1 = \frac{3+\sqrt{3}}{6}$ and $\mu_2 = \frac{3-\sqrt{3}}{6}$. Therefore

$$\begin{aligned} L_1^2 &= \frac{1}{6}(\mu_1 r^+_{j-\frac{1}{2}} + \mu_2 r^-_{j+\frac{1}{2}}) + \Delta t(\mu_1 c^+_{j-\frac{1}{2}} + \mu_2 c^-_{j+\frac{1}{2}}) q(x_j^1) \\ &= \left(\frac{1}{6}\Phi_{j-\frac{1}{2}} + \Delta t q(x_j^1)\right) \mu_1 c^+_{j-\frac{1}{2}} + \left(\frac{1}{6}\Phi_{j+\frac{1}{2}} + \Delta t q(x_j^1)\right) \mu_2 c^-_{j+\frac{1}{2}}. \end{aligned}$$

then $L_1^2 > 0$ under (2.13). □

Now we study the cells near $\partial\Omega$. For simplicity of presentation, we only demonstrate the results below, and the proof can be obtained exactly the same way in Lemmas 2.1, 2.2 and 2.3 with some minor changes.

Lemma 2.4. *Suppose $r_1^n > 0$ ($c_1^n > 0$), then $\bar{r}_1^{n+1} > 0$ under conditions* (2.13) *and*

$$\alpha > u^+_{\frac{3}{2}}, \quad \lambda \le \frac{\Phi_{\frac{3}{2}}}{6\alpha}, \quad \textit{and} \quad \tilde{\alpha} > \frac{1}{2} D^+_{\frac{3}{2}}, \quad \Lambda \le \min\left\{\frac{\Phi_{\frac{1}{2}}}{3D^-_{\frac{3}{2}}}, \frac{\Phi_{\frac{3}{2}}}{6\tilde{\alpha}}\right\}.$$

Similarly, suppose $r_N^n > 0$ ($c_N^n > 0$), then $\bar{r}_N^{n+1} > 0$ under conditions (2.13) *and*

$$\lambda \le \frac{\Phi_{N-\frac{1}{2}}}{6\left(\alpha - u^+_{N-\frac{1}{2}}\right)}, \quad \textit{and} \quad \tilde{\alpha} > \frac{1}{2} D^-_{N-\frac{1}{2}}, \quad \Lambda \le \min\left\{\frac{\Phi_{N-\frac{1}{2}}}{6\tilde{\alpha}}, \frac{\Phi_{N+\frac{1}{2}}}{3D^+_{N-\frac{1}{2}}}\right\}.$$

Based on the above lemmas, we can state the following theorem.

Theorem 2.1. *Consider the DG scheme* (2.5)-(2.7) *with Euler forward time discretization. Suppose* $r^n > 0$ ($c^n > 0$), *and the parameters* α *and* $\tilde{\alpha}$ *are taken to be*

$$\alpha > \max_{1\le j\le N-1}\left\{u^+_{j+\frac{1}{2}}, 0\right\}, \quad \tilde{\alpha} > \frac{1}{2}D_M,$$

where

$$D_M = \max_{1\le j\le N-1}\left\{D^{\pm}_{j+\frac{1}{2}}\right\}.$$

Then $\bar{r}^{n+1}_j > 0$ *under* (2.13) *and*

$$\lambda \le \min_{1\le j\le N-1}\left\{\frac{\Phi_m}{6\alpha}, \frac{\Phi_{j+\frac{1}{2}}}{6\left(\alpha - u^+_{j+\frac{1}{2}}\right)}\right\}, \quad \Lambda \le \min_{1\le j\le N}\left\{\frac{\Phi_{j-\frac{1}{2}}}{6\tilde{\alpha}+3D^-_{j+\frac{1}{2}}}, \frac{\Phi_{j+\frac{1}{2}}}{6\tilde{\alpha}+3D^+_{j-\frac{1}{2}}}\right\},$$

where

$$\Phi_m = \min_{1\le j\le N-1}\Phi_{j+\frac{1}{2}}.$$

Proof. For $j = 0$ and N, the results was proved in Lemma 2.4. For $j = 1, 2, \cdots, N-1$, by (2.10), we only need H^c_j, H^d_j and H^s_j to be positive, which follow directly from Lemmas 2.1, 2.2 and 2.3, respectively. □

The above theorem guarantees the positivity of $\bar{r}^{n+1}$. However, we still need to show $\bar{r}^{n+1} \le \bar{\Phi}$, and the result is given below.

Theorem 2.2. *Suppose the conditions in Theorem 2.1 are satisfied. Moreover, we assume* $0 \le r^n \le \Phi$ *and the flux pair* $(\widehat{uc}, \hat{u})$ *are consistent, then* $0 \le \bar{r}^{n+1} \le \bar{\Phi}$, *under another condition*

$$\Delta t \le \frac{1}{6z_2p_M},$$

where p_M *was given in* (2.14).

Proof. Since the fluxes $\widehat{uc}$ and $\hat{u}$ are consistent, then $\widehat{uc} + \widehat{uc_2} = \hat{u}$, where $c_2 = 1 - c$. Take $\xi = \zeta$ in (2.5), and subtract (2.7) from (2.5) to obtain

$$\begin{aligned}(r_{2t}, \zeta) = (uc_2 - D(u)c_{2x}, \zeta_x) &+ \sum_{j=1}^{N-1}\widehat{uc_2}[\zeta]_{j+\frac{1}{2}} - \sum_{j=1}^{N-1}\{D(u)c_{2x}\}[\zeta]_{j+\frac{1}{2}} \\ &- \sum_{j=1}^{N-1}\{D(u)\zeta_x\}[c_2]_{j+\frac{1}{2}} - \sum_{j=1}^{N-1}\frac{\tilde{\alpha}}{h}[c_2][\zeta]_{j+\frac{1}{2}} + (\tilde{c}_2 q - r_2 z_2 p_t, \zeta), \end{aligned} \tag{2.15}$$

where $r_2 = \Phi - r$ and $\tilde{c_2} = 1 - \tilde{c}$. It is easy to check that (2.15) is exactly (2.7) with r, c and z_1 replaced by r_2, c_2 and z_2, respectively. Following the same analyses in Theorem 2.1, we can show that $\bar{r}^{n+1}_2 > 0$ under the conditions given in this theorem, which further implies $\bar{r}^{n+1} < \bar{\Phi}$. □

With Theorems 2.2, the numerical cell average $\bar{r}^n$ we constructed is between 0 and $\bar{\Phi}$. However, the numerical approximation r^n may be negative or larger than Φ. Therefore, we also need to apply some limiter to r^n and the procedure is given below. For simplicity, we drop the superscript n in all the numerical approximations.

1. Set up a small number $\epsilon = 10^{-13}$.

2. If $\bar{r} \le \epsilon$, take $r = \bar{r}$. If $\bar{r}_2 \le \epsilon$, take $r = \Phi - \bar{r}_2$. Then skip the following steps.

3. Define $m_j = \min\{r^+_{j-\frac{1}{2}}, r^-_{j+\frac{1}{2}}\}$. If $m_j \ge 0$ then proceed to the next step. Otherwise, without loss of generality, we assume $r^+_{j-\frac{1}{2}} < 0$, then define

$$\tilde{r}^+_{j-\frac{1}{2}} = \epsilon, \quad \tilde{r}^-_{j+\frac{1}{2}} = r^-_{j+\frac{1}{2}} - \epsilon + r^+_{j-\frac{1}{2}}$$

   and use $\tilde{r}^+_{j-\frac{1}{2}}$ and $\tilde{r}^-_{j+\frac{1}{2}}$ to construct a new approximation, also denoted as r.

4. Apply the previous step for $r_2 = \Phi - r$ and construct a new approximation r.

Remark 2.3. *The above algorithm keeps the second-order accuracy. In step 3, we assume* $r^+_{j-\frac{1}{2}} < 0$ *and* $0 \le \Phi_{j-\frac{1}{2}} - r^+_{j-\frac{1}{2}} \le Ch^2$*, then it is very easy to check* $0 \le \Phi_{j-\frac{1}{2}} - \tilde{r}^+_{j-\frac{1}{2}} \le Ch^2$.

Remark 2.4. *For an n-component fluid* ($n \ge 3$)*, we denote the concentration of the i'th component to be* c_i*, and let* $r_i = \phi c_i$*. Follow the same analysis in Theorem 2.2, we can show* $\bar{r}_i \ge 0, i = 1, 2, \cdots, n$ *and* $\sum_{i=1}^n \bar{r}_i = \bar{\Phi}$*. However, the limiter is not easy to construct. We will work on this problem in the future.*

Remark 2.5. *The extension to high-order schemes is not straightforward as demonstrated in [49]. Recently, Chen [4] introduced the third-order maximum-principle preserving direct DG methods for convection-diffusion equations which can be applied to this problem. Another approach is to applied the flux limiter [34, 33]. However, the accuracy was demonstrated by numerical experiments only.*

2.3 High order time discretizations

All the previous analyses are based on first-order Euler forward time discretization. We can also use strong stability preserving (SSP) high-order time discretizations to solve the ODE system $\mathbf{w}_t = \mathbf{L}\mathbf{w}$. More details of these time discretizations can be found in [25, 24, 14]. In this paper, we use the third-order SSP Runge-Kutta method [25]

$$\begin{aligned} \mathbf{w}^{(1)} &= \mathbf{w}^n + \Delta t \mathbf{L}(\mathbf{w}^n), \\ \mathbf{w}^{(2)} &= \frac{3}{4}\mathbf{w}^n + \frac{1}{4}\left(\mathbf{w}^{(1)} + \Delta t \mathbf{L}(\mathbf{w}^{(1)})\right), \\ \mathbf{w}^{n+1} &= \frac{1}{3}\mathbf{w}^n + \frac{2}{3}\left(\mathbf{w}^{(2)} + \Delta t \mathbf{L}(\mathbf{w}^{(2)})\right), \end{aligned} \tag{2.16}$$

and the third order SSP multi-step method [24]

$$\mathbf{w}^{n+1} = \frac{16}{27}\left(\mathbf{w}^n + 3\Delta t \mathbf{L}(\mathbf{w}^n)\right) + \frac{11}{27}\left(\mathbf{w}^{n-3} + \frac{12}{11}\Delta t \mathbf{L}(\mathbf{w}^{n-3})\right). \tag{2.17}$$

Since an SSP time discretization is a convex combination of Euler forward and the algorithm is to construct positive r and $\Phi - r$, hence by using the limiter designed in section 2.2, the numerical solution obtained from the full scheme is also physically relevant.

3 Numerical algorithm in two space dimensions

In this section, we proceed to construct the bound-preserving DG scheme for the compressible miscible displacement in porous media in two space dimensions.

3.1 The DG scheme

We first explain the notations that will be used in this section. Denote $\Omega = [0, 2\pi] \times [0, 2\pi]$ to be the computational domain. In this paper, we consider rectangular meshes only, and the techniques for triangular meshes will be discussed in the future. Let $0 = x_{\frac{1}{2}} < \cdots < x_{N_x+\frac{1}{2}} = 2\pi$ and $0 = y_{\frac{1}{2}} < \cdots < y_{N_y+\frac{1}{2}} = 2\pi$ be the grid points in x and y directions, respectively. Define $I_i = (x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}})$ and $J_j = (y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}})$. Let $K_{ij} = I_i \times J_j$, $i = 1, \cdots, N_x,\ j = 1, \cdots, N_y$, be a partition of Ω and denote $\Omega_h = \{K_{ij}\}$. For simplicity, if not otherwise stated, we always use K to denote the cell. We use Γ for all the cell interfaces, and $\Gamma_0 = \Gamma \setminus \partial\Omega$ for all the interior ones. For any $e \in \Gamma$, denote $|e|$ to be the length of e. Moreover, we define $\mathbf{n}_e = \mathbf{n}_x = (1, 0)$ if e is parallel to the y-axis while $\mathbf{n}_e = \mathbf{n}_y = (0, 1)$ if e is parallel to the x-axis. Further more, we denote $\partial\Omega_+ = \{e \in \partial\Omega : \mathbf{n}_e = \mathbf{n}\}$, where $\mathbf{n}$ is the unit outer normal of $\partial\Omega$, and $\partial\Omega_- = \partial\Omega \setminus \partial\Omega_+$. The mesh sizes in the x and y directions are given as $\Delta x_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$ and $\Delta y_j = y_{j+\frac{1}{2}} - y_{j-\frac{1}{2}}$, respectively. For simplicity, we assume uniform meshes and denote $\Delta x = \Delta x_i$ and $\Delta y = \Delta y_j$. However, this assumption is not essential. Following [49], we use second-order DG scheme. For rectangular meshes, it is impossible to construct a continuous piecewise linear interpolation of ϕ by using P^1 polynomials. Therefore, the finite element space is chosen as

$$W_h = \{z : z|_K \in Q^1(K), \forall K \in \Omega_h\},$$

where $Q^1(K)$ denotes the space of tensor product of linear polynomials in K. Given $v \in W_h$, we denote $v^+_{i-\frac{1}{2},j}$, $v^-_{i+\frac{1}{2},j}$, $v^+_{i,j-\frac{1}{2}}$, $v^-_{i,j+\frac{1}{2}}$ to be the traces of v on the four edges of K_{ij}, respectively, and use $[v] = v^+ - v^-$ and $\{v\} = \frac{1}{2}(v^+ + v^-)$ as the jump and average of v at the cell interfaces, respectively. For simplicity, for any $e \in \partial\Omega_-$, we define $v^-|_e = 0$. Similarly, for any $e \in \partial\Omega_+$, we define $v^+|_e = 0$. Moreover, we denote $\Phi \in W_h$ to be an interpolation of ϕ such that in each cell K, $\Phi = \phi$ at the four corners, and we denote this interpolation operator to be $\mathcal{I}_1$. It is easy to see that Φ is a globally continuous function on Ω and, for simplicity, define $\Phi_{i+\frac{1}{2},j+\frac{1}{2}} = \Phi(x_{i+\frac{1}{2}}, y_{j+\frac{1}{2}}) = \phi(x_{i+\frac{1}{2}}, y_{j+\frac{1}{2}})$ for all $i = 0, \cdots, N_x,\ j = 0, \cdots, N_y$.

We also use $p, c, \mathbf{u}$ as the numerical approximations, then the DG scheme is to find $p, r \in W_h$ and

$\mathbf{u} \in \mathbf{W}_h = W_h \times W_h$ such that for any $\xi, \zeta \in W_h$ and $\boldsymbol{\eta} \in \mathbf{W}_h$,

$$(\tilde{d}(r)p_t, \xi) = (\mathbf{u}, \nabla\xi) + \sum_{e\in\Gamma_0} \int_e \hat{\mathbf{u}}[\xi] \cdot \mathbf{n}_e ds + (q, \xi), \tag{3.1}$$

$$(a(c)\mathbf{u}, \boldsymbol{\eta}) = (p, \nabla \cdot \boldsymbol{\eta}) + \sum_{e\in\Gamma} \int_e \hat{p}[\boldsymbol{\eta} \cdot \mathbf{n}_e] ds, \tag{3.2}$$

$$\begin{aligned}(r_t, \zeta) = &(\mathbf{u}c - \mathbf{D}(\mathbf{u})\nabla c, \nabla\zeta) + (\tilde{c}q - rz_1p_t, \zeta) + \sum_{e\in\Gamma_0} \int_e \widehat{\mathbf{u}c} \cdot \mathbf{n}_e[\zeta] ds \\ &- \sum_{e\in\Gamma_0} \int_e \left(\{\mathbf{D}(\mathbf{u})\nabla c \cdot \mathbf{n}_e\} [\zeta] + \{\mathbf{D}(\mathbf{u})\nabla\zeta \cdot \mathbf{n}_e\} [c] + \frac{\tilde{\alpha}}{|e|}[c][\zeta] \right) ds,\end{aligned} \tag{3.3}$$

where

$$c = \mathcal{I}_1 \left\{ \frac{r}{\Phi} \right\}, \qquad \tilde{d}(r) = z_1 r + z_2(\Phi - r), \qquad (u, v) = \int_\Omega uv dx dy.$$

In (3.1)-(3.3), $\hat{p}$, $\hat{\mathbf{u}}$ and $\widehat{\mathbf{u}c}$ are the numerical fluxes. We also use alternating fluxes for the diffusion terms, and for any $e \in \Gamma_0$

$$\hat{\mathbf{u}}|_e = \mathbf{u}^+|_e, \quad \hat{p}|_e = p^-|_e,$$

and on $\partial\Omega$, we take

$$\hat{p}|_e = p^-|_e, \ \forall\ e \in \partial\Omega_+, \quad \hat{p}|_e = p^+|_e, \ \forall\ e \in \partial\Omega_-.$$

For the convection term, we use

$$\widehat{\mathbf{u}c} = \mathbf{u}^+ c^+ - \alpha[c], \tag{3.4}$$

which is consistent with $\hat{\mathbf{u}}$. Here α and $\tilde{\alpha}$ are two positive constants to be chosen by the bound-preserving technique. Similar to Remark 2.2, there are some other choices of the numerical fluxes, yet we will not discuss them in detail.

3.2 Bound-preserving technique

In this section, we apply the bound-preserving technique to the DG scheme, and construct physically relevant numerical approximations. We consider Euler-forward time discretization only, and the high-order ones were discussed in Section 2.3. For simplicity of presentation, we only discuss the techniques for cells away from $\partial\Omega$, while the boundary cells can be analyzed following the same lines. We use o_{ij} for the numerical approximation o in K_{ij} and the cell average is $\bar{o}_{ij}$. Moreover, we use o^n to represent the solution o at time level n.

In (3.3), we take $\zeta = 1$ in K_{ij} to obtain the equation satisfied by the cell average of r,

$$\bar{r}_{ij}^{n+1} = H^c(r, \mathbf{u}, c) + H_x^d(r, \mathbf{u}, c) + H_y^d(r, \mathbf{u}, c) + H^s(r, \tilde{c}, q, z_1p_t), \tag{3.5}$$

where

$$
\begin{aligned}
H^c(r,\mathbf{u},c) &= \frac{1}{3}\bar{r}_{ij}^n + \lambda\left(\int_{J_j} \widehat{u_1 c}_{i-\frac{1}{2},j} - \widehat{u_1 c}_{i+\frac{1}{2},j}dy + \int_{I_i} \widehat{u_2 c}_{i,j-\frac{1}{2}} - \widehat{u_2 c}_{i,j+\frac{1}{2}}dx\right),\\
H_x^d(r,\mathbf{u},c) &= \frac{1}{6}\bar{r}_{ij}^n - \lambda\left(\int_{J_j} \{D_{11}c_x + D_{12}c_y\}_{i-\frac{1}{2},j} + \frac{\tilde{\alpha}}{|J_j|}[c]_{i-\frac{1}{2},j}\right.\\
&\quad \left. - \{D_{11}c_x + D_{12}c_y\}_{i+\frac{1}{2},j} - \frac{\tilde{\alpha}}{|J_j|}[c]_{i+\frac{1}{2},j}dy\right),\\
H_y^d(r,\mathbf{u},c) &= \frac{1}{6}\bar{r}_{ij}^n - \lambda\left(\int_{I_i} \{D_{21}c_x + D_{22}c_y\}_{i,j-\frac{1}{2}} + \frac{\tilde{\alpha}}{|I_i|}[c]_{i,j-\frac{1}{2}}\right.\\
&\quad \left. - \{D_{21}c_x + D_{22}c_y\}_{i,j+\frac{1}{2}} - \frac{\tilde{\alpha}}{|I_i|}[c]_{i,j+\frac{1}{2}}dy\right),\\
H^s(r,\tilde{c},q,z_1 p_t) &= \frac{1}{3}\bar{r}_{ij}^n + \Delta t\overline{\tilde{c}q - r z_1 p_t},
\end{aligned}
$$

with $\lambda = \frac{\Delta t}{\Delta x \Delta y}$ and $\mathbf{u} = (u_1, u_2)^T$. To continue, we use 2-point Gaussian quadratures to approximate the integrals given above. The Gaussian quadrature points on I_i and J_j are denoted by $\{x_i^1, x_i^2\}$ and $\left\{y_j^1, y_j^2\right\}$, respectively. The corresponding weights on the interval $[-\frac{1}{2}, \frac{1}{2}]$ are denoted as w_1 and w_2. Denote $\lambda_1 = \frac{\Delta t}{\Delta x}$ and $\lambda_2 = \frac{\Delta t}{\Delta y}$, then

$$
H^c(r,\mathbf{u},c) = \frac{1}{3}\bar{r}_{ij}^n + \lambda_1 \sum_{\beta=1}^{2} w_\beta \left[\widehat{u_1 c}_{i-\frac{1}{2},j,\beta} - \widehat{u_1 c}_{i+\frac{1}{2},j,\beta}\right] + \lambda_2 \sum_{\beta=1}^{2} w_\beta \left[\widehat{u_2 c}_{i,j-\frac{1}{2},\beta} - \widehat{u_2 c}_{i,j+\frac{1}{2},\beta}\right],
$$

where $\widehat{u_1 c}_{i-\frac{1}{2},j,\beta} = \widehat{u_1 c}_{i-\frac{1}{2},j}(y_j^\beta)$ is a point value in the Gaussian quadrature on the edge $e = \{x_{i-\frac{1}{2}}\} \times J_j$. Likewise for the other point values. As the general treatment, we rewrite the cell average on the right hand side as

$$
\bar{r}_{ij}^n = \sum_{\beta=1}^{2} \frac{w_\beta}{2}\left(r^+_{i-\frac{1}{2},j,\beta} + r^-_{i+\frac{1}{2},j,\beta}\right) = \sum_{\beta=1}^{2} \frac{w_\beta}{2}\left(r^+_{i,j-\frac{1}{2},\beta} + r^-_{i,j+\frac{1}{2},\beta}\right).
$$

Based on the above,

$$
\begin{aligned}
H^c(r,\mathbf{u},c) &= \sum_{\beta=1}^{2} w_\beta \left[\frac{\lambda_1}{6(\lambda_1+\lambda_2)}\left(r^+_{i-\frac{1}{2},j,\beta} + r^-_{i+\frac{1}{2},j,\beta}\right) + \lambda_1\left(\widehat{u_1 c}_{i-\frac{1}{2},j,\beta} - \widehat{u_1 c}_{i+\frac{1}{2},j,\beta}\right)\right]\\
&\quad + \sum_{\beta=1}^{2} w_\beta \left[\frac{\lambda_2}{6(\lambda_1+\lambda_2)}\left(r^+_{i,j-\frac{1}{2},\beta} + r^-_{i,j+\frac{1}{2},\beta}\right) + \lambda_2\left(\widehat{u_2 c}_{i,j-\frac{1}{2},\beta} - \widehat{u_2 c}_{i,j+\frac{1}{2},\beta}\right)\right].
\end{aligned}
$$

Following the same proof in Lemma 2.1 with some minor changes, we have

Lemma 3.1. *Suppose $r^n > 0$ ($c^n > 0$) and the parameter α is taken to be*

$$
\alpha > \max_{\substack{1 \le i \le N_x - 1,\\ 1 \le j \le N_y - 1,\\ \beta = 1,2}} \{{u_1}^+_{i+\frac{1}{2},j,\beta}, {u_2}^+_{i,j+\frac{1}{2},\beta}, 0\}, \tag{3.6}
$$

then $H^c(r, \mathbf{u}, c) > 0$ *under the condition*

$$\frac{\Delta t}{\Delta x} + \frac{\Delta t}{\Delta y} \leq \min\left\{\frac{\Phi_m}{6\alpha}, B_1, B_2\right\} \tag{3.7}$$

where

$$B_1 = \min_{\substack{1 \leq i \leq N_x - 1 \\ 1 \leq j \leq N_y \\ \beta = 1,2}} \frac{\Phi_{i+\frac{1}{2}, j\pm\frac{1}{2}}}{6\left(\alpha - u^+_{i+\frac{1}{2},j,\beta}\right)}, \quad B_2 = \min_{\substack{1 \leq i \leq N_x \\ 1 \leq j \leq N_y - 1 \\ \beta = 1,2}} \frac{\Phi_{i\pm\frac{1}{2}, j+\frac{1}{2}}}{6\left(\alpha - u^+_{i,j+\frac{1}{2},\beta}\right)},$$

with

$$\Phi_m = \min_{\substack{0 \leq i \leq N_x \\ 0 \leq j \leq N_y}} \Phi_{i+\frac{1}{2}, j+\frac{1}{2}}.$$

For H^d_x and H^d_y, the analyses are similar.

Lemma 3.2. *Suppose* $r^n > 0$ $(c^n > 0)$*, then* $H^d_x(r, \mathbf{u}, c) > 0$ *under the conditions*

$$\tilde{\alpha} \geq \frac{\Delta y}{2\Delta x} D^M_{11} + \sqrt{3} D^M_{12}, \tag{3.8}$$

$$D^M_{11}\Lambda_1 + 2(\tilde{\alpha} + D^M_{12})\lambda \leq \frac{1}{12}\Phi_m, \tag{3.9}$$

where

$$D^M_{11} = \max_{(x,y)\in\Omega} D_{11}(\mathbf{u})(x,y), \quad D^M_{12} = \max_{(x,y)\in\Omega} |D_{12}(\mathbf{u})(x,y)|, \quad \Lambda_1 = \frac{\Delta t}{\Delta x^2}.$$

Similarly, we have $H^d_y(r, \mathbf{u}, c) > 0$ *if*

$$\tilde{\alpha} \geq \frac{\Delta x}{2\Delta y} D^M_{22} + \sqrt{3} D^M_{21}, \tag{3.10}$$

$$D^M_{22}\Lambda_2 + 2(\tilde{\alpha} + D^M_{21})\lambda \leq \frac{1}{12}\Phi_m, \tag{3.11}$$

where

$$D^M_{22} = \max_{(x,y)\in\Omega} D_{22}(\mathbf{u})(x,y), \quad D^M_{21} = \max_{(x,y)\in\Omega} |D_{21}(\mathbf{u})(x,y)|, \quad \Lambda_2 = \frac{\Delta t}{\Delta y^2}.$$

Proof. We only prove for H^d_x. Moreover, we simplify the subscript of the point value in a Gaussian quadrature, and denote $c^+_{i-\frac{1}{2},\beta}$ for $c^+_{i-\frac{1}{2},j,\beta}$. Likewise for the others. Notice the fact that, c is linear along each axis. Therefore,

$$c_x{}^-_{i-\frac{1}{2},\beta} = \frac{1}{\Delta x}\left(c^-_{i-\frac{1}{2},\beta} - c^+_{i-\frac{3}{2},\beta}\right), \quad c_x{}^+_{i-\frac{1}{2},\beta} = \frac{1}{\Delta x}\left(c^-_{i+\frac{1}{2},\beta} - c^+_{i-\frac{1}{2},\beta}\right),$$

$$c_y{}^-_{i-\frac{1}{2},1} = \frac{\sqrt{3}}{\Delta y}\left(c^-_{i-\frac{1}{2},2} - c^-_{i-\frac{1}{2},1}\right) = c_y{}^-_{i-\frac{1}{2},2}, \quad c_y{}^+_{i-\frac{1}{2},1} = \frac{\sqrt{3}}{\Delta y}\left(c^+_{i-\frac{1}{2},2} - c^+_{i-\frac{1}{2},1}\right) = c_y{}^+_{i-\frac{1}{2},2}.$$

We also use 2-point Gaussian quadratures to approximate the integral, then

$$
\begin{aligned}
H_x^d(r,u,c) = & \frac{1}{6}\bar{r}_{ij}^n - \frac{\lambda_1}{2}\sum_{\beta=1}^{2} w_\beta \left(\frac{{D_{11}}^-_{i-\frac{1}{2},\beta}}{\Delta x}\left(c^-_{i-\frac{1}{2},\beta} - c^+_{i-\frac{3}{2},\beta}\right) + \frac{{D_{11}}^+_{i-\frac{1}{2},\beta}}{\Delta x}\left(c^-_{i+\frac{1}{2},\beta} - c^+_{i-\frac{1}{2},\beta}\right)\right) \\
& - \frac{\sqrt{3}\lambda_1}{2}\sum_{\beta=1}^{2} w_\beta \left(\frac{{D_{12}}^-_{i-\frac{1}{2},\beta}}{\Delta y}\left(c^-_{i-\frac{1}{2},2} - c^-_{i-\frac{1}{2},1}\right) + \frac{{D_{12}}^+_{i-\frac{1}{2},\beta}}{\Delta y}\left(c^+_{i-\frac{1}{2},2} - c^+_{i-\frac{1}{2},1}\right)\right) \\
& + \frac{\lambda_1}{2}\sum_{\beta=1}^{2} w_\beta \left(\frac{{D_{11}}^-_{i+\frac{1}{2},\beta}}{\Delta x}\left(c^-_{i+\frac{1}{2},\beta} - c^+_{i-\frac{1}{2},\beta}\right) + \frac{{D_{11}}^+_{i+\frac{1}{2},\beta}}{\Delta x}\left(c^-_{i+\frac{3}{2},\beta} - c^+_{i+\frac{1}{2},\beta}\right)\right) \\
& + \frac{\sqrt{3}\lambda_1}{2}\sum_{\beta=1}^{2} w_\beta \left(\frac{{D_{12}}^-_{i+\frac{1}{2},\beta}}{\Delta y}\left(c^-_{i+\frac{1}{2},2} - c^-_{i+\frac{1}{2},1}\right) + \frac{{D_{12}}^+_{i+\frac{1}{2},\beta}}{\Delta y}\left(c^+_{i+\frac{1}{2},2} - c^+_{i+\frac{1}{2},1}\right)\right) \\
& - \lambda_1\sum_{\beta=1}^{2} w_\beta \frac{\tilde{\alpha}}{\Delta y}\left(c^+_{i-\frac{1}{2},\beta} - c^-_{i-\frac{1}{2},\beta}\right) + \lambda_1\sum_{\beta=1}^{2} w_\beta \frac{\tilde{\alpha}}{\Delta y}\left(c^+_{i+\frac{1}{2},\beta} - c^-_{i+\frac{1}{2},\beta}\right) \\
= & \frac{1}{6}\bar{r}_{ij}^n + \Lambda_1\sum_{\beta=1}^{2}\frac{w_\beta}{2}{D_{11}}^-_{i-\frac{1}{2},\beta}c^+_{i-\frac{3}{2},\beta} + \Lambda_1\sum_{\beta=1}^{2}\frac{w_\beta}{2}{D_{11}}^+_{i+\frac{1}{2},\beta}c^-_{i+\frac{3}{2},\beta} \\
& + \sum_{\beta=1}^{2}\frac{w_\beta}{2}\left(\tau^-_{i-\frac{1}{2},\beta}c^-_{i-\frac{1}{2},\beta} + \tau^+_{i-\frac{1}{2},\beta}c^+_{i-\frac{1}{2},\beta} + \tau^-_{i+\frac{1}{2},\beta}c^-_{i+\frac{1}{2},\beta} + \tau^+_{i+\frac{1}{2},\beta}c^+_{i+\frac{1}{2},\beta}\right),
\end{aligned}
$$

where

$$
\begin{aligned}
\tau^-_{i-\frac{1}{2},1} &= -\Lambda_1{D_{11}}^-_{i-\frac{1}{2},1} + \sqrt{3}\lambda{D_{12}}^-_{i-\frac{1}{2},1} + \sqrt{3}\lambda{D_{12}}^-_{i-\frac{1}{2},2} + 2\lambda\tilde{\alpha}, \\
\tau^-_{i-\frac{1}{2},2} &= -\Lambda_1{D_{11}}^-_{i-\frac{1}{2},2} - \sqrt{3}\lambda{D_{12}}^-_{i-\frac{1}{2},1} - \sqrt{3}\lambda{D_{12}}^-_{i-\frac{1}{2},2} + 2\lambda\tilde{\alpha}, \\
\tau^+_{i-\frac{1}{2},1} &= \Lambda_1{D_{11}}^+_{i-\frac{1}{2},1} + \sqrt{3}\lambda{D_{12}}^+_{i-\frac{1}{2},1} + \sqrt{3}\lambda{D_{12}}^+_{i-\frac{1}{2},2} - \Lambda_1{D_{11}}^-_{i+\frac{1}{2},1} - 2\lambda\tilde{\alpha}, \\
\tau^+_{i-\frac{1}{2},2} &= \Lambda_1{D_{11}}^+_{i-\frac{1}{2},2} - \sqrt{3}\lambda{D_{12}}^+_{i-\frac{1}{2},1} - \sqrt{3}\lambda{D_{12}}^+_{i-\frac{1}{2},2} - \Lambda_1{D_{11}}^-_{i+\frac{1}{2},2} - 2\lambda\tilde{\alpha}, \\
\tau^-_{i+\frac{1}{2},1} &= -\Lambda_1{D_{11}}^+_{i-\frac{1}{2},1} - \sqrt{3}\lambda{D_{12}}^-_{i+\frac{1}{2},1} - \sqrt{3}\lambda{D_{12}}^-_{i+\frac{1}{2},2} + \Lambda_1{D_{11}}^-_{i+\frac{1}{2},1} - 2\lambda\tilde{\alpha}, \\
\tau^-_{i+\frac{1}{2},2} &= -\Lambda_1{D_{11}}^+_{i-\frac{1}{2},2} + \sqrt{3}\lambda{D_{12}}^-_{i+\frac{1}{2},1} + \sqrt{3}\lambda{D_{12}}^-_{i+\frac{1}{2},2} + \Lambda_1{D_{11}}^-_{i+\frac{1}{2},2} - 2\lambda\tilde{\alpha}, \\
\tau^+_{i+\frac{1}{2},1} &= -\Lambda_1{D_{11}}^+_{i+\frac{1}{2},1} - \sqrt{3}\lambda{D_{12}}^+_{i+\frac{1}{2},1} - \sqrt{3}\lambda{D_{12}}^+_{i+\frac{1}{2},2} + 2\lambda\tilde{\alpha}, \\
\tau^+_{i+\frac{1}{2},2} &= -\Lambda_1{D_{11}}^+_{i+\frac{1}{2},2} + \sqrt{3}\lambda{D_{12}}^+_{i+\frac{1}{2},1} + \sqrt{3}\lambda{D_{12}}^+_{i+\frac{1}{2},2} + 2\lambda\tilde{\alpha}.
\end{aligned}
$$

Since $D_{11} > 0$, then we only need to consider the terms on the second line in the last step. Firstly, if we take

$$\tilde{\alpha} \geq \frac{\Delta y}{2\Delta x}D_{11}^M + \sqrt{3}D_{12}^M,$$

where

$$D_{11}^M = \max_{(x,y)\in\Omega} D_{11}(\mathbf{u})(x,y), \quad D_{12}^M = \max_{(x,y)\in\Omega} |D_{12}(\mathbf{u})(x,y)|.$$

then $\tau^-_{i-\frac{1}{2},\beta} > 0$ and $\tau^+_{i+\frac{1}{2},\beta} > 0$. For all the other $\tau' s$, they are related to the point values of c in cell K_{ij}. It is easy to see

$$c_{i\pm\frac{1}{2},1} = \mu_1 c_{i\pm\frac{1}{2},j-\frac{1}{2}} + \mu_2 c_{i\pm\frac{1}{2},j+\frac{1}{2}}, \quad c_{i\pm\frac{1}{2},2} = \mu_2 c_{i\pm\frac{1}{2},j-\frac{1}{2}} + \mu_1 c_{i\pm\frac{1}{2},j+\frac{1}{2}},$$

where $\mu_1 = \frac{3+\sqrt{3}}{6}$, $\mu_2 = \frac{3-\sqrt{3}}{6}$ and the point values of c mentioned above are all the ones chosen in cell K_{ij}. In this section, if not otherwise stated, this is the default definition of element in W_h along the traces in K_{ij}. With the definition above, we can write

$$\bar{r}_{ij}^n = \frac{(c\Phi)_{i-\frac{1}{2},j-\frac{1}{2}} + (c\Phi)_{i-\frac{1}{2},j+\frac{1}{2}} + (c\Phi)_{i+\frac{1}{2},j-\frac{1}{2}} + (c\Phi)_{i+\frac{1}{2},j+\frac{1}{2}}}{4},$$

which further yields

$$\begin{aligned}
&\frac{1}{6}\bar{r}_{ij}^n + \sum_{\beta=1}^{2}\frac{w_\beta}{2}\tau^+_{i-\frac{1}{2},\beta}c^+_{i-\frac{1}{2},\beta} + \sum_{\beta=1}^{2}\frac{w_\beta}{2}\tau^-_{i+\frac{1}{2},\beta}c^-_{i+\frac{1}{2},\beta} \\
=\ &\frac{1}{2}A_1 c_{i-\frac{1}{2},j-\frac{1}{2}} + \frac{1}{2}A_2 c_{i-\frac{1}{2},j+\frac{1}{2}} + \frac{1}{2}A_3 c_{i+\frac{1}{2},j-\frac{1}{2}} + \frac{1}{2}A_4 c_{i+\frac{1}{2},j+\frac{1}{2}},
\end{aligned}$$

where

$$\begin{aligned}
A_1 &= \frac{1}{12}\Phi_{i-\frac{1}{2},j-\frac{1}{2}} + w_1\mu_1\tau^+_{i-\frac{1}{2},1} + w_2\mu_2\tau^+_{i-\frac{1}{2},2}, \qquad A_2 = \frac{1}{12}\Phi_{i-\frac{1}{2},j+\frac{1}{2}} + w_1\mu_2\tau^+_{i-\frac{1}{2},1} + w_2\mu_1\tau^+_{i-\frac{1}{2},2}, \\
A_3 &= \frac{1}{12}\Phi_{i+\frac{1}{2},j-\frac{1}{2}} + w_1\mu_1\tau^-_{i+\frac{1}{2},1} + w_2\mu_2\tau^-_{i+\frac{1}{2},2}, \qquad A_4 = \frac{1}{12}\Phi_{i+\frac{1}{2},j+\frac{1}{2}} + w_1\mu_2\tau^-_{i+\frac{1}{2},1} + w_2\mu_1\tau^-_{i+\frac{1}{2},2}.
\end{aligned}$$

We want all the $A_i's$ to be positive, and only prove for A_1, since the other three can be analyzed by the same lines. It is easy to check that

$$\begin{aligned}
A_1 =\ &\frac{1}{12}\Phi_{i-\frac{1}{2},j-\frac{1}{2}} + \sum_{\beta=1}^{2} w_\beta\mu_\beta\left(\Lambda_1 {D11}^+_{i-\frac{1}{2},\beta} - \Lambda_1 {D11}^-_{i+\frac{1}{2},\beta} - 2\lambda\tilde{\alpha}\right) \\
&+(w_1\mu_1 - w_2\mu_2)\sqrt{3}\lambda\left({D12}^+_{i-\frac{1}{2},1} + {D12}^+_{i-\frac{1}{2},2}\right) \\
\geq\ &\frac{1}{12}\Phi_{i-\frac{1}{2},j-\frac{1}{2}} - 2\lambda\tilde{\alpha} - D_{11}^M\Lambda_1 - 2D_{12}^M\lambda.
\end{aligned}$$

Therefore, by taking

$$D_{11}^M\Lambda_1 + 2(\tilde{\alpha} + D_{12}^M)\lambda \leq \frac{1}{12}\Phi_{i-\frac{1}{2},j-\frac{1}{2}},$$

we have $A_1 > 0$. □

For H^s, the result is similar to Lemma 2.3, so we skip the proof and state the result below.

Lemma 3.3. *Suppose $c^n > 0$ and $r^n > 0$ then $H^s(r, \mathbf{u}, c) > 0$ under the condition*

$$\Delta t \leq \frac{1}{6z_1 p_M}, \quad \textit{and} \quad \Delta t \leq \min_{i,j}\frac{\Phi_{ij}^m}{6\max_{\beta,\gamma}\left\{-q(x_i^\beta, y_j^\gamma), 0\right\}}, \tag{3.12}$$

where

$$p_M = \max_{\beta,\gamma=1,2} p_t(x_i^\beta, y_j^\gamma), \quad \Phi_{ij}^m = \min\{\Phi_{i-\frac{1}{2},j-\frac{1}{2}}, \Phi_{i-\frac{1}{2},j+\frac{1}{2}}, \Phi_{i+\frac{1}{2},j-\frac{1}{2}}, \Phi_{i+\frac{1}{2},j+\frac{1}{2}}\}. \tag{3.13}$$

Based on the above three lemmas, we can state the following theorem.

Theorem 3.1. *Suppose $r^n > 0$ ($c^n > 0$), and the parameters α and $\tilde{\alpha}$ satisfy* (3.6) *and* (3.8), (3.10), *respectively. Then $\bar{r}_j^{n+1} > 0$ under the conditions* (3.7), (3.9), (3.11) *and* (3.12).

Following the same proof in Theorem 2.2 with some minor changes, we have the following one.

Theorem 3.2. *Suppose $0 \le r^n \le \Phi$ and the conditions in Theorem 3.1 are satisfied. Moreover, if the fluxes $\widehat{\mathbf{u}c}$ and $\hat{\mathbf{u}}$ are consistent, then $\bar{r}^{n+1} \le \bar{\Phi}$, under the condition*

$$\Delta t \le \frac{1}{6 z_2 p_M} \tag{3.14}$$

where p_M was given in (3.13).

With Theorem 3.2, we can construct physically relevant numerical cell averages $\bar{r}^n$. However, the numerical approximation r^n may be negative or larger than Φ. Therefore, we also need to apply limiters to r^n and the procedure is given in the following steps. For simplicity, we drop the superscript n in all the numerical approximations

1. Set up a small number $\epsilon = 10^{-13}$.

2. If $\bar{r} < \epsilon$, we take $r = \bar{r}$. If $\bar{\Phi} - \bar{r} < \epsilon$, we take $r = \Phi - \bar{\Phi} + \bar{r}$. Then skip the following steps.

3. Compute $m_j = \min_{(x,y)\in K_{ij}} r(x,y)$. If $m_j < 0$, then define the four corners of K_{ij} to be $\mathbf{x}_\ell$, $\ell = 1,2,3,4$. Compute
$$s = \frac{\bar{r}}{\sum_{\ell=1}^4 (r(\mathbf{x}_\ell) + |r(\mathbf{x}_\ell)|)},$$
then the values of the modified r at the corners are given as
$$\tilde{r}(\mathbf{x}_i) = \begin{cases} 0, & \text{if} \quad r(\mathbf{x}_i) < 0, \\ s r(\mathbf{x}_i), & \text{if} \quad r(\mathbf{x}_i) > 0. \end{cases}$$
Finally, we use $\tilde{r}(\mathbf{x}_i)$ to construct a new approximation, also denoted as r.

4. Apply the previous step for $r_2 = \Phi - r$ and construct a new approximation r.

Remark 3.1. *The algorithm above can be extended to triangular meshes following [49], but the finite element space should be made up of piecewise P^1 polynomials. This is because Φ, the projection of ϕ, is different. In general, we require Φ to be continuous. For rectangular meshes, there are four degrees of freedom (DOF) to be determined, and we have to use Q^1 polynomials. For triangular meshes, only three DOFs available, hence P^1 polynomials is necessary.*

Remark 3.2. *The bound-preserving technique can also be extended to compressible miscible displacements in three space dimensions following the same analysis above with some minor changes. So we skip the details here.*

4 Numerical example

In this section, we provide numerical examples to illustrate the accuracy and capability of the method. The time discretization is given as the third-order SSP Runge-Kutta method [14].

4.1 One space dimension

In this subsection, we consider problem (2.2)-(2.4) subject to boundary condition (2.1) on the computational domain $[0, 2\pi]$. If not otherwise stated, we choose $N = 80$. In the first example, we would like to construct an analytical solution and test the accuracy of the bound-preserving DG scheme.

Example 4.1. *We choose the initial condition as*

$$c(x,0) = \frac{1}{2}(1-\cos(x)), \qquad p(x,0) = \cos(x) - 1.$$

The source parameters q and $\tilde{c}$ are taken as

$$q(x,t) = e^{-t}, \quad \tilde{c} = \frac{1}{2}(e^{-\gamma t}(\sin^2(x) - \cos(x)) + 1).$$

Moreover, we also choose other parameters as

$$z_1 = z_2 = 1, \quad \phi(x) = 1, \quad D(u) = \gamma, \quad \kappa(x,y) = \mu(c) = 1.$$

It is easy to see that the exact solutions are

$$c(x,t) = \frac{1}{2}(1 - e^{-\gamma t}\cos(x)), \qquad p(x,t) = e^{-t}(\cos(x) - 1).$$

We choose $\Delta t = 0.05\Delta x^2$ and compute up to $T = 1$. We take $\gamma = 10^{-5}$ such that the exact solution c is very close to 0 for $t > 0$. Therefore, the bound-preserving limiter enact frequently. We compute the L^∞-norm of the error between the numerical and exact solutions of c, and the results are given in Table 1. From the table, we can observe second-order accuracy of the DG scheme with and without

	with limiter		no limiter	
N	L^∞ error	order	L^∞ error	order
20	4.02e-3	–	3.21e-3	–
40	1.02e-3	1.98	8.15e-4	1.98
80	2.57e-4	1.99	2.07e-4	2.00
160	6.41e-5	2.00	5.07e-5	2.00

Table 1: Example 4.1: accuracy test for the second-order DG methods with and without the bound-preserving limiter.

the bound-preserving limiter. Therefore, the bound-preserving technique does not kill the accuracy. This is quite different from the TVD-like schemes, which has only first order accuracy in L^∞-norm. Next, we test the effect of the bound-preserving limiter and solve the following example.

Example 4.2. *We chose the initial condition as*

$$c(x,0) = \frac{1}{2}(\cos(x) + 1), \qquad p(x,0) = -\gamma\cos(x).$$

Other parameters are chosen as

$$q(x,t) = 0, \ z_1 = 0.35, \ z_2 = \mu(c) = \kappa(x) = 1, \ \phi(x) = \frac{1}{4}(3 + \cos(x)), \ D(u) = 0.$$

In this example, we use $\Delta t = 0.01\Delta x^2$ and compute up to $T = 0.1$. Since we take $D(u) = 0$, the diffusion term cannot provide any stability to the numerical scheme. In the numerical experiments, we consider $\gamma = 1, 2, 3, 4$ and apply the bound-preserving limiters. The numerical approximations of c are given in the left panel of Figure 1. From the figure, we can observe that the larger the γ, the

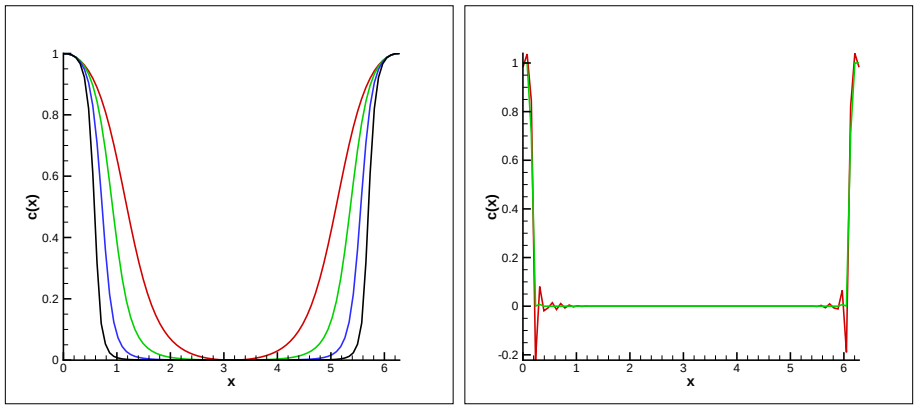

Figure 1: Example 4.2: Left panel shows numerical approximations of c for γ=1 (red), 2 (green), 3 (blue), 4 (black) with bound-preserving limiter. Right panel shows numerical approximations of c for γ=10 with (green) and without (red) bound-preserving limiter. Other parameters are taken as $t = 0.1$ and $N = 80$.

larger the gradient in the numerical approximation. Next, we take $\gamma = 10$ to test the effect of the bound-preserving limiter. In the right panel of Figure 1, the green and red curves are the numerical solutions obtained with and without bound-preserving technique, respectively. It is easy to see that, with the special technique, the green curve is not oscillatory and the numerical approximation lies between 0 and 1. This example clearly demonstrate that our algorithm can also be applied to problems with the present of vacuum.

In the previous two examples, the numerical solutions with and without bound-preserving limiters look similar. However, in the following example, we can observe significant differences. Finally, we consider the necessity of the bound-preserving limiter

Example 4.3. *We choose the initial condition as*

$$c(x,0) = \begin{cases} 1, & x < 1 \\ 0, & x > 1 \end{cases}, \qquad p(x,0) = \begin{cases} 5, & x < 1 \\ 0, & x > 1 \end{cases}.$$

Other parameters are chosen as

$$q(x,t) = 0,\ z_1 = 0.1,\ \kappa(x) = \mu(c) = z_2 = 1,\ \phi(x) = 1,\ D(u) = 0,$$

We compute up to $T = 1$. with $\Delta t = 0.001\Delta x^2$ to reduce the time error. We solve the problem with the bound-preserving limiter and the result is given in Figure 2. We can see that the numerical approximation lies between 0 and 1. Moreover, we also solve the problem without the bound-preserving limiter and the numerical approximation blows up at $t \approx 0.19$. This is because, near $x = 1$, the gradient of c is large, which further yields strong oscillations in the numerical approximations. Therefore, the value of $d(c) = 1 - 0.9c$ might be small or even negative, leading to ill-posed problems. We also choose a smaller time step, say $\Delta t = 0.0001\Delta x^2$, and the numerical approximations blow up at $t \approx 0.106$.

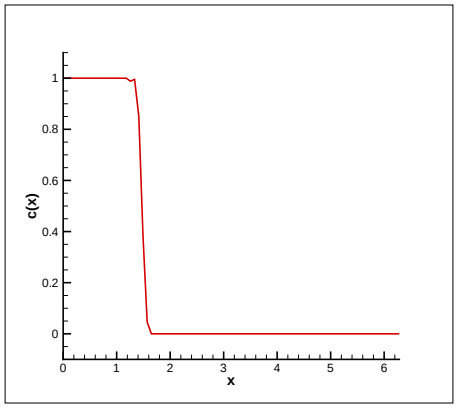

Figure 2: Example 4.3: Numerical approximations of c at $t = 1$.

4.2 Two space dimensions

In this subsection, we consider (1.1) subject to boundary condition (1.3). We solve the problems on the computational domain $\Omega = [0, 2\pi] \times [0, 2\pi]$. If not otherwise stated, we take $N_x = N_y = N$. We first construct an analytical solution and test the accuracy of the bound-preserving DG method.

Example 4.4. *We choose the initial condition as*

$$c(x,0) = \frac{1}{2}(1 - \cos(x)\cos(y)), \qquad p(x,0) = \cos(x)\cos(y) - 1.$$

The source parameters q and $\tilde{c}$ are taken as

$$q(x,t) = 2e^{-2t}, \quad \tilde{c} = \frac{1}{2}\left(e^{-2\gamma t}\left(\frac{1}{2}\sin^2(x)\cos^2(y) + \frac{1}{2}\cos^2(x)\sin^2(y) - \cos(x)\cos(y)\right) + 1\right).$$

Moreover, we also choose other parameters as

$$z_1 = z_2 = 1, \quad \phi(x) = \kappa(x,y) = \mu(c) = 1, \quad \mathbf{D}(\mathbf{u}) = \begin{pmatrix} \gamma & 0 \\ 0 & \gamma \end{pmatrix}.$$

It is easy to see that the exact solutions are

$$c(x,y,t) = \frac{1}{2}(1 - e^{-2\gamma t}\cos(x)\cos(y)), \qquad p(x,y,t) = e^{-2t}(\cos(x)\cos(y) - 1).$$

We choose $\Delta t = 0.02\min\{\Delta x^2, \Delta y^2\}$ with final time $T = 0.01$. We take $\gamma = 10^{-3}$ such that the exact solution c is very close to 0 for $t > 0$. We compute the L^∞-norm of the error between the numerical and exact solution, and results are given in Table 2. From the table, we can observe second-order accuracy of the DG scheme with and without bound-preserving limiter. Therefore, the limiter does not kill the accuracy.

Example 4.5. *We choose the initial condition as*

$$c(x,0) = \begin{cases} 1, & 0 < x < 1,\ 0 < y < 1 \\ 0, & otherwise \end{cases}, \qquad p(x,0) = \begin{cases} 5, & 0 < x < 1,\ 0 < y < 1 \\ 0, & otherwise \end{cases}.$$

Other parameters are taken as

$$q(x,y,t) = 0,\ z_1 = 0.1,\ z_2 = 1,\ \phi(x) = 1,\ \mathbf{D}(\mathbf{u}) = \mathbf{0}.$$

	with limiter		no limiter	
N	L^∞ error	order	L^∞ error	order
20	1.04e-2	–	1.00e-2	–
40	2.64e-3	1.97	2.53e-3	1.99
80	6.77e-4	1.96	6.36e-4	1.99
160	1.75e-4	1.96	1.61e-4	1.99

Table 2: Example 4.4: accuracy test at $T = 0.1$ for the second-order DG methods with and without the bound-preserving limiter.

We compute the numerical approximations of c at $T = 0.1$ and $T = 0.5$, with $\Delta t = 0.001 \min\{\Delta x^2, \Delta y^2\}$ and the bound-preserving limiter. The results are given in Figure 3. We can see that the numerical approximation lies between 0 and 1. However, without the bound-preserving limiter, the numerical

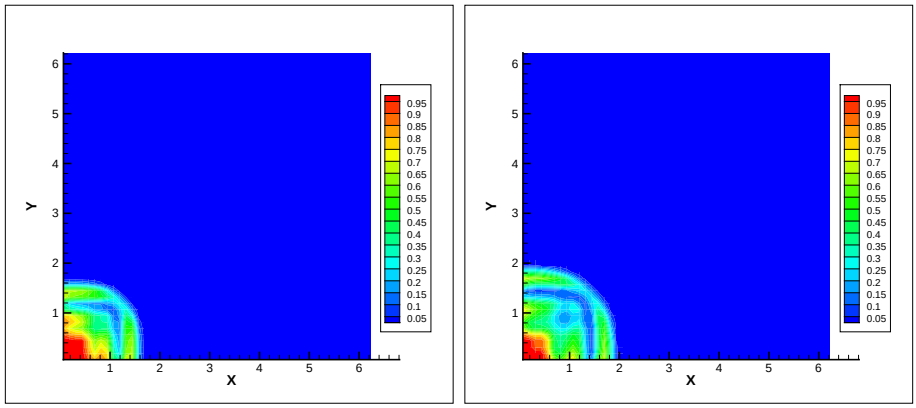

Figure 3: Example 4.5: Numerical approximations of c at $t = 0.1$ (left) and $t = 0.5$ (right).

approximation blows up at $t \approx 1.89 \times 10^{-3}$ due to the ill-posedness of the problems.

Example 4.6. *We choose the initial condition as*

$$c(x, 0) = 0.5, \qquad p(x, 0) = 0.$$

Other parameters are taken as

$$z_1 = 0.4, \ z_2 = 0.6, \ \phi(x) = 1, \ \mathbf{D}(\mathbf{u}) = \begin{pmatrix} |u| & 0 \\ 0 & |u| \end{pmatrix}.$$

The injection well is located at $(2\pi, 2\pi)$ *with* $\tilde{c} = 1$ *while the production well is located at* $(0, 0)$.

We choose $\Delta t = 0.01 \min\{\Delta x^2, \Delta y^2\}$ and compute the numerical approximation of c at t=1, 10 and 100. The results are given in Figure 4. From the figure, we can observe that all the numerical approximations are between 0 and 1. Therefore, the bound-preserving technique also works for the 2D problems.

5 Concluding remarks

In this paper, we apply DG methods to compressible miscible displacement problem in porous media. The bound-preserving technique has been applied to the problems in one and two space dimensions. In

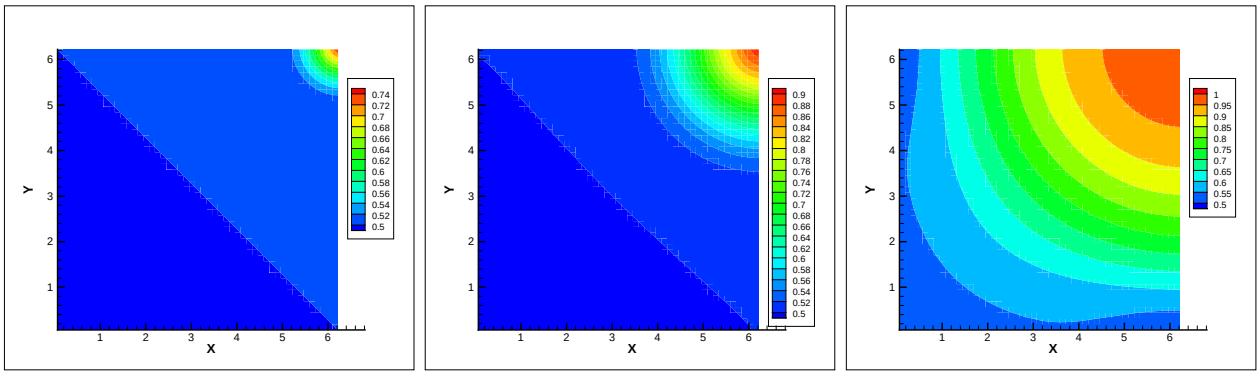

Figure 4: Example 4.6: Numerical approximations of c at $t = 1$ (left) $t = 10$ (middle) $t = 100$ (right) with $N_x = N_y = 50$ and bound-preserving limiter.

the future, we will apply the technique to triangular meshes and consider multi-species fluid mixtures.

Reference

[1] S. Bartels, M. Jensen and R. Müller, *Discontinuous Galerkin finite element convergence for incompressible miscible displacement problem of low regularity,* SIAM Journal on Numerical Analysis, 47 (2009), 3720-3743.

[2] Peter Bastian, *A fully-coupled discontinuous Galerkin method for two-phase flow in porous media with discontinuous capillary pressure,* Computational Geosciences, 18 (2014), 779-796.

[3] H.-Z. Chen, H. Wang, *An optimal-order error estimate on an H^1-Galerkin mixed method for a nonlinear parabolic equation in porous medium flow,* Numerical Methods for Partial Differential Equations, 26 (2010), 188-205.

[4] Z. Chen, H. Huang and J. Yan, *Third order Maximum-principle-satisfying direct discontinuous Galerkin methods for time dependent convection diffusion equations on unstructured triangular meshes,* Journal of Computational Physics, 308 (2016), 198-217.

[5] S.-H. Chou and Q. Li, *Mixed finite element methods for compressible miscible displacement in porous media,* Mathematics of Computation, 57 (1991), 507-527.

[6] M. Cui, *A combined mixed and discontinuous Galerkin method for compressible miscible displacement problem in porous media,* Journal of Computational and Applied Mathematics, 198 (2007), 19-34.

[7] M. Cui, *Analysis of a semidiscrete discontinuous Galerkin scheme for compressible miscible displacement problem,* Journal of Computational and Applied Mathematics, 214 (2008), 617-636.

[8] A. Dedner and R. Klöfkorn, *A Generic Stabilization Approach for Higher Order Discontinuous Galerkin Methods for Convection Dominated Problems,* Journal of Scientific Computing, 47 (2011), 365-388.

[9] J. Douglas, Jr., R.E. Ewing and M.F. Wheeler, *A time-discretization procedure for a mixed finite element approximation of miscible displacement in porous media,* R.A.I.R.O. Analyse numérique, 17 (1983), 249-256.

[10] J. Douglas, Jr., R.E. Ewing and M.F. Wheeler, *The approximation of the pressure by a mixed method in the simulation of miscible displacement,* R.A.I.R.O. Analyse numérique, 17 (1983), 17-33.

[11] J. Douglas, Jr. and J. Roberts, *Numerical methods for a model for compressible miscible displacement in porous media,* Mathematics of Computation, 41 (1983), 441-459.

[12] A. Ern, I. Mozolevski, L. Schuh, *Discontinuous Galerkin approximation of two-phase flows in heterogeneous porous media with discontinuous capillary pressures,* Computer Methods in Applied Mechanics and Engineering, 199 (2010), 1491-1501.

[13] A. Ern, I. Mozolevski, L. Schuh, *Accurate velocity reconstruction for Discontinuous Galerkin approximations of two-phase porous media flows,* Comptes Rendus Mathematique, 347 (2009), 551554.

[14] S. Gottlieb, C.-W. Shu and E. Tadmor, *Strong stability-preserving high-order time discretization methods*, SIAM Review, 43 (2001), 89-112.

[15] H. Guo and Q. Zhang, *Error analysis of the semi-discrete local discontinuous Galerkin method for compressible miscible displacement problem in porous media*, Applied Mathematics and Computation, 259 (2015), 88-105.

[16] H. Guo, Q. Zhang and Y. Yang, *A combined mixed finite element method and local discontinuous Galerkin method for miscible displacement problem in porous media,* Science China Mathematics, 57 (2014), 2301-2320.

[17] L. Guo and Y. Yang, *Positivity-preserving high-order local discontinuous Galerkin method for parabolic equations with blow-up solutions,* Journal of Computational Physics, 289 (2015), 181-195.

[18] L. Krivodonova, J. Xin, J.-F. Remacle, N. Chevaugeon and J. E. Flaherty, *Shock detection and limiting with discontinuous Galerkin methods for hyperbolic conservation laws,* Applied Numerical Mathematics, 48 (2004), 323-338.

[19] S. Kumar, *A mixed and discontinuous Galerkin finite volume element method for incompressible miscible displacement problems in porous media,* Numerical Methods for Partial Differential Equations, 28 (2012), 1354-1381.

[20] X. Li, C.-W. Shu and Y. Yang, *Local discontinuous Galerkin methods for Keller-Segel chemotaxis model,* submitted to Journal of Scientific Computing.

[21] N. Ma, D. Yang and T. Lu, *L^2-norm error bounds of characteristics collocation method for compressible miscible displacement in porous media,* International Journal of Numerical Analysis and Modeling, 2 Supp (2005), 28-42.

[22] T. Qin, C.-W. Shu and Y. Yang, *Bound-preserving discontinuous Galerkin methods for relativistic hydrodynamics,* Journal of Computational Physics, 315 (2016), 323-347.

[23] B. Rivière, *Discontinuous Galerkin finite element methods for solving the miscible displacement problem in porous media,* Ph.D. Thesis, The University of Texas at Austin, 2000.

[24] C.-W. Shu, *Total-variation-diminishing time discretizations*, SIAM Journal on Scientific and Statistical Computing, 9 (1988), 1073-1084.

[25] C.-W. Shu and S. Osher, *Efficient implementation of essentially non-oscillatory shock-capturing schemes*, Journal of Computational Physics, 77 (1988), 439-471.

[26] S. Sun, B. Rivière and M.F. Wheeler, *A combined mixed finite element and discontinuous Galerkin method for miscible displacement problem in porous media,* Recent Progress in Computational and Applied PDEs, Tony Chan et al. (Eds.), Kluwer Academic Publishers, Plenum Press, Dordrecht, NewYork, 2002, 323-351.

[27] S. Sun and M.F. Wheeler, *Discontinuous Galerkin methods for coupled flow and reactive transport problems,* Applied Numerical Mathematics, 52 (2005), 273-298.

[28] S. Sun, M.F. Wheeler, *Symmetric and nonsymmetric discontinuous Galerkin methods for reactive transport in porous media,* SIAM Journal on Numerical Analysis, 43 (2005), 195-219.

[29] H. Wang, D. Liang, R.E. Ewing, S.L. Lyons and G. Qin, *An approximation to miscible fluid flows in porous media with point sources and sinks by an Eulerian-Lagrangian localized adjoint method and mixed finite element methods,* SIAM Journal on Scentific Computing, 22 (2000), 561-581.

[30] H. Wang, D. Liang, R.E. Ewing, S.L. Lyons and G. Qin, *An accurate approximation to compressible flow in porous media with wells,* Numerical Treatment of Multiphase Flows in Porous Media, Lecture Notes in Physics, 552 (2000), 324-332.

[31] M. F. Wheeler and B. L. Darlow, *Interiori penalty Galerkin methods for miscible displacement problems in porous media,* Computational Methods in Nonlinear Mechanics, North-Holland, Amsterdam, 1980, 458-506.

[32] Y. Xing, X. Zhang and C.-W. Shu, *Positivity preserving high order well balanced discontinuous Galerkin methods for the shallow water equations,* Advances inWater Resources, 33 (2010), 1476-1493.

[33] T. Xiong, J.-M. Qiu and Z. Xu, *High order maximum-principle-preserving discontinuous Galerkin method for convection-diffusion equations,* SIAM Journal on Scientific Computing, 37 (2015), A583-A608.

[34] Z. Xu, *Parametrized maximum principle preserving flux limiters for high order schemes solving hyperbolic conservation laws: one-dimensional scalar problem,* Mathematics of Computation, 83 (2014), 310-331.

[35] D. Yang *A splitting positive definite mixed element method for miscible displacement of compressible flow in porous media,* Numerical Methods for Partial Differential Equations, 17 (2001), 229-249.

[36] J. Yang and Y. Chen *A priori error estimates of a combined mixed finite element and discontinuous Galerkin method for compressible miscible displalcement with molecular diffusion and dispersion,* Journal of Computational Mathematics, 28 (2010), 1005-1022.

[37] J. Yang and Y. Chen , *A priori error analysis of a discontinuous Galerkin approximation for a kind of compressible miscible displacement problems,* Science China Mathematics, 53 (2010), 2679-2696.

[38] J. Yang *A posteriori error of a discontinuous Galerkin scheme for compressible miscible displacement problems with molecular diffusion and dispersion,* International Journal for Nnumerical Methods in Fluids 65 (2011) 781-797.

[39] J. Yang and Y. Chen *Superconvergence of a combined mixed finite element and discontinuous Galerkin method for a compressible miscible displacement problem*, Acta Mathematicae Applicatae Sinica, 27 (2011), 481-494.

[40] J. Yang, Y. Chen and Z. Xiong, *Superconvergence of a full-discrete combined mixed finite element and discontinuous Galerkin method for a compressible miscible displacement problem,* Numerical Methods for Partial Differential Equations, 29 (2013), 1801-1820.

[41] Y. Yang and C.-W. Shu, *Discontinuous Galerkin method for hyperbolic equations involving δ-singularities: negative-order norm error estimates and applications,* Numerische Mathematik, 124, (2013), 753-781.

[42] Y. Yang, D. Wei and C.-W. Shu, *Discontinuous Galerkin method for Krause's consensus models and pressureless Euler equations,* Journal of Computational Physics, 252 (2013), 109-127.

[43] Y. Yuan *The characteristic finite difference fractional steps methods for compressible two-phase displacement problem,* Science in China Series A: Mathematics, 42 (1999), 48-57.

[44] Y. Yuan, *The upwind finite difference fractional steps mthods for two-phase compressible flow in porous media,* Numerical Methods for Partial Differential Equations, 19 (2003), 67-88.

[45] Y. Yuan, *The modified upwind finite difference fractional steps method for compressible two-phase displacement problem,* Acta Mathematicae Applicatae Sinica, 20 (2004), 381-396.

[46] X. Zhang and C.-W. Shu, *On maximum-principle-satisfying high order schemes for scalar conservation laws*, Journal of Computational Physics, 229 (2010), 3091-3120.

[47] X. Zhang and C.-W. Shu, *On positivity preserving high order discontinuous Galerkin schemes for compressible Euler equations on rectangular meshes*, Journal of Computational Physics, 229 (2010), 8918-8934.

[48] X. Zhang and C.-W. Shu, *Positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations with source terms*, Journal of Computational Physics, 230 (2011), 1238-1248.

[49] Y. Zhang, X. Zhang and C.-W. Shu, *Maximum-principle-satisfying second order discontinuous Galerkin schemes for convection-diffusion equations on triangular meshes,* Journal of Computational Physics, 234 (2013), 295-316.