

High-order bound-preserving discontinuous Galerkin methods for multicomponent chemically reacting flows [☆]



Jie Du ^{a,b}, Yang Yang ^{c,*}

^a Yau Mathematical Sciences Center, Tsinghua University, Beijing, 100084, PR China

^b Yanqi Lake Beijing Institute of Mathematical Sciences and Applications, Beijing 101408, PR China

^c Department of Mathematical Sciences, Michigan Technological University, Houghton, MI 49931, USA

ARTICLE INFO

Article history:

Received 26 February 2022

Received in revised form 14 July 2022

Accepted 15 August 2022

Available online xxxx

Keywords:

Bound-preserving

Discontinuous Galerkin methods

Conservative time integration

Multicomponent chemically reacting flows

ABSTRACT

In this paper, we design high-order bound-preserving discontinuous Galerkin (DG) methods for multicomponent chemically reacting flows. In this problem, the density and pressure are positive and the mass fractions are between 0 and 1. There are three main difficulties. First of all, it is not easy to construct high-order positivity-preserving schemes for convection-diffusion equations. In this paper, we design a special penalty term to the diffusion term and construct the positivity-preserving flux for the system. The proposed idea is locally conservative, high-order accurate and easy to implement. Secondly, the positivity-preserving technique cannot preserve the upper bound 1 of the mass fractions. To bridge this gap, we apply the positivity-preserving technique to each mass fraction and develop consistent numerical fluxes in the system and conservative time integrations to preserve the summation of the mass fractions to be 1. Therefore, each mass fraction would be between 0 and 1. Finally, most previous bound-preserving DG methods are based on Euler forward time discretization. However, due to the rapid reaction rates, the target system may contain stiff sources, leading to restricted time step sizes. To fix this and preserve conservative property, we apply the conservative modified exponential Runge-Kutta method. The method is third-order accurate and keeps the summation of the mass fractions to be 1. Numerical experiments will be given to demonstrate the good performance of the proposed schemes.

© 2022 Elsevier Inc. All rights reserved.

1. Introduction

The multicomponent chemically reacting flows have been widely used to simulate inertial confinement fusion, high-speed combustion, supernovae explosion and cavitation bubble clouds. The effective numerical methods may not be easy to design due to the appearance of the strong oscillations resulting from the high-order shock-capturing numerical methods. The oscillations may send some positive quantities negative, leading to the blow-up of the numerical simulations. There are many works discussing numerical methods for multicomponent chemically reacting flows in the literature, such as the low-dissipation method [19], the ghost fluid method [20], the entropy-stable hybrid method [31], and the interface-

[☆] The first author was supported by National Key R&D Program of China (Grant No. 2021YFA0719200) and Tsinghua University Initiative Scientific Research Program. The second author was supported by NSF grant DMS-1818467.

* Corresponding author.

E-mail addresses: jdu@tsinghua.edu.cn (J. Du), yyang7@mtu.edu (Y. Yang).

capturing methods [24], etc. Moreover, adaptive methods [35,3] and moving mesh methods [34] were also discussed to reduce the computational cost. In addition to the above, the discontinuous Galerkin (DG) methods were also successfully applied to multicomponent chemically reacting flows [29,30,25,2]. However, no previous works focused on the bound-preserving techniques, and the most commonly used technique to preserve the bounds is to apply a cut-off operator, leading to the loss of the mass. In this paper, we apply high-order DG methods for multicomponent chemically reacting flows and construct the bound-preserving technique to obtain positive density and pressure. Moreover, the mass fractions are also designed to be between 0 and 1.

The DG method, first introduced in [38] in the framework of neutron linear transport, becomes one of the most popular numerical methods for convection-diffusion equations due to its high-order accuracy and flexibility on h-p adaptivity and on complex geometry. Because of the local structure of the DG scheme, limiters can easily be applied to preserve some physical properties. One of the major achievements in this direction is the high-order bound-preserving technique. There were several works discussing the bound-preserving techniques for convection-diffusion equations in the past several years. In [49], Zhang and Shu first constructed the genuinely high-order maximum-principle-preserving (MPP) DG schemes for scalar conservation laws. Later, the high-order positivity-preserving (PP) DG schemes for compressible Euler equations were developed in [50,52,51] to preserve the positivity of the density and pressure. Subsequently, the idea was further extended to other hyperbolic equations and systems [47,54,37] to obtain the L^1 stability of the schemes. For convection-diffusion equations, the second-order MPP DG methods were given in [53]. However, the idea can hardly be extended to construct high-order MPP schemes. Later, the third-order MPP DG schemes were developed in [5,9] based on the direct DG methods [27] and the local DG (LDG) methods on overlapping meshes [12]. Besides the above, the flux limiters provide another direction in constructing high-order MPP and PP schemes for convection-diffusion equations, see [45,18] as an incomplete list. In addition, a special high-order PP DG method for compressible Navier-Stokes equations was given in [48], where an artificially designed penalty term was added to the numerical fluxes of the diffusion term.

Recently, the PP technique has been applied to inviscid compressible reacting flows in [42], where the density, pressure and mass fractions are designed to be positive. However, the scheme could not theoretically guarantee the upper bound 1 of the mass fractions. The time integration proposed in [42] was the strong-stability-preserving (SSP) explicit Runge-Kutta (RK)/multistep methods [15,39,40], hence the time step due to the stiff source can be limited. Moreover, the modified Patankar RK methods were also applied to preserve the mass conservation and the positivity of the density and mass fractions [22,23]. However, such an idea cannot theoretically yield positive pressure for multicomponent flows. To the best knowledge of the authors, no previous works discussing the bound-preserving technique for multicomponent chemically reacting flows governed by Navier-Stokes equations are available. There are three main difficulties:

1. The high-order PP spatial discretization. Most of the previous works discussing bound-preserving DG methods for convection-diffusion equations are not easily extendable to multicomponent chemically reacting flows. In this paper, we extend the idea proposed in [48] and construct a new penalty term to preserve the positivity of the density, pressure, and the mass fractions. The proposed scheme is locally conservative, keeps the high-order accuracy and is easy to implement.
2. The high-order bound-preserving technique. The PP technique cannot preserve the upper bound 1 of the mass fractions. To fix this gap, we apply the PP technique to each component and construct conservative time integrations to preserve the summation of the mass fractions to be 1. Therefore, each mass fraction would be between 0 and 1. This idea was first introduced in [17,6,46,14] for DG methods for problems in petroleum engineering, and further extended to finite difference methods in [28,16]. The key point is to design special consistent fluxes such that the summation of the schemes for the mass fractions is that for the density function.
3. The system may contain stiff source terms. The bound-preserving technique to be constructed in this paper only works for explicit SSP time integrations. However, the traditional SSP RK methods may require limited time step sizes. To eliminate the constraint, we adopt the modified exponential RK methods [21]. The basic idea is to add an artificial source to remove the stiffness and use the exponential factor to restore the decay property. Unfortunately, with the exponential factor, the conservative property may not be satisfied. To fix this gap, we designed the conservative modified exponential RK (CMERK) methods [8,10,11]. The CMERK method is up to third-order accurate in time, conservative, sign-preserving and steady-state-preserving.

With the PP technique and the CMERK time discretization, we can theoretically prove that the proposed scheme preserves all the physical bounds.

The rest of the work is organized as follows. In Section 2, we discuss the governing equations and the conservative property of the system. In Section 3, we construct the DG spatial discretization and the CMERK methods for problems in one space dimension. The bound-preserving techniques will then be developed in Sections 4. Problems in two space dimensions will be discussed in Section 5. Some numerical experiments will be given in Section 6 to demonstrate the good performance of the proposed scheme. We will end in Section 7 with some concluding remarks.

2. Governing equations

In this section, we give the formulation of the governing equations. For the brevity of this paper, we start with the two dimensional model directly. The one dimensional version is straightforward to obtain. The Navier-Stokes equations for multi-species flow with chemical reactions in two space dimensions [13,19,26,36] read

$$\mathbf{w}_t + \mathbf{f}_x^a + \mathbf{g}_y^a = \mathbf{f}_x^d + \mathbf{g}_y^d + \mathbf{s}, \quad (2.1)$$

where the primitive variables and the advection fluxes are given by

$$\mathbf{w} = \begin{pmatrix} \rho \\ \rho u \\ \rho v \\ E \\ r_1 \\ \vdots \\ r_{N_s} \end{pmatrix}, \quad \mathbf{f}^a = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ u(E + p) \\ r_1 u \\ \vdots \\ r_{N_s} u \end{pmatrix}, \quad \mathbf{g}^a = \begin{pmatrix} \rho v \\ \rho uv \\ \rho v^2 + p \\ v(E + p) \\ r_1 v \\ \vdots \\ r_{N_s} v \end{pmatrix},$$

and the diffusion fluxes and the source term are given by

$$\mathbf{f}^d = \begin{pmatrix} 0 \\ \tau_{xx} \\ \tau_{xy} \\ u\tau_{xx} + v\tau_{xy} - q_x \\ -r_1 u_1 \\ \vdots \\ -r_{N_s} u_{N_s} \end{pmatrix}, \quad \mathbf{g}^d = \begin{pmatrix} 0 \\ \tau_{yx} \\ \tau_{yy} \\ u\tau_{yx} + v\tau_{yy} - q_y \\ -r_1 v_1 \\ \vdots \\ -r_{N_s} v_{N_s} \end{pmatrix}, \quad \mathbf{s} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \omega_1 \\ \vdots \\ \omega_{N_s} \end{pmatrix}.$$

Here ρ , u , v , E and p are the total density, the velocity in x direction, the velocity in y direction, the total energy per unit column and the pressure, respectively. The stress tensor and the diffusion heat flux vector are represented by

$$\tau = \begin{pmatrix} \tau_{xx} & \tau_{xy} \\ \tau_{yx} & \tau_{yy} \end{pmatrix}, \quad \mathbf{q} = \begin{pmatrix} q_x \\ q_y \end{pmatrix},$$

respectively. The total number of chemical species is denoted as N_s . We further denote $m := N_s + 4$ and hence we have $\mathbf{w} \in \mathbb{R}^m$. We use r_i , u_i , v_i and ω_i to denote the density, the diffusion velocity in x direction, the diffusion velocity in y direction and the mass production rate of the i th ($i = 1, \dots, N_s$) species, respectively. Also, we let $r_i = \rho Y_i$ and hence Y_i is the mass fraction of the i th species. Physically, the mass fractions satisfy

$$0 \leq Y_i \leq 1, \quad i = 1, \dots, N_s, \quad \sum_{i=1}^{N_s} Y_i = 1. \quad (2.2)$$

Notice that the detailed formulations of the pressure function, chemical reaction rate, diffusion velocity, stress tensor and diffusion heat flux are not unique based on different hypothesis. The discussions in this paper do not rely on the specific definitions. The key point is that the governing equations (2.1) should satisfy the conservative property, which will be used to design the bound preserving technique. In Sections 2.1-2.3, we will briefly show the formulations adopted in our numerical examples. In Section 2.4, we will define and discuss the conservative property.

2.1. Energy and pressure

The total energy E is the sum of the potential energy and kinetic energy and can be written as

$$E = \rho \sum_{i=1}^{N_s} Y_i h_i - p + \frac{1}{2} \rho (u^2 + v^2), \quad (2.3)$$

where

$$h_i(T) = \Delta h_{f,i}^{T_0} + \int_{T_0}^T c_{p,i}(s) ds$$

is the enthalpy per unit mass of the i th species. Here $\Delta h_{f,i}^{T_0}$ and $c_{p,i}$ are the enthalpy of formation at reference temperature T_0 and constant-pressure specific heat of species i , and T is the temperature of the mixture. The enthalpy of formation of each specific species is a standard constant and can be found in the JANAF (Joint Army Navy NASA Air Force) Thermochemical Tables [41]. For a calorically perfect gas, the specific heat is a constant [1] and can also be taken from the JANAF table. For a thermally perfect gas in which the specific heat is a function of the temperature [1], one can use the NASA's polynomial expressions [33].

The equation of state for multi-species flow is

$$p = \rho R_u T \sum_{i=1}^{N_s} \frac{Y_i}{W_i}, \quad (2.4)$$

where W_i is the molecular weight of species i and R_u is the universal gas constant. By combining equations (2.3) and (2.4), we can solve for the temperature T from the primitive variables and thus can further compute the value of pressure p .

2.2. Diffusion fluxes

The diffusion fluxes $\mathbf{f}^d$ and $\mathbf{g}^d$ account for mass diffusion, viscosity and heat conduction. The species mass diffusion flux of species i is denoted as

$$\mathbf{J}_i = \rho Y_i \mathbf{V}_i,$$

where $\mathbf{V}_i = [u_i, v_i]^T$ is the diffusion velocity of species i . Following the formulations used in [19], we first compute the diffusion velocity as

$$\hat{\mathbf{V}}_i = -\frac{D_{i,mix}}{X_i} (\nabla X_i + (X_i - Y_i) \nabla (\ln p)),$$

and then apply the following correction procedure recommended in [7] to ensure mass conservation

$$\mathbf{V}_i = \hat{\mathbf{V}}_i - \sum_{j=1}^{N_s} Y_j \hat{\mathbf{V}}_j, \quad (2.5)$$

where

$$X_i = \frac{Y_i}{W_i} / \sum_{j=1}^{N_s} \frac{Y_j}{W_j}$$

is the mole fraction of species i and

$$D_{i,mix} = \frac{1 - Y_i}{\sum_{j \neq i}^{N_s} X_j / D_{ij}}$$

is the mixture-averaged diffusion coefficient for species i relative to the rest of the multicomponent mixture. Here D_{ij} is the binary diffusion coefficient defined for species j with respect to species i [26].

The stress tensor can be computed by

$$\tau_{xx} = \frac{2}{3} \eta (2u_x - v_y), \quad \tau_{xy} = \tau_{yx} = \eta (u_y + v_x), \quad \tau_{yy} = \frac{2}{3} \eta (2v_y - u_x),$$

where η is the mixture viscosity. The thermal diffusion vector can be computed by

$$\mathbf{q} = \sum_{i=1}^{N_s} h_i \mathbf{J}_i - \lambda \nabla T,$$

where $\mathbf{J}_i$ is the species mass diffusion flux of species i and λ is the thermal conductivity of the mixture. For a single species i , we denote its viscosity and thermal conductivity [26] by η_i and λ_i , respectively, which can be determined if we know the temperature T . For a mixture, we can find the viscosity by using the Wilke formula [44,4]

$$\eta = \sum_{i=1}^{N_s} \frac{X_i \eta_i}{\sum_{j=1}^{N_s} X_j \phi_{ij}},$$

where

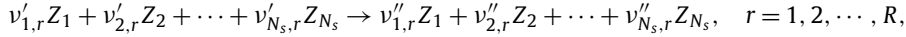
$$\phi_{ij} = \frac{1}{\sqrt{8}} \left(\frac{1}{\sqrt{1 + \frac{W_i}{W_j}}} \right) \left(1 + \sqrt{\frac{\eta_i}{\eta_j}} \sqrt{\frac{W_j}{W_i}} \right)^2.$$

We can write the mixture thermal conductivity using a combination averaging formula due to Mathur [32].

$$\lambda = \frac{1}{2} \left(\sum_{i=1}^{N_s} X_i \lambda_i + \frac{1}{\sum_{i=1}^{N_s} \frac{X_i}{\lambda_i}} \right).$$

2.3. Stiff source terms

The source term ω_i describes the chemical reactions. We consider R reactions of the form



where Z_i is the chemical symbol of species i , $\nu'_{i,r}$ and $\nu''_{i,r}$ are the stoichiometric coefficients of the reactants and products, respectively, of the i th species in the r th reaction. For reversible reactions, we rewrite the reverse reaction as a forward reaction and add it to the system. The mass rate of production of the i th species can be written as

$$\omega_i = W_i \sum_{r=1}^R (\nu''_{i,r} - \nu'_{i,r}) \left[k_r(T) \prod_{j=1}^{N_s} [X_j]^{\nu'_{j,r}} \right], \quad i = 1, 2, \dots, N_s,$$

where $[X_j] = \frac{\rho Y_j}{W_j}$ is the molar concentration of Z_j . The chemical reaction rate $k_r(T)$ is given as

$$k_r(T) = A_r T^{\beta_r} \exp \frac{E_{a,r}}{R_u T},$$

where A_r , β_r and $E_{a,r}$ are the pre-exponential factor, the temperature exponent and the activation energy, respectively, for the r th reaction. In general, the reaction speed of the chemical species is extremely fast, leading to stiff source terms in the model system.

2.4. Conservative property

As we mentioned above, the discussions in this paper can be applied to generic forms of the pressure, stress tensor, heat flux, chemical reaction rate and so on. The key point is that the model should satisfy the following conservative property.

Definition 2.1. Considering the governing equations (2.1), if there exists a constant vector $\mathbf{v} \in \mathbb{R}^m$ such that the solution $\mathbf{w}$ satisfies

$$\frac{d}{dt}(\mathbf{w} \cdot \mathbf{v}) = 0, \quad (2.6)$$

then we say the model problem is conservative. That is to say, $\mathbf{w} \cdot \mathbf{v}$ is a conserved quantity and its value remains unchanged during the time evolution.

To achieve the conservative property, one key point is to apply the correction procedure in (2.5) for the diffusion velocity, and hence can check that

$$\sum_{i=1}^{N_s} Y_i \mathbf{V}_i = \mathbf{0}, \quad (2.7)$$

as long as we have $\sum_{i=1}^{N_s} Y_i = 1$. Moreover, it is easy to check that

$$\sum_{i=1}^{N_s} \omega_i = 0.$$

By introducing a constant vector $\mathbf{v} = [1, 0, 0, 0, -1, \dots, -1] \in \mathbb{R}^m$, system (2.1) satisfies

$$\mathbf{f}^a \cdot \mathbf{v} = \mathbf{g}^a \cdot \mathbf{v} = \mathbf{f}^d \cdot \mathbf{v} = \mathbf{g}^d \cdot \mathbf{v} = \mathbf{s} \cdot \mathbf{v} = 0. \quad (2.8)$$

Taking dot product with $\mathbf{v}$ on both sides of (2.1), we can obtain the conservative property (2.6), and the conserved quantity is

$$\mathbf{w} \cdot \mathbf{v} = \rho - \rho \sum_{i=1}^{N_s} Y_i.$$

We assume that initially we have $\mathbf{w} \cdot \mathbf{v} = 0$, i.e., $\sum_{i=1}^{N_s} Y_i = 1$ and then this relation remains valid for latter time.

In this paper, we will construct suitable numerical schemes that can preserve the conservative property, which will help design the bound-preserving technique. Also, one may have noticed that the summation of the last N_s equations in (2.1) is exactly the first equation in this system as long as the conservative property is satisfied. Hence, one of these equations is redundant. We still use (2.1) to illustrate the bound-preserving technique and conservative property. However, in practice we can replace the last equation in (2.1) by $\sum_{i=1}^{N_s} Y_i = 1$ to save computational cost.

3. Numerical schemes in one dimension

In this section, we introduce the numerical schemes for solving the one dimensional Navier-Stokes system for multi-species flow with chemical reactions. In this case, the governing equations have the following form:

$$\mathbf{w}_t + \mathbf{f}_x^a = \mathbf{f}_x^d + \mathbf{s}, \quad (3.1)$$

where the primitive variables, the advection and diffusion fluxes and the source term are given by

$$\mathbf{w} = \begin{pmatrix} \rho \\ \rho u \\ E \\ r_1 \\ \vdots \\ r_{N_s} \end{pmatrix}, \quad \mathbf{f}^a = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ u(E + p) \\ r_1 u \\ \vdots \\ r_{N_s} u \end{pmatrix}, \quad \mathbf{f}^d = \begin{pmatrix} 0 \\ \tau_{xx} \\ u\tau_{xx} - q_x \\ -r_1 u_1 \\ \vdots \\ -r_{N_s} u_{N_s} \end{pmatrix}, \quad \mathbf{s} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \omega_1 \\ \vdots \\ \omega_{N_s} \end{pmatrix}.$$

Here $\mathbf{w} \in \mathbb{R}^m$ with $m = N_s + 3$.

In Section 3.1, we will use the LDG method to solve our problem in space and obtain the semi-discrete schemes. After that, we march in time by using the CMERK method to deal with the stiff source terms.

3.1. DG schemes in one dimension

The diffusion flux $\mathbf{f}^d$ is dependent on the derivative of conserved variables $\mathbf{w}$. Hence, we introduce an auxiliary variable $\mathbf{q} := \mathbf{w}_x$ and then rewrite the model (3.1) into the following system

$$\mathbf{w}_t = -\mathbf{f}^a(\mathbf{w})_x + \mathbf{f}^d(\mathbf{w}, \mathbf{q})_x + \mathbf{s},$$

$$\mathbf{q} = \mathbf{w}_x.$$

Let $\Omega_h = \{I_i, i = 1, \dots, N\}$ be a partition of the computational domain Ω , where $I_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$. We define the finite element space V_h^k as

$$V_h^k = \{\psi : \psi|_{I_i} \in P^k(I_i), i = 1, \dots, N\},$$

where $P^k(I_i)$ denotes the set of polynomials of degree up to k in cell I_i . For simplicity, we still use the notations $\mathbf{w}$ and $\mathbf{q}$ as the numerical approximations.

Following [48], we regard $\mathbf{f} = \mathbf{f}^a - \mathbf{f}^d$ as a single flux and formally treat $\mathbf{f}_x$ as a convection term. Then the local DG scheme is to find $\mathbf{w} \in [V_h^k]^m$ and $\mathbf{q} \in [V_h^k]^m$, such that for any test functions $\psi \in [V_h^k]^m$ and $\phi \in [V_h^k]^m$ and any cell $I_i \in \Omega_h$, we have

$$\int_{I_i} \mathbf{w}_t \cdot \psi \, dx = \int_{I_i} \mathbf{f} \cdot \psi_x \, dx - \hat{\mathbf{f}}_{i+\frac{1}{2}} \cdot \psi_{i+\frac{1}{2}}^- + \hat{\mathbf{f}}_{i-\frac{1}{2}} \cdot \psi_{i-\frac{1}{2}}^+ + \int_{I_i} \mathbf{s} \cdot \psi \, dx, \quad (3.2)$$

$$\int_{I_i} \mathbf{q} \cdot \phi \, dx = - \int_{I_i} \mathbf{w} \cdot \phi_x \, dx + \hat{\mathbf{w}}_{i+\frac{1}{2}} \cdot \phi_{i+\frac{1}{2}}^- - \hat{\mathbf{w}}_{i-\frac{1}{2}} \cdot \phi_{i-\frac{1}{2}}^+, \quad (3.3)$$

where $\psi_{i+\frac{1}{2}}^-$ and $\psi_{i+\frac{1}{2}}^+$ are values of ψ at the point $x_{i+\frac{1}{2}}$ taken from the left and the right respectively. Central flux can be used for $\mathbf{w}$:

$$\hat{\mathbf{w}}_{i+\frac{1}{2}} = \frac{1}{2}(\mathbf{w}_{i+\frac{1}{2}}^- + \mathbf{w}_{i+\frac{1}{2}}^+).$$

The flux $\hat{\mathbf{f}}_{i+\frac{1}{2}}$ is taken as the following Lax-Friedrichs type positivity-preserving flux

$$\begin{aligned} & \hat{\mathbf{f}}_{i+\frac{1}{2}}(\mathbf{w}_{i+\frac{1}{2}}^-, \mathbf{q}_{i+\frac{1}{2}}^-, \mathbf{w}_{i+\frac{1}{2}}^+, \mathbf{q}_{i+\frac{1}{2}}^+) \\ &= \frac{1}{2} [\mathbf{f}(\mathbf{w}_{i+\frac{1}{2}}^-, \mathbf{q}_{i+\frac{1}{2}}^-) + \mathbf{f}(\mathbf{w}_{i+\frac{1}{2}}^+, \mathbf{q}_{i+\frac{1}{2}}^+) - \beta_{i+\frac{1}{2}}(\mathbf{w}_{i+\frac{1}{2}}^+ - \mathbf{w}_{i+\frac{1}{2}}^-)], \end{aligned} \quad (3.4)$$

where

$$\beta_{i+\frac{1}{2}} > \max_{\mathbf{w}_{i+\frac{1}{2}}^-, \mathbf{q}_{i+\frac{1}{2}}^-, \mathbf{w}_{i+\frac{1}{2}}^+, \mathbf{q}_{i+\frac{1}{2}}^+} \left\{ \left| u + \frac{b}{2\rho P(\mathbf{w})} \right| + \frac{\sqrt{D}}{2\rho P(\mathbf{w})}, |u + u_n|, n = 1, \dots, N_s \right\}, \quad (3.5)$$

with

$$b := \rho q_x - \rho \sum_{n=1}^{N_s} r_n h_n(0) u_n, \quad P(\mathbf{w}) := E - \sum_{i=1}^{N_s} r_i h_i(0) - \frac{1}{2} \rho u^2,$$

and

$$D := b^2 + 2\rho P(\mathbf{w})(p - \tau_{xx})^2.$$

The condition on $\beta_{i+\frac{1}{2}}$ is to guarantee the bound-preserving property and the detailed proof will be discussed in Section 4.

Recall that the physical flux $\mathbf{f}$ in our model satisfies $\mathbf{f} \cdot \mathbf{v} = 0$. If the same property is also satisfied by the numerical flux $\hat{\mathbf{f}}_{i+\frac{1}{2}}$, then we say the numerical flux is consistent.

Definition 3.1. If the numerical flux satisfies $\hat{\mathbf{f}}_{i+\frac{1}{2}} \cdot \mathbf{v} = 0$, then we say $\hat{\mathbf{f}}_{i+\frac{1}{2}}$ is consistent.

It is easy to check that the numerical flux $\hat{\mathbf{f}}_{i+\frac{1}{2}}$ defined in (3.4) is consistent as long as we already have $\mathbf{w} \cdot \mathbf{v} = 0$ at the current time t . As discussed in Section 2.4, the original PDE system is conservative. Next, we will show that the semi-discrete scheme preserves this conservative property.

Theorem 3.1. The semi-discrete scheme (3.2)-(3.3) obtained by the LDG scheme is conservative in the sense that

$$\frac{d}{dt}(\mathbf{w} \cdot \mathbf{v}) = 0,$$

as long as the numerical flux $\hat{\mathbf{f}}_{i+\frac{1}{2}}$ is consistent, where $\mathbf{v} = [1, 0, 0, -1, \dots, -1] \in \mathbb{R}^m$.

Proof. By taking the test function as $\psi = z\mathbf{v}$ with $z \in V_h^k$, (3.2) becomes

$$\int_{I_i} \frac{d}{dt}(\mathbf{w} \cdot \mathbf{v}) z dx = \int_{I_i} (\mathbf{f} \cdot \mathbf{v}) z_x dx - (\hat{\mathbf{f}}_{i+\frac{1}{2}} \cdot \mathbf{v}) z_{i+\frac{1}{2}}^- + (\hat{\mathbf{f}}_{i-\frac{1}{2}} \cdot \mathbf{v}) z_{i-\frac{1}{2}}^+ + \int_{I_i} (\mathbf{s} \cdot \mathbf{v}) z dx.$$

For the original model, we already have $\mathbf{f} \cdot \mathbf{v} = \mathbf{s} \cdot \mathbf{v} = 0$. As long as the numerical flux is consistent, we have

$$\int_{I_i} \frac{d}{dt}(\mathbf{w} \cdot \mathbf{v}) z dx = 0.$$

Since $\frac{d}{dt}(\mathbf{w} \cdot \mathbf{v}) \in V_h^k$ and z can be any function in V_h^k , we have

$$\frac{d}{dt}(\mathbf{w} \cdot \mathbf{v}) = 0. \quad \square$$

3.2. Conservative time integrations

It is easy to solve for $\mathbf{q}$ locally on each cell I_i by using (3.3) and then substitute the results into (3.2). Next, we need to apply suitable time integrations on (3.2) to march in time. For simplicity of the formulation, we introduce the notation

$$f_i(\mathbf{w}, \psi) := \int_{I_i} \mathbf{f} \cdot \psi_x dx - \hat{\mathbf{f}}_{i+\frac{1}{2}} \cdot \psi_{i+\frac{1}{2}}^- + \hat{\mathbf{f}}_{i-\frac{1}{2}} \cdot \psi_{i-\frac{1}{2}}^+, \quad (3.6)$$

and then the semi-discrete scheme (3.2) for solving $\mathbf{w}$ becomes

$$\frac{d}{dt}(\mathbf{w}, \psi)_{I_i} = f_i(\mathbf{w}, \psi) + (\mathbf{s}(\mathbf{w}), \psi)_{I_i}, \quad \forall \psi \in [V_h^k]^m, \quad i = 1, \dots, N,$$

where $(\cdot, \cdot)_{I_i}$ denotes the L_2 inner product on I_i . Since the system has stiff source term, we adopt the following third-order CMERK method [10] to get the fully discretized scheme:

$$\begin{aligned} (\mathbf{w}^{(1)}, \psi)_{I_i} &= \left[\alpha_{10}(\mathbf{w}^n, \psi)_{I_i} + \beta_{10}\Delta t f_i(\mathbf{w}^n, \psi) + \beta_{10}\Delta t (s(\mathbf{w}^n) + \mu \mathbf{w}^n, \psi)_{I_i} \right] / A_1, \\ (\mathbf{w}^{(2)}, \psi)_{I_i} &= \left[\alpha_{20}(\mathbf{w}^n, \psi)_{I_i} + \beta_{20}\Delta t f_i(\mathbf{w}^n, \psi) + \beta_{20}\Delta t (s(\mathbf{w}^n) + \mu \mathbf{w}^n, \psi)_{I_i} \right] / A_2 \\ &\quad + e^{\beta_{10}\mu\Delta t} \left[\alpha_{21}(\mathbf{w}^{(1)}, \psi)_{I_i} + \beta_{21}\Delta t f_i(\mathbf{w}^{(1)}, \psi) + \beta_{21}\Delta t (s(\mathbf{w}^{(1)}) + \mu \mathbf{w}^{(1)}, \psi)_{I_i} \right] / A_2, \\ (\mathbf{w}^{n+1}, \psi)_{I_i} &= \left[\alpha_{30}(\mathbf{w}^n, \psi)_{I_i} + \beta_{30}\Delta t f_i(\mathbf{w}^n, \psi) + \beta_{30}\Delta t (s(\mathbf{w}^n) + \mu \mathbf{w}^n, \psi)_{I_i} \right] / A_3 \\ &\quad + e^{\beta_{10}\mu\Delta t} \left[\alpha_{31}(\mathbf{w}^{(1)}, \psi)_{I_i} + \beta_{31}\Delta t f_i(\mathbf{w}^{(1)}, \psi) + \beta_{31}\Delta t (s(\mathbf{w}^{(1)}) + \mu \mathbf{w}^{(1)}, \psi)_{I_i} \right] / A_3 \\ &\quad + e^{A\mu\Delta t} \left[\alpha_{32}(\mathbf{w}^{(2)}, \psi)_{I_i} + \beta_{32}\Delta t f_i(\mathbf{w}^{(2)}, \psi) + \beta_{32}\Delta t (s(\mathbf{w}^{(2)}) + \mu \mathbf{w}^{(2)}, \psi)_{I_i} \right] / A_3, \end{aligned}$$

where

$$\begin{aligned} \alpha_{10} &= 1, \quad \beta_{10} = 0.7071933376925014, \\ \alpha_{20} &= 0.6686892933074404, \quad \beta_{20} = 0, \\ \alpha_{21} &= 0.3313107066925596, \quad \beta_{21} = 0.4178047564915065, \\ \alpha_{30} &= 0.3487419430256090, \quad \beta_{30} = 0, \\ \alpha_{31} &= 0.2039576138780898, \quad \beta_{31} = 0, \\ \alpha_{32} &= 0.4473004430963011, \quad \beta_{32} = 0.5640754637100439, \end{aligned}$$

and

$$A = \beta_{20} + \alpha_{21}\beta_{10} + \beta_{21}, \quad (3.7)$$

$$A_1 = \alpha_{10} + \beta_{10}\mu\Delta t, \quad (3.8)$$

$$A_2 = \alpha_{20} + \beta_{20}\mu\Delta t + e^{\beta_{10}\mu\Delta t}(\alpha_{21} + \beta_{21}\mu\Delta t), \quad (3.9)$$

$$A_3 = \alpha_{30} + \beta_{30}\mu\Delta t + e^{\beta_{10}\mu\Delta t}(\alpha_{31} + \beta_{31}\mu\Delta t) + e^{A\mu\Delta t}(\alpha_{32} + \beta_{32}\mu\Delta t). \quad (3.10)$$

Recall that the semi-discrete scheme preserves the conservative property of the original PDE. Next, we will show that the above fully discrete scheme is also conservative.

Theorem 3.2. *The fully discrete scheme is conservative in the sense that if $\mathbf{w}^n \cdot \mathbf{v} = 0$, then we have*

$$\mathbf{w}^{n+1} \cdot \mathbf{v} = \mathbf{w}^n \cdot \mathbf{v} = 0.$$

In other words, if we have $\sum_{i=1}^{N_s} Y_i = 1$ at time level n , then this equation is still true at time level $n + 1$.

Proof. By taking the test function as $\psi = z\mathbf{v}$ with $z \in V_h^k$ and using the fact that $\mathbf{s} \cdot \mathbf{v} = 0$, we have

$$(\mathbf{w}^{(1)} \cdot \mathbf{v}, z)_{I_i} = \left[\alpha_{10}(\mathbf{w}^n \cdot \mathbf{v}, z)_{I_i} + \beta_{10}\Delta t f_i(\mathbf{w}^n, z\mathbf{v}) + \beta_{10}\Delta t (\mu \mathbf{w}^n \cdot \mathbf{v}, z)_{I_i} \right] / A_1.$$

By using the definition of f_i in (3.6) and the fact that $\mathbf{f} \cdot \mathbf{v} = \hat{\mathbf{f}}_{i+\frac{1}{2}} \cdot \mathbf{v} = 0$, we know that $f_i(\mathbf{w}^n, z\mathbf{v}) = 0$. Hence, we have

$$(\mathbf{w}^{(1)} \cdot \mathbf{v}, z)_{I_i} = \frac{\alpha_{10} + \beta_{10}\mu\Delta t}{A_1} (\mathbf{w}^n \cdot \mathbf{v}, z)_{I_i}.$$

Since $A_1 = \alpha_{10} + \beta_{10}\mu\Delta t$, we obtain

$$(\mathbf{w}^{(1)} \cdot \mathbf{v}, z)_{I_i} = (\mathbf{w}^n \cdot \mathbf{v}, z)_{I_i}.$$

Since z is an arbitrary function in V_h^k , we have

$$\mathbf{w}^{(1)} \cdot \mathbf{v} = \mathbf{w}^n \cdot \mathbf{v}.$$

Similarly, one can prove that

$$\mathbf{w}^{n+1} \cdot \mathbf{v} = \mathbf{w}^{(2)} \cdot \mathbf{v} = \mathbf{w}^{(1)} \cdot \mathbf{v} = \mathbf{w}^n \cdot \mathbf{v}.$$

The idea of proof is the same and hence we omit the details. $\square$

In our numerical simulation, we take the initial data which satisfies $\mathbf{w}^0 \cdot \mathbf{v} = 0$. Then our numerical results always preserve the condition that $\sum_{i=1}^{N_s} Y_i = 1$ at each time level.

Remark 3.1. In CMERK scheme, if μ is large, the exponential terms would be extremely large. We can apply Taylor's expansion to all the exponential functions, e.g.

$$e^{A\mu\Delta t} \approx [1 - A\mu\Delta t + \frac{1}{2}(A\mu\Delta t)^2 - \frac{1}{6}(A\mu\Delta t)^3 + \frac{1}{24}(A\mu\Delta t)^4]^{-1}.$$

This trick keeps the conservative property. Numerical experiments demonstrated that this trick can lead to better numerical approximations. For more details, we refer to [10].

4. Bound preserving technique in one dimension

We discuss the bound preserving technique in this section. In particular, we want to preserve the positivity of the total density ρ and the pressure p , and also preserve the two bounds 0 and 1 of each mass fraction Y_i , $i = 1, \dots, N_s$.

4.1. The admissible set

For simplicity, we consider the calorically perfect gas for which the specific heats are constants. In this case, the one dimensional version of Equation (2.3) becomes

$$\begin{aligned} E &= \sum_{i=1}^{N_s} r_i (\Delta h_{f,i}^{T_0} + c_{p,i}(T - T_0)) - p + \frac{1}{2}\rho u^2 \\ &= \sum_{i=1}^{N_s} r_i h_i(0) + \rho T \sum_{i=1}^{N_s} Y_i c_{p,i} - p + \frac{1}{2}\rho u^2, \end{aligned}$$

where $h_i(0) = \Delta h_{f,i}^{T_0} - c_{p,i}T_0$ is the enthalpy of formation at 0K which is a constant for species i . Substituting (2.4) into the above equation, we can get

$$E = \sum_{i=1}^{N_s} r_i h_i(0) + T \sum_{i=1}^{N_s} r_i c_{v,i} + \frac{1}{2}\rho u^2,$$

where

$$c_{v,i} = c_{p,i} - \frac{R_u}{W_i}$$

is the specific heat at constant volume for species i . Thus, we can solve for the temperature T as:

$$T = [E - \sum_{i=1}^{N_s} r_i h_i(0) - \frac{1}{2}\rho u^2] / \sum_{i=1}^{N_s} (\rho Y_i c_{v,i}).$$

Substituting this equation into (2.4), we obtain

$$p = (\gamma - 1) [E - \sum_{i=1}^{N_s} r_i h_i(0) - \frac{1}{2}\rho u^2],$$

where

$$\gamma(\mathbf{w}) = \frac{\sum_{i=1}^{N_s} Y_i c_{p,i}}{\sum_{i=1}^{N_s} Y_i c_{v,i}}$$

is the ratio of specific heats for the mixture. It is easy to check that

$$\gamma(\mathbf{w}) - 1 = \frac{\sum_{i=1}^{N_s} Y_i R_u / W_i}{\sum_{i=1}^{N_s} Y_i c_{v,i}} > 0$$

as long as $Y_i \geq 0, i = 1, \dots, N_s$ and hence

$$p \geq 0 \iff P(\mathbf{w}) = E - \sum_{i=1}^{N_s} r_i h_i(0) - \frac{1}{2} \rho u^2 \geq 0. \quad (4.1)$$

Moreover, as discussed in the last section, the numerical results obtained by our scheme always satisfy

$$\sum_{i=1}^{N_s} Y_i = 1 \quad (4.2)$$

at each time level. Hence, we only need to preserve the lower bound 0 of each mass fraction and guarantee that (4.2) will not be harmed after applying the bound preserving technique, and then the upper bound 1 of each $Y_i, i = 1, \dots, N_s$ will be preserved automatically. Also, the requirement $Y_i \geq 0$ can be replaced by $r_i \geq 0$ as long as the total density $\rho \geq 0$. Hence, we define the set of admissible states as

$$G = \left\{ \mathbf{w} = \begin{pmatrix} \rho \\ \rho u \\ E \\ r_1 \\ \vdots \\ r_{N_s} \end{pmatrix} : \rho \geq 0, \quad P(\mathbf{w}) \geq 0, \quad r_i \geq 0, \quad i = 1, \dots, N_s \right\}.$$

One can check that P is a concave function with respect to $\mathbf{w}$ and therefore G is a convex set. Next, we will construct conservative bound-preserving methods such that the numerical solutions belong to the set of admissible states G .

4.2. Bound-preserving of the cell averages

Taking the test function as $\psi = \mathbf{e}_l, l = 1, \dots, m$, respectively, where $\mathbf{e}_l \in \mathbb{R}^m$ is a constant vector with the l th component being 1 and other components being 0, the fully discrete scheme lead to the following equations for computing the cell averages of the solutions

$$\bar{\mathbf{w}}_i^{(1)} = \left[\alpha_{10} \mathbf{C}_i^n \left(\frac{\beta_{10}}{\alpha_{10}} \Delta t \right) + \beta_{10} \mu \Delta t \mathbf{D}_i^n \right] / A_1, \quad (4.3)$$

$$\begin{aligned} \bar{\mathbf{w}}_i^{(2)} &= \left[\alpha_{20} \mathbf{C}_i^n \left(\frac{\beta_{20}}{\alpha_{20}} \Delta t \right) + \beta_{20} \mu \Delta t \mathbf{D}_i^n \right] / A_2 \\ &+ e^{\beta_{10} \mu \Delta t} \left[\alpha_{21} \mathbf{C}_i^{(1)} \left(\frac{\beta_{21}}{\alpha_{21}} \Delta t \right) + \beta_{21} \mu \Delta t \mathbf{D}_i^{(1)} \right] / A_2, \end{aligned} \quad (4.4)$$

$$\begin{aligned} \bar{\mathbf{w}}_i^{n+1} &= \left[\alpha_{30} \mathbf{C}_i^n \left(\frac{\beta_{30}}{\alpha_{30}} \Delta t \right) + \beta_{30} \mu \Delta t \mathbf{D}_i^n \right] / A_3 \\ &+ e^{\beta_{10} \mu \Delta t} \left[\alpha_{31} \mathbf{C}_i^{(1)} \left(\frac{\beta_{31}}{\alpha_{31}} \Delta t \right) + \beta_{31} \mu \Delta t \mathbf{D}_i^{(1)} \right] / A_3 \\ &+ e^{A \mu \Delta t} \left[\alpha_{32} \mathbf{C}_i^{(2)} \left(\frac{\beta_{32}}{\alpha_{32}} \Delta t \right) + \beta_{32} \mu \Delta t \mathbf{D}_i^{(2)} \right] / A_3, \end{aligned} \quad (4.5)$$

where

$$\mathbf{C}_i^p(\Delta t) := \bar{\mathbf{w}}_i^p - \frac{\Delta t}{\Delta x_i} (\hat{\mathbf{f}}_{i+\frac{1}{2}}^p - \hat{\mathbf{f}}_{i-\frac{1}{2}}^p), \quad \mathbf{D}_i^p = \frac{1}{\mu} \bar{\mathbf{s}}_i^p + \bar{\mathbf{w}}_i^p, \quad p = n, (1), (2), \quad (4.6)$$

with

$$\bar{\mathbf{w}}_i^p = \frac{1}{\Delta x_i} \int_{I_i} \mathbf{w}^p dx, \quad \bar{\mathbf{s}}_i^p = \frac{1}{\Delta x_i} \int_{I_i} \mathbf{s}(\mathbf{w}^p) dx$$

being the cell averages of $\mathbf{w}$ and $\mathbf{s}(\mathbf{w})$ on I_i , respectively. By using the definition of A_1, A_2 and A_3 in (3.8)-(3.10), we know that each stage of our scheme (4.3)-(4.5) is just a convex combination of the basic structures $\mathbf{C}$ and $\mathbf{D}$. Next, assuming that

$\bar{\mathbf{w}}_i^p \in G$, we aim to find sufficient conditions such that $\mathbf{C}_i^p(\Delta t) \in G$ and $\mathbf{D}_i^p \in G$ for $p = n, (1), (2)$, respectively. For simplicity, we omit the superscript p in the following discussions.

We consider $\mathbf{C}_i(\Delta t)$ first. We use the idea in [48], but still need some recomputations since the model in [48] does not contain multiple species and reactions. Before we look into the details of the computation of $\mathbf{C}_i(\Delta t)$, we need the following lemma.

Lemma 4.1. Consider any $\mathbf{w} = (\rho, \rho u, E, r_1, \dots, r_{N_s})^T \in G$ and

$$\mathbf{f} = \mathbf{f}^a - \mathbf{f}^d = \begin{pmatrix} \rho u \\ \rho u^2 + p - \tau_{xx} \\ u(E + p) - u\tau_{xx} + q_x \\ r_1(u + u_1) \\ \vdots \\ r_{N_s}(u + u_{N_s}) \end{pmatrix}$$

where p, τ_{xx}, q_x and $u_i, i = 1, \dots, N_s$ are not necessarily dependent on $\mathbf{w}$. Then we have $\mathbf{w} \pm \beta^{-1}\mathbf{f} \in G$ if and only if

$$\beta \geq \max \left\{ \left| u + \frac{b}{2\rho P(\mathbf{w})} \right| + \frac{\sqrt{D}}{2\rho P(\mathbf{w})}, |u + u_i|, i = 1, \dots, N_s \right\},$$

where

$$b = \rho q_x - \rho \sum_{i=1}^{N_s} r_i h_i(0) u_i, \quad D = b^2 + 2\rho P(\mathbf{w})(p - \tau_{xx})^2.$$

Proof. First we have

$$\mathbf{w} \pm \beta^{-1}\mathbf{f} = \begin{pmatrix} \rho \pm \beta^{-1}\rho u \\ \rho u \pm \beta^{-1}(\rho u^2 + p - \tau_{xx}) \\ E \pm \beta^{-1}(u(E + p) - u\tau_{xx} + q_x) \\ r_1 \pm \beta^{-1}r_1 \tilde{u}_1 \\ \vdots \\ r_{N_s} \pm \beta^{-1}r_{N_s} \tilde{u}_{N_s} \end{pmatrix},$$

where $\tilde{u}_i = u + u_i$. In order to have $\mathbf{w} \pm \beta^{-1}\mathbf{f} \in G$, we require

$$\begin{aligned} \rho \pm \beta^{-1}\rho u &\geq 0, \\ r_i \pm \beta^{-1}r_i \tilde{u}_i &\geq 0, \quad i = 1, \dots, N_s \\ E \pm \beta^{-1}(u(E + p) - u\tau_{xx} + q_x) - \sum_{i=1}^{N_s} (r_i \pm \beta^{-1}r_i \tilde{u}_i) h_i(0) \\ &\quad - \frac{1}{2}(\rho u \pm \beta^{-1}(\rho u^2 + p - \tau_{xx}))^2 / (\rho \pm \beta^{-1}\rho u) \geq 0 \end{aligned}$$

Since $\beta > 0$, by multiplying β on both sides, the above conditions are equivalent to

$$\bar{\beta}\rho \geq 0, \quad \bar{\beta}_i r_i \geq 0, \quad i = 1, \dots, N_s \quad (4.7)$$

$$\bar{\beta}^2 \rho E \pm \rho(u p - u\tau_{xx} + q_x) \bar{\beta} - \bar{\beta}\rho \sum_{i=1}^{N_s} \bar{\beta}_i r_i h_i(0) - \frac{1}{2}(\bar{\beta}\rho u \pm (p - \tau_{xx}))^2 \geq 0, \quad (4.8)$$

where $\bar{\beta} = \beta \pm u$ and $\bar{\beta}_i = \beta \pm \tilde{u}_i$. Then (4.7) is satisfied if and only if

$$\bar{\beta} \geq 0, \quad \bar{\beta}_i \geq 0,$$

which means

$$\beta \geq |u|, \quad \beta \geq |u + u_i|. \quad (4.9)$$

Moreover, (4.8) becomes

$$\rho P(\mathbf{w})\bar{\beta}^2 \pm \rho \left(q_x - \sum_{i=1}^{N_s} r_i h_i(0) u_i \right) \bar{\beta} - \frac{1}{2}(p - \tau_{xx})^2 \geq 0, \quad (4.10)$$

where $P(\mathbf{w})$ is defined in (4.1). Substituting $\bar{\beta} = \beta \pm u$ into (4.10), we obtain

$$\rho P(\mathbf{w})\beta^2 \pm (2\rho u P(\mathbf{w}) + b)\beta + c \geq 0, \quad (4.11)$$

with

$$c = \rho u^2 P(\mathbf{w}) + bu - \frac{1}{2}(p - \tau_{xx})^2,$$

which are two quadratic forms of β . Since $\mathbf{w} \in G$, we have $\rho P(\mathbf{w}) \geq 0$ and we can check that

$$\begin{aligned} D &= (2\rho u P(\mathbf{w}) + b)^2 - 4\rho P(\mathbf{w})c \\ &= b^2 + 2\rho P(\mathbf{w})(p - \tau_{xx})^2 \geq 0. \end{aligned}$$

Hence, either quadratic form is opening up and has at least one root. Then the four roots for two quadratic equations are

$$\begin{aligned} \beta_{1,2} &= -u + \frac{1}{2\rho P(\mathbf{w})}(-b \pm \sqrt{D}), \\ \beta_{3,4} &= u + \frac{1}{2\rho P(\mathbf{w})}(b \pm \sqrt{D}). \end{aligned}$$

Hence, (4.9) together with (4.11) are satisfied if and only if

$$\beta \geq \max \left\{ \left| u + \frac{b}{2\rho P(\mathbf{w})} \right| + \frac{\sqrt{D}}{2\rho P(\mathbf{w})}, |u + u_i|, i = 1, \dots, N_s \right\}. \quad \square$$

Now we try to find sufficient conditions such that $\mathbf{C}_i(\Delta t) \in G$. Let L be the smallest integer satisfying $2L - 3 \geq k$. We consider an L -point Legendre Gauss-Lobatto quadrature on I_i and denote the quadrature points as

$$\hat{S}_i = \{x_{i-\frac{1}{2}} = \hat{x}_i^1, \hat{x}_i^2, \dots, \hat{x}_i^{L-1}, \hat{x}_i^L = x_{i+\frac{1}{2}}\}.$$

Moreover, we denote the quadrature weights on $[-\frac{1}{2}, \frac{1}{2}]$ as $\hat{\omega}_l$ such that $\sum_{l=1}^L \hat{\omega}_l = 1$ and let $\hat{\omega} = \hat{\omega}_1 = \hat{\omega}_L = \frac{1}{L(L-1)}$. Then, we can get the following theorem.

Theorem 4.1. Suppose

$$\mathbf{w}_{i \pm \frac{1}{2}}^\pm \in G, \quad \mathbf{w}(\hat{x}_i^\ell) \in G \quad (\ell = 2, \dots, L-1), \quad (4.12)$$

then under the CFL condition

$$\frac{\Delta t}{\Delta x} \max_i \beta_{i+\frac{1}{2}} \leq \hat{\omega} = \frac{1}{L(L-1)},$$

we have $\mathbf{C}_i(\Delta t) \in G$ in (4.6). Moreover, the conclusion is also valid if (4.12) is replaced by a weaker condition

$$\mathbf{w}_{i \pm \frac{1}{2}}^\pm \in G, \quad \sum_{\ell=2}^{L-1} \frac{\hat{\omega}_\ell}{1 - 2\hat{\omega}} \mathbf{w}(\hat{x}_i^\ell) \in G. \quad (4.13)$$

Proof. Using the L -point Legendre Gauss-Lobatto quadrature rule on I_i and the definition of the numerical flux (3.4), we obtain

$$\begin{aligned} \mathbf{C}_i(\Delta t) &= \left(\hat{\omega}_1 - \frac{\beta_{i-\frac{1}{2}}}{2} \frac{\Delta t}{\Delta x_i} \right) \left(\mathbf{w}_{i-\frac{1}{2}}^+ + \frac{\Delta t}{2\Delta x_i} \left(\hat{\omega}_1 - \frac{\beta_{i-\frac{1}{2}}}{2} \frac{\Delta t}{\Delta x_i} \right)^{-1} \mathbf{f}(\mathbf{w}_{i-\frac{1}{2}}^+, \mathbf{q}_{i-\frac{1}{2}}^+) \right) \\ &\quad + \left(\hat{\omega}_L - \frac{\beta_{i+\frac{1}{2}}}{2} \frac{\Delta t}{\Delta x_i} \right) \left(\mathbf{w}_{i+\frac{1}{2}}^- - \frac{\Delta t}{2\Delta x_i} \left(\hat{\omega}_L - \frac{\beta_{i+\frac{1}{2}}}{2} \frac{\Delta t}{\Delta x_i} \right)^{-1} \mathbf{f}(\mathbf{w}_{i+\frac{1}{2}}^-, \mathbf{q}_{i+\frac{1}{2}}^-) \right) \\ &\quad + \frac{\beta_{i-\frac{1}{2}}}{2} \frac{\Delta t}{\Delta x_i} \left(\mathbf{w}_{i-\frac{1}{2}}^- + \beta_{i-\frac{1}{2}}^{-1} \mathbf{f}(\mathbf{w}_{i-\frac{1}{2}}^-, \mathbf{q}_{i-\frac{1}{2}}^-) \right) \end{aligned}$$

$$\begin{aligned}
& + \frac{\beta_{i+\frac{1}{2}}}{2} \frac{\Delta t}{\Delta x_i} \left(\mathbf{w}_{i+\frac{1}{2}}^+ - \beta_{i+\frac{1}{2}}^{-1} \mathbf{f} \left(\mathbf{w}_{i+\frac{1}{2}}^+, \mathbf{q}_{i+\frac{1}{2}}^+ \right) \right) \\
& + \sum_{\ell=2}^{L-1} \hat{\omega}_\ell \mathbf{w}(\hat{x}_i^\ell)
\end{aligned}$$

It is easy to check

$$\frac{\Delta t}{2\Delta x_i} \left(\hat{\omega}_1 - \frac{\beta_{i-\frac{1}{2}}}{2} \frac{\Delta t}{\Delta x_i} \right)^{-1} \leq \beta_{i-\frac{1}{2}}^{-1} \quad \text{and} \quad \frac{\Delta t}{2\Delta x_i} \left(\hat{\omega}_L - \frac{\beta_{i+\frac{1}{2}}}{2} \frac{\Delta t}{\Delta x_i} \right)^{-1} \leq \beta_{i+\frac{1}{2}}^{-1}$$

under the CFL condition $\frac{\Delta t}{\Delta x} \max_i \beta_{i+\frac{1}{2}} \leq \hat{\omega}$. Then by using Lemma 4.1, and the condition in (3.5), we have

$$\begin{aligned}
\mathbf{w}_{i-\frac{1}{2}}^+ \in G & \implies \mathbf{w}_{i-\frac{1}{2}}^+ + \frac{\Delta t}{2\Delta x_i} \left(\hat{\omega}_1 - \frac{\beta_{i-\frac{1}{2}}}{2} \frac{\Delta t}{\Delta x_i} \right)^{-1} \mathbf{f} \left(\mathbf{w}_{i-\frac{1}{2}}^+, \mathbf{q}_{i-\frac{1}{2}}^+ \right) \in G, \\
\mathbf{w}_{i+\frac{1}{2}}^- \in G & \implies \mathbf{w}_{i+\frac{1}{2}}^- - \frac{\Delta t}{2\Delta x_i} \left(\hat{\omega}_L - \frac{\beta_{i+\frac{1}{2}}}{2} \frac{\Delta t}{\Delta x_i} \right)^{-1} \mathbf{f} \left(\mathbf{w}_{i+\frac{1}{2}}^-, \mathbf{q}_{i+\frac{1}{2}}^- \right) \in G, \\
\mathbf{w}_{i-\frac{1}{2}}^- \in G & \implies \mathbf{w}_{i-\frac{1}{2}}^- + \beta_{i-\frac{1}{2}}^{-1} \mathbf{f} \left(\mathbf{w}_{i-\frac{1}{2}}^-, \mathbf{q}_{i-\frac{1}{2}}^- \right) \in G, \\
\mathbf{w}_{i+\frac{1}{2}}^+ \in G & \implies \mathbf{w}_{i+\frac{1}{2}}^+ - \beta_{i+\frac{1}{2}}^{-1} \mathbf{f} \left(\mathbf{w}_{i+\frac{1}{2}}^+, \mathbf{q}_{i+\frac{1}{2}}^+ \right) \in G.
\end{aligned}$$

Then $\mathbf{C}_i(\Delta t)$ is a convex combination of elements in G and hence we have $\mathbf{C}_i(\Delta t) \in G$. For more details of the weaker condition, we refer to [48]. $\square$

Next, we consider the condition for $\mathbf{D}_i \in G$. We adopt the M -point Gauss quadrature rule on I_i and denote the set of quadrature points on I_i as

$$\tilde{S}_i = \{x_i^1, x_i^2, \dots, x_i^M\}.$$

In addition, we denote

$$\begin{aligned}
\mathbf{w}(x_i^\alpha) &= (\rho_i^\alpha, (\rho u)_i^\alpha, E_i^\alpha, (r_1)_i^\alpha, \dots, (r_{N_s})_i^\alpha)^T, \\
\mathbf{s}(\mathbf{w}(x_i^\alpha)) &= (0, 0, 0, (\omega_1)_i^\alpha, \dots, (\omega_{N_s})_i^\alpha)^T,
\end{aligned}$$

for $\alpha = 1, \dots, M$. Moreover, we denote the quadrature weights on $[-\frac{1}{2}, \frac{1}{2}]$ as c_α , $\alpha = 1, \dots, M$. We choose M large enough such that $\tilde{\mathbf{s}}_i$ can be approximated accurately. In the numerical experiments, we choose $M = k + 1$. Then we can prove the following theorem.

Theorem 4.2. Suppose

$$\mathbf{w}(x_i^\alpha) \in G, \quad \alpha = 1, \dots, M,$$

then under the condition

$$\mu \geq \max_{\alpha=1, \dots, M} \left\{ \sum_{n=1}^{N_s} \frac{(\omega_n)_i^\alpha h_n(0)}{P(\mathbf{w}(x_i^\alpha))}, \max_{n=1, \dots, N_s} -\frac{(\omega_n)_i^\alpha}{(r_n)_i^\alpha}, 0 \right\}, \quad (4.14)$$

we have $\mathbf{D}_i = \frac{1}{\mu} \tilde{\mathbf{s}}_i + \bar{\mathbf{w}}_i \in G$.

Proof. By using the M -point Gauss quadrature rule on I_i , we have

$$\mathbf{D}_i = \frac{1}{\mu} \tilde{\mathbf{s}}_i + \bar{\mathbf{w}}_i = \sum_{\alpha=1}^M c_\alpha \left[\frac{1}{\mu} \mathbf{s}(\mathbf{w}(x_i^\alpha)) + \mathbf{w}(x_i^\alpha) \right]. \quad (4.15)$$

We need to find sufficient conditions such that $\frac{1}{\mu} \mathbf{s}(\mathbf{w}(x_i^\alpha)) + \mathbf{w}(x_i^\alpha) \in G$ for each fixed point x_i^α . Notice that

$$\frac{1}{\mu} \mathbf{s}(\mathbf{w}(x_i^\alpha)) + \mathbf{w}(x_i^\alpha) = \begin{pmatrix} \rho_i^\alpha \\ (\rho u)_i^\alpha \\ E_i^\alpha \\ (r_1)_i^\alpha + \frac{(\omega_1)_i^\alpha}{\mu} \\ \vdots \\ (r_{N_s})_i^\alpha + \frac{(\omega_{N_s})_i^\alpha}{\mu} \end{pmatrix}.$$

Since $\mathbf{w}(x_i^\alpha) \in G$, then we have $\rho_i^\alpha \geq 0$. Next, we require

$$(r_n)_i^\alpha + \frac{(\omega_n)_i^\alpha}{\mu} \geq 0, \quad n = 1, \dots, N_s,$$

$$P\left(\frac{1}{\mu} \mathbf{s}(x_i^\alpha) + \mathbf{w}(x_i^\alpha)\right) = P(\mathbf{w}(x_i^\alpha)) - \sum_{n=1}^{N_s} \frac{(\omega_n)_i^\alpha}{\mu} h_n(0) \geq 0,$$

which yield

$$\mu \geq -\frac{(\omega_n)_i^\alpha}{(r_n)_i^\alpha}, \quad n = 1, \dots, N_s,$$

$$\mu \geq \sum_{n=1}^{N_s} \frac{(\omega_n)_i^\alpha h_n(0)}{P(\mathbf{w}(x_i^\alpha))},$$

and hence we get (4.14). Then (4.15) shows that $\mathbf{D}_i$ is a convex combination of elements in G , and hence we have $\mathbf{D}_i \in G$. $\square$

Now we denote $S_i = \hat{S}_i \cup \tilde{S}_i$. Recall that the cell average at each time stage in (4.3)–(4.5) is a convex combination of the basic structures of $\mathbf{C}$ and $\mathbf{D}$. Then combining Theorems 4.1 and 4.2, we get the following one. The proof is straightforward and we omit it.

Theorem 4.3. Suppose $\mathbf{w}(x) \in G$, $\forall x \in S_i$, $i = 1, \dots, N$ and

$$\mu \geq \max_{i=1, \dots, N} \max_{\alpha=1, \dots, M} \left\{ \sum_{n=1}^{N_s} \frac{(\omega_n)_i^\alpha h_n(0)}{P(\mathbf{w}(x_i^\alpha))}, \max_{n=1, \dots, N_s} -\frac{(\omega_n)_i^\alpha}{(r_n)_i^\alpha}, 0 \right\} \quad (4.16)$$

for $\mathbf{w} = \mathbf{w}^n$, $\mathbf{w}^{(1)}$, $\mathbf{w}^{(2)}$, then we have $\bar{\mathbf{w}}_i^{n+1} \in G$ under the CFL condition

$$\frac{\Delta t}{\Delta x} \max_i \beta_{i+\frac{1}{2}} \leq \min\left\{ \frac{\alpha_{10}}{\beta_{10}}, \frac{\alpha_{20}}{\beta_{20}}, \frac{\alpha_{21}}{\beta_{21}}, \frac{\alpha_{30}}{\beta_{30}}, \frac{\alpha_{31}}{\beta_{31}}, \frac{\alpha_{32}}{\beta_{32}} \right\} \hat{\omega}. \quad (4.17)$$

4.3. Bound preserving limiter

At each time stage, we have $\bar{\mathbf{w}}_i \in G$. Next, we proceed to discuss the bound-preserving limiter on each cell I_i and replace the original polynomial $\mathbf{w}(x)$ with a new one $\mathbf{w}^{new}(x)$, such that $\mathbf{w}^{new}(x) \in G$, $\forall x \in S_i$. The full algorithm can be found in [10]. For the completeness of this paper, we review the main steps as follows. For more details and discussions, we refer to [10].

1. Set a small number $\epsilon = 10^{-13}$. If $\bar{\rho} > \epsilon$, then we proceed to the next step. Otherwise, we take $\mathbf{w}^{new} = \bar{\mathbf{w}}$ and skip the following steps.
2. Modify the total density. Compute

$$\rho_{min} = \min_{x \in S_i} \rho(x).$$

If $\rho_{min} < 0$, then we take

$$\hat{\rho} = \bar{\rho} + \theta(\rho - \bar{\rho}), \quad \hat{r}_n = \bar{r}_n + \theta(r_n - \bar{r}_n), \quad n = 1, \dots, N_s$$

with

$$\theta = \frac{\bar{\rho} - \epsilon}{\bar{\rho} - \rho_{min}}.$$

3. Modify the mass fraction. For $1 \leq n \leq N_s$, define $Q_n = \{x \in S_i : \hat{r}_n(x) \leq 0\}$. Take

$$\tilde{r}_n = \hat{r}_n + \theta \left(\frac{\tilde{r}_n}{\bar{\rho}} \hat{\rho} - \hat{r}_n \right), \quad n = 1, \dots, N_s$$

with

$$\theta = \max_{1 \leq n \leq N_s} \max_{x \in Q_n} \left\{ \frac{-\hat{r}_n(x) \bar{\rho}}{\tilde{r}_n \hat{\rho}(x) - \hat{r}_n(x) \bar{\rho}}, 0 \right\}.$$

4. Modify the pressure. Denote $\tilde{\mathbf{w}} = (\hat{\rho}, \rho u, E, \tilde{r}_1, \dots, \tilde{r}_{N_s})^T$. For each $x \in S_i$, if $\tilde{\mathbf{w}}(x) \in G$, then take $\theta_x = 1$. Otherwise, take

$$\theta_x = \frac{P(\tilde{\mathbf{w}})}{P(\tilde{\mathbf{w}}) - P(\tilde{\mathbf{w}}(x))}.$$

Then we compute

$$\mathbf{w}^{new} = \tilde{\mathbf{w}} + \theta(\tilde{\mathbf{w}} - \tilde{\mathbf{w}}), \quad \theta = \min_{x \in S_i} \theta_x.$$

5. Two dimensional problems

In this section, we consider the two dimensional governing equations (2.1). Now we have $m = N_s + 4$ and $\mathbf{w} \in \mathbb{R}^m$. We will show the numerical schemes as well as the conservative property in Section 5.1. Then we give the bound-preserving technique in Section 5.2.

5.1. Numerical schemes

The diffusion fluxes $\mathbf{f}^d$ and $\mathbf{g}^d$ are dependent on the derivatives of conserved variables $\mathbf{w}$. Similar to the one dimensional case, we first introduce an auxiliary variable $\mathbf{Q} := \nabla \mathbf{w}$, where $\nabla \mathbf{w} \in \mathbb{R}^{m \times 2}$ is the row-wise gradient of $\mathbf{w}$, and then rewrite the model (2.1) into the following system

$$\mathbf{w}_t = -\text{div} \mathbf{F}^a(\mathbf{w}) + \text{div} \mathbf{F}^d(\mathbf{w}, \mathbf{Q}) + \mathbf{s},$$

$$\mathbf{Q} = \nabla \mathbf{w},$$

where $\mathbf{F}^a := [\mathbf{f}^a, \mathbf{g}^a] \in \mathbb{R}^{m \times 2}$ and $\mathbf{F}^d := [\mathbf{f}^d, \mathbf{g}^d] \in \mathbb{R}^{m \times 2}$. For any $\phi \in \mathbb{R}^{m \times 2}$, we let ϕ_i be the i th row of ϕ and define the divergence as $\text{div} \phi = [\text{div} \phi_1, \dots, \text{div} \phi_m]^T$. Let $\Omega_h = \{K\}$ be a partition of the computational domain Ω with polygonal cells. We define the finite element space V_h^k as

$$V_h^k = \{\psi : \psi|_K \in P^k(K), \forall K \in \Omega_h\}.$$

For simplicity, we still use the notations $\mathbf{w}$ and $\mathbf{Q}$ as the numerical approximations.

Following the idea in [48], we regard $\mathbf{F} = \mathbf{F}^a - \mathbf{F}^d$ as a single flux and formally treat $\text{div} \mathbf{F}$ as a convection term. Then the LDG scheme is to find $\mathbf{w} \in [V_h^k]^m$ and $\mathbf{Q} \in [V_h^k]^{m \times 2}$, such that for any test functions $\psi \in [V_h^k]^m$ and $\phi \in [V_h^k]^{m \times 2}$ and any cell $K \in \Omega_h$, we have

$$\iint_K \mathbf{w}_t \cdot \psi \, dV = \iint_K \mathbf{F} : \nabla \psi \, dV - \int_{\partial K} \hat{\mathbf{F}}^n \cdot \psi \, ds + \iint_K \mathbf{s} \cdot \psi \, dV, \quad (5.1)$$

$$\iint_K \mathbf{Q} : \phi \, dV = - \iint_K \mathbf{w} \cdot \text{div} \phi \, dV + \int_{\partial K} \hat{\mathbf{w}} \cdot (\phi \mathbf{n}) \, ds, \quad (5.2)$$

where $\mathbf{n}$ is the unit outer normal on ∂K . Central flux can be used for $\mathbf{w}$:

$$\hat{\mathbf{w}}|_e = \frac{1}{2}(\mathbf{w}^{int} + \mathbf{w}^{ext}),$$

where $e \in \partial K$ is an edge of the cell K , $\mathbf{w}^{int}$ and $\mathbf{w}^{ext}$ denote the approximations to $\mathbf{w}$ on e taken from interior and exterior of K , respectively. The numerical flux for $\mathbf{F}$ is taken as the Lax-Friedrichs type positivity-preserving flux

$$\hat{\mathbf{F}}^n(\mathbf{w}^{int}, \mathbf{Q}^{int}, \mathbf{w}^{ext}, \mathbf{Q}^{ext})|_e = \frac{1}{2}[\mathbf{F}(\mathbf{w}^{int}, \mathbf{Q}^{int})\mathbf{n} + \mathbf{F}(\mathbf{w}^{ext}, \mathbf{Q}^{ext})\mathbf{n} - \beta_e(\mathbf{w}^{ext} - \mathbf{w}^{int})], \quad (5.3)$$

where

$$\beta_e \geq \max \left\{ \left| \mathbf{V} \cdot \mathbf{n} + \frac{b}{2\rho P(\mathbf{w})} \right| + \frac{\sqrt{D}}{2\rho P(\mathbf{w})}, |\mathbf{V} + \mathbf{V}_i| \cdot \mathbf{n}|, i = 1, \dots, N_s \right\} \quad (5.4)$$

with

$$\mathbf{V} = [u, v]^T, \quad b = \rho(\mathbf{q} \cdot \mathbf{n}) - \rho \sum_{i=1}^{N_s} r_i h_i(0)(\mathbf{V}_i \cdot \mathbf{n}), \quad D = b^2 + 2\rho P(\mathbf{w}) \|p\mathbf{n} - \tau\mathbf{n}\|^2.$$

The maximum in (5.4) is taken over all $\mathbf{w}^{int}$, $\mathbf{Q}^{int}$, $\mathbf{w}^{ext}$ and $\mathbf{Q}^{ext}$ along the edge e . The condition on β_e is to guarantee the bound-preserving property.

In the above scheme, the numerical flux $\hat{\mathbf{F}}^n$ is an approximation to $\mathbf{F}\mathbf{n}$ on the edge of each cell. By using (2.8), we know that $\mathbf{F}\mathbf{n}$ satisfies

$$(\mathbf{F}\mathbf{n}) \cdot \mathbf{v} = \mathbf{v}^T [\mathbf{f}^a - \mathbf{f}^d, \mathbf{g}^a - \mathbf{g}^d] \mathbf{n} = 0, \quad (5.5)$$

where $\mathbf{v} = [1, 0, 0, 0, -1, \dots, -1] \in \mathbb{R}^m$. Then we clearly know that $\hat{\mathbf{F}}^n$ preserves the same property and have the following theorem.

Theorem 5.1. *The numerical flux $\hat{\mathbf{F}}^n$ is consistent in the sense that $\hat{\mathbf{F}}^n \cdot \mathbf{v} = 0$.*

Now, it is easy to prove that the semi-discrete scheme in two dimensions preserves the conservative property of the original model. The idea of proof is almost the same as that for the one dimensional case, and hence we omit it.

Theorem 5.2. *The semi-discrete scheme (5.1)-(5.2) is conservative in the sense that*

$$\frac{d}{dt}(\mathbf{w} \cdot \mathbf{v}) = 0,$$

as long as the numerical flux $\hat{\mathbf{F}}^n$ is consistent.

Next, we apply the third-order CMERK method [10] in time and obtain:

$$(\mathbf{w}^{(1)}, \psi)_K = [\alpha_{10}(\mathbf{w}^n, \psi)_K + \beta_{10}\Delta t F_K(\mathbf{w}^n, \psi) + \beta_{10}\Delta t(s(\mathbf{w}^n) + \mu\mathbf{w}^n, \psi)_K] / A_1, \quad (5.6)$$

$$\begin{aligned} (\mathbf{w}^{(2)}, \psi)_K &= [\alpha_{20}(\mathbf{w}^n, \psi)_K + \beta_{20}\Delta t F_K(\mathbf{w}^n, \psi) + \beta_{20}\Delta t(s(\mathbf{w}^n) + \mu\mathbf{w}^n, \psi)_K] / A_2 \\ &\quad + \frac{e^{\beta_{10}\mu\Delta t}}{A_2} \left[\alpha_{21}(\mathbf{w}^{(1)}, \psi)_K + \beta_{21}\Delta t F_K(\mathbf{w}^{(1)}, \psi) + \beta_{21}\Delta t(s(\mathbf{w}^{(1)}) + \mu\mathbf{w}^{(1)}, \psi)_K \right], \end{aligned} \quad (5.7)$$

$$\begin{aligned} (\mathbf{w}^{n+1}, \psi)_K &= [\alpha_{30}(\mathbf{w}^n, \psi)_K + \beta_{30}\Delta t F_K(\mathbf{w}^n, \psi) + \beta_{30}\Delta t(s(\mathbf{w}^n) + \mu\mathbf{w}^n, \psi)_K] / A_3 \\ &\quad + \frac{e^{\beta_{10}\mu\Delta t}}{A_3} \left[\alpha_{31}(\mathbf{w}^{(1)}, \psi)_K + \beta_{31}\Delta t F_K(\mathbf{w}^{(1)}, \psi) + \beta_{31}\Delta t(s(\mathbf{w}^{(1)}) + \mu\mathbf{w}^{(1)}, \psi)_K \right] \\ &\quad + \frac{e^{A\mu\Delta t}}{A_3} \left[\alpha_{32}(\mathbf{w}^{(2)}, \psi)_K + \beta_{32}\Delta t F_K(\mathbf{w}^{(2)}, \psi) + \beta_{32}\Delta t(s(\mathbf{w}^{(2)}) + \mu\mathbf{w}^{(2)}, \psi)_K \right], \end{aligned} \quad (5.8)$$

where

$$F_K(\mathbf{w}, \psi) := \iint_K \mathbf{F} : \nabla \psi \, dV - \int_{\partial K} \hat{\mathbf{F}}^n \cdot \psi \, ds,$$

and $(\cdot, \cdot)_K$ denotes the L_2 inner product on K .

As in the one dimensional case, it is easy to prove that the above fully discrete scheme is also conservative. Then our numerical results always preserve the condition that $\sum_{i=1}^{N_s} Y_i = 1$ at each time level.

Theorem 5.3. *The fully discrete scheme (5.6)-(5.8) is conservative in the sense that if $\mathbf{w}^n \cdot \mathbf{v} = 0$, then we have*

$$\mathbf{w}^{n+1} \cdot \mathbf{v} = \mathbf{w}^n \cdot \mathbf{v} = 0.$$

In other words, if we have $\sum_{i=1}^{N_s} Y_i = 1$ at time level n , then this equation is still valid at time level $n+1$.

5.2. Bound preserving technique

Similar to the one dimensional case, we define the admissible set to be

$$G = \left\{ \mathbf{w} = \begin{pmatrix} \rho \\ \rho u \\ \rho v \\ E \\ r_1 \\ \vdots \\ r_{N_s} \end{pmatrix} : \rho \geq 0, \quad P(\mathbf{w}) \geq 0, \quad r_i \geq 0, \quad i = 1, \dots, N_s \right\},$$

where $P(\mathbf{w})$ is redefined as

$$P(\mathbf{w}) := E - \sum_{i=1}^{N_s} r_i h_i(0) - \frac{1}{2} \rho (u^2 + v^2). \quad (5.9)$$

Once the lower bound of each mass fraction is preserved, the conservative property indicates that the upper bound is also preserved.

By taking suitable test functions in the fully discrete scheme (5.6)-(5.8), we can get the equations for computing the cell average

$$\bar{\mathbf{w}}_K = \frac{1}{|K|} \iint_K \mathbf{w} dV,$$

where $|K|$ is the area of the cell K . The resulting equations are similar to (4.3)-(4.5). One only need to replace the operators C_i and D_i with

$$\mathbf{C}_K(\Delta t) := \bar{\mathbf{w}}_K - \frac{\Delta t}{|K|} \int_{\partial K} \hat{\mathbf{F}}^n ds, \quad \mathbf{D}_K = \frac{1}{\mu} \bar{\mathbf{s}}_K + \bar{\mathbf{w}}_K,$$

respectively. Next, we need to find sufficient conditions such that $\mathbf{C}_K(\Delta t) \in G$ and $\mathbf{D}_K \in G$, respectively. Similar to Lemma 4.1, we have the following lemma for the two dimensional problem.

Lemma 5.1. Suppose $\mathbf{w} = (\rho, \rho \mathbf{V}^T, E, r_1, \dots, r_{N_s})^T \in G$ and

$$\mathbf{F} = [\mathbf{f}^a - \mathbf{f}^d, \mathbf{g}^a - \mathbf{g}^d] = \begin{pmatrix} \rho \mathbf{V}^T \\ \rho \mathbf{V} \otimes \mathbf{V} + p \mathbb{I} - \tau \\ (E + p) \mathbf{V}^T - \mathbf{V}^T \tau + \mathbf{q}^T \\ r_1 (\mathbf{V} + \mathbf{V}_1)^T \\ \vdots \\ r_{N_s} (\mathbf{V} + \mathbf{V}_{N_s})^T \end{pmatrix}$$

where $\mathbf{V} = [u, v]^T$ and $\mathbb{I}$ is the unit tensor. Here $p, \tau, \mathbf{q}$ and $\mathbf{V}_i, i = 1, \dots, N_s$ are not necessarily dependent on $\mathbf{w}$. For any unit vector $\mathbf{n}$, we denote $\tilde{v} = \mathbf{V} \cdot \mathbf{n}, \tilde{q} = \mathbf{q} \cdot \mathbf{n}, \tilde{\tau} = \tau \mathbf{n} \cdot \mathbf{n}$ and $\tilde{v}_i = \mathbf{V}_i \cdot \mathbf{n}$ for $i = 1, \dots, N_s$. Then we have $\mathbf{w} \pm \beta^{-1} \mathbf{F} \mathbf{n} \in G$ if and only if

$$\beta \geq \max \left\{ |\tilde{v} + \frac{b}{2\rho P(\mathbf{w})}| + \frac{\sqrt{D}}{2\rho P(\mathbf{w})}, |\tilde{v} + \tilde{v}_i|, i = 1, \dots, N_s \right\}$$

where

$$b = \rho \tilde{q} - \rho \sum_{i=1}^{N_s} r_i h_i(0) \tilde{v}_i, \quad D = b^2 + 2\rho P(\mathbf{w}) \|\mathbf{p} \mathbf{n} - \tilde{\tau}\|^2.$$

Proof. First we have

$$\mathbf{w} \pm \beta^{-1} \mathbf{F} \mathbf{n} = \begin{pmatrix} \rho \pm \beta^{-1} \rho \tilde{v} \\ \rho \mathbf{V} \pm \beta^{-1} (\rho \tilde{v} \mathbf{V} + \mathbf{p} \mathbf{n} - \tilde{\tau}) \\ E \pm \beta^{-1} ((E + p) \tilde{v} - \mathbf{V} \cdot \tilde{\tau} + \tilde{q}) \\ r_1 \pm \beta^{-1} r_1 \tilde{u}_1 \\ \vdots \\ r_{N_s} \pm \beta^{-1} r_{N_s} \tilde{u}_{N_s} \end{pmatrix},$$

where $\tilde{u}_i = \tilde{v} + \tilde{v}_i$. In order to obtain $\mathbf{w} \pm \beta^{-1} \mathbf{F} \mathbf{n} \in G$, we require

$$\begin{aligned} \rho \pm \beta^{-1} \rho \tilde{v} &\geq 0, \\ r_i \pm \beta^{-1} r_i \tilde{u}_i &\geq 0, \quad i = 1, \dots, N_s \\ E \pm \beta^{-1} ((E + p) \tilde{v} - \mathbf{V} \cdot \tilde{\tau} + \tilde{q}) - \sum_{i=1}^{N_s} (r_i \pm \beta^{-1} r_i \tilde{u}_i) h_i(0) \\ &- \frac{1}{2} \|\rho \mathbf{V} \pm \beta^{-1} (\rho \tilde{v} \mathbf{V} + p \mathbf{n} - \tilde{\tau})\|^2 / (\rho \pm \beta^{-1} \rho \tilde{v}) \geq 0 \end{aligned}$$

Since $\beta > 0$, by multiplying β on both sides, the above conditions are equivalent to

$$(\beta \pm \tilde{v}) \rho \geq 0, \quad (5.10)$$

$$(\beta \pm \tilde{u}_i) r_i \geq 0, \quad i = 1, \dots, N_s, \quad (5.11)$$

$$\rho P(\mathbf{w}) \beta^2 \pm (2\rho \tilde{v} P(\mathbf{w}) + b) \beta + c \geq 0, \quad (5.12)$$

with

$$c = \rho \tilde{v}^2 P(\mathbf{w}) + b \tilde{v} - \frac{1}{2} \|p \mathbf{n} - \tilde{\tau}\|^2.$$

Notice that the conditions (5.10)-(5.12) are similar to (4.7) and (4.10). Following the same idea of proof as in Lemma 4.1, we can obtain the condition for β . $\square$

Now we try to find sufficient conditions such that $\mathbf{C}_K(\Delta t) \in G$. We assume that the number of edges of K is E and denote the i th edge as e_i . For integrals along each edge e_i , we use the I -point Gauss quadrature and denote the ν -th quadrature point as $\mathbf{x}_{\nu,i}$. Let b_ν ($\nu = 1, \dots, I$) denote the I -point Gauss quadrature weights on $[-\frac{1}{2}, \frac{1}{2}]$. Then we have

$$\mathbf{C}_K(\Delta t) = \bar{\mathbf{w}}_K - \frac{\Delta t}{|K|} \sum_{i=1}^E |e_i| \sum_{\nu=1}^I b_\nu \hat{\mathbf{F}}^n(\mathbf{w}_{\nu,i}^{int}, \mathbf{Q}_{\nu,i}^{int}, \mathbf{w}_{\nu,i}^{ext}, \mathbf{Q}_{\nu,i}^{ext}), \quad (5.13)$$

where $\mathbf{w}_{\nu,i}^{int}$ and $\mathbf{w}_{\nu,i}^{ext}$ are the numerical approximations of $\mathbf{w}$ at $\mathbf{x}_{\nu,i}$ from interior and exterior of K , respectively. As in the one dimensional case, we need an accurate enough quadrature on K to represent the cell average $\bar{\mathbf{w}}_K$ as a convex combination of certain point values. The quadrature points should include $\mathbf{x}_{\nu,i}$ for all ν and i . We denote the set of quadrature points as $\hat{S}_K$. Assuming that the total number of quadrature points is J , we denote the remaining points as $\mathbf{x}_\lambda$ ($\lambda = EI + 1, \dots, J$). Moreover, let $a_{\nu,i}$ and a_λ denote the corresponding normalized quadrature weights so that $\sum_{i=1}^E \sum_{\nu=1}^I a_{\nu,i} + \sum_{\lambda=EI+1}^J a_\lambda = 1$. Then, we have the following theorem.

Theorem 5.4. A sufficient condition for $\mathbf{C}_K(\Delta t) \in G$ in (4.6) is

$$\begin{aligned} \mathbf{w}_{\nu,i}^{int}, \mathbf{w}_{\nu,i}^{ext} &\in G, \quad i = 1, \dots, E, \nu = 1, \dots, I, \\ \mathbf{w}(\mathbf{x}_\lambda) &\in G, \quad \lambda = EI + 1, \dots, J, \end{aligned}$$

under the CFL condition

$$\Delta t \frac{|e_i|}{|K|} \max_i \beta_{e_i} \leq \min_{\nu,i} \frac{a_{\nu,i}}{b_\nu}.$$

Proof. By using the J -point quadrature rule on the cell K , we obtain the following cell average decomposition

$$\bar{\mathbf{w}}_K = \frac{1}{|K|} \iint_K \mathbf{w} dV = \sum_{i=1}^E \sum_{\nu=1}^I a_{\nu,i} \mathbf{w}_{\nu,i}^{int} + \sum_{\lambda=EI+1}^J a_\lambda \mathbf{w}(\mathbf{x}_\lambda)$$

Plugging the above decomposition and the definition of the numerical flux (5.3) into (5.13), we obtain

$$\begin{aligned} \mathbf{C}_K(\Delta t) &= \sum_{\lambda=EI+1}^J a_\lambda \mathbf{w}(\mathbf{x}_\lambda) + \sum_{i=1}^E \sum_{\nu=1}^I \Lambda_{\nu,i} \beta_{e_i} [\mathbf{w}_{\nu,i}^{ext} - \beta_{e_i}^{-1} \mathbf{F}(\mathbf{w}_{\nu,i}^{ext}, \mathbf{Q}_{\nu,i}^{ext}) \mathbf{n}_i] \\ &\quad + \sum_{i=1}^E \sum_{\nu=1}^I (a_{\nu,i} - \Lambda_{\nu,i} \beta_{e_i}) [\mathbf{w}_{\nu,i}^{int} - \Lambda_{\nu,i} (a_{\nu,i} - \Lambda_{\nu,i} \beta_{e_i})^{-1} \mathbf{F}(\mathbf{w}_{\nu,i}^{int}, \mathbf{Q}_{\nu,i}^{int}) \mathbf{n}_i] \end{aligned}$$

where $\Lambda_{v,i} = \frac{1}{2} \Delta t \frac{|e_i|}{|K|} b_v$. Notice that $\Lambda_{v,i}(a_{v,i} - \Lambda_{v,i}\beta_{e_i})^{-1} \leq \beta_{e_i}^{-1}$ if and only if $0 \leq \Delta t \frac{|e_i|}{|K|} \beta_{e_i} \leq \frac{a_{v,i}}{b_v}$. By using Lemma 5.1, under the CFL condition $\Delta t \frac{|e_i|}{|K|} \max_i \beta_{e_i} \leq \min_{v,i} \frac{a_{v,i}}{b_v}$ and the condition in (5.4), we have

$$\begin{aligned} \mathbf{w}_{v,i}^{ext} \in G &\implies \mathbf{w}_{v,i}^{ext} - \beta_{e_i}^{-1} \mathbf{F}(\mathbf{w}_{v,i}^{ext}, \mathbf{Q}_{v,i}^{ext}) \mathbf{n}_i \in G, \\ \mathbf{w}_{v,i}^{int} \in G &\implies \mathbf{w}_{v,i}^{int} - \Lambda_{v,i}(a_{v,i} - \Lambda_{v,i}\beta_{e_i})^{-1} \mathbf{F}(\mathbf{w}_{v,i}^{int}, \mathbf{Q}_{v,i}^{int}) \mathbf{n}_i \in G. \end{aligned}$$

Then $\mathbf{C}_K(\Delta t)$ is a convex combination of elements in G and hence $\mathbf{C}_K(\Delta t) \in G$. $\square$

Remark 5.1. The J point quadrature rule on the cell K is not unique. For rectangular mesh with lengths of sides being Δx and Δy , one can use tensor products of Gauss quadrature in one direction and Gauss-Lobatto quadrature in another direction. Then the CFL condition becomes $\Delta t(\frac{1}{\Delta x} + \frac{1}{\Delta y}) \max_i \beta_{e_i} \leq \frac{1}{P(P-1)}$, where $P = \lceil (k+3)/2 \rceil$. For more details and quadrature on other type of cells, we refer to [48]. For triangular meshes, the detailed construction of the quadrature points has been discussed in [52].

Next, we consider the condition for $\mathbf{D}_K \in G$. We adopt an accurate enough Gauss quadrature rule on the cell K and denote the set of quadrature points as $\tilde{S}_K$. Also, we denote the density and mass rate of production of species n at a certain point $\mathbf{x}$ as $r_n(\mathbf{x})$ and $\omega_n(\mathbf{x})$, respectively. Then we can prove the following theorem. The proof is similar to Theorem 4.2 and hence we omit it.

Theorem 5.5. Suppose

$$\mathbf{w}(\mathbf{x}) \in G, \quad \forall \mathbf{x} \in \tilde{S}_K,$$

then under the condition

$$\mu \geq \max_{\mathbf{x} \in \tilde{S}_K} \left\{ \sum_{n=1}^{N_s} \frac{\omega_n(\mathbf{x}) h_n(0)}{P(\mathbf{w}(\mathbf{x}))}, \max_{n=1, \dots, N_s} -\frac{\omega_n(\mathbf{x})}{r_n(\mathbf{x})}, 0 \right\}, \quad (5.14)$$

we have $\mathbf{D}_K = \frac{1}{\mu} \tilde{\mathbf{S}}_K + \bar{\mathbf{w}}_K \in G$.

Now we denote $S_K = \hat{S}_K \cup \tilde{S}_K$. Recall that the cell average at each time stage is a convex combination of the basic structures $\mathbf{C}$ and $\mathbf{D}$. Then combining Theorems 5.4 and 5.5, we get the following one, whose proof is straightforward hence we omit it.

Theorem 5.6. Suppose $\mathbf{w}(\mathbf{x}) \in G, \forall \mathbf{x} \in S_K, \forall K$ and

$$\mu \geq \max_K \max_{\mathbf{x} \in \tilde{S}_K} \left\{ \sum_{n=1}^{N_s} \frac{\omega_n(\mathbf{x}) h_n(0)}{P(\mathbf{w}(\mathbf{x}))}, \max_{n=1, \dots, N_s} -\frac{\omega_n(\mathbf{x})}{r_n(\mathbf{x})}, 0 \right\}$$

for $\mathbf{w} = \mathbf{w}^n, \mathbf{w}^{(1)}, \mathbf{w}^{(2)}$, then we have $\bar{\mathbf{w}}_K^{n+1} \in G$ under the CFL condition

$$\Delta t \frac{|e_i|}{|K|} \max_i \beta_{e_i} \leq \min \left\{ \frac{\alpha_{10}}{\beta_{10}}, \frac{\alpha_{20}}{\beta_{20}}, \frac{\alpha_{21}}{\beta_{21}}, \frac{\alpha_{30}}{\beta_{30}}, \frac{\alpha_{31}}{\beta_{31}}, \frac{\alpha_{32}}{\beta_{32}} \right\} \min_{v,i} \frac{a_{v,i}}{b_v}.$$

At each time stage, we already have $\bar{\mathbf{w}}_K \in G$. Then, we can apply the bound-preserving limiter on each cell K and replace the original polynomial $\mathbf{w}(x)$ with a new one $\mathbf{w}^{new}(x)$, such that $\mathbf{w}^{new}(x) \in G$ for any $x \in S_K$. The algorithm is the same as that given in Section 4.3. One only need to replace S_i with S_K .

6. Numerical examples

In this section, we test some numerical examples. We expand the exponential terms in CMERK method as demonstrated in Remark 3.1.

For the one-dimensional problem, our scheme is bound-preserving with suitable μ satisfying (4.16) for $\mathbf{w} = \mathbf{w}^n, \mathbf{w}^{(1)}, \mathbf{w}^{(2)}$ under the CFL condition (4.17) with $\beta_{i+\frac{1}{2}}$ defined in (3.5). But it is hard to accurately estimate the values of μ for the inner time stages $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$, given solutions $\mathbf{w}^n$ at time step n . Also, the constraint (4.17) is just a sufficient condition and may result in unnecessarily small time steps. In practice at time level n , we let μ satisfy (4.16) for $\mathbf{w} = \mathbf{w}^n$ and set the time step size as

$$\Delta t = \min \left\{ a \frac{\Delta x}{\beta^*}, b \frac{\Delta x^2}{\eta^*} \right\}, \quad (6.1)$$

where

$$\beta^* = \max_i \max_{\mathbf{w}_{i+\frac{1}{2}}^-, \mathbf{w}_{i+\frac{1}{2}}^+} \left(|u| + \sqrt{\frac{p^2}{2\rho P(\mathbf{w})}} \right),$$

and $\eta^* = \max_{\mathbf{w}} \eta$ is the maximum mixture viscosity. Here a and b are two parameters. Then we restart the computation with the value of μ doubled and the time step size halved when non-physical cell averages emerge in any stage of CMERK. Theorem 4.3 ensures that there will be no endless loops for such a treatment. The recomputation will end at least when μ is large enough to satisfy (4.16) and Δt is small enough to satisfy (4.17). In the numerical examples in this section, very few recomputations are observed. For the two-dimensional problem, we apply the similar treatment. At time level n , we compute μ based on $\mathbf{w}^n$ and set the time step size as

$$\Delta t = \min\{a \frac{\Delta x}{\beta^*}, b \frac{\Delta x^2}{\eta^*}\}, \quad (6.2)$$

where

$$\beta = \max_e \max_{\mathbf{w}^{int}, \mathbf{w}^{ext}} \left(|\mathbf{V} \cdot \mathbf{n}| + \sqrt{\frac{p^2}{2\rho P(\mathbf{w})}} \right)$$

and $\Delta x = \min_K \min_{e \in \partial K} \frac{|K|}{|e|}$.

For all examples, we apply the DG method coupled with the CMERK method in time. Only bound-preserving techniques are added to show the performance of our numerical methods. For Examples 6.2, 6.3 and 6.4, we also show reference solutions which are computed by using our schemes on dense meshes with extra TVD limiters added.

Example 6.1. Accuracy test in 1D

In this example, we consider the one dimensional problem (3.1) and test the accuracy of our scheme. Two species are considered. The first species is H_2 and the second species is O_2 . The computational domain is $[0, 2\pi]$. We use the following initial conditions

$$\begin{aligned} u &= 1 \text{ m/s}, \\ p &= 1 \text{ Pa}, \\ \rho &= 0.1(2 + \sin(x) + \cos(x)) \text{ kg/m}^3, \\ r_1 &= 0.1(1 + \sin(x)) \text{ kg/m}^3, \end{aligned}$$

and periodic boundary conditions. Moreover, we artificially define the source terms as $\omega_1 = -c(r_1)^7$ and $\omega_2 = c(r_1)^7$. The parameter c can be used to adjust the stiffness of the problem. The final time is taken as $T = 0.5$.

We apply the DG method with piecewise P^2 polynomials. The two parameters in the time step (6.1) are taken as $a = 0.1$ and $b = 0.001$. The numerical errors of r_1 with different choices of c are listed in the left part of Table 6.1. Both nonstiff ($c = 100$) and stiff ($c = 10000$) cases are calculated. As shown in the table, we can observe the expected third order of accuracy of our scheme. For this problem, the total density should be nonnegative and the mass fractions should be between 0 and 1. Hence, we further add the bound preserving limiter and the results are listed in the right part of Table 6.1. The percentage of cells that have been modified by the limiter is listed in the last column. By comparing the results with and without limiter, we can see that the limiter does not harm the original high order of accuracy. To test the advantage of the current CMERK methods over the traditional explicit SSP-RK3 time discretizations for stiff problems, we take $c = 10^6$ and compare different methods in Table 6.2. By taking the time step Δt as in (6.1) with a and b defined above, we can observe high order of accuracy for the CMERK method. However, for the RK3 method with the same time step Δt , the code will blow up when Δx is not small enough. As shown in the last column, we need to further reduce the time step size and hence need more computational cost to obtain reasonable results.

Example 6.2. He/N₂ shock tube problem

In this example, we consider a multi-component flow without chemical reactions. The first species is helium (He) and the second species is nitrogen (N₂). This test problem calculates a Riemann problem in a 1 m long tube with initial conditions given by

$$(T, u, p, Y_1, Y_2)(x, 0) = \begin{cases} (300 \text{ K}, 0 \text{ m/s}, 10 \text{ atm}, 1, 0), & x \leq 0.4 \text{ m}, \\ (300 \text{ K}, 0 \text{ m/s}, 1 \text{ atm}, 0, 1), & x > 0.4 \text{ m}. \end{cases}$$

Table 6.1

Accuracy test for the one dimensional problem.

N	Without limiter				With limiter				
	L^2 norm	order	L^∞ norm	order	L^2 norm	order	L^∞ norm	order	percentage
$c=100$									
10	5.54E-04	–	1.64E-03	–	1.34E-03	–	6.64E-03	–	6.88%
20	8.94E-05	2.63	3.59E-04	2.20	1.25E-04	3.43	6.64E-04	3.32	2.19%
40	1.21E-05	2.89	4.65E-05	2.95	1.35E-05	3.21	6.00E-05	3.47	0.55%
80	1.58E-06	2.93	6.36E-06	2.87	1.59E-06	3.08	6.62E-06	3.18	0.21%
160	2.01E-07	2.97	8.05E-07	2.98	2.02E-07	2.98	8.09E-07	3.03	5.39E-2%
320	2.53E-08	2.99	1.01E-07	3.00	2.54E-08	2.99	1.01E-07	3.00	8.58E-3%
$c=10000$									
10	6.19E-04	–	2.10E-03	–	1.44E-03	–	7.47E-03	–	7.50%
20	9.98E-05	2.63	5.44E-04	1.95	1.33E-04	3.44	7.04E-04	3.41	1.41%
40	1.50E-05	2.73	7.04E-05	2.95	1.54E-05	3.11	7.06E-05	3.32	0.43%
80	2.11E-06	2.83	1.09E-05	2.70	2.12E-06	2.86	1.09E-05	2.70	0.21%
160	2.79E-07	2.92	1.48E-06	2.88	2.80E-07	2.92	1.48E-06	2.88	5.39E-2%
320	3.57E-08	2.97	1.90E-07	2.96	3.57E-08	2.97	1.90E-07	2.96	8.58E-3%

Table 6.2Accuracy test for the one dimensional stiff problem with $c = 10^6$.

N	CMERK with Δt		RK3 with Δt		RK3 with $\frac{2}{3}\Delta t$	
	L^2 norm	order	L^2 norm	order	L^2 norm	order
10	1.84E-03	–	NAN	–	6.97E-04	–
20	1.97E-04	3.22	NAN	–	1.52E-04	2.20
40	2.98E-05	2.73	2.83E-05	–	2.77E-05	2.46
80	4.88E-06	2.61	5.75E-06	2.30	4.41E-06	2.65
160	7.38E-07	2.72	9.48E-07	2.60	6.57E-07	2.75
320	1.00E-07	2.88	1.24E-07	2.93	8.98E-08	2.87
640	1.28E-08	2.96	1.54E-08	3.01	1.16E-08	2.95

We use 1000 cells in the computational domain $[0, 1]$. Third order ($k = 2$) DG method is adopted for the spacial discretization. The two parameters in the time step (6.1) are taken as $a = 0.1$ and $b = 0.001$. The solutions at the time $300 \mu s$ are shown in Fig. 6.1, which are consistent with the results shown in [19]. We use black lines to show the reference solutions obtained by using 5000 cells, and use red circles to denote the numerical solutions with 1000 cells. We can see that our method preserves the positivity of the density and pressure, and the two bounds 0 and 1 of each mass fraction. If we do not use the bound-preserving limiter, the code will soon blow up. We compare the mass fraction profiles with and without the limiter at $0.5085 \mu s$ in Fig. 6.2. From the zoom in figure, we can see that the mass fraction Y_2 will become negative if we do not apply the limiter.

Example 6.3. $H_2/O_2/Ar$ shock tube problem

We consider the 1D NS equations for multi-species flow without chemical reactions. Assume that we have a $2/1/7$ molar ratio of $H_2/O_2/Ar$. We run a 1D shock tube problem with initial conditions given by

$$(T, u, p)(x, 0) = \begin{cases} (400 \text{ K}, 0 \text{ m/s}, 8000 \frac{\text{J}}{\text{m}^3}), & x \leq 0.5 \text{ cm}, \\ (1200 \text{ K}, 0 \text{ m/s}, 80000 \frac{\text{J}}{\text{m}^3}), & x > 0.5 \text{ cm}. \end{cases}$$

This is done on a 10 cm domain for a time of $40 \mu s$.

We use 400 cells. Third order ($k = 2$) DG method with bound-preserving limiter is adopted for the spacial discretization. The two parameters in the time step (6.1) are taken as $a = 0.1$ and $b = 0.001$. The solutions are shown in Fig. 6.3. We use red circles to denote the numerical solutions with 400 cells. Also, black lines are the reference solutions obtained by using 5000 cells. We can see that our method preserves the right physical bounds. As in the previous example, if we do not apply the bound-preserving limiter, the code will soon blow up.

Example 6.4. $H_2/O_2/Ar$ with chemical reactions in 1D

In this example, we consider the 1D NS equations for multi-species flow with chemical reactions. Consider a shock hitting a solid wall boundary and reflecting off. After a delay a reaction wave kicks in at the boundary. This reaction wave picks up steam and merges with the shock causing a split into 3 waves. From wall to outflow (left to right) these waves are

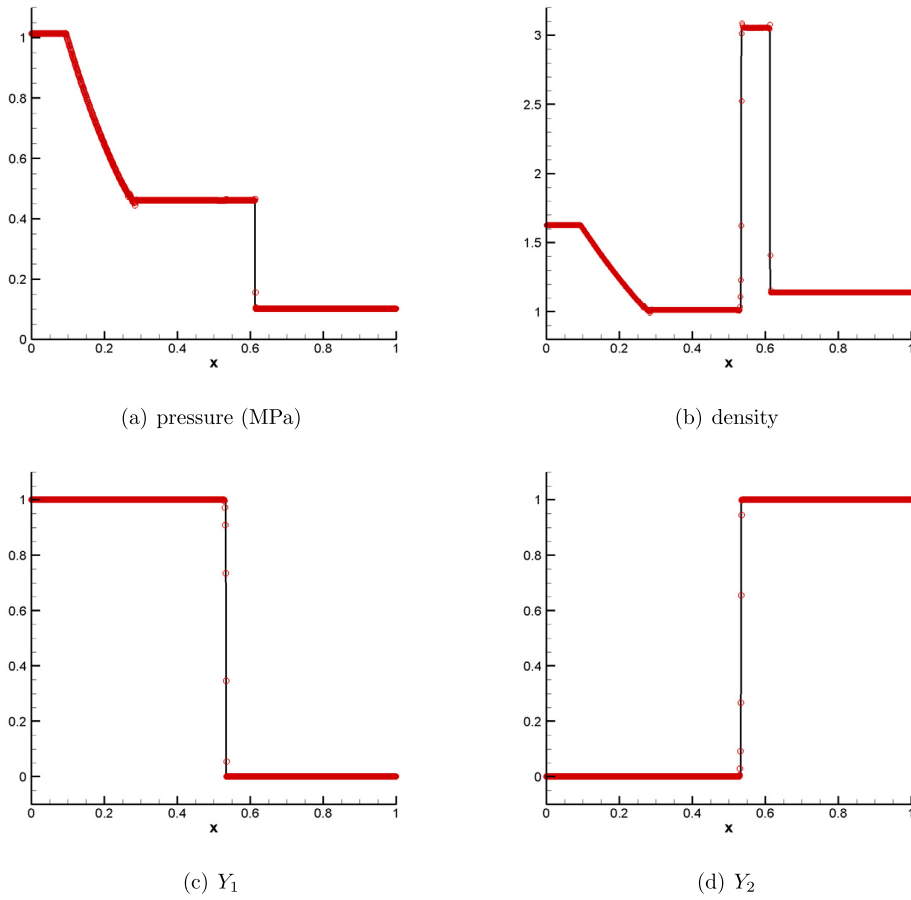


Fig. 6.1. Helium/nitrogen shock tube problem at 300 μ s. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

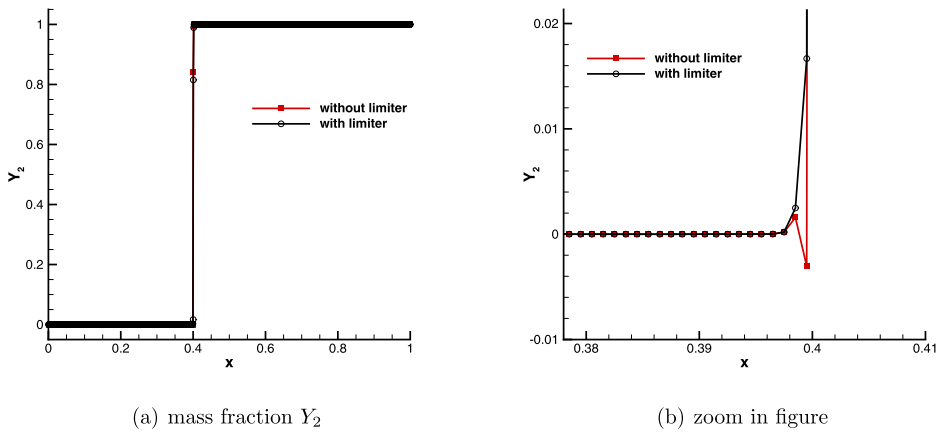


Fig. 6.2. Helium/nitrogen shock tube problem at 0.5085 μ s.

a rarefaction, a contact discontinuity, and a shock. Assume that we have a 2/1/7 molar ratio of $H_2/O_2/Ar$. All gases involved are assumed to be calorically perfect. The reaction mechanism used in this work consisted of 9 species (H , O , H_2 , O_2 , OH , H_2O , HO_2 , H_2O_2 , and Ar) and 37 irreversible reactions [13]. We use the following initial conditions

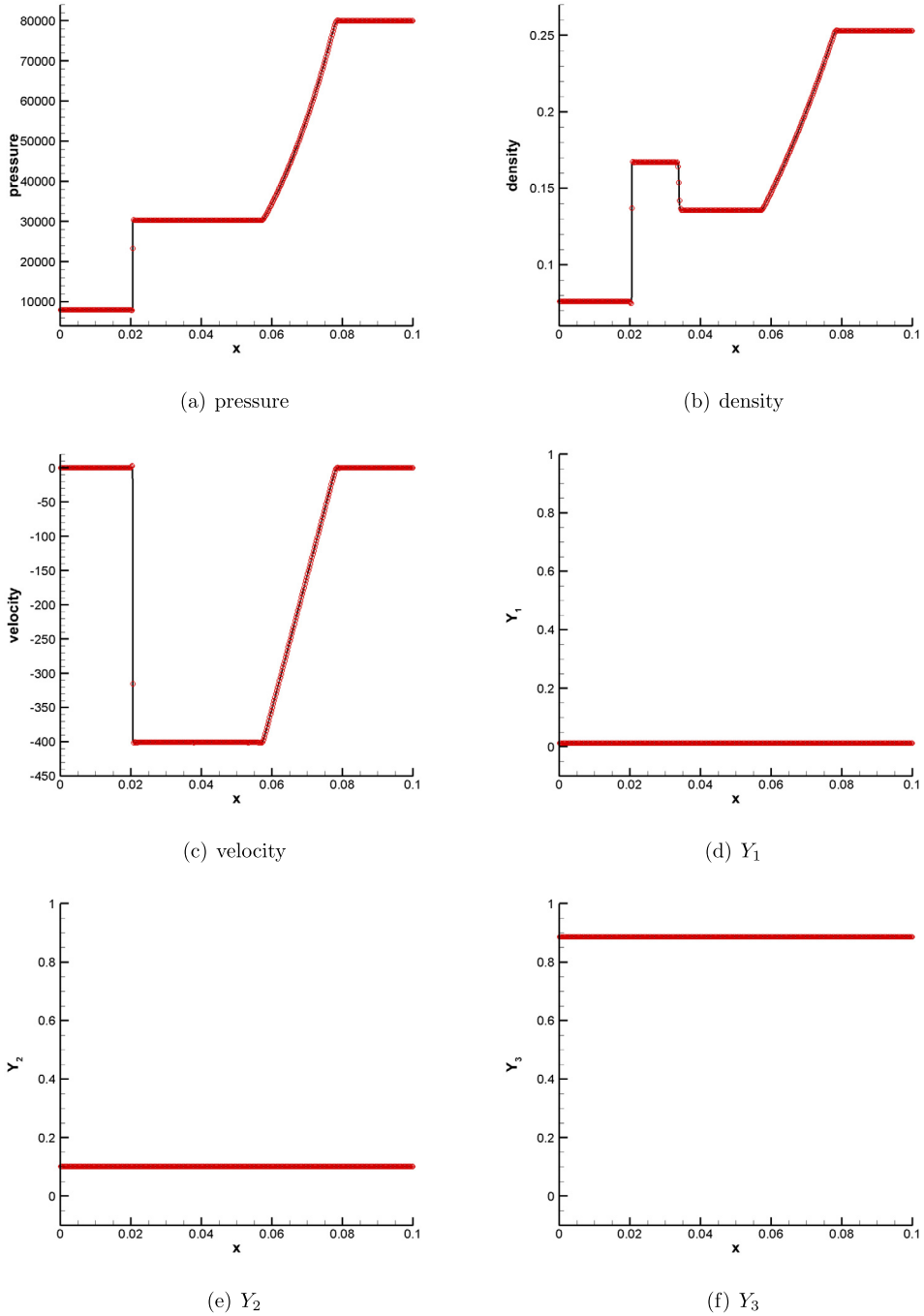


Fig. 6.3. $H_2/O_2/Ar$ shock tube problem at $40 \mu s$.

$$(\rho, u, p)(x, 0) = \begin{cases} (0.072 \text{ kg/m}^3, 0 \text{ m/s}, 7173 \frac{\text{J}}{\text{m}^3}), & x \leq 6 \text{ cm}, \\ (0.18075 \text{ kg/m}^3, -487.34 \text{ m/s}, 35594 \frac{\text{J}}{\text{m}^3}), & x > 6 \text{ cm}. \end{cases}$$

This is done on a 12 cm domain for a time of $190 \mu s$.

We use 400 grid cells. Third order ($k=2$) DG method is adopted for the spacial discretization. The two parameters in the time step (6.1) are taken as $a=0.01$ and $b=0.001$. The numerical results are shown in Fig. 6.4 and Fig. 6.5. We use red circles to denote the numerical solutions with 400 cells. Also, black lines are the reference solutions obtained by using

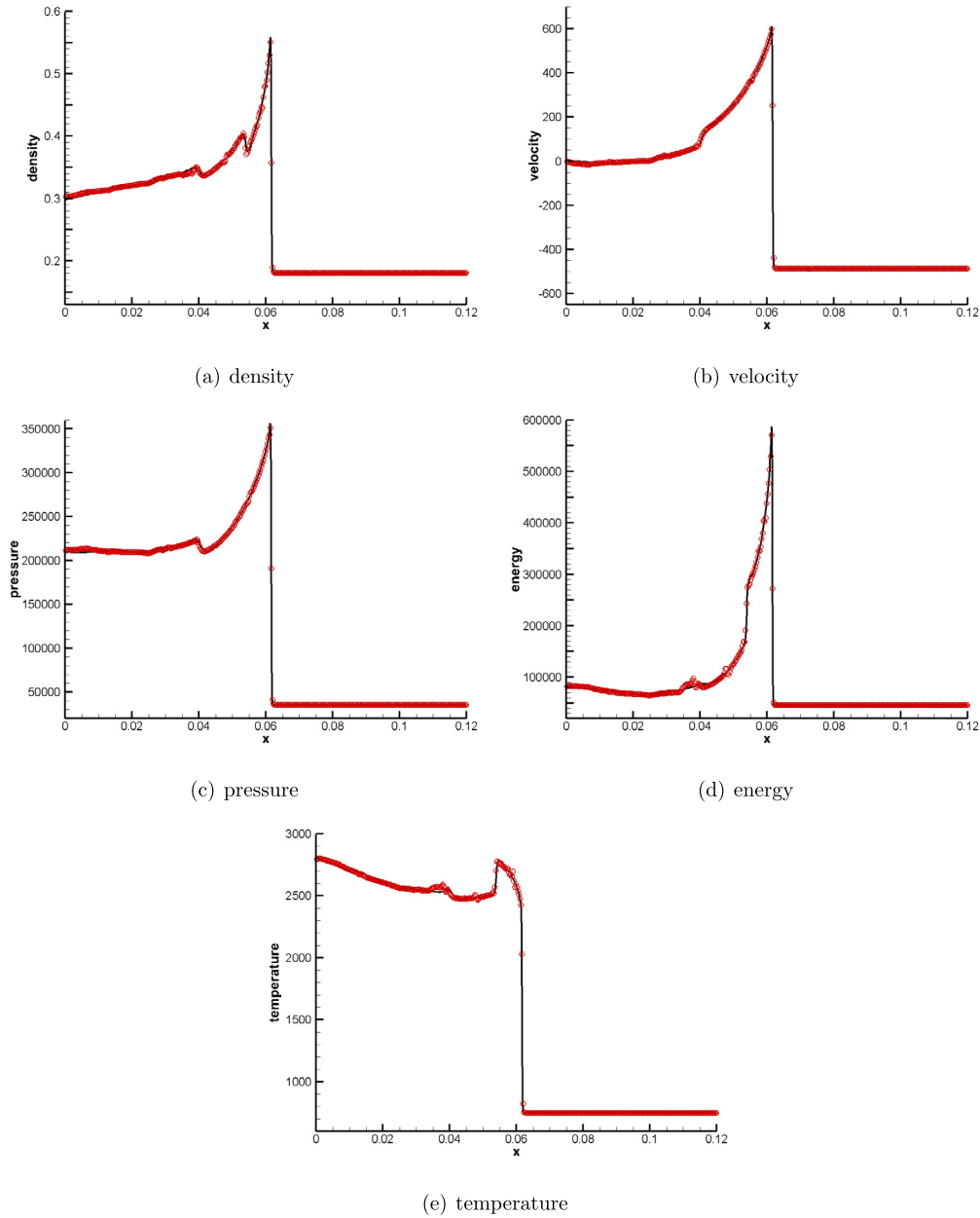


Fig. 6.4. $H_2/O_2/Ar$ with chemical reactions.

1000 cells. We can observe some numerical oscillations since only the bound-preserving limiter is applied. But we can see that our method preserves all the physical bounds.

Example 6.5. Accuracy test in 2D

In this example, we consider the two dimensional problem (2.1) and test the accuracy of our scheme. Two species are considered. The first species is H_2 and the second species is O_2 . The computational domain is $[0, 2\pi] \times [0, 2\pi]$. We use the following initial conditions

$$\begin{aligned}
 u &= 1 \text{ m/s}, \\
 v &= 1 \text{ m/s}, \\
 p &= 1 \text{ Pa}, \\
 \rho &= 0.1(2 + \sin(x + y) + \cos(x + y)) \text{ kg/m}^3,
 \end{aligned}$$

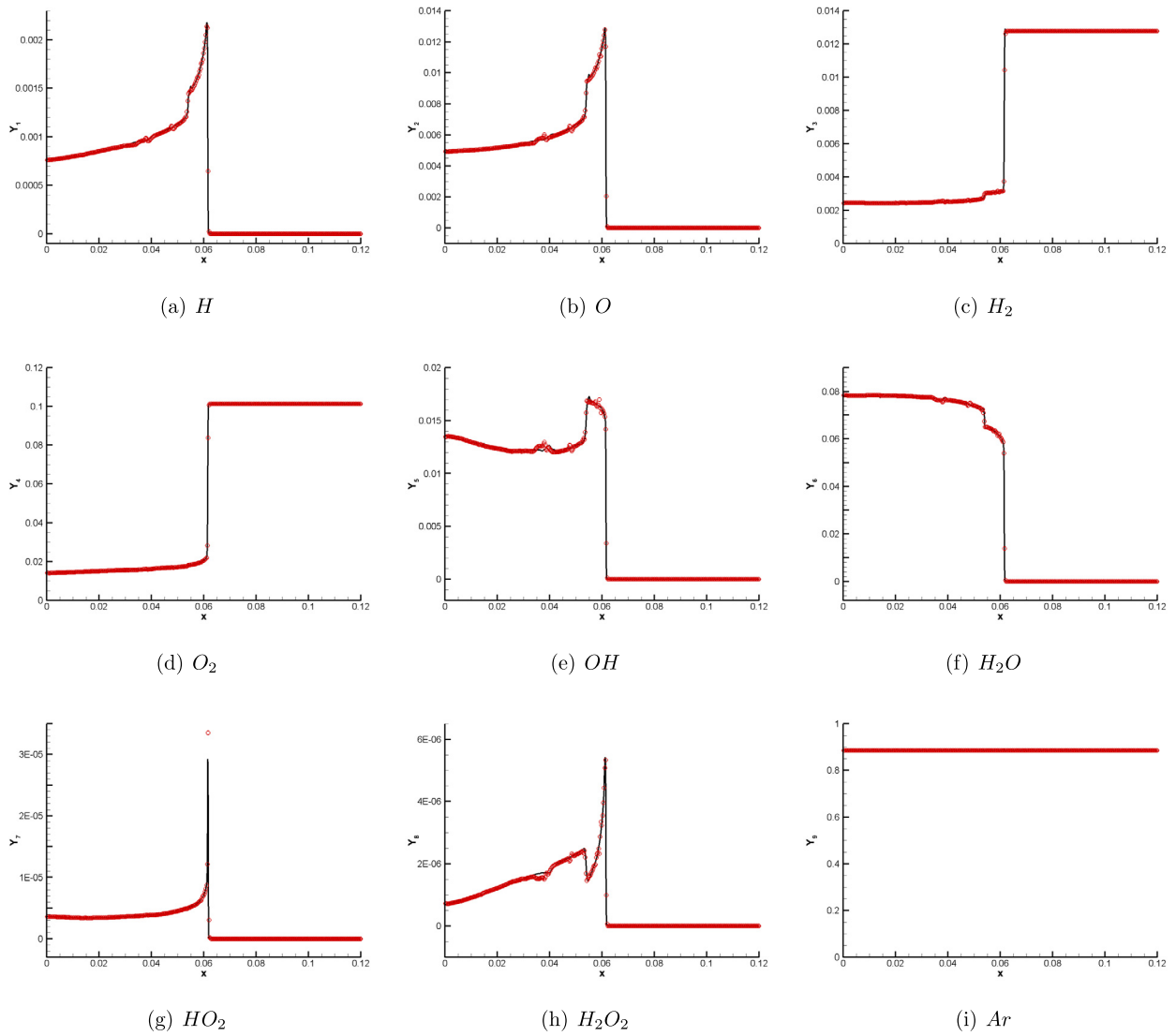


Fig. 6.5. $H_2/O_2/Ar$ with chemical reactions. Mass fractions.

$$r_1 = 0.1(1 + \sin(x + y)) \text{ kg/m}^3,$$

and periodic boundary conditions. Moreover, we define the source terms as $\omega_1 = -c(r_1)^7$ and $\omega_2 = c(r_1)^7$. The parameter c can be used to adjust the stiffness of the problem. The final time is taken as $T = 0.5$.

We apply the DG method with piecewise P^2 polynomials. The two parameters in the time step (6.2) are taken as $a = 0.1$ and $b = 0.001$. The numerical errors of r_1 with different choices of c are listed in the left part of Table 6.3. Both nonstiff ($c = 100$) and stiff ($c = 10000$) cases are calculated. As shown in the table, we can observe the expected third order of accuracy of our scheme. For this problem, the total density should be nonnegative and the mass fractions should be between 0 and 1. Hence, we further add the bound preserving limiter and the results are listed in the right part of Table 6.3. The percentage of cells that have been modified by the limiter is listed in the last column. By comparing the results with and without limiter, we can see that the limiter does not harm the original high order of accuracy.

Example 6.6. $H_2/O_2/Ar$ with chemical reactions in 2D

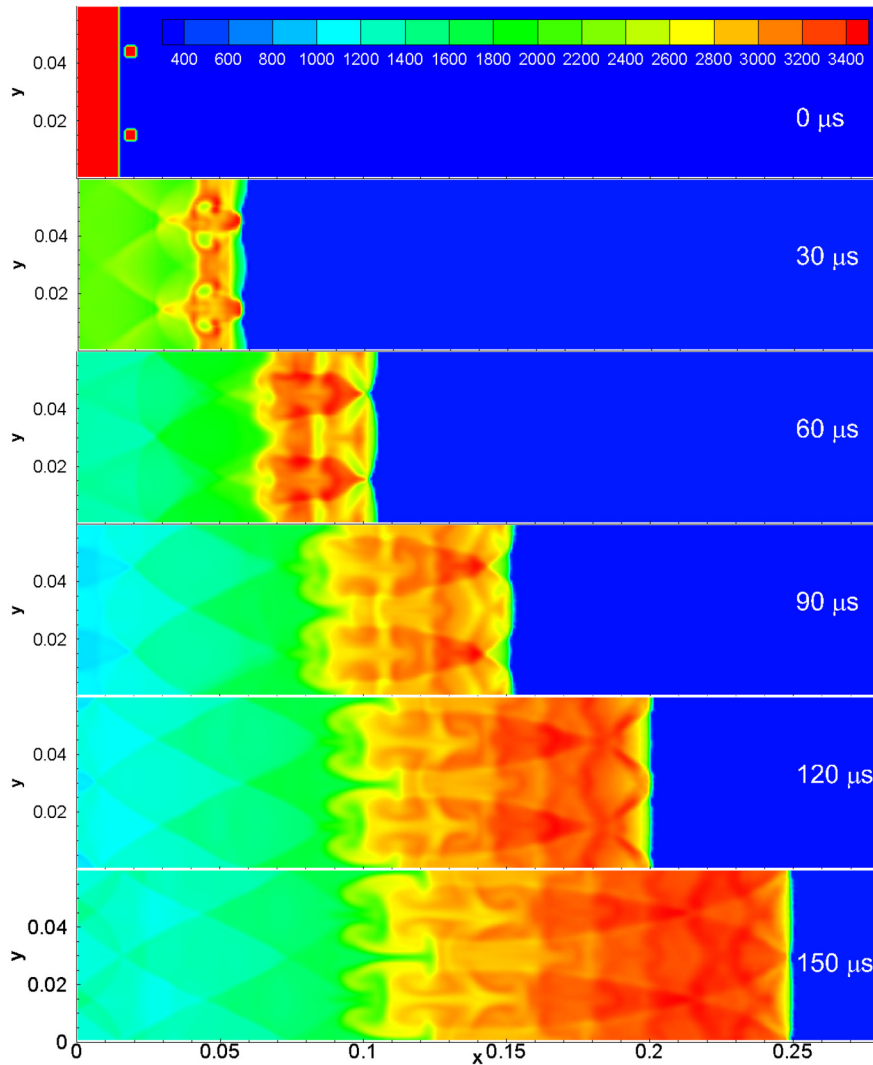
In this example, we consider a two-dimensional hydrogen-oxygen detonation wave diluted in Argon. The computational domain has a channel with height 0.06 m. The four boundaries are simulated as walls. We use the following initial conditions

$$(u, v) = (0, 0) \text{ m/s},$$

Table 6.3

Accuracy test for the two dimensional problem.

N	Without limiter				With limiter				
	L^2 norm	order	L^∞ norm	order	L^2 norm	order	L^∞ norm	order	percentage
$c=100$									
20	3.66E-04	–	2.08E-03	–	8.80E-04	–	2.98E-03	–	6.56%
40	4.47E-05	3.03	2.85E-04	2.86	5.59E-05	3.98	3.11E-04	3.26	3.09%
80	5.53E-06	3.02	3.61E-05	2.98	5.80E-06	3.27	3.85E-05	3.01	0.86%
160	6.92E-07	3.00	4.55E-06	2.99	7.13E-07	3.02	4.82E-06	3.00	0.31%
320	8.66E-08	3.00	5.65E-07	3.01	8.97E-08	2.99	5.99E-07	3.01	5.09E-2%
$c=10000$									
20	5.02E-04	–	2.77E-03	–	9.16E-04	–	3.69E-03	–	3.91%
40	6.37E-05	2.98	5.19E-04	2.41	6.71E-05	3.77	5.25E-04	2.81	1.82%
80	7.96E-06	3.00	6.92E-05	2.91	8.08E-06	3.05	6.92E-05	2.92	0.68%
160	9.94E-07	3.00	8.87E-06	2.96	1.01E-06	3.00	8.87E-06	2.96	0.29%
320	1.24E-07	3.00	1.12E-06	2.99	1.26E-07	2.99	1.12E-06	2.99	4.72E-2%

**Fig. 6.6.** 2D $H_2/O_2/Ar$ with chemical reactions. Temperature.

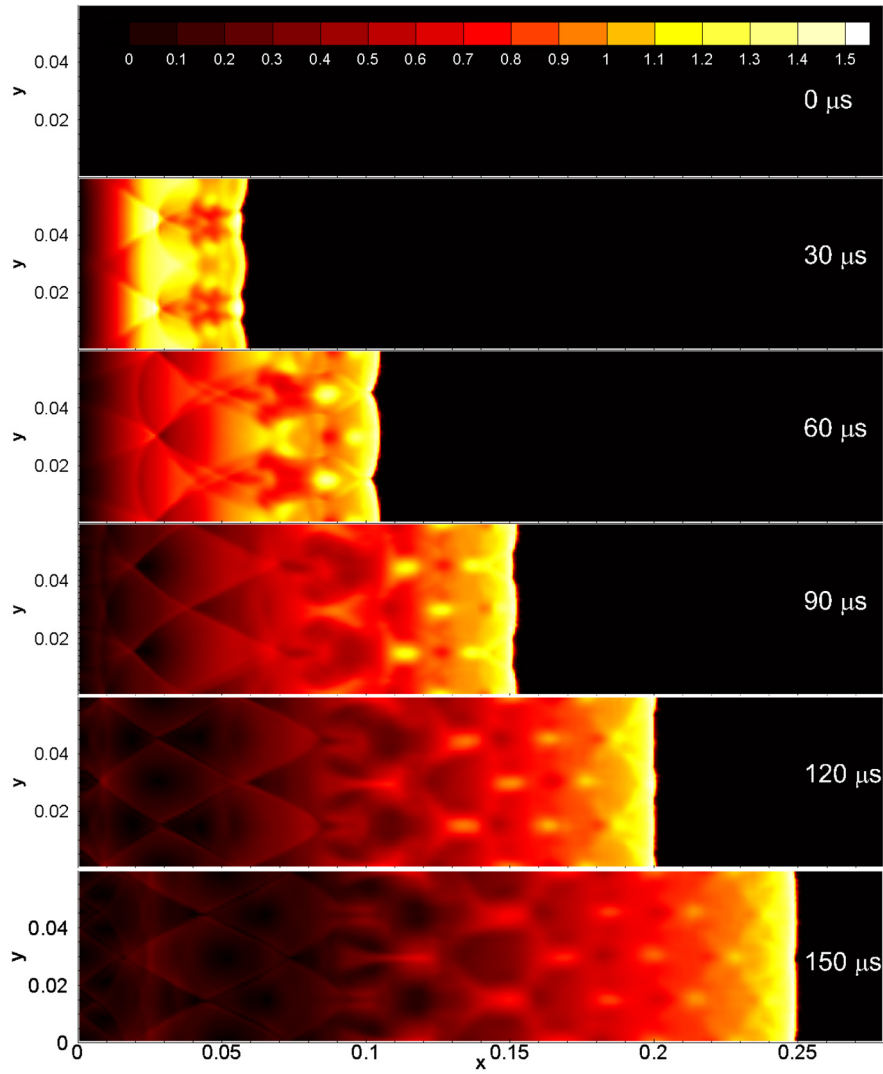


Fig. 6.7. 2D $H_2/O_2/Ar$ with chemical reactions. Mach numbers.

$$Y_{Ar} : Y_{O_2} : Y_{H_2} = 7 : 1 : 2, \quad x > 0.025 \text{ m},$$

$$Y_{Ar} : Y_{O_2} : Y_{H_2} : Y_{OH} = 7 : 1 : 2 : 0.01, \quad 0.015 \text{ m} < x < 0.025 \text{ m},$$

$$Y_{Ar} : Y_{H_2O} : Y_{OH} = 8 : 2 : 0.01, \quad x < 0.015 \text{ m},$$

$$p = \begin{cases} 5.50e5 \text{ Pa}, & x < 0.015 \text{ m}, \\ 6.67e3 \text{ Pa}, & x > 0.015 \text{ m}, \end{cases}$$

$$T = \begin{cases} 298 \text{ K}, & x > 0.025 \text{ m}, \\ 350 \text{ K}, & 0.015 \text{ m} < x < 0.025 \text{ m}, \\ 3500 \text{ K}, & x < 0.015 \text{ m}, \end{cases}$$

with the exception of two additional high pressure and high temperature regions, located within the regions $\sqrt{(x - 0.019)^2 + (y - 0.015)^2} = 0.0025 \text{ m}$ and $\sqrt{(x - 0.01)^2 + (y - 0.044)^2} = 0.0025 \text{ m}$, with conditions

$$(u, v) = (0, 0) \text{ m/s},$$

$$Y_{Ar} : Y_{H_2O} : Y_{OH} = 8 : 2 : 0.01,$$

$$p = 5.5e5 \text{ Pa},$$

$$T = 3500 \text{ K}.$$

The reaction mechanism used in this work consisted of 9 species (H , O , H_2 , O_2 , OH , H_2O , HO_2 , H_2O_2 , and Ar) and 34 irreversible reactions. The detailed chemical kinetics are described by the reaction mechanism of Westbrook [43].

We use rectangular meshes with the cell length as $\Delta x = \Delta y = 0.001$ m. Second order ($k = 1$) DG method is adopted for the spatial discretization. The two parameters in the time step (6.2) are taken as $a = 0.00005$ and $b = 0.01$. Fig. 6.6 shows the temperature solutions at different times. Fig. 6.7 shows the Mach numbers at different times. The detonation is established after the initial shock collides with the two additional high pressure and high temperature regions. The perturbations lead to transverse waves traveling in the vertical directions that reflect off the top and bottom walls. The detonation front progresses through the simulation domain. For higher order spatial discretization, further limiter is needed to remove numerical oscillations and we will explore this in our future work.

7. Conclusion

In this paper, we constructed high-order bound-preserving DG methods for multicomponent chemically reacting flows. The CMERK methods was used for time discretization to preserve the conservative property. Thanks to the proposed scheme, the numerical approximations yield positive density and pressure, and the mass fractions are between 0 and 1.

CRedit authorship contribution statement

Jie Du: Funding acquisition, Investigation, Software, Validation, Writing – original draft. **Yang Yang:** Conceptualization, Funding acquisition, Investigation, Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

- [1] John D. Anderson, Hypersonic and High Temperature Gas Dynamics, McGraw-Hill, 1989.
- [2] K. Bando, M. Sekachev, M. Ihme, Comparison of algorithms for simulating multi-component reacting flows using high-order discontinuous Galerkin methods, AIAA, 2020-1751, 2020.
- [3] G. Billet, R. Abgrall, An adaptive shock-capturing algorithm for solving unsteady reactive flows, Comput. Fluids 32 (2003) 1473–1495.
- [4] R.B. Bird, W.E. Stewart, E.N. Lightfoot, Transport Phenomena, second edition, John Wiley, New York, 2002.
- [5] Z. Chen, H. Huang, J. Yan, Third order maximum-principle-satisfying direct discontinuous Galerkin methods for time dependent convection diffusion equations on unstructured triangular meshes, J. Comput. Phys. 308 (2016) 198–217.
- [6] N. Chuenjarern, Z. Xu, Y. Yang, High-order bound-preserving discontinuous Galerkin methods for compressible miscible displacements in porous media on triangular meshes, J. Comput. Phys. 378 (2019) 110–128.
- [7] T.P. Coffee, J.M. Heimerl, Transport algorithms for premixed, laminar steady-state flames, Combust. Flame 43 (1981) 273–289.
- [8] J. Du, C. Wang, C. Qian, Y. Yang, High-order bound-preserving discontinuous Galerkin methods for stiff multispecies detonation, SIAM J. Sci. Comput. 41 (2019) B250–B273.
- [9] J. Du, Y. Yang, Maximum-principle-preserving third-order local discontinuous Galerkin methods on overlapping meshes, J. Comput. Phys. 377 (2019) 117–141.
- [10] J. Du, Y. Yang, Third-order conservative sign-preserving and steady-state-preserving time integrations and applications in stiff multispecies and multi-reaction detonations, J. Comput. Phys. 395 (2019) 489–510.
- [11] J. Du, Y. Yang, High-order bound-preserving finite difference methods for multispecies and multireaction detonations, Commun. Appl. Math. Comput. (2022), <https://doi.org/10.1007/s42967-020-00117-y>, in press.
- [12] J. Du, Y. Yang, E. Chung, Stability analysis and error estimates of local discontinuous Galerkin methods for convection-diffusion equations on overlapping meshes, BIT Numer. Math. 59 (2019) 853–876.
- [13] R.P. Fedkiw, B. Merriman, S. Osher, Numerical Methods for a Mixture of Thermally Perfect and/or Calorically Perfect Gaseous Species with Chemical Reactions, Lecture Notes in Physics, 1997.
- [14] W. Feng, H. Guo, Z. Xu, Y. Yang, Conservative numerical methods for the reinterpreted discrete fracture model on non-conforming meshes and their applications in contaminant transportation in fractured porous media, Adv. Water Resour. 153 (2021) 103951.
- [15] S. Gottlieb, C.-W. Shu, E. Tadmor, Strong stability-preserving high-order time discretization methods, SIAM Rev. 43 (2001) 89–112.
- [16] H. Guo, X. Liu, Y. Yang, High-order bound-preserving finite difference methods for miscible displacements in porous media, J. Comput. Phys. 406 (2020) 109219.
- [17] H. Guo, Y. Yang, Bound-preserving discontinuous Galerkin method for compressible miscible displacement problem in porous media, SIAM J. Sci. Comput. 39 (2017) A1969–A1990.
- [18] L. Guo, Y. Yang, Positivity-preserving high-order local discontinuous Galerkin method for parabolic equations with blow-up solutions, J. Comput. Phys. 289 (2015) 181–195.
- [19] R.W. Houim, K.K. Kuo, A low-dissipation and time-accurate method for compressible multi-component flow with variable specific heat ratios, J. Comput. Phys. 230 (2011) 8527–8553.
- [20] R.W. Houim, K.K. Kuo, A ghost fluid method for compressible reacting flows with phase change, J. Comput. Phys. 235 (2013) 865–900.

- [21] J. Huang, C.-W. Shu, Bound-preserving modified exponential Runge-Kutta discontinuous Galerkin methods for scalar hyperbolic equations with stiff source terms, *J. Comput. Phys.* 361 (2018) 111–135.
- [22] J. Huang, C.-W. Shu, Positivity-preserving time discretizations for production-destruction equations with applications to non-equilibrium flows, *J. Sci. Comput.* 78 (2019) 1811–1839.
- [23] J. Huang, W. Zhao, C.-W. Shu, A third-order unconditionally positivity-preserving scheme for production-destruction equations with applications to non-equilibrium flows, *J. Sci. Comput.* 79 (2019) 1015–1056.
- [24] E. Johnsen, F. Ham, Preventing numerical errors generated by interface-capturing schemes in compressible multi-material flows, *J. Comput. Phys.* 231 (2012) 5705–5717.
- [25] R. Johnson, A. Kercher, A conservative discontinuous Galerkin discretization for the chemically reacting Navier-Stokes equations 423 (2020) 109826.
- [26] R.J. Kee, M.E. Coltrin, P. Glarborg, *Chemically Reacting Flow: Theory and Practice*, John Wiley & Sons, Hoboken, NJ, 2003.
- [27] H. Liu, J. Yan, The direct discontinuous Galerkin (DDG) methods for diffusion problems, *SIAM J. Numer. Anal.* 47 (2009) 675–698.
- [28] X. Liu, Y. Yang, H. Guo, High-order bound-preserving finite difference methods for incompressible wormhole propagation, *J. Sci. Comput.* 89 (2021) 7.
- [29] Y. Lv, M. Ihme, Discontinuous Galerkin method for multicomponent chemically reacting flows and combustion, *J. Comput. Phys.* 270 (2014) 105–137.
- [30] Y. Lv, M. Ihme, High-order discontinuous Galerkin method for applications to multicomponent and chemically reacting flows, *Acta Mech. Sin.* 33 (2017) 486–499.
- [31] P. Ma, Y. Lv, M. Ihme, An entropy-stable hybrid scheme for simulations of transcritical real-fluid flows, *J. Comput. Phys.* 340 (2017) 330–357.
- [32] S. Mathur, K. Tondon, S.C. Saxena, Thermal conductivity of binary, ternary and quaternary mixtures of rare gases, *Mol. Phys.* 12 (1967) 569–579.
- [33] B.J. McBride, M.J. Zehe, S. Gordon, NASA Glenn coefficients for calculating thermodynamic properties of individual species, NASA/TP-2002-211556, NASA Glenn Research Center, Cleveland, OH, 2002.
- [34] V. Moureau, G. Lartigue, Y. Sommerer, C. Angelberger, O. Colin, T. Poinso, Numerical methods for unsteady compressible multi-component reacting flows on fixed and moving grids, *J. Comput. Phys.* 202 (2005) 710–736.
- [35] P. Movahed, E. Johnsen, A solution-adaptive method for efficient compressible multifluid simulations, with application to the Richtmyer-Meshkov instability, *J. Comput. Phys.* 239 (2013) 166–186.
- [36] T. Poinso, D. Veynante, *Theoretical and Numerical Combustion*, R.T. Edwards, Inc., 2005.
- [37] T. Qin, C.-W. Shu, Y. Yang, Bound-preserving discontinuous Galerkin methods for relativistic hydrodynamics, *J. Comput. Phys.* 315 (2016) 323–347.
- [38] W.H. Reed, T.R. Hill, Triangular mesh methods for the Neutron transport equation, Los Alamos Scientific Laboratory Report LA-UR-73-479, Los Alamos, NM, 1973.
- [39] C.-W. Shu, Total-variation-diminishing time discretizations, *SIAM J. Sci. Stat. Comput.* 9 (1988) 1073–1084.
- [40] C.-W. Shu, S. Osher, Efficient implementation of essentially non-oscillatory shock-capturing schemes, *J. Comput. Phys.* 77 (1988) 439–471.
- [41] D.R. Stall, H. Prophet, JANAF Thermochemical Tables, National Standard Reference Data Series, 1971.
- [42] C. Wang, X. Zhang, C.-W. Shu, J. Ning, Robust high order discontinuous Galerkin schemes for two-dimensional gaseous detonations, *J. Comput. Phys.* 231 (2012) 653–665.
- [43] C.K. Westbrook, Chemical kinetics of hydrocarbon oxidation in gaseous detonations, *Combust. Flame* 46 (1982) 191–210.
- [44] C.R. Wilke, A viscosity equation for gas mixtures, *J. Chem. Phys.* 18 (1950) 517–522.
- [45] T. Xiong, J.-M. Qiu, Z. Xu, High order maximum-principle-preserving discontinuous Galerkin method for convection-diffusion equations, *SIAM J. Comput.* 37 (2015) A583–A608.
- [46] Z. Xu, Y. Yang, H. Guo, High-order bound-preserving discontinuous Galerkin methods for wormhole propagation on triangular meshes, *J. Comput. Phys.* 390 (2019) 323–341.
- [47] Y. Yang, D. Wei, C.-W. Shu, Discontinuous Galerkin method for Krause's consensus models and pressureless Euler equations, *J. Comput. Phys.* 252 (2013) 109–127.
- [48] X. Zhang, On positivity-preserving high order discontinuous Galerkin schemes for compressible Navier-Stokes equations, *J. Comput. Phys.* 328 (2017) 301–343.
- [49] X. Zhang, C.-W. Shu, On maximum-principle-satisfying high order schemes for scalar conservation laws, *J. Comput. Phys.* 229 (2010) 3091–3120.
- [50] X. Zhang, C.-W. Shu, On positivity preserving high order discontinuous Galerkin schemes for compressible Euler equations on rectangular meshes, *J. Comput. Phys.* 229 (2010) 8918–8934.
- [51] X. Zhang, C.-W. Shu, Positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations with source terms, *J. Comput. Phys.* 230 (2011) 1238–1248.
- [52] X. Zhang, Y. Xia, C.-W. Shu, Maximum-principle-satisfying and positivity-preserving high order discontinuous Galerkin schemes for conservation laws on triangular meshes, *J. Sci. Comput.* 50 (2012) 29–32.
- [53] Y. Zhang, X. Zhang, C.-W. Shu, Maximum-principle-satisfying second order discontinuous Galerkin schemes for convection-diffusion equations on triangular meshes, *J. Comput. Phys.* 234 (2013) 295–316.
- [54] X. Zhao, Y. Yang, C. Seyler, A positivity-preserving semi-implicit discontinuous Galerkin scheme for solving extended magnetohydrodynamics equations, *J. Comput. Phys.* 278 (2014) 400–415.