Journal of Computational Physics 395 (2019) 489–510

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

# Third-order conservative sign-preserving and steady-state-preserving time integrations and applications in stiff multispecies and multireaction detonations

Jie Du [a,1], Yang Yang [b,*,2]

[a] Yau Mathematical Sciences Center, Tsinghua University, Beijing, 100084, China
[b] Department of Mathematical Sciences, Michigan Technological University, Houghton, MI 49931, United States of America

ARTICLE INFO

*Article history:*
Received 20 February 2019
Received in revised form 15 June 2019
Accepted 15 June 2019
Available online 19 June 2019

*Keywords:*
Discontinuous Galerkin method
Bound-preserving
Third-order
Stiff source
Detonation

ABSTRACT

In this paper, we develop third-order conservative sign-preserving and steady-state-preserving time integrations and seek their applications in multispecies and multireaction chemical reactive flows. In this problem, the density and pressure are nonnegative, and the mass fraction for the $i$th species, denoted as $z_i$, $1 \le i \le M$, should be between 0 and 1, where $M$ is the total number of species. There are four main difficulties in constructing high-order bound-preserving techniques for multispecies and multireaction detonations. First of all, most of the bound-preserving techniques available are based on Euler forward time integration. Therefore, for problems with stiff source, the time step will be significantly limited. Secondly, the mass fraction does not satisfy a maximum-principle and hence it is not easy to preserve the upper bound 1. Thirdly, in most of the previous works for gaseous denotation, the algorithm relies on second-order Strang splitting methods where the flux and stiff source terms can be solved separately, and the extension to high-order time discretization seems to be complicated. Finally, most of the previous ODE solvers for stiff problems cannot preserve the total mass and the positivity of the numerical approximations at the same time. In this paper, we will construct third-order conservative sign-preserving Rugne-Kutta and multistep methods to overcome all these difficulties. The time integrations do not depend on the Strang splitting, i.e. we do not split the flux and the stiff source terms. Moreover, the time discretization can handle the stiff source with large time step and preserves the steady-state. Numerical experiments will be given to demonstrate the good performance of the bound-preserving technique and the stability of the scheme for problems with stiff source terms.

## 1. Introduction

In this paper, we develop third-order conservative sign-preserving and steady-state-preserving time integrations and construct high-order bound-preserving numerical methods for stiff multispecies and multireaction chemical reactive flows. We investigate the following convection-reaction equation in two space dimensions

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

[1] Research supported by NSFC grant 11801302.

[2] Research supported by NSF grant DMS-1818467.

https://doi.org/10.1016/j.jcp.2019.06.040

$$\rho_t + m_x + n_y = 0, \tag{1.1a}$$

$$m_t + (mu + p)_x + (nu)_y = 0, \tag{1.1b}$$

$$n_t + (mv)_x + (nv + p)_y = 0, \tag{1.1c}$$

$$E_t + ((E + p)u)_x + ((E + p)v)_y = 0, \tag{1.1d}$$

$$(r_1)_t + (mz_1)_x + (nz_1)_y = s_1, \tag{1.1e}$$

$$\cdots$$

$$(r_{M-1})_t + (mz_{M-1})_x + (nz_{M-1})_y = s_{M-1}, \tag{1.1f}$$

where $\rho$, $u$, $v$, $m = \rho u$, $n = \rho v$, $E$ and $p$ are the total density, velocity in $x$ direction, velocity in $y$ direction, momentum in $x$ direction, momentum in $y$ direction, the total energy, and pressure, respectively. $M$ is the total number of chemical species. For $1 \le i \le M$, $r_i = \rho z_i$ with $z_i$ being the mass fraction for the $i$th species, and $\sum_{i=1}^M z_i = 1$. Therefore, we have

$$\sum_{i=1}^{M} r_i = \rho, \tag{1.2}$$

and hence $0 \le z_i \le 1$. The equation of state is given as

$$p = (\gamma - 1)\left(E - \frac{1}{2}\rho(u^2 + v^2) - \rho z_1 q_1 - \cdots - \rho z_M q_M\right),$$

where $q_i$ is the enthalpy of formation for the $i$th species and the temperature is defined as $T = p/\rho$. The $s_i$ given in the source term describes the chemical reactions. We consider R reactions of the form

$$\nu'_{1,r}X_1 + \nu'_{2,r}X_2 + \cdots + \nu'_{M,r}X_M \to \nu''_{1,r}X_1 + \nu''_{2,r}X_2 + \cdots + \nu''_{M,r}X_M, \quad r = 1, 2, \cdots, R,$$

where $\nu'_{i,r}$ and $\nu''_{i,r}$ are the stoichiometric coefficients of the reactants and products, respective, of the $i$th species in the $r$th reaction. For non-equilibrium chemistry, the rate of production of the $i$th species can be written as

$$s_i = M_i \sum_{r=1}^{R} (\nu''_{i,r} - \nu'_{i,r}) \left[ k_r(T) \prod_{j=1}^{M} \left(\frac{r_j}{M_j}\right)^{\nu'_{j,r}} \right], \quad i = 1, 2, \cdots, M,$$

where $M_i$ is the molar mass of the $i$th species. $k_r(T)$, a function of the temperature $T$, indicates the reaction rate. In this paper, we take

$$k_r(T) = \begin{cases} B_r T^{\alpha_r}, & T > T_r, \\ 0, & T \le T_r, \end{cases}$$

where $T_r$ is the ignition temperature for the $r$th reaction, and $B_r$ and $\alpha_r$ are pre-exponential factor and index of temperature, respectively. Moreover, it is easy to check that $\sum_{i=1}^M s_i = 0$. Therefore, using the fact $\sum_{i=1}^M z_i = 1$, we can subtract (1.1e)-(1.1f) from (1.1a) to obtain a new equation

$$(r_M)_t + (mz_M)_x + (nz_M)_y = s_M, \tag{1.3}$$

which is similar to (1.1e)-(1.1f), and this can help us construct the bound-preserving technique.

Numerical simulations for wave propagation in gaseous detonation are essential for minimizing devastating hazards. It is well known that correct ignition process of the mixture could not be predicted by single-step models. Therefore, it is common to use detailed chemical model to reproduce results that agree with the experimental data. Thus, designing an efficient and accurate numerical method is of practical importance. However, due to the complexity of chemical kinetics, the construction of the numerical methods is not an easy task. There are three main difficulties. Firstly, the reaction speed of the chemical species is extremely fast, leading to stiff source terms in the model system, see e.g. [6,18]. Hence, the time step would be significantly limited if some explicit time integrations, such as Euler forward, are applied. Secondly, due to the existence of shocks in the exact solutions, direct numerical simulation may be highly oscillatory near the shocks and send positive density and pressure to be negative. Furthermore, the mass fraction may not be between 0 and 1, either. The physically irrelevant numerical approximations may yield ill-posedness of the problems leading to the blow-up of the numerical simulations. This phenomenon is especially significant for high-order numerical schemes. Therefore, it is very important to develop special bound-preserving techniques to preserve the physical bounds in the numerical simulations. Finally, direct numerical simulations on coarse meshes may yield nonphysical shock waves due to the stiff source, see e.g. [18] for the discussion. In this paper, we will focus on the first two problems and construct suitable high-order bound-preserving numerical schemes. The key step in this technique is to develop suitable high-order time integrations in which the time step restriction depends on the convection term only, not the stiff source term. Therefore, the time step can be large. We will extend the idea to deal with the last problem in the future. For the spacial discretization, we would like to apply the discontinuous Galerkin (DG) method, as it is high-order accurate and uses piecewise polynomials as the numerical approximations and hence is easy to apply limiters.

The DG method, first introduced by Reed and Hill [24] in the framework of neutron linear transport, gained even greater popularity for good stability, high order accuracy, and flexibility on h-p adaptivity and on complex geometry. There were some previous works discussing DG methods in solving gaseous denotation, see [19,20] as an incomplete list. However, neither of them focused on the bound-preserving technique. In the last few years, there were several works focusing on the construction of high-order bound-preserving numerical methods for conservation laws. In [31], genuinely maximum-principle-preserving high-order DG schemes for scalar conservation laws have been constructed. Subsequently, positivity-preserving (PP) high-order DG schemes for compressible Euler equations were given in [32,34]. Later, the technique was applied to other hyperbolic systems, see for example [30,35,23], and the $L^1$ stability was demonstrated. In [33], the authors studied the compressible Euler equations with source terms, and the idea was later extended to gaseous detonation in [28] to preserve the positivity of density, pressure and all the mass fractions except the last one. The PP technique in [28] is based on Euler forward time discretization. The extension to high-order time discretization is based on the strong-stability-preserving (SSP) Runge-Kutta (RK)/multistep methods [8,25,26], which can be written as convex combinations of Euler forwards. It is not easy to extend the idea in [28] to preserve the upper bound 1 for the mass fractions. We will encounter three main difficulties in designing high-order time integrations.

1. The construction of conservative time integrations.
   Most of the previous works that preserve two bounds are based on the maximum-principle-preserving technique, see for example [31,34]. However, the mass fraction $z_i$ does not satisfy a maximum-principle. Therefore, the bound-preserving technique discussed before cannot be applied directly. Recently, one of the authors studied miscible displacements in porous media and constructed a second-order DG scheme that preserves the two bounds 0 and 1 for the volumetric percentage in [9] on rectangular meshes, and the extension to triangular meshes has been given in [5]. In this paper, we follow the ideas given in [9,5] to gaseous detonation to construct high-order DG schemes on general rectangular and triangular meshes. The basic idea is to solve (1.1) and (1.3) together, and apply the PP technique to each $r_i$ (or $z_i$), $i = 1, \cdots, M$. By doing so, the total mass conservation (1.2) might be missing. Therefore, to enforce $\sum_{i=1}^M r_i = \rho$ (or $\sum_{i=1}^M z_i = 1$), we need to choose consistent fluxes (see the Definition 3.1) in the convection term and conservative time integrations that guarantee the total mass conservation. Then with positive $z_i$ and total mass conservation $\sum_{i=1}^M z_i = 1$, the numerical approximation of $z_i$ would be between 0 and 1.
2. The construction of high-order time integration for the stiff source term.
   The time discretization in the analysis in [28,9,5] was chosen as Euler forward method. However, in gaseous detonation, $k_r(T)$ would be a large constant, leading to an extremely stiff source $s_i$. Therefore, by applying the idea in [28,9,5], the time step will be significantly limited. One alternative is to consider backward Euler discretization and derive the PP technique. To the best knowledge of the author, the only work in this direction is given in [22], where the maximum-principle-preserving technique was investigated for hyperbolic equations. However, by using backward Euler method, the scheme is only first-order accurate in time and the idea cannot be extended to high-order methods following [28,9,5] since no high-order SSP RK methods can be written as a convex combination of backward Euler methods [8]. Moreover, due to the time step restriction by the PP technique, any time integration that is the combination of Euler forward and backward Euler, such as Crank-Nicolson method, cannot be applied. Notice that, the time step constraint of the PP technique with Euler forward time discretization is due to the stiffness of the source. Some alternative time integrations such as the integration factor RK method [15], implicit-explicit Runge-Kutta (RK) schemes, see, e.g., [1,10,11,21] and semi-implicit method, where only a portion of the stiff term is implicitly treated, see. e.g. [2–4,36]. However, all the methods given above cannot preserve the positivity of the numerical approximations and the total mass conservation at the same time. Hence they cannot be applied to construct bound-preserving technique for gaseous detonation. Besides the above two methods, in almost all the previous works for gaseous detonation, the splitting methods were applied to separate the convection and the source terms. By doing so, it is possible to apply Euler forward time discretization for the convection term and other suitable ODE solvers for the source term. However, the most commonly used splitting method is the second-order Strang splitting method [27], and the extension to high-order time integration is complicated. Another possible idea to construct the time integration is to apply the modified Patankar-Runge-Kutta scheme blue [16,17,13,14]. However, high-order schemes contain some defects as the fraction used in the trick may have zero denominator with nonzero numerator. Therefore, one has to assume the exact solution to be strictly positive. However, this may not be true as one of the species may not appear initially and will be created during the chemical reaction. Moreover, it is very difficult to preserve the positivity of pressure by using the modified Patankar-Runge-Kutta scheme. Recently, there is a new idea introduced in [12] to solve scalar hyperbolic equations with stiff source terms by using the modified exponential RK/multistep DG methods. The algorithm in [12] is not based on the splitting methods nor the Patankar-Runge-Kutta method. However, the scheme does not preserve the total mass conservation. Hence, it cannot be applied to construct bound-preserving technique in the stiff multispecies detonation.
3. The construction of high-order conservative sign-preserving RK method.

> In [7], the authors have constructed high-order multistep methods to preserve the total mass following [12]. It is well known that the time step in a multistep method is fixed. However, the time step size needed for stability such as positivity is usually dependent on the wave speed, which can be changing very quickly, or even wildly in detonations. So in practice, it is significantly difficult to use the multistep method in [7] for detonation.

In this paper, we would like to construct third-order conservative sign-preserving RK methods that are suitable for the bound-preserving technique given in [7]. We will modify the scheme introduced in [12] to preserve the total mass and demonstrate the sufficient and necessary conditions for the third-order accuracy. Then we will analyze other properties of the proposed schemes, such as the steady-state-preserving, sign-preserving, $A(\frac{\pi}{4})$-stability, decay property. Moreover, we will also prove the same properties for the third-order multistep method discussed in [7]. Then we will use the new time integrations to construct bound-preserving schemes and apply to stiff multispecies and multireaction detonations. The time integrations constructed in this paper are explicit and the time step restriction comes from the convection term only and does not depend on the stiff source. Therefore, the time step can be large. Moreover, it is possible to sufficiently refine the mesh to capture the correct position of the shocks. In this paper, we only discuss the bound-preserving technique on fine meshes and the numerical simulations on coarse meshes will be given in the future. Before we finish the introduction, we would like to summarize the advantages of the proposed scheme. The algorithm

1. is third-order accurate in time and high-order accurate in space;
2. is explicit and can handle stiff source term with relatively large time steps;
3. is not based on the splitting technique nor the Patankar-Runge-Kutta methods;
4. has local mass conservation;
5. preserves the total mass;
6. preserves the bounds, such as the positivity of the density and pressure, and the two bounds 0 and 1 of the mass fractions;

The organization of this paper is as follows. In Section 2, we construct third-order conservative sign-preserving Runge-Kutta and multistep methods, and discuss the properties of the time integrations. In Section 3, we consider DG spatial discretizations. We will demonstrate the new bound-preserving technique and the full algorithm. Numerical experiments will be given in Section 4. We will finish in Section 5 with some conclusion remarks.

## 2. Conservative sign-preserving and steady-state-preserving time integrations

In this section, we proceed to construct and analyze high order time integrations. We consider the following ODE system

$$\mathbf{w}_t = \mathbf{F}(\mathbf{w}) + \frac{1}{\varepsilon}\mathbf{s}(\mathbf{w}), \tag{2.1}$$

where $\varepsilon$ is a positive real number, $\mathbf{w} = (w_1, \cdots, w_\ell)^T$ is the unknown variable, $\mathbf{F} = (f_1, \cdots, f_\ell)^T$ is the spatial discretization of the flux, and $\mathbf{s} = (s_1, \cdots, s_\ell)^T$ is the source term. The problem becomes stiff if $\varepsilon$ is small. Moreover, we make the following three assumptions:

1. The system is conservative: there exists a constant vector $\mathbf{v} \in R^\ell$ such that $\mathbf{v}\cdot\mathbf{F} = \mathbf{v}\cdot\mathbf{s} = 0$. In this case, we can easily obtain
$$\frac{d}{dt}(\mathbf{w}\cdot\mathbf{v}) = 0, \tag{2.2}$$
and hence $\mathbf{v}\cdot\mathbf{w}(t) = \mathbf{v}\cdot\mathbf{w}(0)$ for all $t > 0$.
2. There exists a sufficiently small $\Delta t_E$ such that if $\mathbf{w} \geq 0$ and $\Delta t \leq \Delta t_E$ then
$$\mathbf{w} + \Delta t\mathbf{F}(\mathbf{w}) \geq 0. \tag{2.3}$$
For simplicity, here and below we say a vector is nonnegative if each component in the vector is nonnegative.
3. We write the stiff source term as $\mathbf{s} = \mathbf{p} - \mathbf{d}$, where $\mathbf{p}$ and $\mathbf{d}$ are nonnegative vectors, denoting the production and destruction terms, respectively. Then we assume that
$$\lim_{w_i\to 0}\frac{d_i}{w_i} \text{ exists}, \forall\, i = 1, \cdots, \ell. \tag{2.4}$$

**Remark 2.1.** In the first assumption, we just consider the most general conservative property. Different problems may have different constant vectors $\mathbf{v}$. For the gaseous detonation (1.1a)-(1.1f) together with the ghost equation (1.3), $\mathbf{w} = (\rho, m, n, E, r_1, \cdots, r_M)^T$ and $\mathbf{v}$ takes the special form $(1, 0, 0, 0, -1, \cdots, -1)^T \in R^{M+4}$. In this case, (2.2) becomes

$$\frac{d}{dt}\sum_{i=1}^{M} r_i = \frac{d}{dt}\rho. \tag{2.5}$$

Moreover, for the special case $\mathbf{F} + \frac{1}{\varepsilon}\mathbf{s} = \mathbf{0}$, $\mathbf{v}$ can be any constant vector and (2.2) just indicates the steady-state case $\mathbf{w}_t = \mathbf{0}$.

From now on, we use the notation $\mathbf{w}^n$ to denote the numerical solution at the $n$-th time level. Considering an explicit numerical scheme that uses $\mathbf{w}^{n-p}, \cdots, \mathbf{w}^n$ to compute $\mathbf{w}^{n+1}$ ($p \geq 0$), our goal is to construct a suitable high order numerical scheme that enjoys the following properties:

1. Conservative: $\mathbf{v} \cdot \mathbf{w}^n = \mathbf{v} \cdot \mathbf{w}^0$ for all $n \geq 0$. This is an important property for designing bound preserving technique when applying to the gaseous detonation problem;
2. Sign-preserving: If $\mathbf{w}^0 = \mathbf{w}(0) \geq 0$, then $\mathbf{w}^n \geq 0$ for all $n \geq 0$;
3. Steady-state-preserving: If $\mathbf{w}^{n-p} = \cdots = \mathbf{w}^n = \widehat{\mathbf{w}}$ satisfies $\mathbf{F}(\widehat{\mathbf{w}}) + \frac{1}{\varepsilon}\mathbf{s}(\widehat{\mathbf{w}}) = 0$, then $\mathbf{w}^{n+1} = \widehat{\mathbf{w}}$.

The third-order conservative RK and multistep methods will be discussed in Subsections 2.1 and 2.2, respectively.

**Remark 2.2.** The conservative RK and multistep methods to be discussed in the following two subsections have the strong stability preserving structure, i.e. they can be written as convex combinations of several first-order schemes. Therefore, the properties satisfied by the first-order scheme is also satisfied by the proposed RK and multistep methods. If not otherwise stated, we only consider the first stage in the proofs.

### *2.1. Third-order RK method*

In this subsection, we construct third-order RK methods. For the RK method, time steps can change in different time levels. Hence, for practical problems in which the wave speed changes quickly, Runge-Kutta method can be a good choice. We start with the exponential Runge-Kutta methods constructed in [12], and then make some further modifications to make them to be conservative and high order accurate. Moreover, we will prove sign-preserving property, steady-state-preserving property and A($\frac{\pi}{4}$)-stability of our new schemes.

Following [12], we rewrite (2.1) as

$$\mathbf{w}_t + \mu\mathbf{w} = \mathbf{F}(\mathbf{w}) + \frac{1}{\varepsilon}\mathbf{s}(\mathbf{w}) + \mu\mathbf{w},$$

where $\mu \geq 0$ is a constant to be determined in each time step but may depend on the time level $n$. The above equation further yields

$$(e^{\mu t}\mathbf{w})_t = e^{\mu t}(\mathbf{F}(\mathbf{w}) + \frac{1}{\varepsilon}\mathbf{s}(\mathbf{w}) + \mu\mathbf{w}).$$

Motivated by [12], the general framework of the exponential SSP RK scheme for solving the above equation is

$$\mathbf{w}^{(1)} = e^{-\beta_{10}\mu\Delta t}\left[\alpha_{10}\mathbf{w}^n + \beta_{10}\Delta t\mathbf{F}(\mathbf{w}^n) + \beta_{10}\Delta t(\frac{1}{\varepsilon}\mathbf{s}(\mathbf{w}^n) + \mu\mathbf{w}^n)\right], \tag{2.6}$$

$$\begin{aligned}\mathbf{w}^{(2)} = {} & e^{-A\mu\Delta t}\left[\alpha_{20}\mathbf{w}^n + \beta_{20}\Delta t\mathbf{F}(\mathbf{w}^n) + \beta_{20}\Delta t(\frac{1}{\varepsilon}\mathbf{s}(\mathbf{w}^n) + \mu\mathbf{w}^n)\right] \\ & + e^{(\beta_{10}-A)\mu\Delta t}\left[\alpha_{21}\mathbf{w}^{(1)} + \beta_{21}\Delta t\mathbf{F}(\mathbf{w}^{(1)}) + \beta_{21}\Delta t(\frac{1}{\varepsilon}\mathbf{s}(\mathbf{w}^{(1)}) + \mu\mathbf{w}^{(1)})\right],\end{aligned} \tag{2.7}$$

$$\begin{aligned}\mathbf{w}^{n+1} = {} & e^{-\mu\Delta t}\left[\alpha_{30}\mathbf{w}^n + \beta_{30}\Delta t\mathbf{F}(\mathbf{w}^n) + \beta_{30}\Delta t(\frac{1}{\varepsilon}\mathbf{s}(\mathbf{w}^n) + \mu\mathbf{w}^n)\right] \\ & + e^{(\beta_{10}-1)\mu\Delta t}\left[\alpha_{31}\mathbf{w}^{(1)} + \beta_{31}\Delta t\mathbf{F}(\mathbf{w}^{(1)}) + \beta_{31}\Delta t(\frac{1}{\varepsilon}\mathbf{s}(\mathbf{w}^{(1)}) + \mu\mathbf{w}^{(1)})\right] \\ & + e^{(A-1)\mu\Delta t}\left[\alpha_{32}\mathbf{w}^{(2)} + \beta_{32}\Delta t\mathbf{F}(\mathbf{w}^{(2)}) + \beta_{32}\Delta t(\frac{1}{\varepsilon}\mathbf{s}(\mathbf{w}^{(2)}) + \mu\mathbf{w}^{(2)})\right],\end{aligned} \tag{2.8}$$

where $A = \beta_{20} + \alpha_{21}\beta_{10} + \beta_{21}$, all $\alpha_{ij}$ and $\beta_{ij}$ given above are positive constants to be determined by the order conditions and $\mu$ is a nonnegative constant to be determined by the bound-preserving technique. Take dot product with $\mathbf{v}$ in (2.6)-(2.8) and define $w = \mathbf{v} \cdot \mathbf{w}$ to obtain

$$\begin{aligned}
w^{(1)} &= e^{-\beta_{10}\mu\Delta t}\,[\alpha_{10}+\beta_{10}\mu\Delta t]\,w^n,\\
w^{(2)} &= e^{-A\mu\Delta t}\,[\alpha_{20}+\beta_{20}\mu\Delta t]\,w^n + e^{(\beta_{10}-A)\mu\Delta t}\,[\alpha_{21}+\beta_{21}\mu\Delta t]\,w^{(1)},\\
w^{n+1} &= e^{-\mu\Delta t}\,[\alpha_{30}+\beta_{30}\mu\Delta t]\,w^n + e^{(\beta_{10}-1)\mu\Delta t}\,[\alpha_{31}+\beta_{31}\mu\Delta t]\,w^{(1)}\\
&\quad + e^{(A-1)\mu\Delta t}\,[\alpha_{32}+\beta_{32}\mu\Delta t]\,w^{(2)}.
\end{aligned}$$

It is easy to see that $w^{n+1}\neq w^n$ for $\mu\neq 0$ and hence the scheme (2.6)-(2.8) is not conservative. Therefore, we modify (2.6)-(2.8) and construct

$$\mathbf{w}^{(1)} = \left[\alpha_{10}\mathbf{w}^n + \beta_{10}\Delta t\mathbf{F}(\mathbf{w}^n) + \beta_{10}\Delta t(\frac{1}{\varepsilon}\mathbf{s}(\mathbf{w}^n)+\mu\mathbf{w}^n)\right]/A_1, \tag{2.9}$$

$$\begin{aligned}
\mathbf{w}^{(2)} &= \left[\alpha_{20}\mathbf{w}^n + \beta_{20}\Delta t\mathbf{F}(\mathbf{w}^n) + \beta_{20}\Delta t(\frac{1}{\varepsilon}\mathbf{s}(\mathbf{w}^n)+\mu\mathbf{w}^n)\right]/A_2\\
&\quad + e^{\beta_{10}\mu\Delta t}\left[\alpha_{21}\mathbf{w}^{(1)} + \beta_{21}\Delta t\mathbf{F}(\mathbf{w}^{(1)}) + \beta_{21}\Delta t(\frac{1}{\varepsilon}\mathbf{s}(\mathbf{w}^{(1)})+\mu\mathbf{w}^{(1)})\right]/A_2,
\end{aligned} \tag{2.10}$$

$$\begin{aligned}
\mathbf{w}^{n+1} &= \left[\alpha_{30}\mathbf{w}^n + \beta_{30}\Delta t\mathbf{F}(\mathbf{w}^n) + \beta_{30}\Delta t(\frac{1}{\varepsilon}\mathbf{s}(\mathbf{w}^n)+\mu\mathbf{w}^n)\right]/A_3\\
&\quad + e^{\beta_{10}\mu\Delta t}\left[\alpha_{31}\mathbf{w}^{(1)} + \beta_{31}\Delta t\mathbf{F}(\mathbf{w}^{(1)}) + \beta_{31}\Delta t(\frac{1}{\varepsilon}\mathbf{s}(\mathbf{w}^{(1)})+\mu\mathbf{w}^{(1)})\right]/A_3\\
&\quad + e^{A\mu\Delta t}\left[\alpha_{32}\mathbf{w}^{(2)} + \beta_{32}\Delta t\mathbf{F}(\mathbf{w}^{(2)}) + \beta_{32}\Delta t(\frac{1}{\varepsilon}\mathbf{s}(\mathbf{w}^{(2)})+\mu\mathbf{w}^{(2)})\right]/A_3,
\end{aligned} \tag{2.11}$$

where

$$\begin{aligned}
&A_1 = \alpha_{10}+\beta_{10}\mu\Delta t, \quad A_2 = [\alpha_{20}+\beta_{20}\mu\Delta t] + e^{\beta_{10}\mu\Delta t}\,[\alpha_{21}+\beta_{21}\mu\Delta t],\\
&A_3 = [\alpha_{30}+\beta_{30}\mu\Delta t] + e^{\beta_{10}\mu\Delta t}\,[\alpha_{31}+\beta_{31}\mu\Delta t] + e^{A\mu\Delta t}\,[\alpha_{32}+\beta_{32}\mu\Delta t].
\end{aligned}$$

Taking dot product with $\mathbf{v}$, one can check that this new scheme is conservative for any choice of $\mu$ and we summarize this property in the following theorem.

**Theorem 2.1.** *Consider the ODE system (2.1), the new exponential Runge-Kutta time discretization (2.9)-(2.11) is conservative in the sense that*

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

Next, we consider the accuracy issue. After some basic computations, we can rewrite (2.9)-(2.11) into the three-stage SSP explicit RK scheme in the Shu-Osher form [26] as

$$\begin{aligned}
\mathbf{w}^{(1)} &= \left[\tilde\alpha_{10}\mathbf{w}^n + \tilde\beta_{10}\Delta t\mathbf{F}(\mathbf{w}^n) + \tilde\beta_{10}\Delta t\frac{1}{\varepsilon}\mathbf{s}(\mathbf{w}^n)\right],\\
\mathbf{w}^{(2)} &= \left[\tilde\alpha_{20}\mathbf{w}^n + \tilde\beta_{20}\Delta t\mathbf{F}(\mathbf{w}^n) + \tilde\beta_{20}\Delta t\frac{1}{\varepsilon}\mathbf{s}(\mathbf{w}^n)\right] + \left[\tilde\alpha_{21}\mathbf{w}^{(1)} + \tilde\beta_{21}\Delta t\mathbf{F}(\mathbf{w}^{(1)}) + \tilde\beta_{21}\Delta t\frac{1}{\varepsilon}\mathbf{s}(\mathbf{w}^{(1)})\right],\\
\mathbf{w}^{n+1} &= \left[\tilde\alpha_{30}\mathbf{w}^n + \tilde\beta_{30}\Delta t\mathbf{F}(\mathbf{w}^n) + \tilde\beta_{30}\Delta t\frac{1}{\varepsilon}\mathbf{s}(\mathbf{w}^n)\right] + \left[\tilde\alpha_{31}\mathbf{w}^{(1)} + \tilde\beta_{31}\Delta t\mathbf{F}(\mathbf{w}^{(1)}) + \tilde\beta_{31}\Delta t\frac{1}{\varepsilon}\mathbf{s}(\mathbf{w}^{(1)})\right]\\
&\quad + \left[\tilde\alpha_{32}\mathbf{w}^{(2)} + \tilde\beta_{32}\Delta t\mathbf{F}(\mathbf{w}^{(2)}) + \tilde\beta_{32}\Delta t\frac{1}{\varepsilon}\mathbf{s}(\mathbf{w}^{(2)})\right],
\end{aligned}$$

where

$$\begin{aligned}
&\tilde\alpha_{10} = 1, \quad \tilde\alpha_{20} = (\alpha_{20}+\beta_{20}\mu\Delta t)/A_2, \quad \tilde\alpha_{21} = e^{\beta_{10}\mu\Delta t}(\alpha_{21}+\beta_{21}\mu\Delta t)/A_2,\\
&\tilde\alpha_{30} = (\alpha_{30}+\beta_{30}\mu\Delta t)/A_3, \quad \tilde\alpha_{31} = e^{\beta_{10}\mu\Delta t}(\alpha_{31}+\beta_{31}\mu\Delta t)/A_3, \quad \tilde\alpha_{32} = e^{A\mu\Delta t}(\alpha_{32}+\beta_{32}\mu\Delta t)/A_3,\\
&\tilde\beta_{10} = \beta_{10}/A_1, \quad \tilde\beta_{20} = \beta_{20}/A_2, \quad \tilde\beta_{21} = e^{\beta_{10}\mu\Delta t}\beta_{21}/A_2,\\
&\tilde\beta_{30} = \beta_{30}/A_3, \quad \tilde\beta_{31} = e^{\beta_{10}\mu\Delta t}\beta_{31}/A_3, \quad \tilde\beta_{32} = e^{A\mu\Delta t}\beta_{32}/A_3.
\end{aligned}$$

Following the derivation steps in [26], the sufficient and necessary conditions for third-order accuracy are

$$\tilde{\alpha}_{10} = \tilde{\alpha}_{20} + \tilde{\alpha}_{21} = \tilde{\alpha}_{30} + \tilde{\alpha}_{31} + \tilde{\alpha}_{32} = 1, \tag{2.12}$$

$$\tilde{\alpha}_{31}\tilde{\beta}_{10} + \tilde{\alpha}_{32}A + \tilde{\beta}_{30} + \tilde{\beta}_{31} + \tilde{\beta}_{32} = 1 + \mathcal{O}(\Delta t^3), \tag{2.13}$$

$$\tilde{\beta}_{10}(\tilde{\alpha}_{32}\tilde{\beta}_{21} + \tilde{\beta}_{31}) + \tilde{\beta}_{32}A = \frac{1}{2} + \mathcal{O}(\Delta t^2), \tag{2.14}$$

$$\tilde{\beta}_{10}^2(\tilde{\alpha}_{32}\tilde{\beta}_{21} + \tilde{\beta}_{31}) + \tilde{\beta}_{32}A^2 = \frac{1}{3} + \mathcal{O}(\Delta t), \quad \tilde{\beta}_{10}\tilde{\beta}_{21}\tilde{\beta}_{32} = \frac{1}{6} + \mathcal{O}(\Delta t). \tag{2.15}$$

It is easy to see that (2.12) is satisfied for all $\tilde{\alpha}$ and $\tilde{\beta}$. Without loss of generality, we assume

$$\alpha_{10} = 1, \quad \alpha_{20} + \alpha_{21} = 1, \quad \alpha_{30} + \alpha_{31} + \alpha_{32} = 1. \tag{2.16}$$

Next, we take $\mu = 0$, then we have $A_1 = A_2 = A_3 = 1$ and (2.13)-(2.15) yield

$$\alpha_{31}\beta_{10} + \alpha_{32}A + \beta_{30} + \beta_{31} + \beta_{32} = 1, \tag{2.17}$$

$$\beta_{10}(\alpha_{32}\beta_{21} + \beta_{31}) + \beta_{32}A = \frac{1}{2}, \tag{2.18}$$

$$\beta_{10}^2(\alpha_{32}\beta_{21} + \beta_{31}) + \beta_{32}A^2 = \frac{1}{3}, \quad \beta_{10}\beta_{21}\beta_{32} = \frac{1}{6}. \tag{2.19}$$

(2.16)-(2.19) are necessary order conditions and can be used to derive the sufficient conditions below. For general $\mu > 0$, we apply Taylor's expansion to all the exponential functions in $\tilde{\alpha}$ and $\tilde{\beta}$ and use Mathematica for all the computations. For simplicity, we skip all the complicated and tedious algebra to simplify (2.13)-(2.15), and only demonstrate the results. Under the conditions (2.16)-(2.19), we can verify (2.14) and (2.15) for general $\mu$. Yet (2.13) is not satisfied. To obtain (2.13), we need a new condition

$$\beta_{10}^2\alpha_{31}(1-\beta_{10}) + \alpha_{32}A^2(1-A) + \beta_{10}^2\alpha_{21}\alpha_{32}(A-\beta_{10}) = 2\beta_{10}\beta_{21}\alpha_{32}(1-A). \tag{2.20}$$

Now, we have finished constructing the sufficient and necessary conditions for third-order accuracy and we can demonstrate the following theorem.

**Theorem 2.2.** *Consider the ODE system (2.1), the exponential Runge-Kutta time discretization (2.9)-(2.11) is third-order accurate if and only if the conditions (2.16)-(2.20) are satisfied.*

It is easy to check that the following parameters satisfy all the conditions in the above theorem:

$$\begin{aligned} &\alpha_{10} = 1, \quad \beta_{10} = \frac{2}{3}, \quad \alpha_{20} = \frac{7}{8}, \quad \beta_{20} = \frac{1}{12}, \quad \alpha_{21} = \frac{1}{8}, \quad \beta_{21} = \frac{1}{2}, \\ &\alpha_{30} = \frac{1}{2}, \quad \beta_{30} = \frac{1}{12}, \quad \alpha_{31} = \frac{1}{6}, \quad \beta_{31} = \frac{1}{12}, \quad \alpha_{32} = \frac{1}{3}, \quad \beta_{32} = \frac{1}{2}. \end{aligned} \tag{2.21}$$

**Remark 2.3.** In (2.9)-(2.11), we can apply Taylor's expansion to all the exponential functions following [7], e.g.

$$e^{A\mu\Delta t} \approx \left[1 - A\mu\Delta t + \frac{1}{2}(A\mu\Delta t)^2 - \frac{1}{6}(A\mu\Delta t)^3 + \frac{1}{24}(A\mu\Delta t)^4\right]^{-1}. \tag{2.22}$$

As demonstrated in [12], if $\mu$ is large, $e^{A\mu\Delta t}$ would be an extremely large number. Numerical experiments demonstrated that with the trick (2.22), we can obtain better numerical approximations. Moreover, such a trick keeps the properties to be discussed below, i.e. the scheme is also sign-preserving, steady-state-preserving and has $A(\frac{\pi}{4})$-stability. The proofs would be basically the same, so we skip them.

However, the above choice of the parameters may not be the best one we want. Before we seek the best choice, we would like to demonstrate the sign-preserving property of the conservative third-order RK scheme. The result is given in the following theorem.

**Theorem 2.3.** *Consider the ODE system (2.1) with the flux* **F** *and the stiff source satisfying (2.3) and (2.4), respectively. The scheme (2.9)-(2.11) is sign-preserving: If* $\mathbf{w}^n \geq 0$*, then we have* $\mathbf{w}^{(1)} \geq 0$ *under the conditions*

$$\mu \geq \frac{1}{\varepsilon}\max_{1\leq i\leq l}\left\{-\frac{s_i}{w_i}(\mathbf{w}^n), 0\right\} \quad \text{and} \quad \Delta t \leq \frac{\alpha_{10}}{\beta_{10}}\Delta t_E.$$

*In addition to the above conditions, if*

$$\mu \geq \frac{1}{\varepsilon} \max_{1\leq i\leq \ell}\{-\frac{s_i}{w_i}(\mathbf{w}^{(1)}), -\frac{s_i}{w_i}(\mathbf{w}^{(2)})\}, \quad \textit{and} \quad \Delta t \leq \zeta \Delta t_E,$$

*where*

$$\zeta = \min\{\frac{\alpha_{10}}{\beta_{10}}, \frac{\alpha_{20}}{\beta_{20}}, \frac{\alpha_{21}}{\beta_{21}}, \frac{\alpha_{30}}{\beta_{30}}, \frac{\alpha_{31}}{\beta_{31}}, \frac{\alpha_{32}}{\beta_{32}}\},$$

*then* $\mathbf{w}^{n+1} \geq 0$.

**Proof.** We only prove $\mathbf{w}^{(1)} \geq 0$, since the proof for positive $\mathbf{w}^{n+1}$ can be obtained following the same lines. The conditions $\mathbf{w}^n \geq \mathbf{0}$ and $\mu \geq \frac{1}{\varepsilon}\max_{1\leq i\leq l}\left\{-\frac{s_i}{w_i}(\mathbf{w}^n), 0\right\}$ imply

$$\frac{1}{\varepsilon}\mathbf{s}(\mathbf{w}^n) + \mu\mathbf{w}^n \geq \mathbf{0}. \tag{2.23}$$

Moreover if we choose $\Delta t \leq \frac{\alpha_{10}}{\beta_{10}}\Delta t_E$, then

$$\alpha_{10}\mathbf{w}^n + \beta_{10}\Delta t\mathbf{F}(\mathbf{w}^n) = \alpha_{10}\left(\mathbf{w}^n + \frac{\beta_{10}}{\alpha_{10}}\Delta t\mathbf{F}(\mathbf{w}^n)\right) \geq 0, \tag{2.24}$$

where we have used the basic property (2.3) and the fact that all $\alpha_{ij}$ and $\beta_{ij}$ are positive constants. Combining (2.23) and (2.24), and using the fact that $A_1 > 0$, we have

$$\mathbf{w}^{(1)} = \left[\alpha_{10}\mathbf{w}^n + \beta_{10}\Delta t\mathbf{F}(\mathbf{w}^n) + \beta_{10}\Delta t(\frac{1}{\varepsilon}\mathbf{s}(\mathbf{w}^n) + \mu\mathbf{w}^n)\right]/A_1 \geq 0. \quad \square$$

Notice that if we write $\mathbf{s} = \mathbf{p} - \mathbf{d}$, where $\mathbf{p}$ and $\mathbf{d}$ are the production and destruction terms, respectively, then the requirement $\mu \geq \frac{1}{\varepsilon}\max_{0\leq i\leq l}\left\{-\frac{s_i}{w_i}, 0\right\}$ becomes

$$\mu \geq \frac{1}{\varepsilon}\max_{0\leq i\leq l}\left\{\frac{d_i - p_i}{w_i}, 0\right\}.$$

Since we have assumed that $\lim_{w_i\to 0}\frac{d_i}{w_i}$ exists, we are able to obtain a suitable $\mu$. In practice, we can also take $\mu = \frac{1}{\varepsilon}\max_{0\leq i\leq l}\left\{\frac{d_i}{w_i}, 0\right\}$.

Based on the above theorem, we would like the value of $\zeta$ to be large. One can check that $\zeta = 0.25$ if we choose the parameters in (2.21). We further solve for nonnegative $\alpha$ and $\beta$ satisfying the order conditions (2.16)-(2.20) such that $\zeta$ is maximized. We use the MATLAB function *fmincon* to solve this optimization problem and take (2.21) as the initial data for iteration. The optimal coefficients are

$$\begin{aligned}
&\alpha_{10} = 1, \quad \beta_{10} = 0.7071933376925014,\\
&\alpha_{20} = 0.6686892933074404, \quad \beta_{20} = 0,\\
&\alpha_{21} = 0.3313107066925596, \quad \beta_{21} = 0.4178047564915065,\\
&\alpha_{30} = 0.3487419430256090, \quad \beta_{30} = 0,\\
&\alpha_{31} = 0.2039576138780898, \quad \beta_{31} = 0,\\
&\alpha_{32} = 0.4473004430963011, \quad \beta_{32} = 0.5640754637100439,
\end{aligned} \tag{2.25}$$

with

$$\zeta = \min\{\frac{\alpha_{10}}{\beta_{10}}, \frac{\alpha_{20}}{\beta_{20}}, \frac{\alpha_{21}}{\beta_{21}}, \frac{\alpha_{30}}{\beta_{30}}, \frac{\alpha_{31}}{\beta_{31}}, \frac{\alpha_{32}}{\beta_{32}}\} = 0.7929797388491311.$$

The above results may be locally optimal and it is difficult to solve the global optimization problem.

Besides the above, we can also show that the conservative time integration (2.9)-(2.11) is steady-state-preserving.

**Theorem 2.4.** *Consider the ODE system* (2.1) *and the time integration is given as* (2.9)-(2.11)*. The scheme is steady-state-preserving, namely, if* $\mathbf{w}^n = \widehat{\mathbf{w}}$ *satisfies* $\mathbf{F}(\widehat{\mathbf{w}}) + \frac{1}{\varepsilon}\mathbf{s}(\widehat{\mathbf{w}}) = \mathbf{0}$*, then* $\mathbf{w}^{n+1} = \widehat{\mathbf{w}}$*.*

**Proof.** Take $\mathbf{w}^n = \widehat{\mathbf{w}}$ in (2.9), then

$$\begin{aligned}
\mathbf{w}^{(1)} &= \left[\alpha_{10}\widehat{\mathbf{w}} + \beta_{10}\Delta t\mathbf{F}(\widehat{\mathbf{w}}) + \beta_{10}\Delta t(\frac{1}{\varepsilon}\mathbf{s}(\widehat{\mathbf{w}}) + \mu\widehat{\mathbf{w}})\right]/A_1 \\
&= \left[(\alpha_{10} + \beta_{10}\mu\Delta t)\widehat{\mathbf{w}} + \beta_{10}\Delta t\left(\mathbf{F}(\widehat{\mathbf{w}}) + \frac{1}{\varepsilon}\mathbf{s}(\widehat{\mathbf{w}})\right)\right]/A_1 \\
&= A_1\widehat{\mathbf{w}}/A_1 \\
&= \widehat{\mathbf{w}}.
\end{aligned}$$

Following the same analysis above, we can use the fact that $\mathbf{w}^n = \mathbf{w}^{(1)} = \widehat{\mathbf{w}}$ to show that $\mathbf{w}^{(2)} = \widehat{\mathbf{w}}$ in (2.10), which further yields $\mathbf{w}^{n+1} = \widehat{\mathbf{w}}$. So we skip the details here. □

Finally, we discuss the $A(\frac{\pi}{4})$-stability of the scheme and the result is given below.

**Theorem 2.5.** *Consider the scalar ODE $w_t = \lambda w$ ($F = 0$, $\frac{1}{\varepsilon}s(w) = \lambda w$), where $\lambda \in C$ is a constant with $Re\lambda < 0$. We can rewrite the time integration (2.9)-(2.11) as*

$$w^{n+1} = R(z)w^n,$$

*where $z = \lambda\Delta t$ and $R(z)$ is a function of $z$. Then the scheme satisfies*

1. *$A(\alpha)$-stability with $\alpha = \pi/4$: If $\mu \geq -Re\lambda$ and $Rez \leq -|Imz|$ (i.e. $Re\lambda \leq -|Im\lambda|$), then*

$$|R(z)| \leq 1.$$

2. *Decay property: If we take $\mu = -Re\lambda$, then $R(z) \to 0$ as $Rez \to -\infty$.*

**Proof.** We first consider the $A(\alpha)$-stability. Since $Re\lambda \leq -|Im\lambda|$, then $|Re\lambda| = -Re\lambda \geq |Im\lambda|$. From (2.9), we have

$$w^{(1)} = \frac{\alpha_{10}w^n + \beta_{10}\Delta t(\lambda w^n + \mu w^n)}{\alpha_{10} + \beta_{10}\mu\Delta t} = \frac{\alpha_{10} + \beta_{10}\mu\Delta t + \beta_{10}\lambda\Delta t}{\alpha_{10} + \beta_{10}\mu\Delta t}w^n := R^{(1)}w^n,$$

then

$$R^{(1)} = \frac{\alpha_{10} + \beta_{10}(\mu + Re\lambda)\Delta t + i\beta_{10}Im\lambda\Delta t}{\alpha_{10} + \beta_{10}\mu\Delta t}$$

which further yields

$$\begin{aligned}
|R^{(1)}|^2 &= \frac{(\alpha_{10} + \beta_{10}(\mu + Re\lambda)\Delta t)^2 + (\beta_{10}Im\lambda\Delta t)^2}{(\alpha_{10} + \beta_{10}\mu\Delta t)^2} \\
&\leq \frac{(\alpha_{10} + \beta_{10}(\mu + Re\lambda)\Delta t)^2 + (-\beta_{10}Re\lambda\Delta t)^2}{(\alpha_{10} + \beta_{10}(\mu + Re\lambda)\Delta t + (-\beta_{10}Re\lambda\Delta t))^2} \\
&\leq 1,
\end{aligned}$$

where in the last step we use the fact that both $a_{10} + \beta_{10}(\mu + Re\lambda)\Delta t$ and $-\beta_{10}Re\lambda\Delta t$ are nonnegative real numbers. Hence we have proved $|w^{(1)}| \leq |w^n|$. Applying the same analysis above to (2.10) and (2.11) we can obtain $w^{(2)} = R^{(2)}w^n$ and $w^{n+1} = Rw^n$ with

$$\begin{aligned}
R^{(2)} &= \frac{\alpha_{20} + \beta_{20}(\mu + \lambda)\Delta t + e^{\beta_{10}\mu\Delta t}(\alpha_{21} + \beta_{21}(\mu + \lambda)\Delta t)R^{(1)}}{\alpha_{20} + \beta_{20}\mu\Delta t + e^{\beta_{10}\mu\Delta t}(\alpha_{21} + \beta_{21}\mu\Delta t)}, \\
R &= \frac{\alpha_{30} + \beta_{30}(\mu + \lambda)\Delta t + e^{\beta_{10}\mu\Delta t}(\alpha_{31} + \beta_{31}(\mu + \lambda)\Delta t)R^{(1)} + e^{A\mu\Delta t}(\alpha_{32} + \beta_{32}(\mu + \lambda)\Delta t)R^{(2)}}{\alpha_{30} + \beta_{30}\mu\Delta t + e^{\beta_{10}\mu\Delta t}(\alpha_{31} + \beta_{31}\mu\Delta t) + e^{A\mu\Delta t}(\alpha_{32} + \beta_{32}\mu\Delta t)}.
\end{aligned}$$

Then

$$R^{(2)} = \nu_1 + \nu_2\frac{-Im\lambda}{Re\lambda}i + \nu_3R^{(1)} + \nu_4\frac{-Im\lambda}{Re\lambda}iR^{(1)},$$

where

$$\begin{aligned}
\nu_1 &= \frac{\alpha_{20} + \beta_{20}(\mu + Re\lambda)\Delta t}{A_2}, \quad \nu_2 = -\frac{\beta_{20}Re\lambda\Delta t}{A_2}, \\
\nu_3 &= e^{\beta_{10}\mu\Delta t}\frac{\alpha_{21} + \beta_{21}(\mu + Re\lambda)\Delta t}{A_2}, \quad \nu_4 = -e^{\beta_{10}\mu\Delta t}\frac{\beta_{21}Re\lambda\Delta t}{A_2}.
\end{aligned}$$

It is easy to see that $\nu_i \geq 0$, $i = 1, \cdots, 4$ and $\sum_{i=1}^4 \nu_i = 1$. Therefore, $R^{(2)}$ can be written as a convex combination of points in the unit circle centered at the origin in the complex plane. Therefore, $|R^{(2)}| \leq 1$. Following the same analysis with some minor changes, we can also obtain that $|R| \leq 1$, hence we skip it. Now we finish the proof of part 1.

In part 2, we take $\mu = -Re\lambda$, then $Rez = -\mu\Delta t$. Since $\beta_{10} \neq 0$, we have

$$R^{(1)} = \frac{\alpha_{10} + i\beta_{10} Imz}{\alpha_{10} - \beta_{10} Rez} \to 0 \quad \text{as} \quad Rez \to -\infty.$$

Similarly, since $\beta_{21} \neq 0$, we can obtain that

$$R^{(2)} = \frac{\alpha_{20} + i\beta_{20} Imz + e^{-\beta_{10} Rez}(\alpha_{21} + i\beta_{21} Imz) R^{(1)}}{\alpha_{20} - \beta_{20} Rez + e^{-\beta_{10} Rez}(\alpha_{21} - \beta_{21} Rez)} \to 0 \quad \text{as} \quad Rez \to -\infty.$$

Following the same analysis with some minor changes, we can prove $R \to 0$ as $Rez \to -\infty$. □

**Remark 2.4.** In the above theorem, we require $\mu \geq -Re\lambda$ which is exactly the same as that given in Theorem 2.3. In the decay property, though we assume $\mu = -Re\lambda$, the conclusion is still valid if we take $\mu = \mathcal{O}(-Re\lambda)$. However, if $\mu \gg -Re\lambda$, the conclusion may not be true.

**Remark 2.5.** In this paper, we consider third-order RK scheme only, and the fourth-order scheme will be discussed in the future. There are two main difficulties. The first one is how to find the sufficient and necessary conditions for the fourth-order accuracy. The second one is to find the parameters $\alpha_{ij}$ and $\beta_{ij}$ based on the accuracy condition.

In this subsection, we proved several properties of the proposed time integration. However, due to the conservative requirement of the system, it is very difficulty to obtain the asymptotic-preserving (AP) property for general systems. However, following the analysis in [12], we can obtain the weak AP property. So we skip the proof and only demonstrate the result in the following statement.

**Proposition 2.1.** *Consider the scalar ODE $u_t = \frac{1}{\varepsilon} s(u)$ with s(0)=0 and $s'(u) \leq 0$. Assuming that $s(u) \neq 0$ for $u \neq 0$ and $\beta_{30} > 0$, then the new RK3 scheme (2.9)-(2.11) with modification (2.22) is AP in the weak sense: for any $\varepsilon > 0$ and any initial value $u_0$, and $\Delta t \gg \varepsilon$, there exists an integer $N_\varepsilon \geq 1$ (independent of $\Delta t$), such that*

$$s(u^n) = \mathcal{O}(\varepsilon), n \geq N_\varepsilon.$$

We can see that (2.21) satisfies the condition given in the proposition but (2.25) does not. However, $\beta_{30} > 0$ is only a sufficient condition. Numerical experiments also demonstrate convergence of the numerical approximations in the stiff regime by using (2.25).

### *2.2. Multistep method*

In this subsection, we proceed to analyze the multistep method. For problems in which the wave speed changes slowly, multistep method can be used as an alternative to the RK method.

In [7], we have used the SSP exponential multistep methods with some modifications to discretize (2.1). The scheme is given as

$$\begin{aligned}\mathbf{w}^{n+1} = {} & A_3^1 \left[ \mathbf{w}^n + 3\Delta t \mathbf{F}(\mathbf{w}^n) + 3\Delta t \left( \frac{1}{\varepsilon} \mathbf{s}(\mathbf{w}^n) + \mu \mathbf{w}^n \right) \right] \\ & + A_3^2 \left[ \mathbf{w}^{n-3} + \frac{12}{11} \Delta t \mathbf{F}(\mathbf{w}^{n-3}) + \frac{12}{11} \Delta t \left( \frac{1}{\varepsilon} \mathbf{s}(\mathbf{w}^{n-3}) + \mu \mathbf{w}^{n-3} \right) \right],\end{aligned} \tag{2.26}$$

where

$$A_3^1 = \frac{16}{27} \frac{1 - \mu\Delta t + \frac{1}{2}(\mu\Delta t)^2 - \frac{1}{6}(\mu\Delta t)^3 + \frac{1}{24}(\mu\Delta t)^4}{1 - \frac{2}{3}(\mu\Delta t)^4 + \frac{130}{27}(\mu\Delta t)^5},$$

$$A_3^2 = \frac{11}{27} \frac{1 - 4\mu\Delta t + 8(\mu\Delta t)^2 - \frac{32}{3}(\mu\Delta t)^3 + \frac{32}{3}(\mu\Delta t)^4}{1 - \frac{2}{3}(\mu\Delta t)^4 + \frac{130}{27}(\mu\Delta t)^5}.$$

It is easy to verify the following facts via direct computation:

$$A_3^1 \geq 0, \quad A_3^2 \geq 0, \quad A_3^1(1 + 3\mu\Delta t) + A_3^2(1 + \frac{12}{11}\mu\Delta t) = 1. \tag{2.27}$$

In [7], we have proved that this scheme is conservative and third-order accurate. In this paper, we will continue to prove that the scheme is also sign-preserving and steady state-preserving. We state these properties in the following theorems.

**Theorem 2.6.** *Consider the ODE system (2.1), the multistep time integration (2.26) is third-order accurate and conservative. Moreover, the scheme is sign-preserving. If $\mathbf{w}^n \geq 0$ and $\mathbf{w}^{n-3} \geq 0$, then we have $\mathbf{w}^{n+1} \geq 0$ under the conditions*

$$\mu \geq \frac{1}{\varepsilon} \sup_{1\leq i\leq \ell} \{-\frac{s_i}{w_i}(\mathbf{w}^n), -\frac{s_i}{w_i}(\mathbf{w}^{n-3})\} \quad \textit{and} \quad \Delta t \leq \frac{1}{3}\Delta t_E.$$

**Proof.** We only need to prove the sign-preserving property. Following the proof of Theorem 2.3, we have

$$\frac{1}{\varepsilon}\mathbf{s}(\mathbf{w}^n) + \mu\mathbf{w}^n \geq \mathbf{0}, \quad \frac{1}{\varepsilon}\mathbf{s}(\mathbf{w}^{n-3}) + \mu\mathbf{w}^{n-3} \geq \mathbf{0}. \tag{2.28}$$

Moreover if we choose $\Delta t \leq \frac{1}{3}\Delta t_E$, then

$$\mathbf{w}^n + 3\Delta t\mathbf{F}(\mathbf{w}^n) \geq 0, \qquad \mathbf{w}^{n-3} + \frac{12}{11}\Delta t\mathbf{F}(\mathbf{w}^{n-3}) \geq 0, \tag{2.29}$$

where we have used the basic property (2.3). Combining the above equations, we have

$$\begin{aligned}\mathbf{w}^{n+1} &= A_3^1\left[\mathbf{w}^n + 3\Delta t\mathbf{F}(\mathbf{w}^n) + 3\Delta t\left(\frac{1}{\varepsilon}\mathbf{s}(\mathbf{w}^n) + \mu\mathbf{w}^n\right)\right] \\ &\quad + A_3^2\left[\mathbf{w}^{n-3} + \frac{12}{11}\Delta t\mathbf{F}(\mathbf{w}^{n-3}) + \frac{12}{11}\Delta t\left(\frac{1}{\varepsilon}\mathbf{s}(\mathbf{w}^{n-3}) + \mu\mathbf{w}^{n-3}\right)\right] \geq 0. \quad \square\end{aligned} \tag{2.30}$$

Next, we will show that the scheme (2.26) is steady-state-preserving.

**Theorem 2.7.** *Consider the ODE system (2.1), the multistep time integration (2.26) is steady-state-preserving, namely, if $\mathbf{w}^{n-3} = \mathbf{w}^n = \widehat{\mathbf{w}}$ satisfies $\mathbf{F}(\widehat{\mathbf{w}}) + \frac{1}{\varepsilon}\mathbf{s}(\widehat{\mathbf{w}}) = \mathbf{0}$, then $\mathbf{w}^{n+1} = \widehat{\mathbf{w}}$.*

**Proof.** Taking $\mathbf{w}^{n-3} = \mathbf{w}^n = \widehat{\mathbf{w}}$, we have

$$\begin{aligned}\mathbf{w}^{n+1} &= A_3^1\left[\widehat{\mathbf{w}} + 3\Delta t\mathbf{F}(\widehat{\mathbf{w}}) + 3\Delta t\left(\frac{1}{\varepsilon}\mathbf{s}(\widehat{\mathbf{w}}) + \mu\widehat{\mathbf{w}}\right)\right] \\ &\quad + A_3^2\left[\widehat{\mathbf{w}} + \frac{12}{11}\Delta t\mathbf{F}(\widehat{\mathbf{w}}) + \frac{12}{11}\Delta t\left(\frac{1}{\varepsilon}\mathbf{s}(\widehat{\mathbf{w}}) + \mu\widehat{\mathbf{w}}\right)\right] \\ &= A_3^1[\widehat{\mathbf{w}} + 3\Delta t\mu\widehat{\mathbf{w}}] + A_3^2\left[\widehat{\mathbf{w}} + \frac{12}{11}\Delta t\mu\widehat{\mathbf{w}}\right] \\ &= \left[A_3^1(1+3\mu\Delta t) + A_3^2(1+\frac{12}{11}\mu\Delta t)\right]\widehat{\mathbf{w}} = \widehat{\mathbf{w}}. \quad \square\end{aligned}$$

Finally, we will investigate the region of absolute stability of the multistep method (2.26).

**Theorem 2.8.** *Consider the scalar ODE $w_t = \lambda w$ ($F = 0$, $\frac{1}{\varepsilon}s(w) = \lambda w$), where $\lambda \in C$ is a constant with $Re\lambda < 0$. Define $w^n$ to be the numerical approximation at time level n. Then, the time integration (2.26) satisfies*

1. *$A(\frac{\pi}{4})$-stability: If $\mu \geq -Re\lambda$ then the region of absolute stability contains $\{z : |Rez \leq -|Imz|\}$, where $z = \lambda\Delta t$.*
2. *Decay property: If we take $\mu = -Re\lambda$, then $|w^n| \to 0$ as $Rez \to -\infty$.*

**Proof.** It is easy to check that

$$w^{n+1} = A_3^1[1 + 3\Delta t(\mu+\lambda)]\,w^n + A_3^2\left[1 + \frac{12}{11}\Delta t(\mu+\lambda)\right]w^{n-3}.$$

The characteristic equation is

$$\xi^4 = A_3^1[1 + 3\Delta t(\mu+\lambda)]\,\xi^3 + A_3^2\left[1 + \frac{12}{11}\Delta t(\mu+\lambda)\right]. \tag{2.31}$$

We assume $Rez \le -|Imz|$ (i.e. $Re\lambda \le -|Im\lambda|$) and first show that all the roots of (2.31) are $\le 1$ in modulus. If false, there exists $\eta$ with $|\eta| > 1$ to be a solution to (2.31), then

$$|\eta^3| < |\eta^4| \le |A_3^1 [1 + 3\Delta t(\mu + \lambda)]\, \eta^3| + \left| A_3^2 \left[ 1 + \frac{12}{11} \Delta t(\mu + \lambda) \right] \right|,$$

which further implies

$$\begin{aligned} 1 &< A_3^1 |1 + 3\Delta t(\mu + \lambda)| + A_3^2 \left| 1 + \frac{12}{11} \Delta t(\mu + \lambda) \right| \\ &\le A_3^1 (1 + 3\Delta t(\mu + Re\lambda) + |3\Delta t Im\lambda|) + A_3^2 \left( 1 + \frac{12}{11} \Delta t(\mu + Re\lambda) + \frac{12}{11} \Delta t |Im\lambda| \right) \\ &\le A_3^1 (1 + 3\Delta t(\mu + Re\lambda) - 3\Delta t Re\lambda) + A_3^2 \left( 1 + \frac{12}{11} \Delta t(\mu + Re\lambda) - \frac{12}{11} \Delta t Re\lambda \right) \\ &= A_3^1 (1 + 3\Delta t\mu) + A_3^2 \left( 1 + \frac{12}{11} \Delta t \mu \right) \\ &= 1, \end{aligned}$$

which is a contradiction.

Next, we will show that if $\eta$ is a solution to (2.31) with $|\eta| = 1$, then $\eta$ is a single solution. If false, suppose $\eta$ with $|\eta| = 1$ is not a single solution to (2.31). Take derivative of (2.31) with respect to $\xi$ to obtain

$$4\xi^3 = 3A_3^1 [1 + 3\Delta t(\mu + \lambda)]\, \xi^2. \tag{2.32}$$

Then $\eta$ is a solution to (2.32). Hence

$$4\eta = 3A_3^1 [1 + 3\Delta t(\mu + \lambda)],$$

which further yields

$$\begin{aligned} 4 &= 3A_3^1 |1 + 3\Delta t(\mu + \lambda)| \\ &\le 3A_3^1 (1 + 3\Delta t(\mu + Re\lambda) + 3\Delta t |Im\lambda|) \\ &\le 3A_3^1 (1 + 3\Delta t(\mu + Re\lambda) - 3\Delta t Re\lambda) \\ &= 3A_3^1 (1 + 3\Delta t\mu) \\ &\le 3. \end{aligned}$$

Now, we find a contradiction. Therefore, if $\eta$ is a solution to (2.31), then $|\eta| \le 1$. Moreover, if $|\eta| = 1$, then $\eta$ is a single solution. We finish the proof of part 1.

Assume $\mu = -Re\lambda$, then (2.31) can be written as

$$\xi^4 = A_3^1 [1 + 3i\Delta t Im\lambda]\, \xi^3 + A_3^2 \left[ 1 + \frac{12}{11} i\Delta t Im\lambda \right].$$

It is easy to see that $A_3^1 \to 0$ and $A_3^2 \to 0$ as $Rez \to -\infty$. Hence $\xi \to 0$ as $Rez \to -\infty$ and we finish the proof. □

## 3. Applications in non-equilibrium stiff multispecies and multireaction detonations

In this section, we aim to solve the multispecies and multireaction detonations problem (1.1). We will simply review the DG method for spacial discretization in Section 3.1 and use the new ODE solvers constructed in this paper for the time discretizations. For simplicity, we only discuss RK methods in this section. Moreover, we will also construct the bound-preserving technique in Section 3.2. We will demonstrate that the conservative property of the new time integrations is essential for the bound-preserving technique. For simplicity, we define $\mathbf{v} = (-1, 0, 0, 0, 1, \cdots, 1)^T \in R^{M+4}$ throughout this section.

### 3.1. DG methods coupled with the conservative RK methods

We rewrite (1.1) into the form of

$$\mathbf{w}_t + \mathbf{f}(\mathbf{w})_x + \mathbf{g}(\mathbf{w})_y = \mathbf{s}(\mathbf{w}), \tag{3.1}$$

where

$$\begin{aligned}\mathbf{w} &= (\rho, m, n, E, \rho z_1, \cdots, \rho z_{M-1})^T,\\ \mathbf{f}(\mathbf{w}) &= (m, mu+p, mv, (E+p)u, mz_1, \cdots, mz_{M-1})^T,\\ \mathbf{g}(\mathbf{w}) &= (n, nu, nv+p, (E+p)v, nz_1, \cdots, nz_{M-1})^T,\\ \mathbf{s}(\mathbf{w}) &= (0,0,0,0,s_1,\cdots,s_{M-1})^T.\end{aligned}$$

We first consider the spacial discretization. Let $\Omega_h = \{K\}$ be a quasi-uniform partition of the computational domain $\Omega$ with rectangular or triangular elements. We define the finite element space $V_h^k$ as

$$V_h^k = \left\{ z : z\big|_K \in P^k(K), \forall K \in \Omega_h \right\},$$

where $P^k(K)$ denotes the set of polynomials of degree up to $k$ in cell $K$. For simplicity, we also use the notation $\mathbf{w}$ as the numerical approximations. The DG scheme is to find $\mathbf{w} \in \mathbf{V}_h = [V_h^k]^{M+3}$ such that for any test functions $\boldsymbol{\xi} \in \mathbf{V}_h$ and $K \in \Omega_h$ we have

$$\int_K \mathbf{w}_t \cdot \boldsymbol{\xi}\, d\mathbf{x} = \int_K \mathbf{F}(\mathbf{w}) \cdot \nabla \boldsymbol{\xi}\, d\mathbf{x} - \int_{\partial K} \mathbf{H}(\mathbf{w}^{int}, \mathbf{w}^{ext}, \boldsymbol{\nu}) \cdot \boldsymbol{\xi} ds + \int_K \mathbf{s}(\mathbf{w}) \cdot \boldsymbol{\xi}\, d\mathbf{x}, \tag{3.2}$$

where $\mathbf{F} = \langle \mathbf{f}, \mathbf{g} \rangle$ and $\boldsymbol{\nu}$ is the unit outer normal of $\partial K$ in cell $K$. Here, $\mathbf{w}^{int}$ and $\mathbf{w}^{ext}$ are the values of $\mathbf{w}$ on the edge $\partial K$ obtained from the interior and the exterior of $K$, respectively, and $\mathbf{H}(\mathbf{w}^{int}, \mathbf{w}^{ext}, \boldsymbol{\nu})$ is the numerical flux. In this paper, we consider Lax-Friedrichs flux and

$$\mathbf{H}(\mathbf{w}_1, \mathbf{w}_2, \boldsymbol{\nu}) = \frac{1}{2}\left[\mathbf{F}(\mathbf{w}_1) \cdot \boldsymbol{\nu} + \mathbf{F}(\mathbf{w}_2) \cdot \boldsymbol{\nu} - \alpha(\mathbf{w}_2 - \mathbf{w}_1)\right], \quad \alpha = \| |\langle u, v \rangle| + c \|_\infty, \tag{3.3}$$

where $c = \sqrt{\frac{\gamma p}{\rho}}$ is the sound speed.

**Definition 3.1.** *We say the elements in the numerical flux* $\mathbf{H} = (h_\rho, h_m, h_n, h_E, h_1, \cdots, h_{M-1})^T$ *are consistent if* $h_\rho = h_i$ *if we take* $z_i = 1$ *for all* $1 \le i \le M-1$.

The elements in the numerical flux $\mathbf{H}$ in (3.3) are consistent and

$$\begin{aligned} h_\rho(\mathbf{w}_1, \mathbf{w}_2, \boldsymbol{\nu}) &= \frac{1}{2}\left[\mathbf{F}_\rho(\mathbf{w}_1) \cdot \boldsymbol{\nu} + \mathbf{F}_\rho(\mathbf{w}_2) \cdot \boldsymbol{\nu} - \alpha(\rho_2 - \rho_1)\right],\\ h_i(\mathbf{w}_1, \mathbf{w}_2, \boldsymbol{\nu}) &= \frac{1}{2}\left[\mathbf{F}_i(\mathbf{w}_1) \cdot \boldsymbol{\nu} + \mathbf{F}_i(\mathbf{w}_2) \cdot \boldsymbol{\nu} - \alpha(r_{i2} - r_{i1})\right], \quad i = 1, 2, \cdots, M-1,\end{aligned}$$

where $\mathbf{F}_\rho = (m, n)$ and $\mathbf{F}_i = (mz_i, nz_i)$. Define $h_M = h_\rho - \sum_{i=1}^{M-1} h_i$, and we can obtain

$$h_M(\mathbf{w}_1, \mathbf{w}_2, \boldsymbol{\nu}) = \frac{1}{2}\left[\mathbf{F}_M(\mathbf{w}_1) \cdot \boldsymbol{\nu} + \mathbf{F}_M(\mathbf{w}_2) \cdot \boldsymbol{\nu} - \alpha(r_{M2} - r_{M1})\right],$$

with

$$\mathbf{F}_M = (mz_M, nz_M).$$

Moreover, we can define

$$\tilde{\mathbf{H}} = (\mathbf{H}^T, h_M)^T, \quad \tilde{\mathbf{F}} = (\mathbf{F}^T, \mathbf{F}_M)^T, \quad \tilde{\mathbf{s}} = (\mathbf{s}^T, s_M)^T, \quad \tilde{\mathbf{w}} = (\mathbf{w}^T, r_M)^T,$$

then it is easy to see that $\mathbf{v} \cdot \tilde{\mathbf{H}} = \mathbf{v} \cdot \tilde{\mathbf{F}} = \mathbf{v} \cdot \tilde{\mathbf{s}} = \mathbf{v} \cdot \tilde{\mathbf{w}} = 0$ Then (3.2) together with the "hidden" condition that $\mathbf{v} \cdot \tilde{\mathbf{w}} = 0$ due to the total mass conservation is equivalent to

$$\int_K \tilde{\mathbf{w}}_t \cdot \tilde{\xi}\, d\mathbf{x} = \int_K \tilde{\mathbf{F}}(\tilde{\mathbf{w}}) \cdot \nabla \tilde{\xi}\, d\mathbf{x} - \int_{\partial K} \tilde{\mathbf{H}}(\tilde{\mathbf{w}}^{int}, \tilde{\mathbf{w}}^{ext}, \boldsymbol{\nu}) \cdot \tilde{\xi} ds + \int_K \tilde{\mathbf{s}}(\tilde{\mathbf{w}}) \cdot \tilde{\xi}\, d\mathbf{x}, \tag{3.4}$$

where $\tilde{\boldsymbol{\xi}} \in [V_h^k]^{M+4}$. We can see that (3.4) is in the weak form of (2.1). Then instead of forcing the total mass conservation explicitly by letting $r_M = \rho - \sum_{i=1}^{M-1} r_i$, we analyze the equivalent system (3.4). In fact, the last equation in (3.4) is used for solving $r_M$ and it is a numerical scheme for the hidden equation (1.3). The total mass conservation can be obtained by using the conservative time integration introduced in Section 2.

### 3.2. Bound-preserving technique

In this subsection, we develop the bound-preserving technique for the gaseous detonation. We consider the new equivalent DG scheme (3.4). For simplicity, we omit the tilde in the scheme, and use $\mathbf{o}$ for $\tilde{\mathbf{o}}$ with $o = w, F, H, \xi, s$ in this subsection.

The convex admissible set of solutions in [7] is defined as

$$G = \left\{ \mathbf{w} = \begin{pmatrix} \rho \\ m \\ n \\ E \\ r_1 \\ \cdots \\ r_{M-1} \end{pmatrix}, \rho > 0, p > 0, z_1 > 0, \cdots, z_M > 0, \sum_{i=1}^{M} z_i = 1 \right\}.$$

We aim to obtain the numerical approximations that lie in $G$.

We first introduce some notations. For each cell $K \in \Omega_h$, we need to define a set of quadrature points, denoted as $S_K$, which will be used in the bound-preserving technique. We first consider the rectangular cell. Let $K_{ij} = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$ be the $(i, j)$-th cell. For simplicity, we assume uniform meshes and denote $\Delta x$ and $\Delta y$ as the mesh sizes in the $x$ and $y$ directions, respectively. However, this assumption is not essential. We choose $L \geq k+1$ and use $p_i^x = \left\{x_i^\beta : \beta = 1, \cdots, L\right\}$ and $p_j^y = \left\{y_j^\beta : \beta = 1, \cdots, L\right\}$ to denote the Gauss quadrature points on $\left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right]$ and $\left[y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}\right]$, respectively. Moreover, we use $\hat{p}_i^x = \left\{\hat{x}_i^\alpha : \alpha = 0, \cdots, \hat{L}\right\}$ and $\hat{p}_j^y = \left\{\hat{y}_j^\alpha : \alpha = 0, \cdots, \hat{L}\right\}$ to denote the Gauss-Lobatto quadrature points on $\left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right]$ and $\left[y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}\right]$, respectively, with $2\hat{L} - 1 \geq k$. Also, we denote $\hat{w}_\alpha$ as the corresponding Gauss-Lobatto quadrature weights on the interval $\left[-\frac{1}{2}, \frac{1}{2}\right]$. As in [28], we denote $S_{K_{i,j}}$ as

$$S_{K_{i,j}} = (p_i^x \otimes \hat{p}_j^y) \cup (\hat{p}_i^x \otimes p_j^y) \cup (p_i^x \otimes p_j^y). \tag{3.5}$$

Next, we consider the triangular cell $K$. We use $\ell_K^i$ $(i = 1, 2, 3)$ to denote the length of its three edges $e_K^i$ $(i = 1, 2, 3)$. We consult the quadrature introduced in [34], where the quadrature points are given in the barycentric coordinates as

$$\begin{aligned} S_K = \Bigg\{ & \left(\frac{1}{2} + z^\beta, (\frac{1}{2} + \hat{z}^\alpha)(\frac{1}{2} - z^\beta), (\frac{1}{2} - \hat{z}^\alpha)(\frac{1}{2} - z^\beta)\right), \\ & \left((\frac{1}{2} - \hat{z}^\alpha)(\frac{1}{2} - z^\beta), \frac{1}{2} + z^\beta, (\frac{1}{2} + \hat{z}^\alpha)(\frac{1}{2} - z^\beta)\right), \\ & \left((\frac{1}{2} + \hat{z}^\alpha)(\frac{1}{2} - z^\beta), (\frac{1}{2} - \hat{z}^\alpha)(\frac{1}{2} - z^\beta), \frac{1}{2} + z^\beta\right), \\ & \alpha = 0, \cdots, \hat{L}, \ \beta = 1, \cdots, L \Bigg\}, \end{aligned} \tag{3.6}$$

where $\hat{z}^\alpha$ $(\alpha = 0, \cdots, \hat{L})$ and $z^\beta$ $(\beta = 1, \cdots, L)$ are the Gauss-Lobatto and Gaussian quadrature points on the reference interval $[-\frac{1}{2}, \frac{1}{2}]$, respectively.

For the time discretization, we adopt the new conservative three-stage RK method designed in (2.9)-(2.11). Specially, if we take the test function as 1 in each component of the DG scheme (3.4), then we get

$$\frac{d}{dt}\bar{\mathbf{w}}_K = -\frac{1}{|K|}\int_{\partial K} \mathbf{H}(\mathbf{w}^{int}, \mathbf{w}^{ext}, \boldsymbol{\nu})ds + \frac{1}{|K|}\int_K \mathbf{s}(\mathbf{w})d\mathbf{x}, \tag{3.7}$$

where $\bar{\mathbf{w}}_K = \frac{1}{|K|}\int_K \mathbf{w}d\mathbf{x}$ is the cell average of $\mathbf{w}$ in cell $K$. We further denote

$$\mathbf{G}(\mathbf{w}) = -\frac{1}{|K|}\int_{\partial K} \mathbf{H}(\mathbf{w}^{int}, \mathbf{w}^{ext}, \boldsymbol{\nu})ds \quad \text{and} \quad \bar{\mathbf{s}}(\mathbf{w}) = \frac{1}{|K|}\int_K \mathbf{s}(\mathbf{w})d\mathbf{x},$$

then we obtain the following formulations to compute the cell averages in each stage of RK:

$$\bar{\mathbf{w}}^{(1)} = \left[\alpha_{10}\bar{\mathbf{w}}^n + \beta_{10}\Delta t\mathbf{G}(\mathbf{w}^n) + \beta_{10}\Delta t(\bar{\mathbf{s}}(\mathbf{w}^n) + \mu\bar{\mathbf{w}}^n)\right]/A_1, \tag{3.8}$$

$$\begin{aligned}\bar{\mathbf{w}}^{(2)} &= \left[\alpha_{20}\bar{\mathbf{w}}^n + \beta_{20}\Delta t\mathbf{G}(\mathbf{w}^n) + \beta_{20}\Delta t(\bar{\mathbf{s}}(\mathbf{w}^n) + \mu\bar{\mathbf{w}}^n)\right]/A_2 \\ &\quad + e^{\beta_{10}\mu\Delta t}\left[\alpha_{21}\bar{\mathbf{w}}^{(1)} + \beta_{21}\Delta t\mathbf{G}(\mathbf{w}^{(1)}) + \beta_{21}\Delta t(\bar{\mathbf{s}}(\mathbf{w}^{(1)}) + \mu\bar{\mathbf{w}}^{(1)})\right]/A_2,\end{aligned} \tag{3.9}$$

$$\begin{aligned}\bar{\mathbf{w}}^{n+1} &= \left[\alpha_{30}\bar{\mathbf{w}}^n + \beta_{30}\Delta t\mathbf{G}(\mathbf{w}^n) + \beta_{30}\Delta t(\bar{\mathbf{s}}(\mathbf{w}^n) + \mu\bar{\mathbf{w}}^n)\right]/A_3 \\ &\quad + e^{\beta_{10}\mu\Delta t}\left[\alpha_{31}\bar{\mathbf{w}}^{(1)} + \beta_{31}\Delta t\mathbf{G}(\mathbf{w}^{(1)}) + \beta_{31}\Delta t(\bar{\mathbf{s}}(\mathbf{w}^{(1)}) + \mu\bar{\mathbf{w}}^{(1)})\right]/A_3 \\ &\quad + e^{A\mu\Delta t}\left[\alpha_{32}\bar{\mathbf{w}}^{(2)} + \beta_{32}\Delta t\mathbf{G}(\mathbf{w}^{(2)}) + \beta_{32}\Delta t(\bar{\mathbf{s}}(\mathbf{w}^{(2)}) + \mu\bar{\mathbf{w}}^{(2)})\right]/A_3.\end{aligned} \tag{3.10}$$

Before we state the main theorem, we would like to demonstrate the following lemma whose proof has been given in Theorem 3.3 and Lemma 4.1 in [7].

**Lemma 3.1.** *Consider the DG scheme (3.4) (or the equivalent scheme (3.2)). If $\mathbf{w} \in G$ for all $(x, y) \in S$, where $S$ is defined in (3.5) and (3.6) for rectangular and triangular meshes, respectively. Then we have $\bar{\mathbf{w}} + \Delta t\mathbf{G}(\mathbf{w}) \in G$ under the condition $\Delta t \le \Delta\tilde{t}$, where $\Delta\tilde{t}$ satisfies*

$$\alpha(\frac{\Delta\tilde{t}}{\Delta x} + \frac{\Delta\tilde{t}}{\Delta y}) \le \hat{\omega}_1 \tag{3.11}$$

*for rectangular meshes, and satisfies*

$$\alpha\frac{\Delta\tilde{t}}{|K|}\sum_{i=1}^{3}\ell_K^i \le \frac{2}{3}\hat{\omega}_1 \tag{3.12}$$

*for triangular meshes. Moreover, if we take*

$$\mu \ge \max_{1\le i\le M}\left\{-\frac{s_i}{r_i}, \frac{\sum_{j=1}^{M}s_jq_j}{p}, 0\right\}, \tag{3.13}$$

*then $\frac{1}{\mu}(\bar{\mathbf{s}}(\mathbf{w}) + \mu\bar{\mathbf{w}}) \in G$.*

**Remark 3.1.** The sufficient condition for $\mu$ in Theorem 2.3 is different from that in Lemma 3.1 because we only consider the positivity-preserving of $\mathbf{w}$ in Theorem 2.3. In Lemma 3.1, we also include the positivity-preserving of the pressure following the analysis in [7].

Now, we can state the main theorem for bound-preserving technique.

**Theorem 3.1.** *Consider the DG scheme (3.4) (or the equivalent scheme (3.2)) coupled with the three-stage RK method (2.9)-(2.11), where $\mu$ satisfies (3.13) for $\mathbf{w} = \mathbf{w}^n, \mathbf{w}^{(1)}, \mathbf{w}^{(2)}$. If $\mathbf{w}^n(\mathbf{x})$, $\mathbf{w}^{(1)}(\mathbf{x})$, $\mathbf{w}^{(2)}(\mathbf{x}) \in G$ for all $\mathbf{x} \in S_K$ on each $K \in \Omega_h$, where $S_K$ is defined in (3.5) and (3.6) for rectangular and triangular meshes, respectively. Then we have $\bar{\mathbf{w}}^{n+1} \in G$ under the condition*

$$\Delta t \le \min\{\frac{\alpha_{10}}{\beta_{10}}, \frac{\alpha_{20}}{\beta_{20}}, \frac{\alpha_{21}}{\beta_{21}}, \frac{\alpha_{30}}{\beta_{30}}, \frac{\alpha_{31}}{\beta_{31}}, \frac{\alpha_{32}}{\beta_{32}}\}\Delta\tilde{t},$$

*where $\Delta\tilde{t}$ satisfies (3.11) and (3.12) for rectangular and triangular meshes, respectively.*

**Proof.** For simplicity, we consider the first stage of the RK method only and prove $\bar{\mathbf{w}}^{(1)} \in G$. After simple computations, we get

$$\bar{\mathbf{w}}^{(1)} = [\alpha_{10}\mathbf{R}_1 + \beta_{10}\mu\Delta t\mathbf{R}_2]/A_1,$$

where

$$\mathbf{R}_1 = \bar{\mathbf{w}}^n + \frac{\beta_{10}}{\alpha_{10}}\Delta t\mathbf{G}(\mathbf{w}^n) \qquad \text{and} \qquad \mathbf{R}_2 = \frac{1}{\mu}(\bar{\mathbf{s}}(\mathbf{w}^n) + \mu\bar{\mathbf{w}}^n).$$

In Lemma 3.1, we have proved $\mathbf{R}_1 \in G$ under the condition $\Delta t \le \frac{\alpha_{10}}{\beta_{10}}\Delta\tilde{t}$ and $\mathbf{R}_2 \in G$ under the condition (3.13). Recall that the time integration is conservative and $A_1 = \alpha_{10} + \beta_{10}\mu\Delta t$, then $\bar{\mathbf{w}}^{(1)}$ is a convex combination of $\mathbf{R}_1$ and $\mathbf{R}_2$. Since $G$ is a convex set, we have $\bar{\mathbf{w}}^{(1)} \in G$.

Following the same analysis above, we can also prove that $\bar{\mathbf{w}}^{(2)}, \bar{\mathbf{w}}^{\mathbf{n+1}} \in G$. □

**Remark 3.2.** In the theorem above, the value of $\mu$ is not easy to compute. In practice, at time level $n$, we want $\mu$ satisfies (3.13) for $\mathbf{w} = \mathbf{w}^n$ and choose a suitably small $\Delta t$. We monitor $\mathbf{w}^{(1)}$, $\mathbf{w}^{(2)}$ and $\mathbf{w}^{n+1}$. If one of the above three values is not in $\tilde{G}$, we will double the value of $\mu$ and halve the value of $\Delta t$, then restart the time integration at time level $n$. After we have reached the next time level, we reset the values of $\mu$ and $\Delta t$.

Based on the above theorem, we can construct physically relevant numerical cell averages $\bar{\mathbf{w}}$. However, the numerical approximations $\mathbf{w}$ may be out of the bounds. Hence, we need to apply suitable limiters to $\mathbf{w}^{(1)}$, $\mathbf{w}^{(2)}$ and $\mathbf{w}^{n+1}$, and construct physically relevant numerical approximations in each RK stage. The full algorithm on each fixed element $K$ is given below:

1. Set a small number $\epsilon = 10^{-13}$.
2. If $\bar{\rho} > \epsilon$, then we proceed to the next step. Otherwise, $\mathbf{w}$ is identified as the approximation to vacuum, then we take $\mathbf{w} = \bar{\mathbf{w}}$, and skip the following steps.
3. We modify the density $\rho$ first. Compute

$$\rho_{min} = \min_{(x,y)\in S_K} \rho(x, y).$$

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

$$\hat{\rho} = \overline{\rho} + \theta\left(\rho - \overline{\rho}\right), \qquad \hat{r}_i = \overline{r}_i + \theta\left(r_i - \overline{r}_i\right), \quad i = 1, \cdots, M-1,$$

   with

$$\theta = \frac{\overline{\rho} - \varepsilon}{\overline{\rho} - \rho_{min}}.$$

   Here we implicitly modify $\hat{r}_M = \overline{r}_M + \theta\left(r_M - \overline{r}_M\right)$ to keep $\sum_{i=1}^{M} \hat{r}_i = \hat{\rho}$. For simplicity, we can also take $\hat{r}_i = r_i$, $i = 1, \cdots, M-1$ and implicitly modify $\hat{r}_M = \hat{\rho} - \sum_{i=1}^{M-1} \hat{r}_i$.
4. Modify the mass fraction. For $1 \le i \le M$, define $\hat{S}_i = \{(x, y) \in S_K : \hat{r}_i(x, y) \le 0\}$. Take

$$\tilde{r}_i = \hat{r}_i + \theta\left(\frac{\bar{r}_i}{\bar{\rho}}\hat{\rho} - \hat{r}_i\right), 1 \le i \le M-1, \quad \theta = \max_{1\le i\le M} \max_{(x,y)\in \hat{S}_i} \left\{\frac{-\hat{r}_i(x,y)\bar{\rho}}{\bar{r}_i\hat{\rho}(x,y) - \hat{r}_i(x,y)\bar{\rho}}, 0\right\}. \tag{3.14}$$

5. Modify the pressure. Denote $\tilde{\mathbf{w}} = (\hat{\rho}, m, n, E, \tilde{r}_1, \cdots, \tilde{r}_{M-1})^T$. For each $\mathbf{x} \in S$, if $\tilde{\mathbf{w}}(\mathbf{x}) \in \tilde{G}$, then take $\theta_{\mathbf{x}} = 1$. Otherwise, take

$$\theta_{\mathbf{x}} = \frac{p(\overline{\mathbf{w}})}{p(\overline{\mathbf{w}}) - p(\tilde{\mathbf{w}}(\mathbf{x}))}.$$

   Then, we use

$$\mathbf{w}^{new} = \overline{\mathbf{w}} + \theta(\tilde{\mathbf{w}} - \overline{\mathbf{w}}), \quad \theta = \min_{\mathbf{x}\in S_K} \theta_{\mathbf{x}},$$

   as the new DG approximation. The proof for $p(\mathbf{w}^{new}) \ge 0$ can be found in [28].

## 4. Numerical examples

We use the third-order conservative Runge-Kutta method with the choice of optimal coefficients. Also, we expand the exponential terms by (2.22) as demonstrated in Remark 2.3.

**Example 4.1** *(Accuracy test for the ODE solver).* We first test the stability and accuracy of the ODE solver, and study the following problem:

$$u'(t) = -cu^7, \qquad u(0) = u_0,$$

where $c$ is a parameter that we can adjust. The problem becomes stiff as $c$ increases. The exact solution is

$$u(t) = u_0(6ctu_0^6 + 1)^{-1/6}.$$

We take the final time to be $t = 0.5$ and denote the total number of time steps as $N_t$.

**Table 4.1**
Accuracy test for the new RK method with $u_0 = 0.1$ and $c = 10000$.

| $N_t$ | $L^\infty$ norm | Order |
|---|---|---|
| 2 | 5.90E-11 | – |
| 4 | 7.36E-12 | 3.00 |
| 8 | 9.18E-13 | 3.00 |
| 16 | 1.15E-13 | 3.00 |
| 32 | 1.44E-14 | 2.99 |
| 64 | 1.78E-15 | 3.02 |

**Table 4.2**
Accuracy test for the new RK method with $u_0 = 1$.

| $N_t$ | $c = 1$ | | $c = 100$ | | $c = 10000$ | |
|---|---|---|---|---|---|---|
| | $L^\infty$ norm | Order | $L^\infty$ norm | Order | $L^\infty$ norm | Order |
| | Without expansion (2.22) | | | | | |
| 20 | 2.04E-06 | – | 2.67E-02 | – | 1.74E-01 | – |
| 40 | 2.51E-07 | 3.02 | 1.27E-03 | 4.39 | 1.68E-01 | 0.05 |
| 80 | 3.08E-08 | 3.03 | 1.00E-04 | 3.67 | 1.57E-01 | 0.10 |
| 160 | 3.85E-09 | 3.00 | 2.00E-05 | 2.32 | 1.36E-01 | 0.21 |
| 320 | 4.82E-10 | 3.00 | 3.32E-06 | 2.59 | 9.65E-02 | 0.49 |
| | With expansion (2.22) | | | | | |
| 20 | 2.04E-06 | – | 1.79E-02 | – | 8.20E-01 | – |
| 40 | 2.51E-07 | 3.02 | 1.25E-03 | 3.84 | 8.16E-01 | 0.01 |
| 80 | 3.08E-08 | 3.03 | 9.99E-05 | 3.65 | 3.91E-02 | 4.38 |
| 160 | 3.85E-09 | 3.00 | 2.00E-05 | 2.32 | 1.19E-03 | 5.04 |
| 320 | 4.82E-10 | 3.00 | 3.32E-06 | 2.59 | 2.18E-04 | 2.45 |

We first take $u_0 = 0.1$ with $c = 10000$. Numerical results for the 3rd order conservative RK method are listed in Table 4.1. The initial condition is well-prepared, and we can observe optimal convergence rates. Next, we take $u_0 = 1$, the results are given in Table 4.2. For this problem, the initial condition is not well-prepared, and we can observe optimal convergence rate if the problem is not stiff, e.g. $c = 1$. If the problem is stiff, e.g. $c = 10000$, with the exact exponential term, the method may not converge at the expected rate. However, by applying the Taylor expansion (2.22), we can observe better convergence rates.

**Example 4.2** *(Steady-state-preserving test for the ODE solver).* We consider the following scalar ODE:

$$u'(t) = 1 - k|u|u,$$

where $k$ is a positive real number. This problem has one equilibrium point $u^* = 1/\sqrt{k}$, and its exact solution is given by

$$u(t) = \begin{cases} \frac{1}{\sqrt{k}} coth(\sqrt{k}t + coth^{-1}(\sqrt{k}u(0))), & u(0) \geqslant u^*, \\ \frac{1}{\sqrt{k}} tan(\sqrt{k}t + tan^{-1}(\sqrt{k}u(0))), & u(0) < 0 \text{ and } t < -\frac{tan^{-1}(\sqrt{k}u(0))}{\sqrt{k}}, \\ \frac{1}{\sqrt{k}} tanh(\sqrt{k}t + tanh^{-1}(\sqrt{k}u(0))), & \text{otherwise.} \end{cases}$$

We take the final time as $t = 0.05$ and also denote the total number of time steps as $N_t$.

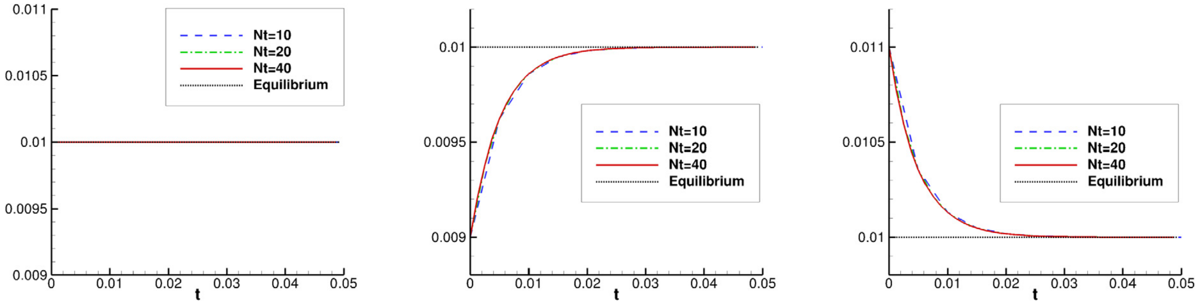
We take $k = 10,000$, which corresponds to the equilibrium point $u^* = 0.01$. We consider three different initial values,

$$(a)\ u(0) = u^*, \qquad (b)\ u(0) = 0.9u^*, \qquad (c)\ u(0) = 1.1u^*.$$

The numerical results computed with different $N_t$ are plotted in Fig. 4.1. As one can see, when $u$ is initially at the equilibrium (case (a)), our method preserves the steady state exactly as expected. In cases (b) and (c), our method also accurately captures and preserves the exact equilibrium as the time increases.

**Example 4.3** *(Accuracy test for 2D system).* From now on, we consider the two dimensional reactive Euler equations. In this example, we consider periodic boundary condition and take $u = v = 1$ and $p = 0$ in the exact solution. We choose $M = 2$ and the source is given as $s_1 = -cr_1^7$. Hence, we need to solve the following system

$$\begin{cases} \rho_t + \rho_x + \rho_y = 0, \\ (r_1)_t + (r_1)_x + (r_1)_y = -c(r_1)^7, \end{cases} \quad (x, y) \in [0, 2\pi]^2.$$



(a) $u(0) = u^*$ (b) $u(0) = 0.9u^*$ (c) $u(0) = 1.1u^*$

**Fig. 4.1.** Convergence toward the equilibrium.

**Table 4.3**
Accuracy test for the two dimensional problem.

| N | Without limiter | | | | With limiter | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $L^2$ norm | Order | $L^\infty$ norm | Order | $L^2$ norm | Order | $L^\infty$ norm | Order | Percentage |
| | $c = 100$, CFL $= 0.05$ | | | | | | | | |
| 10 | 6.41E-04 | – | 3.45E-03 | – | 1.34E-03 | – | 6.21E-03 | – | 19.38% |
| 20 | 8.07E-05 | 2.99 | 4.33E-04 | 2.99 | 8.50E-05 | 3.98 | 4.65E-04 | 3.74 | 5.36% |
| 40 | 1.01E-05 | 3.00 | 5.41E-05 | 3.00 | 1.02E-05 | 3.06 | 5.63E-05 | 3.05 | 2.38% |
| 80 | 1.26E-06 | 3.00 | 6.75E-06 | 3.00 | 1.27E-06 | 3.00 | 7.02E-06 | 3.00 | 0.85% |
| 160 | 1.58E-07 | 3.00 | 8.44E-07 | 3.00 | 1.62E-07 | 2.97 | 8.79E-07 | 3.00 | 0.30% |
| | $c = 10000$, CFL $= 0.06$ | | | | | | | | |
| 10 | 7.20E-04 | – | 4.15E-03 | – | 1.35E-03 | – | 6.76E-03 | – | 22.14% |
| 20 | 9.32E-05 | 2.95 | 6.07E-04 | 2.77 | 9.53E-05 | 3.82 | 6.11E-04 | 3.47 | 4.57% |
| 40 | 1.17E-05 | 2.99 | 7.67E-05 | 2.98 | 1.18E-05 | 3.01 | 7.67E-05 | 2.99 | 2.01% |
| 80 | 1.47E-06 | 3.00 | 9.63E-06 | 2.99 | 1.49E-06 | 2.99 | 9.63E-06 | 2.99 | 0.71% |
| 160 | 1.84E-07 | 3.00 | 1.20E-06 | 3.00 | 1.91E-07 | 2.97 | 1.20E-06 | 3.00 | 0.29% |

The initial conditions are given as $\rho(x, y, 0) = 0.1\big(2 + sin(x + y) + cos(x + y)\big)$ and $r_1(x, y, 0) = 0.1\big(1 + sin(x + y)\big)$, respectively. For this problem, the total density $\rho$ should be non-negative and the mass fraction $r_1/\rho$ should be between 0 and 1.

We use piecewise $P^2$ polynomials coupled with third-order Runge-Kutta discretization and take the final time to be $t = 0.5$. Numerical errors with different $c$ are given in the left column of Table 4.3. We can again observe the expected high order of accuracy of our scheme. We further add the limiter to preserve the lower bound of $\rho$ and the two bounds of $r_1/\rho$, and show the results in the right part of the error table. 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 4.4** *(A 2D detonation wave with 4 species and 1 reaction).* In this example, we test a 2D reacting model with four species and one reaction. A prototype reaction for this model is

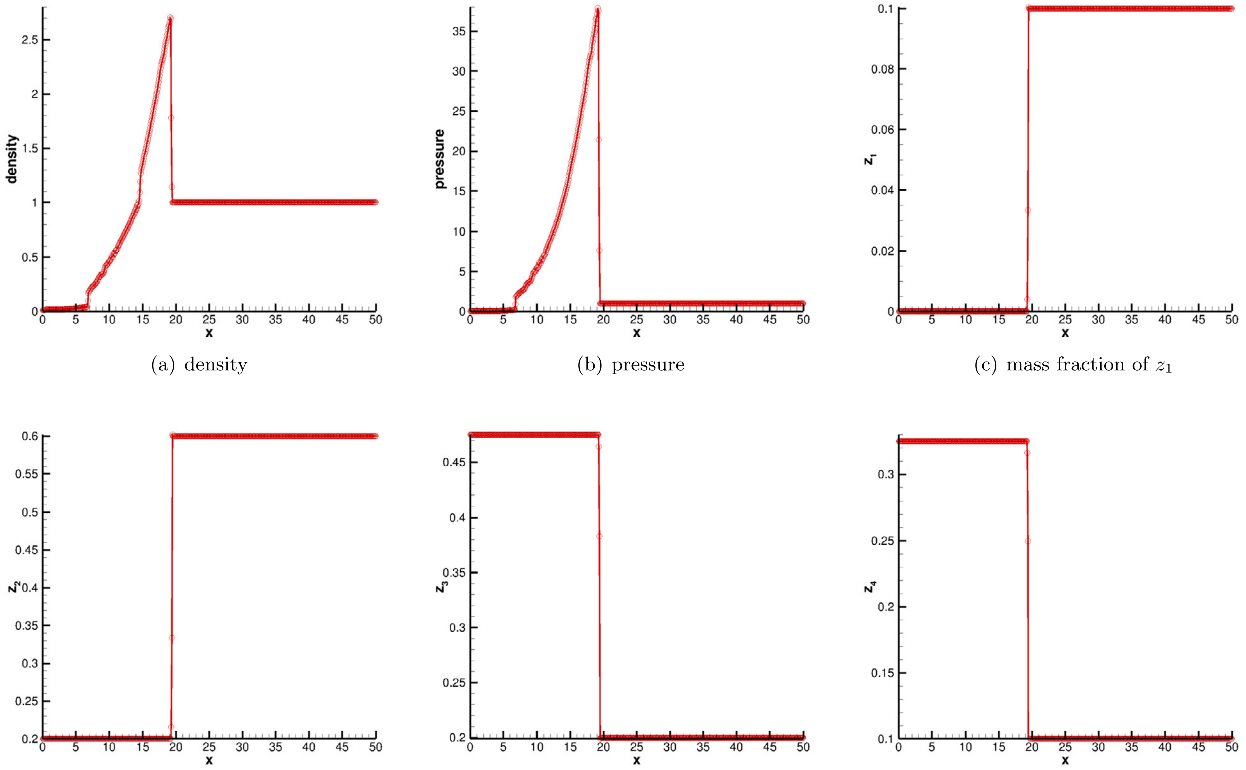
$$CH_4 + 2O_2 \to CO_2 + 2H_2O.$$

The parameters are $T_1 = 2$, $B_1 = 10^6$, $\alpha_1 = 0$, $q_1 = 200$, $q_2 = 0$, $q_3 = 0$, $q_4 = 0$, $M_1 = 16$, $M_2 = 32$, $M_3 = 44$, $M_4 = 18$. The initial values consist of totally burnt gas inside of a circle with radius 10 and totally unburnt gas everywhere outside this circle. The set up is as follows

$$(\rho, u, v, p, z_1, z_2, z_3, z_4)(x, y, 0) = \begin{cases} (2, 10x/r, 10y/r, 40, 0, 0.2, 0.475, 0.325), & r \leqslant 10, \\ (1, 0, 0, 1, 0.1, 0.6, 0.2, 0.1), & r > 10, \end{cases}$$

where $r = \sqrt{x^2 + y^2}$. The computational domain is $[0, 50] \times [0, 50]$. This is a radially symmetric problem and the detonation front is circular. The boundary conditions are solid-wall boundary conditions on the left and lower boundaries and outflow boundary conditions on the right and upper boundaries.

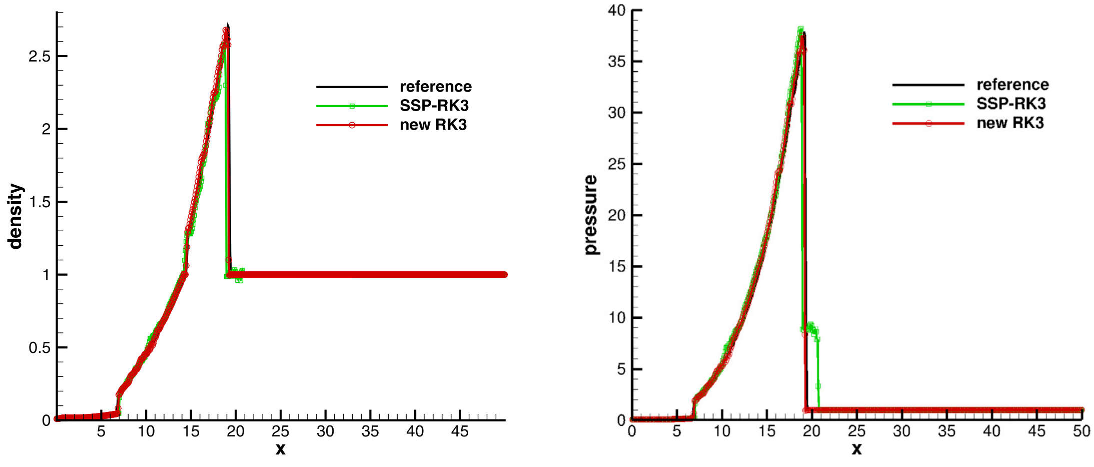
We test the 3rd order conservative Runge-Kutta method with piecewise $P^2$ polynomials. We first take $CFL = 0.1$ and refine the meshes to match the correct positions of the shocks with those given in [29]. Fig. 4.2 shows the one dimensional cuts of pressure, density and mass fractions along the line $x = y$ at $t = 2$ by taking $N_x = N_y = 600$. We can see that our



(a) density (b) pressure (c) mass fraction of $z_1$

(d) mass fraction of $z_2$ (e) mass fraction of $z_3$ (f) mass fraction of $z_4$

**Fig. 4.2.** Numerical solutions of Example 4.4 along the line $x = y$ at $t = 2$ with $N_x = N_y = 600$ and $CFL = 0.1$.

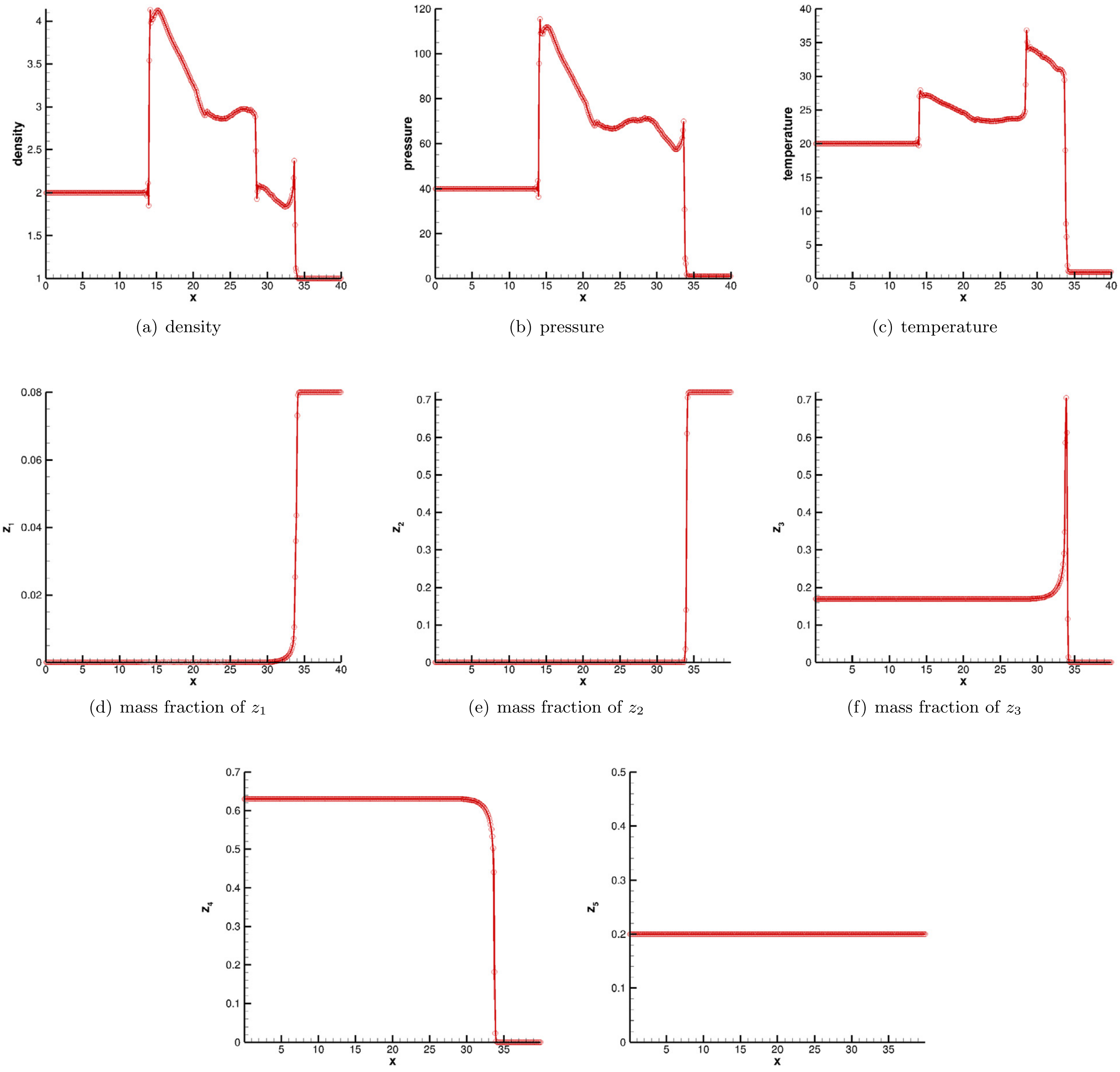


(a) density, $N_x = 600$, $CFL = 0.2$ (b) pressure, $N_x = 600$, $CFL = 0.2$

**Fig. 4.3.** Example 4.4. Comparison of different RK methods. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

scheme captures the detonations well and preserves the positivity of the density and pressure, and the two bounds 0 and 1 of each mass fraction.

Next, we treat the above numerical solutions as reference solutions and compare our method with the traditional SSP-RK3 method [26]. To avoid the technique of subcell resolution [28], we also take $N_x = N_y = 600$. As we can see in Figs. 4.3(a) and 4.3(b), $CFL = 0.2$ is already enough for our scheme to obtain the correct solutions. However, the traditional SSP-RK3 method will blow up when $CFL = 0.2$. In this case, if we restart the time integration by halving the time step each time when we detect solutions outside the physical bounds, we can obtain the green curves in Figs. 4.3(a) and 4.3(b). We observe that the majority time steps will be restarted and the computational cost increases. However, there are still some spurious



(a) density

(b) pressure

(c) temperature

(d) mass fraction of $z_1$

(e) mass fraction of $z_2$

(f) mass fraction of $z_3$

(g) mass fraction of $z_4$

(h) mass fraction of $z_5$

**Fig. 4.4.** Numerical solutions of Example 4.5 at $t = 2$.

waves and an even denser mesh is needed. Hence, the computational cost is bigger than our method in order to reach correct shock locations.

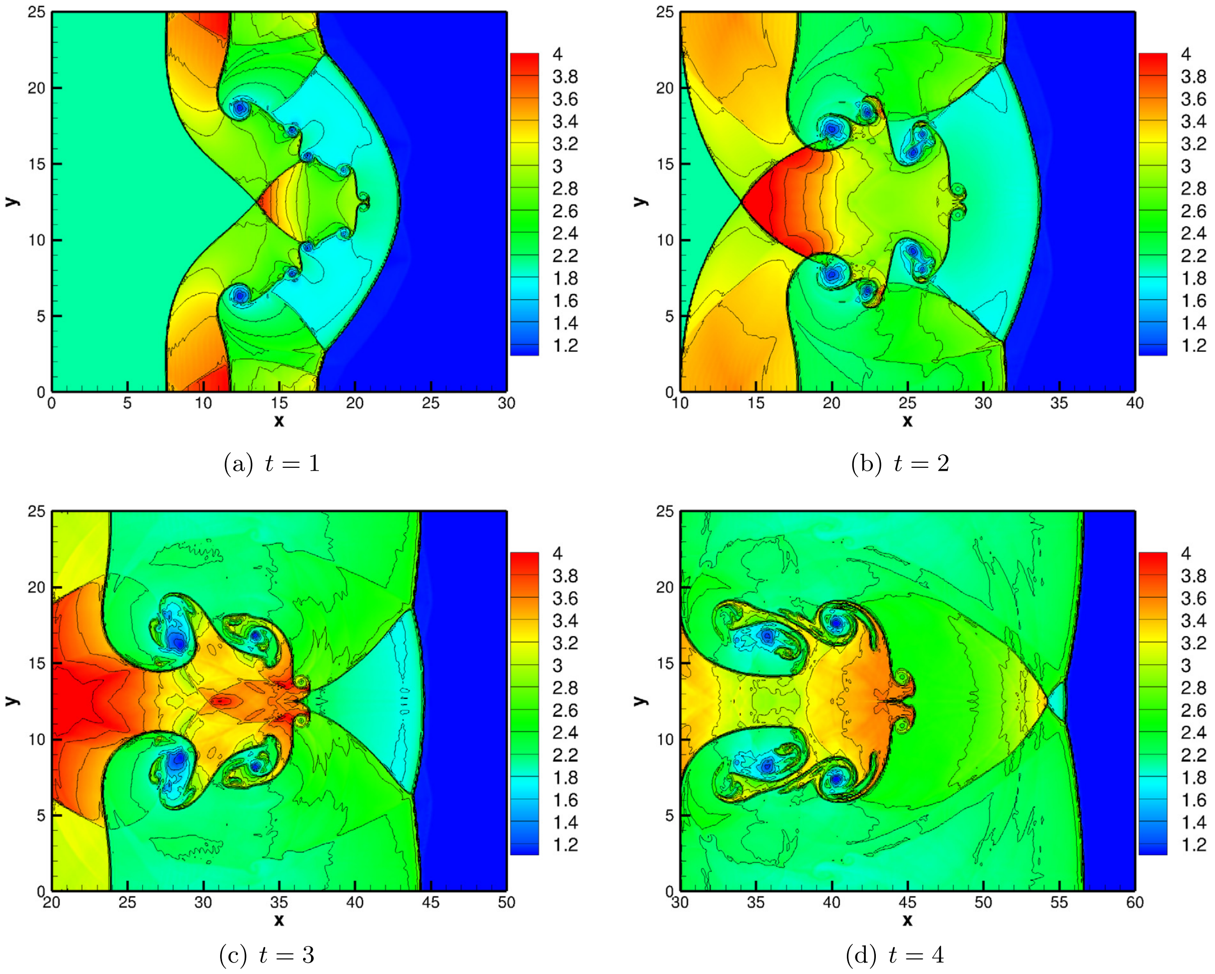
**Example 4.5** *(A 2D detonation wave with 5 species and 2 reactions).* The third 2D example is the 2D reacting model with 5 species and 2 reactions. Consider

$$H_2 + O_2 \to 2OH, \; 2OH + H_2 \to 2H_2O,$$

where $N_2$ appearing as a catalyst. The parameters are $T_1 = 2$, $T_2 = 10$, $B_1 = B_2 = 10^6$, $\alpha_1 = \alpha_2 = 0$, $q_1 = 0$, $q_2 = 0$, $q_3 = -20$, $q_4 = -100$, $q_5 = 0$, $M_1 = 2$, $M_2 = 32$, $M_3 = 17$, $M_4 = 18$, $M_5 = 28$. The initial values are given by

$$(\rho, u, v, p, z_1, z_2, z_3, z_4, z_5)(x, y, 0) = \begin{cases} (2, 10, 0, 40, 0, 0, 0.17, 0.63, 0.2), & x \leqslant \xi(y), \\ (1, 0, 0, 1, 0.08, 0.72, 0, 0, 0.2), & x > \xi(y), \end{cases}$$

where



(a) $t = 1$

(b) $t = 2$

(c) $t = 3$

(d) $t = 4$

**Fig. 4.5.** Density plots of Example 4.5 at different times.

$$\xi(y) = \begin{cases} 12.5 - |y - 12.5|, & |y - 12.5| \leqslant 7.5, \\ 5, & |y - 12.5| > 7.5. \end{cases}$$

The computational domain is $[0, 100] \times [0, 25]$. The inflow boundary conditions are used on the left boundary and the outflow boundary conditions are used on the right boundary. The top and bottom boundaries are solid walls. One important feature of this solution is the appearance of triple points, which travel along the detonation front in the transverse direction and reflect from the upper and lower walls, forming a cellular pattern. Behind the detonation front, there is a strong shock.

We take $N_x = 1000$, $N_y = 251$ and $CFL = 0.1$. We first show the solutions at the 1D cross section $y = 12.5$ at $t = 2$ in Fig. 4.4. Since at $t = 2$ the flow has not touched $x = 40$, the results are computed on the cutoff computational domain $[0, 40] \times [0, 25]$. We can see that there are some oscillations, but the main purpose of our work is not to control oscillations. It is easy to see from the pressure, temperature and mass fraction results that there are no spurious waves and our scheme preserves the bounds. The density contours at different times are shown in Fig. 4.5.

## 5. Conclusion

In this paper, we have introduced the high-order conservative bound-preserving DG methods for stiff multispecies detonation. A new explicit Runge-Kutta time integration has been constructed. Numerical experiments demonstrated the good performance of the scheme.

## References

[1] U.M. Ascher, S.J. Ruuth, R.J. Spiteri, Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations, Appl. Numer. Math. 25 (1997) 151–167.

[2] V. Casulli, Semi-implicit finite difference methods for the two-dimensional shallow water equations, J. Comput. Phys. 86 (1990) 56–74.

[3] L. Cea, M.E. Vázquez-Cendón, Unstructured finite volume discretisation of bed friction and convective flux in solute transport models linked to the shallow water equations, J. Comput. Phys. 231 (2012) 3317–3339.

[4] A. Chertock, S. Cui, A. Kurganov, T. Wu, Steady state and sign preserving semi-implicit Runge-Kutta methods for ODEs with stiff damping term, SIAM J. Numer. Anal. 53 (2015) 2008–2029.

[5] 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.

[6] J.F. Clarke, S. Karni, J.J. Quirk, P.L. Roe, L.G. Simmonds, E.F. Toro, Numerical computation of two-dimensional unsteady detonation waves in high energy solids, J. Comput. Phys. 106 (1993) 215–233.

[7] 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.

[8] S. Gottlieb, C.-W. Shu, E. Tadmor, Strong stability-preserving high-order time discretization methods, SIAM Rev. 43 (2001) 89–112.

[9] 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.

[10] I. Higueras, N. Happenhofer, O. Koch, F. Kupka, Optimized strong stability preserving IMEX Runge-Kutta methods, J. Comput. Appl. Math. 272 (2014) 116–140.

[11] W. Hundsdorfer, S.J. Ruuth, IMEX extensions of linear multistep methods with general monotonicity and boundedness properties, J. Comput. Phys. 225 (2007) 2016–2042.

[12] 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.

[13] 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.

[14] 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.

[15] L. Isherwood, z. Grant, S. Gottlieb, Strong stability preserving integrating factor Runge Kutta methods, SIAM J. Numer. Anal. 56 (2018) 3276–3307.

[16] S. Kopecz, A. Meister, On order conditions for modified Patankar-Runge-Kutta schemes, Appl. Numer. Math. 123 (2018) 159–179.

[17] S. Kopecz, A. Meister, Unconditionally positive and conservative third order modified Patankar-Runge-Kutta discretizations of production-destruction systems, BIT Numer. Math. 58 (2018) 691–728.

[18] R.J. LeVeque, H.C. Yee, A study of numerical methods for hyperbolic conservation laws with stiff source terms, J. Comput. Phys. 86 (1990) 187–210.

[19] Y. Lv, M. Ihme, Discontinuous Galerkin method for multicomponent chemically reacting flows and combustion, J. Comput. Phys. 270 (2014) 105–137.

[20] Y. Lv, M. Ihme, High-order discontinuous Galerkin method for applications to multicomponent and chemically reacting flows, Acta Mech. Sin. 33 (2017) 486–499.

[21] L. Pareschi, G. Russo, Implicit-explicit Runge-Kutta schemes and applications to hyperbolic systems with relaxation, J. Sci. Comput. 25 (2005) 129–155.

[22] T. Qin, C.-W. Shu, Implicit positivity-preserving high order discontinuous Galerkin methods for conservation laws, SIAM J. Sci. Comput. 40 (2018) A81–A107.

[23] T. Qin, C.-W. Shu, Y. Yang, Bound-preserving discontinuous Galerkin methods for relativistic hydrodynamics, J. Comput. Phys. 315 (2016) 323–347.

[24] W.H. Reed, T.R. Hill, Triangular Mesh Methods for the Neutron Transport Equation, Report LA-UR-73-479, Los Alamos Scientific Laboratory, Los Alamos, NM, 1973.

[25] C.-W. Shu, Total-variation-diminishing time discretizations, SIAM J. Sci. Stat. Comput. 9 (1988) 1073–1084.

[26] C.-W. Shu, S. Osher, Efficient implementation of essentially non-oscillatory shock-capturing schemes, J. Comput. Phys. 77 (1988) 439–471.

[27] G. Strang, On the construction and comparison of difference schemes, SIAM J. Numer. Anal. 5 (1968) 506–517.

[28] 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.

[29] W. Wang, C.-W. Shu, H.C. Yee, D.V. Kotov, B. Sjögreen, High order finite difference methods with subcell resolution for stiff multispecies detonation capturing, Commun. Comput. Phys. 17 (2015) 317–336.

[30] 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.

[31] X. Zhang, C.-W. Shu, On maximum-principle-satisfying high order schemes for scalar conservation laws, J. Comput. Phys. 229 (2010) 3091–3120.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] X. Zhong, Additive semi-implicit Runge-Kutta methods for computing high-speed nonequilibrium reactive flows, J. Comput. Phys. 128 (1996) 19–31.