On positivity preserving high order discontinuous Galerkin schemes for compressible Euler equations on rectangular meshes[1]

Xiangxiong Zhang[2] and Chi-Wang Shu[3]

Abstract

We construct uniformly high order accurate discontinuous Galerkin (DG) schemes which preserve positivity of density and pressure for Euler equations of compressible gas dynamics. The same framework also applies to high order accurate finite volume (e.g. essentially non-oscillatory (ENO) or weighted ENO (WENO)) schemes. Motivated by [20, 26], a general framework, for arbitrary order of accuracy, is established to construct a positivity preserving limiter for the finite volume and DG methods with first order Euler forward time discretization solving one dimensional compressible Euler equations. The limiter can be proven to maintain high order accuracy and is easy to implement. Strong stability preserving (SSP) high order time discretizations will keep the positivity property. Following the idea in [26], we extend this framework to higher dimensions on rectangular meshes in a straightforward way. Numerical tests for the third order DG method are reported to demonstrate the effectiveness of the methods.

AMS subject classification: 65M60, 76N15

Keywords: hyperbolic conservation laws; discontinuous Galerkin method; positivity preserving; high order accuracy; compressible Euler equations; gas dynamics; finite volume scheme; essentially non-oscillatory scheme; weighted essentially non-oscillatory scheme

[1]Research supported by AFOSR grant FA9550-09-1-0126 and NSF grant DMS-0809086.

[2]Department of Mathematics, Brown University, Providence, RI 02912. E-mail: zhangxx@dam.brown.edu

[3]Division of Applied Mathematics, Brown University, Providence, RI 02912. E-mail: shu@dam.brown.edu

1 Introduction

In this paper we are interested in constructing high order accurate schemes for solving hyperbolic conservation law systems. For scalar conservation laws, the entropy solution is total variation diminishing (TVD), which is a desired property for numerical solutions. While traditional TVD schemes (e.g. [10]) measure the total variation by that of the cell averages or grid values, leading to a necessary degeneracy of accuracy to first order at smooth extrema [18], genuinely high order TVD schemes can be constructed for one dimensional scalar conservation laws [21, 27] by measuring the total variation of the reconstruction polynomials. For multi-dimensional scalar conservation laws, it is difficult to enforce the TVD property for a high order scheme, however it is reasonable to insist on a strict maximum principle, which is satisfied by the entropy solution. Genuinely high order accurate finite volume and discontinuous Galerkin (DG) schemes which satisfy a strict maximum principle have been constructed recently in [26].

For hyperbolic conservation law systems, the entropy solutions in general satisfy neither the TVD property nor the maximum principle. In this paper we are mainly interested in the Euler equations for the perfect gas, the one dimensional version being given by

$$\mathbf{w}_t + \mathbf{f}(\mathbf{w})_x = 0, \quad t \geq 0, \ x \in \mathbb{R}, \tag{1.1}$$

$$\mathbf{w} = \begin{pmatrix} \rho \\ m \\ E \end{pmatrix}, \quad \mathbf{f}(\mathbf{w}) = \begin{pmatrix} m \\ \rho u^2 + p \\ (E+p)u \end{pmatrix} \tag{1.2}$$

with

$$m = \rho u, \quad E = \frac{1}{2}\rho u^2 + \rho e, \quad p = (\gamma - 1)\rho e,$$

where ρ is the density, u is the velocity, m is the momentum, E is the total energy, p is the pressure, e is the internal energy, and $\gamma > 1$ is a constant ($\gamma = 1.4$ for the air). The speed of sound is given by $c = \sqrt{\gamma p/\rho}$ and the three eigenvalues of the Jacobian $\mathbf{f}'(\mathbf{w})$ are $u - c$, u and $u + c$. Physically, the density ρ and the pressure p should both be positive. We are interested in positivity-preserving high order schemes, which maintain the positivity of

density and pressure at time level $n+1$, provided that they are positive at time level n. The techniques developed in this paper can be considered as generalizations of the maximum-principle-satisfying limiters in [26] and the positivity-preserving schemes in [20]. We remark that failure of preserving positivity of density or pressure may cause blow-ups of the numerical algorithm, for example, for low density problems in computing blast waves, and low pressure problems in the computing high Mach number astrophysical jets [9]. We also remark that most commonly used high order numerical schemes for solving hyperbolic conservation law systems, including, among others, the Runge-Kutta discontinuous Galerkin (RKDG) method with a total variation bounded (TVB) limiter [2, 4], the essentially non-oscillatory (ENO) finite volume and finite difference schemes [11, 25], and the weighted ENO (WENO) finite volume and finite difference schemes [17, 13], do not in general satisfy the positivity property for Euler equations automatically.

We now consider the Euler equations (1.1) in more detail. Let $p(\mathbf{w}) = (\gamma-1)(E-\frac{1}{2}\frac{m^2}{\rho})$ be the pressure function. It can be easily verified that p is a concave function of $\mathbf{w} = (\rho, m, E)^T$ if $\rho \geq 0$. For $\mathbf{w}_1 = (\rho_1, m_1, E_1)^T$ and $\mathbf{w}_2 = (\rho_2, m_2, E_2)^T$, Jensen's inequality implies, for $0 \leq s \leq 1$,

$$p\left(s\mathbf{w}_1 + (1-s)\mathbf{w}_2\right) \geq sp\left(\mathbf{w}_1\right) + (1-s)p\left(\mathbf{w}_2\right), \quad \text{if} \quad \rho_1 \geq 0, \quad \rho_2 \geq 0. \tag{1.3}$$

Define the set of admissible states by

$$G = \left\{ \mathbf{w} = \begin{pmatrix} \rho \\ m \\ E \end{pmatrix} \middle| \, \rho > 0 \quad \text{and} \quad p = (\gamma-1)\left(E - \frac{1}{2}\frac{m^2}{\rho}\right) > 0 \right\},$$

then G is a convex set. If the density or pressure becomes negative, the system (1.1) will be non-hyperbolic and thus the initial value problem will be ill-posed.

We are interested in schemes for (1.1) producing the numerical solutions in the admissible set G. We start with a first order finite volume scheme

$$\mathbf{w}_j^{n+1} = \mathbf{w}_j^n - \lambda[\mathbf{h}(\mathbf{w}_j^n, \mathbf{w}_{j+1}^n) - \mathbf{h}(\mathbf{w}_{j-1}^n, \mathbf{w}_j^n)], \tag{1.4}$$

where $\mathbf{h}(\cdot,\cdot)$ is a numerical flux, n refers to the time step and j to the spatial cell (we assume uniform mesh size only for simplicity), and $\lambda = \frac{\Delta t}{\Delta x}$ is the ratio of time and space mesh sizes. $\mathbf{w}_j^n$ is the approximation to the cell average of the exact solution $\mathbf{v}(x,t)$ in the cell $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$, or the point value of the exact solution $\mathbf{v}(x,t)$ at x_j, at time level n. The scheme (1.4) and its numerical flux $\mathbf{h}(\cdot,\cdot)$ are called positivity preserving, if the numerical solution $\mathbf{w}_j^n$ being in the set G for all j implies the solution $\mathbf{w}_j^{n+1}$ being also in the set G. This is usually achieved under a standard CFL condition

$$\lambda \parallel (|u| + c) \parallel_\infty \leq \alpha_0. \tag{1.5}$$

Examples of positivity preserving fluxes include the Godunov flux [6], the Lax-Friedrichs flux [20], the Boltzmann type flux [19], and the Harten-Lax-van Leer flux [12].

We now consider a general high order finite volume scheme, or the scheme satisfied by the cell averages of a DG method solving (1.1), which has the following form

$$\overline{\mathbf{w}}_j^{n+1} = \overline{\mathbf{w}}_j^n - \lambda \left[\mathbf{h}\left(\mathbf{w}_{j+\frac{1}{2}}^-, \mathbf{w}_{j+\frac{1}{2}}^+ \right) - \mathbf{h}\left(\mathbf{w}_{j-\frac{1}{2}}^-, \mathbf{w}_{j-\frac{1}{2}}^+ \right) \right], \tag{1.6}$$

where $\mathbf{h}$ is a positivity preserving flux under the CFL condition (1.5), $\overline{\mathbf{w}}_j^n$ is the approximation to the cell average of the exact solution $\mathbf{v}(x,t)$ in the cell $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$ at time level n, and $\mathbf{w}_{j+\frac{1}{2}}^-$, $\mathbf{w}_{j+\frac{1}{2}}^+$ are the high order approximations of the point values $\mathbf{v}(x_{j+\frac{1}{2}}, t^n)$ within the cells I_j and I_{j+1} respectively. These values are either reconstructed from the cell averages $\overline{\mathbf{w}}_j^n$ in a finite volume method or read directly from the evolved polynomials in a DG method. We assume that there is a polynomial vector $\mathbf{q}_j(x) = (\rho_j(x), m_j(x), E_j(x))^T$ (either reconstructed in a finite volume method or evolved in a DG method) with degree k, where $k \geq 1$, defined on I_j such that $\overline{\mathbf{w}}_j^n$ is the cell average of $\mathbf{q}_j(x)$ on I_j, $\mathbf{w}_{j-\frac{1}{2}}^+ = \mathbf{q}_j(x_{j-\frac{1}{2}})$ and $\mathbf{w}_{j+\frac{1}{2}}^- = \mathbf{q}_j(x_{j+\frac{1}{2}})$.

A general framework to construct a high order positivity preserving finite volume scheme for the Euler equations was introduced in [20], in which a sufficient condition for the solution $\overline{\mathbf{w}}_j^{n+1}$ of (1.6) to be in the set G is that, all the nodal values $\mathbf{w}_{j+\frac{1}{2}}^{\pm}$ and $\overline{\mathbf{w}}_j^{n+1} - a(\mathbf{w}_{j+\frac{1}{2}}^- + \mathbf{w}_{j-\frac{1}{2}}^+)$ are in the set G under the CFL condition

$$\lambda \parallel (|u| + c) \parallel_\infty \leq a\alpha_0,$$

where $a \in (0, 1]$ is a constant. Strong stability preserving (SSP) high order Runge-Kutta [25] and multi-step [24] time discretization will keep the positivity since G is convex.

It is reasonable to require and easy to enforce the positivity of the point values $\mathbf{w}^{\pm}_{j+\frac{1}{2}}$. However, it is more difficult to enforce the positivity of $\overline{\mathbf{w}}^{n+1}_j - a(\mathbf{w}^{-}_{j+\frac{1}{2}} + \mathbf{w}^{+}_{j-\frac{1}{2}})$ without destroying accuracy for an arbitrary high order scheme. We refer to [20] for more discussions on this point. In this paper, we provide a similar sufficient condition, which is however much easier to enforce. We need the N-point Legendre Gauss-Lobatto quadrature rule on the interval $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$, which is exact for the integral of polynomials of degree up to $2N-3$. We would need to choose N such that $2N-3 \geq k$. Denote these quadrature points on I_j as

$$S_j = \{x_{j-\frac{1}{2}} = \widehat{x}^1_j, \widehat{x}^2_j, \cdots, \widehat{x}^{N-1}_j, \widehat{x}^N_j = x_{j+\frac{1}{2}}\}. \tag{1.7}$$

We will prove that a sufficient condition for $\overline{\mathbf{w}}^{n+1}_j \in G$ is simply $\mathbf{q}_j(\widehat{x}^\alpha_j) \in G$ for $\alpha = 1, 2, \cdots, N$, under a suitable CFL condition. The same type of the linear scaling limiter used in [26] can enforce this sufficient condition without destroying accuracy. This limiter is also very easy to implement. Furthermore, we provide a straightforward extension of this result to arbitrary high order two-dimensional schemes on rectangular meshes.

The main conclusion of this paper is, by adding a positivity preserving limiter which will be specified later to a high order accurate finite volume scheme or a discontinuous Galerkin scheme solving one or multi-dimensional Euler equations, with the time evolution by a SSP Runge-Kutta or multi-step method, we obtain a uniformly high order accurate scheme preserving the positivity in the sense that the density and pressure of the cell averages are always positive if they are positive initially.

The paper is organized as follows: we first prove the positivity result for schemes in one space dimension in Section 2. In Section 3, we show a straightforward extension to two space dimensions on rectangular meshes. In section 4, numerical tests for the third order DG method will be shown. Concluding remarks are given in Section 5.

2 Positivity preserving high order schemes in one dimension

2.1 A sufficient condition

We consider the first order Euler forward time discretization (1.6); higher order time discretization will be discussed later. Let $\widehat{w}_\alpha$ be the Legendre Gauss-Lobatto quadrature weights for the interval $[-\frac{1}{2}, \frac{1}{2}]$ such that $\sum_{\alpha=1}^{N} \widehat{w}_\alpha = 1$, with $2N-3 \geq k$. Motivated by the approach in [20, 26], our first result is

Theorem 2.1. For a finite volume scheme or the scheme satisfied by the cell averages of a DG method (1.6), if $\mathbf{q}_j(\widehat{x}_j^\alpha) \in G$ for all j and α, then $\overline{\mathbf{w}}_j^{n+1} \in G$ under the CFL condition

$$\lambda \parallel (|u| + c) \parallel_\infty \leq \widehat{w}_1 \alpha_0. \tag{2.1}$$

Proof: The exactness of the quadrature rule for polynomials of degree k implies

$$\overline{\mathbf{w}}_j^n = \frac{1}{\Delta x} \int_{I_j} \mathbf{q}_j(x) dx = \sum_{\alpha=1}^{N} \widehat{w}_\alpha \mathbf{q}_j(\widehat{x}_j^\alpha).$$

By adding and subtracting $\mathbf{h}\left(\mathbf{w}^+_{j-\frac{1}{2}}, \mathbf{w}^-_{j+\frac{1}{2}}\right)$, the scheme (1.6) becomes

$$\begin{aligned}
\overline{\mathbf{w}}_j^{n+1} &= \sum_{\alpha=1}^{N} \widehat{w}_\alpha \mathbf{q}_j(\widehat{x}_j^\alpha) - \lambda \left[\mathbf{h}\left(\mathbf{w}^-_{j+\frac{1}{2}}, \mathbf{w}^+_{j+\frac{1}{2}}\right) - \mathbf{h}\left(\mathbf{w}^+_{j-\frac{1}{2}}, \mathbf{w}^-_{j+\frac{1}{2}}\right)\right. \\
&\quad \left. + \mathbf{h}\left(\mathbf{w}^+_{j-\frac{1}{2}}, \mathbf{w}^-_{j+\frac{1}{2}}\right) - \mathbf{h}\left(\mathbf{w}^-_{j-\frac{1}{2}}, \mathbf{w}^+_{j-\frac{1}{2}}\right)\right] \\
&= \sum_{\alpha=2}^{N-1} \widehat{w}_\alpha \mathbf{q}_j(\widehat{x}_j^\alpha) + \widehat{w}_N \left(\mathbf{w}^-_{j+\frac{1}{2}} - \frac{\lambda}{\widehat{w}_N}\left[\mathbf{h}\left(\mathbf{w}^-_{j+\frac{1}{2}}, \mathbf{w}^+_{j+\frac{1}{2}}\right) - \mathbf{h}\left(\mathbf{w}^+_{j-\frac{1}{2}}, \mathbf{w}^-_{j+\frac{1}{2}}\right)\right]\right) \\
&\quad + \widehat{w}_1 \left(\mathbf{w}^+_{j-\frac{1}{2}} - \frac{\lambda}{\widehat{w}_1}\left[\mathbf{h}\left(\mathbf{w}^+_{j-\frac{1}{2}}, \mathbf{w}^-_{j+\frac{1}{2}}\right) - \mathbf{h}\left(\mathbf{w}^-_{j-\frac{1}{2}}, \mathbf{w}^+_{j-\frac{1}{2}}\right)\right]\right) \\
&= \sum_{\alpha=2}^{N-1} \widehat{w}_\alpha \mathbf{q}_j(\widehat{x}_j^\alpha) + \widehat{w}_N \mathbf{H}_N + \widehat{w}_1 \mathbf{H}_1,
\end{aligned}$$

where

$$\mathbf{H}_1 = \mathbf{w}^+_{j-\frac{1}{2}} - \frac{\lambda}{\widehat{w}_1}\left[\mathbf{h}\left(\mathbf{w}^+_{j-\frac{1}{2}}, \mathbf{w}^-_{j+\frac{1}{2}}\right) - \mathbf{h}\left(\mathbf{w}^-_{j-\frac{1}{2}}, \mathbf{w}^+_{j-\frac{1}{2}}\right)\right] \tag{2.2}$$

$$\mathbf{H}_N = \mathbf{w}^-_{j+\frac{1}{2}} - \frac{\lambda}{\widehat{w}_N}\left[\mathbf{h}\left(\mathbf{w}^-_{j+\frac{1}{2}}, \mathbf{w}^+_{j+\frac{1}{2}}\right) - \mathbf{h}\left(\mathbf{w}^+_{j-\frac{1}{2}}, \mathbf{w}^-_{j+\frac{1}{2}}\right)\right]. \tag{2.3}$$

Notice that (2.2) and (2.3) are both of the type (1.4), and $\widehat{w}_1 = \widehat{w}_N$, therefore $\mathbf{H_1}$ and $\mathbf{H_2}$ are in the set G under the CFL condition (2.1). Now, it is easy to conclude that $\overline{\mathbf{w}}_j^{n+1}$ is in G, since it is a convex combination of elements in G. ■

Remark 2.2. Here we only discuss Euler forward. Strong stability preserving high order Runge-Kutta [25] and multi-step [24] time discretization will keep the validity of Theorem 2.1 since G is convex.

Remark 2.3. From the proof of Theorem 2.1, we can see that any type of quadrature rule will work as long as the quadrature points include the two cell ends and the quadrature is exact for polynomials of degree k. It would appear that there is a possibility to achieve a larger CFL number if we can find a better quadrature in the sense that $\widehat{w}_1$ is larger. However, for $k = 2, 3$, we have verified that the Gauss-Lobatto quadrature is the best choice.

Remark 2.4. For the Lax-Friedrichs flux

$$\mathbf{h}(\mathbf{u}, \mathbf{v}) = \frac{1}{2}\left[\mathbf{f}(\mathbf{u}) + \mathbf{f}(\mathbf{v}) - a_0(\mathbf{v} - \mathbf{u})\right],$$

where $a_0 = \| \left(|u| + c\right) \|_\infty$, the CFL condition (1.5) was proven for $\lambda a_0 \le \frac{1}{2}$ in [20] by proving the numerical solution of the first order Lax-Friedrichs scheme to be the cell average of the exact solution. Here, we prove that (1.5) for the Lax-Friedrichs flux can be relaxed to $\lambda a_0 \le 1$. The Lax-Friedrichs scheme can be written as

$$\begin{aligned}
\mathbf{w}_j^{n+1} &= \mathbf{w}_j^n - \lambda[\mathbf{h}(\mathbf{w}_j^n, \mathbf{w}_{j+1}^n) - \mathbf{h}(\mathbf{w}_{j-1}^n, \mathbf{w}_j^n)] \\
&= (1 - \lambda a_0)\mathbf{w}_j^n + \frac{\lambda a_0}{2}[\mathbf{w}_{j+1}^n - \frac{1}{a_0}\mathbf{f}(\mathbf{w}_{j+1}^n)] + \frac{\lambda a_0}{2}[\mathbf{w}_{j-1}^n + \frac{1}{a_0}\mathbf{f}(\mathbf{w}_{j-1}^n)]
\end{aligned}$$

Assume $\mathbf{w}_j^n$, $\mathbf{w}_{j-1}^n$ and $\mathbf{w}_{j+1}^n$ are in the set G, we want to show $\mathbf{w}_j^{n+1} \in G$ under the CFL $\lambda a_0 \le 1$. Notice that $\mathbf{w}_j^{n+1}$ is a convex combination of the three vectors $\mathbf{w}_j^n$, $\mathbf{w}_{j+1}^n - \frac{1}{a_0}\mathbf{f}(\mathbf{w}_{j+1}^n)$ and $\mathbf{w}_{j-1}^n + \frac{1}{a_0}\mathbf{f}(\mathbf{w}_{j-1}^n)$, we only need to show $\mathbf{w}_{j-1}^n + \frac{1}{a_0}\mathbf{f}(\mathbf{w}_{j-1}^n)$ and $\mathbf{w}_{j+1}^n - \frac{1}{a_0}\mathbf{f}(\mathbf{w}_{j+1}^n)$ are in

the set G. It is easy to check that the first components of the both vectors are positive. The only nontrivial part is to check the positivity of the "pressure". For simplicity, we drop the subscripts and superscripts, i.e., we prove $\mathbf{w} \pm \frac{1}{a_0}\mathbf{f}(\mathbf{w}) \in G$ if $\mathbf{w} \in G$. Let $p = \frac{1}{\gamma-1}(E - \frac{1}{2}\frac{m^2}{\rho})$ and $u = m/\rho$. By a direct calculation, we have

$$\begin{aligned} p\left(\mathbf{w} \pm \frac{1}{a_0}\mathbf{f}(\mathbf{w})\right) &= p\left[\left((1 \pm \frac{u}{a_0})\rho, (1 \pm \frac{u}{a_0})m - \frac{p}{a_0}, (1 \pm \frac{u}{a_0})E - \frac{p}{a_0}\right)^T\right] \\ &= \left(1 - \frac{p}{\rho}\frac{\gamma-1}{2(a_0 \pm u)^2}\right)\left(1 \pm \frac{u}{a_0}\right)p \end{aligned}$$

Therefore,

$$\begin{aligned} p\left(\mathbf{w} \pm \frac{1}{a_0}\mathbf{f}(\mathbf{w})\right) > 0 &\iff \frac{p}{\rho}\frac{\gamma-1}{2(a_0 \pm u)^2} < 1 \\ &\iff \gamma\frac{p}{\rho} < \frac{2\gamma}{\gamma-1}(a_0 \pm u)^2 \\ &\iff \sqrt{\gamma\frac{p}{\rho}} < \sqrt{\frac{2\gamma}{\gamma-1}}(a_0 \pm u) \end{aligned}$$

Since $c = \sqrt{\gamma p/\rho}$ and $a_0 = \| (|u| + c) \|_\infty$, we have $p(\mathbf{w} \pm \frac{1}{a_0}\mathbf{f}(\mathbf{w})) > 0$.

The CFL condition (2.1) using the Lax-Friedrichs flux and the Gauss-Lobatto quadrature points for $k = 2, 3, 4, 5$ are listed in Table 2.1. We note that these conditions are comparable with and only slightly more restrictive than the standard CFL conditions for linear stability of discontinuous Galerkin methods [5].

Table 2.1: The CFL condition (2.1) of Lax-Friedrichs flux for $2 \le k \le 5$ and the Gauss-Lobatto quadrature points on $[-\frac{1}{2}, \frac{1}{2}]$.

k	CFL	quadrature points on $[-\frac{1}{2}, \frac{1}{2}]$
2	$\lambda a_0 \le \frac{1}{6}$	$\{-\frac{1}{2}, 0, \frac{1}{2}\}$
3	$\lambda a_0 \le \frac{1}{6}$	$\{-\frac{1}{2}, 0, \frac{1}{2}\}$
4	$\lambda a_0 \le \frac{1}{12}$	$\{-\frac{1}{2}, -\frac{1}{\sqrt{20}}, \frac{1}{\sqrt{20}}, \frac{1}{2}\}$
5	$\lambda a_0 \le \frac{1}{12}$	$\{-\frac{1}{2}, -\frac{1}{\sqrt{20}}, \frac{1}{\sqrt{20}}, \frac{1}{2}\}$

2.2 A limiter to enforce the sufficient condition

Given the vector of approximation polynomials $\mathbf{q}_j(x) = (\rho_j(x), m_j(x), E_j(x))^T$, either reconstructed for a finite volume scheme or evolved for a DG scheme, with its cell average $\overline{\mathbf{w}}_j^n = (\overline{\rho}_j^n, \overline{m}_j^n, \overline{E}_j^n)^T \in G$, we would like to modify $\mathbf{q}_j(x)$ into $\widetilde{\mathbf{q}}_j(x)$ such that it satisfies

- Accuracy: For smooth solutions, the limiter does not destroy accuracy

$$\| \widetilde{\mathbf{q}}_j(x) - \mathbf{q}_j(x) \| = O(\Delta x^{k+1}), \quad \forall x \in I_j,$$

  where $\| \cdot \|$ denotes the Euclidean norm.

- Positivity: $\widetilde{\mathbf{q}}_j(\widehat{x}_j^\alpha) \in G$ for $\alpha = 1, 2, \cdots, N$.

- Conservativity:

$$\frac{1}{\Delta x} \int_{I_j} \widetilde{\mathbf{q}}_j(x) dx = \overline{\mathbf{w}}_j^n.$$

Define $\overline{p}_j^n = (\gamma - 1)\left(\overline{E}_j^n - \frac{1}{2}(\overline{m}_j^n)^2/\overline{\rho}_j^n\right)$. Then $\overline{\rho}_j^n > 0$ and $\overline{p}_j^n > 0$ for all j. Assume there exists a small number $\varepsilon > 0$ such that $\overline{\rho}_j^n \geq \varepsilon$ and $\overline{p}_j^n \geq \varepsilon$ for all j. For example, we can take $\varepsilon = 10^{-13}$ in the computation.

The first step is to limit the density. Replace $\rho_j(x)$ by

$$\widehat{\rho}_j(x) = \theta_1(\rho_j(x) - \overline{\rho}_j^n) + \overline{\rho}_j^n, \tag{2.4}$$

where

$$\theta_1 = \min\left\{\frac{\overline{\rho}_j^n - \varepsilon}{\overline{\rho}_j^n - \rho_{\min}}, 1\right\}, \quad \rho_{\min} = \min_\alpha \rho_j(\widehat{x}_j^\alpha). \tag{2.5}$$

Then the cell average of $\widehat{\rho}_j(x)$ over I_j is still $\overline{\rho}_j^n$ and $\widehat{\rho}_j(\widehat{x}_j^\alpha) \geq \varepsilon$ for all α. The accuracy of $\widehat{\rho}_j(x)$ can be proven following the same lines as in [26].

The second step is to enforce the positivity of the pressure. We need to introduce some notations. Let $\widehat{\mathbf{q}}_j(x) = (\widehat{\rho}_j(x), m_j(x), E_j(x))^T$ and $\widehat{\mathbf{q}}_j^\alpha$ denote $\widehat{\mathbf{q}}_j(\widehat{x}_j^\alpha)$. Define

$$G^\varepsilon = \left\{ \mathbf{w} = \begin{pmatrix} \rho \\ m \\ E \end{pmatrix} \middle| \rho \geq \varepsilon \quad \text{and} \quad p = (\gamma - 1)\left(E - \frac{1}{2}\frac{m^2}{\rho}\right) \geq \varepsilon \right\}, \tag{2.6}$$

$$\partial G^\varepsilon = \left\{ \mathbf{w} = \begin{pmatrix} \rho \\ m \\ E \end{pmatrix} \middle| \rho \geq \varepsilon \quad \text{and} \quad p = (\gamma - 1)\left(E - \frac{1}{2}\frac{m^2}{\rho}\right) = \varepsilon \right\}, \tag{2.7}$$

and

$$\mathbf{s}^\alpha(t) = (1 - t)\overline{\mathbf{w}}_j^n + t\widehat{\mathbf{q}}_j(\widehat{x}_j^\alpha), \quad 0 \leq t \leq 1. \tag{2.8}$$

G^ε is a convex set thanks to (1.3). ∂G^ε in (2.7) is a surface which contains part of the boundary of G^ε. $\mathbf{s}^\alpha(t)$ in (2.8) is the straight line passing through the two points $\overline{\mathbf{w}}_j^n$ and $\widehat{\mathbf{q}}_j(\widehat{x}_j^\alpha)$.

If $\widehat{\mathbf{q}}_j^\alpha$ lies outside of G^ε, namely $p\left(\widehat{\mathbf{q}}_j^\alpha\right) < \varepsilon$, then there exists an intersection point of the straight line $\mathbf{s}^\alpha(t)$ and the surface ∂G^ε. Let $\mathbf{s}_\varepsilon^\alpha$ denote this intersection, then $\mathbf{s}_\varepsilon^\alpha = \mathbf{s}^\alpha(t_\varepsilon^\alpha)$ for some $t_\varepsilon^\alpha \in [0,1]$ satisfying $p\left(\mathbf{s}^\alpha(t_\varepsilon^\alpha)\right) = \varepsilon$. We will abuse the notation and let $\mathbf{s}_\varepsilon^\alpha = \widehat{\mathbf{q}}_j^\alpha$ if $p\left(\widehat{\mathbf{q}}_j^\alpha\right) \in G^\varepsilon$. So we have

$$\mathbf{s}_\varepsilon^\alpha = \begin{cases} \mathbf{s}^\alpha(t_\varepsilon^\alpha), & \text{if} \quad p\left(\widehat{\mathbf{q}}_j^\alpha\right) < \varepsilon \\ \widehat{\mathbf{q}}_j^\alpha, & \text{if} \quad p\left(\widehat{\mathbf{q}}_j^\alpha\right) \geq \varepsilon \end{cases} \tag{2.9}$$

We consider the following new vector of polynomials

$$\widetilde{\mathbf{q}}_j(x) = \theta_2\left(\widehat{\mathbf{q}}_j(x) - \overline{\mathbf{w}}_j^n\right) + \overline{\mathbf{w}}_j^n, \tag{2.10}$$

with

$$\theta_2 = \min_{\alpha=1,2,\cdots,N} \frac{\parallel \mathbf{s}_\varepsilon^\alpha - \overline{\mathbf{w}}_j^n \parallel}{\parallel \widehat{\mathbf{q}}_j^\alpha - \overline{\mathbf{w}}_j^n \parallel}. \tag{2.11}$$

It is easy to see that the cell average of $\widetilde{\mathbf{q}}_j(x)$ over I_j is $\overline{\mathbf{w}}_j^n$. Next we would like to show the following lemma.

Lemma 2.5. The $\widetilde{\mathbf{q}}_j(x)$ defined in (2.10) and (2.11) satisfies $\widetilde{\mathbf{q}}_j(\widehat{x}_j^\alpha) \in G^\varepsilon \subset G$ for all α.

Proof: Notice that $\widetilde{\mathbf{q}}_j(x)$ is actually a convex combination of $\widehat{\mathbf{q}}_j(x)$ and $\overline{\mathbf{w}}_j^n$, so the density of $\widetilde{\mathbf{q}}_j(\widehat{x}_j^\alpha)$ is no less than ε. For the same reason, $p\left(\widetilde{\mathbf{q}}_j(\widehat{x}_j^\alpha)\right) \geq \varepsilon$ if $p\left(\widehat{\mathbf{q}}_j(\widehat{x}_j^\alpha)\right) \geq \varepsilon$.

If $p\left(\widehat{\mathbf{q}}_j(\widehat{x}_j^\alpha)\right) < \varepsilon$, then $p\left(\mathbf{s}_\varepsilon^\alpha\right) = \varepsilon$ and

$$\begin{aligned} \widetilde{\mathbf{q}}_j(\widehat{x}_j^\alpha) &= \theta_2\left(\widehat{\mathbf{q}}_j(\widehat{x}_j^\alpha) - \overline{\mathbf{w}}_j^n\right) + \overline{\mathbf{w}}_j^n \\ &= \frac{\theta_2}{t_\varepsilon^\alpha}\left[t_\varepsilon^\alpha\left(\widehat{\mathbf{q}}_j(\widehat{x}_j^\alpha) - \overline{\mathbf{w}}_j^n\right) + \overline{\mathbf{w}}_j^n\right] + \left(1 - \frac{\theta_2}{t_\varepsilon^\alpha}\right)\overline{\mathbf{w}}_j^n \\ &= \frac{\theta_2}{t_\varepsilon^\alpha}\mathbf{s}_\varepsilon^\alpha + \left(1 - \frac{\theta_2}{t_\varepsilon^\alpha}\right)\overline{\mathbf{w}}_j^n, \end{aligned}$$

Notice that $\frac{\|\mathbf{s}_\varepsilon^\alpha - \overline{\mathbf{w}}_j^n\|}{\|\widehat{\mathbf{q}}_j^\alpha - \overline{\mathbf{w}}_j^n\|} = t_\varepsilon^\alpha$. Therefore, (2.11) implies $\theta_2 \leq t_\varepsilon^\alpha$. So $\widetilde{\mathbf{q}}_j(\widehat{x}_j^\alpha)$ is a convex combination of $\mathbf{s}_\varepsilon^\alpha$ and $\overline{\mathbf{w}}_j^n$, and thus $p\left(\widetilde{\mathbf{q}}_j(\widehat{x}_j^\alpha)\right) \geq \varepsilon$. ■

Finally, we need to show the limiter (2.10) and (2.11) does not destroy accuracy when $\mathbf{q}_j(x)$ approximates a smooth solution. Define $d(\mathbf{z}, G) = \min_{\mathbf{w}\in G} \parallel \mathbf{z} - \mathbf{w} \parallel$. Assume the exact solution $\mathbf{v}(x, t^n)$ is smooth and $d(\mathbf{v}(x, t^n), G^\varepsilon) \geq M$, $\forall x$, for some constant $M > 0$.

It suffices to show $\theta_2 = 1 + O(\Delta x^{k+1})$. If $\theta_2 < 1$, then $\theta_2 = \parallel \mathbf{s}_\varepsilon^\beta - \overline{\mathbf{w}}_j^n \parallel / \parallel \widehat{\mathbf{q}}_j^\beta - \overline{\mathbf{w}}_j^n \parallel$ for some β where $\mathbf{s}_\varepsilon^\beta$ is the intersection of the straight line and the surface.

Since $\overline{\mathbf{w}}_j^n$ is a $(k+1)$-th order approximation to the cell average of $\mathbf{v}(x, t^n)$, we have $d(\overline{\mathbf{w}}_j^n, G^\varepsilon) \geq M + O(\Delta x^{k+1}) \geq \frac{M}{2}$ if Δx is small enough. We can also assume the overshoot $\parallel \mathbf{s}_\varepsilon^\beta - \widehat{\mathbf{q}}_j^\beta \parallel = O(\Delta x^{k+1})$ since $\widehat{\mathbf{q}}_j^\beta$ is a $(k+1)$-th order approximation to a point in G^ε.

Thus,

$$\begin{aligned} |1 - \theta_2| &= 1 - \theta_2 \\ &= 1 - \frac{\parallel \mathbf{s}_\varepsilon^\beta - \overline{\mathbf{w}}_j^n \parallel}{\parallel \widehat{\mathbf{q}}_j^\beta - \overline{\mathbf{w}}_j^n \parallel} \\ &= \frac{\parallel \mathbf{s}_\varepsilon^\beta - \widehat{\mathbf{q}}_j^\beta \parallel}{\parallel \widehat{\mathbf{q}}_j^\beta - \overline{\mathbf{w}}_j^n \parallel} \\ &\leq \frac{\parallel \mathbf{s}_\varepsilon^\beta - \widehat{\mathbf{q}}_j^\beta \parallel}{d(\overline{\mathbf{w}}_j^n, G^\varepsilon)} \\ &= O(\Delta x^{k+1}), \end{aligned}$$

where the third equality is due to the fact that $\widehat{\mathbf{q}}_j^\beta$, $\mathbf{s}_\varepsilon^\beta$ and $\overline{\mathbf{w}}_j^n$ lie on the same line.

Therefore, the limiting process (2.4), (2.5), (2.10) and (2.11) returns $\widetilde{\mathbf{q}}_j(x)$ satisfying the accuracy, positivity and conservativity.

2.3 Implementation for the DG method

At time level n, assuming the DG polynomial in cell I_j is $\mathbf{q}_j(x) = (\rho_j(x), m_j(x), E_j(x))^T$ with degree k , and the cell average of $\mathbf{q}_j(x)$ is $\overline{\mathbf{w}}_j^n = \left(\overline{\rho}_j^n, \overline{m}_j^n, \overline{E}_j^n\right)^T \in G$, then the algorithm flowchart of our algorithm for the Euler forward is

- Set up a small number $\varepsilon = \min_j\{10^{-13}, \overline{\rho}_j^n, p(\overline{\mathbf{w}}_j^n)\}$.
- In each cell, modify the density first: evaluate $\min_{\alpha=1,\cdots,N} \rho_j(\widehat{x}_j^\alpha)$ and get $\widehat{\rho}_j(x)$ by (2.4) and (2.5), set $\widehat{\mathbf{q}}_j(x) = (\widehat{\rho}_j(x), m_j(x), E_j(x))^T$.

- Then modify the pressure: let $\widehat{\mathbf{q}}_j^\alpha$ denote $\widehat{\mathbf{q}}_j(x_j^\alpha)$, for each α, if $p(\widehat{\mathbf{q}}_j^\alpha) < \varepsilon$, then solve the following quadratic equation for t_ε^α,

$$p\left[(1-t_\varepsilon^\alpha)\overline{\mathbf{w}}_j^n + t_\varepsilon^\alpha \widehat{\mathbf{q}}_j(\widehat{x}_j^\alpha)\right] = \varepsilon. \tag{2.12}$$

  If $p(\widehat{\mathbf{q}}_j^\alpha) \geq \varepsilon$, then set $t_\varepsilon^\alpha = 1$. θ_2 in (2.11) is mathematically equivalent to $\theta_2 = \min_{\alpha=1,\cdots,N} t_\varepsilon^\alpha$. Get $\widetilde{\mathbf{q}}_j(x)$ by (2.10).

- Use $\widetilde{\mathbf{q}}_j(x)$ instead of $\mathbf{q}_j(x)$ in the DG scheme with Euler forward in time under the CFL condition (2.1). ■

For SSP high order time discretizations, we need to use the limiter in each stage for a Runge-Kutta method or in each step for a multistep method.

Remark 2.6. The implementation for a finite volume method is similar, but it will be a little bit more complicated for WENO since there are only nodal values but no polynomials in each cell after WENO reconstruction. One way to implement the limiter is to construct polynomials using the nodal values and cell averages, see [26] for details. We are also exploring other, simpler ways to implement this positivity preserving limiter for WENO finite volume schemes. These implementation details and numerical tests will be reported elsewhere.

Remark 2.7. Theoretically, there is a complication regarding the CFL condition (2.1) for a Runge-Kutta time discretization. Consider the third order SSP Runge-Kutta method. To enforce (2.1) rigorously, we need to get an accurate estimation of $\| (|u| + c) \|_\infty$ for all the three stages based only on the numerical solution at time level n, which is highly nontrivial mathematically. In practice, we can simply multiply a factor, for example 2 to 3, to the quantity $\| (|u| + c) \|_\infty$ of $\mathbf{w}^n$, as an estimation for all the stages. Although this is a rough estimation, it works well for us to choose a time step satisfying (2.1) in all the examples in Section 4. To be more efficient, we could implement this more stringent CFL condition only when a preliminary calculation to the next time step produces negative density or pressure. This complication does not exist if we use a SSP multi-step time discretization.

3 Positivity preserving high order schemes in two dimensions

3.1 A sufficient condition

In this section we extend our result to finite volume or DG schemes of $(k+1)$-th order accuracy on rectangular meshes solving two dimensional Euler equations

$$\mathbf{w}_t + \mathbf{f}(\mathbf{w})_x + \mathbf{g}(\mathbf{w})_y = 0, \quad t \geq 0, (x,y) \in \mathbb{R}^2, \tag{3.1}$$

$$\mathbf{w} = \begin{pmatrix} \rho \\ m \\ n \\ E \end{pmatrix}, \quad \mathbf{f}(\mathbf{w}) = \begin{pmatrix} m \\ \rho u^2 + p \\ \rho uv \\ (E+p)u \end{pmatrix}, \quad \mathbf{g}(\mathbf{w}) = \begin{pmatrix} n \\ \rho uv \\ \rho v^2 + p \\ (E+p)v \end{pmatrix} \tag{3.2}$$

with

$$m = \rho u, \quad n = \rho v, \quad E = \frac{1}{2}\rho u^2 + \frac{1}{2}\rho v^2 + \rho e, \quad p = (\gamma - 1)\rho e,$$

where ρ is the density, u is the velocity in x direction, v is the velocity in y direction, m and n are the momenta, E is the total energy, p is the pressure, e is the internal energy. The speed of sound is given by $c = \sqrt{\gamma p/\rho}$. The eigenvalues of the Jacobian $\mathbf{f}'(\mathbf{w})$ are $u-c$, u, u and $u+c$ and the eigenvalues of the Jacobian $\mathbf{g}'(\mathbf{w})$ are $v-c$, v, v and $v+c$. The pressure function p is still concave with respect to $\mathbf{w}$ if $\rho \geq 0$ and the set of admissible states

$$G = \left\{ \mathbf{w} = \begin{pmatrix} \rho \\ m \\ n \\ E \end{pmatrix} \middle| \rho > 0 \quad \text{and} \quad p = (\gamma-1)\left(E - \frac{1}{2}\frac{m^2}{\rho} - \frac{1}{2}\frac{n^2}{\rho}\right) > 0 \right\}$$

is still convex.

For simplicity we assume we have a uniform rectangular mesh. At time level n, we have a vector of approximation polynomials of degree k, $\mathbf{q}_{ij}(x,y) = (\rho_{ij}(x,y), m_{ij}(x,y), n_{ij}(x,y), E_{ij}(x,y))^T$ with the cell average $\overline{\mathbf{w}}_{ij}^n = \left(\overline{\rho}_{ij}^n, \overline{m}_{ij}^n, \overline{n}_{ij}^n, \overline{E}_{ij}^n\right)^T$ on the (i,j) cell $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$.

Let $\mathbf{w}^+_{i-\frac{1}{2},j}(y), \mathbf{w}^-_{i+\frac{1}{2},j}(y)$, $\mathbf{w}^+_{i,j-\frac{1}{2}}(x)$, $\mathbf{w}^-_{i,j+\frac{1}{2}}(x)$ denote the traces of $\mathbf{q}_{ij}(x,y)$ on the four edges respectively, see Figure 3.1. All of the traces are vectors of single variable polynomials of degree k.

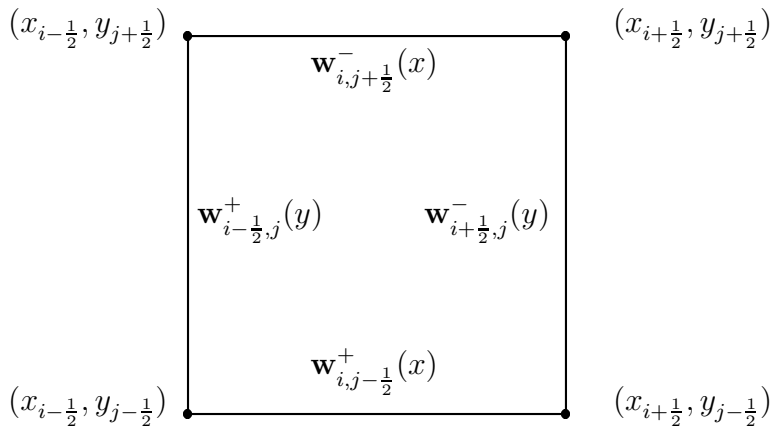


Figure 3.1: The traces of $\mathbf{q}_{ij}(x, y)$.

We only discuss Euler forward in time here. A finite volume scheme or the scheme satisfied by the cell averages of a DG method for (3.1) on a rectangular mesh can be written as

$$\begin{aligned}\overline{\mathbf{w}}_{ij}^{n+1} &= \overline{\mathbf{w}}_{ij}^{n} - \frac{\Delta t}{\Delta x \Delta y}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \mathbf{h}_1\left[\mathbf{w}^-_{i+\frac{1}{2},j}(y), \mathbf{w}^+_{i+\frac{1}{2},j}(y)\right] - \mathbf{h}_1\left[\mathbf{w}^-_{i-\frac{1}{2},j}(y), \mathbf{w}^+_{i-\frac{1}{2},j}(y)\right] dy \\ &\quad -\frac{\Delta t}{\Delta x \Delta y}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathbf{h}_2\left[\mathbf{w}^-_{i,j+\frac{1}{2}}(x), \mathbf{w}^+_{i,j+\frac{1}{2}}(x)\right] - \mathbf{h}_2\left[\mathbf{w}^-_{i,j-\frac{1}{2}}(x), \mathbf{w}^+_{i,j-\frac{1}{2}}(x)\right] dx, \quad (3.3)\end{aligned}$$

where $\mathbf{h}_1(\cdot,\cdot)$ and $\mathbf{h}_2(\cdot,\cdot)$ are one dimensional numerical fluxes. We will use the Lax-Friedrichs flux in this section as an example:

$$\mathbf{h}_1(\mathbf{u}, \mathbf{v}) = \frac{1}{2}\left[\mathbf{f}(\mathbf{u}) + \mathbf{f}(\mathbf{v}) - a_1(\mathbf{v} - \mathbf{u})\right], \quad a_1 = \| \left(|u| + c\right) \|_\infty$$

$$\mathbf{h}_2(\mathbf{u}, \mathbf{v}) = \frac{1}{2}\left[\mathbf{g}(\mathbf{u}) + \mathbf{g}(\mathbf{v}) - a_2(\mathbf{v} - \mathbf{u})\right], \quad a_2 = \| \left(|v| + c\right) \|_\infty$$

The integrals in (3.3) can be approximated by quadratures with sufficient accuracy. Let us assume that we use a Gauss quadrature with L points, which is exact for single variable polynomials of degree $2k + 1$ (see [1] for an analysis of the requirement of the numerical quadrature for the accuracy of the DG solution). We assume

$$S_i^x = \{x_i^\beta : \beta = 1, \cdots, L\} \quad (3.4)$$

denote the Gauss quadrature points on $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$, and

$$S_j^y = \{y_j^\beta : \beta = 1, \cdots, L\} \quad (3.5)$$

denote the Gauss quadrature points on $[y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$. For instance, $(x_{i-\frac{1}{2}}, y_j^\beta)$ $(\beta = 1, \cdots, L)$ are the Gauss quadrature points on the left edge of the (i,j) cell. The subscript β will denote the values at the Gauss quadrature points, for instance, $\mathbf{w}^+_{i-\frac{1}{2},\beta} = \mathbf{w}^+_{i-\frac{1}{2},j}(y_j^\beta)$. Also, w_β denotes the corresponding quadrature weight on interval $[-\frac{1}{2}, \frac{1}{2}]$, so that $\sum_{\beta=1}^L w_\beta = 1$. We will still need to use the Gauss-Lobatto quadrature rule, and we distinguish the two quadrature rules by adding hats to the Gauss-Lobatto points, i.e.,

$$\widehat{S}_i^x = \{\widehat{x}_i^\alpha : \alpha = 1, \cdots, N\} \tag{3.6}$$

will denote the Gauss-Lobatto quadrature points on $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$, and

$$\widehat{S}_j^y = \{\widehat{y}_j^\alpha : \alpha = 1, \cdots, N\} \tag{3.7}$$

will denote the Gauss-Lobatto quadrature points on $[y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$. Subscripts or superscripts β and γ will be used only for Gauss quadrature points and α only for Gauss-Lobatto points.

Then the scheme (3.3) becomes

$$\begin{aligned}
\overline{\mathbf{w}}_{ij}^{n+1} &= \overline{\mathbf{w}}_{ij}^n - \frac{\Delta t}{\Delta x \Delta y} \sum_{\beta=1}^L \left[\mathbf{h}_1 \left(\mathbf{w}^-_{i+\frac{1}{2},\beta}, \mathbf{w}^+_{i+\frac{1}{2},\beta} \right) - \mathbf{h}_1 \left(\mathbf{w}^-_{i-\frac{1}{2},\beta}, \mathbf{w}^+_{i-\frac{1}{2},\beta} \right) \right] w_\beta \Delta y \\
&\quad - \frac{\Delta t}{\Delta x \Delta y} \sum_{\beta=1}^L \left[\mathbf{h}_2 \left(\mathbf{w}^-_{\beta,j+\frac{1}{2}}, \mathbf{w}^+_{\beta,j+\frac{1}{2}} \right) - \mathbf{h}_2 \left(\mathbf{w}^-_{\beta,j-\frac{1}{2}}, \mathbf{w}^+_{\beta,j-\frac{1}{2}} \right) \right] w_\beta \Delta x \\
&= \overline{\mathbf{w}}_{ij}^n - \lambda_1 \sum_{\beta=1}^L w_\beta \left[\mathbf{h}_1 \left(\mathbf{w}^-_{i+\frac{1}{2},\beta}, \mathbf{w}^+_{i+\frac{1}{2},\beta} \right) - \mathbf{h}_1 \left(\mathbf{w}^-_{i-\frac{1}{2},\beta}, \mathbf{w}^+_{i-\frac{1}{2},\beta} \right) \right] \\
&\quad - \lambda_2 \sum_{\beta=1}^L w_\beta \left[\mathbf{h}_2 \left(\mathbf{w}^-_{\beta,j+\frac{1}{2}}, \mathbf{w}^+_{\beta,j+\frac{1}{2}} \right) - \mathbf{h}_2 \left(\mathbf{w}^-_{\beta,j-\frac{1}{2}}, \mathbf{w}^+_{\beta,j-\frac{1}{2}} \right) \right],
\end{aligned} \tag{3.8}$$

where $\lambda_1 = \frac{\Delta t}{\Delta x}$ and $\lambda_2 = \frac{\Delta t}{\Delta y}$.

Consider the quadrature rule for $\mathbf{q}_{ij}(x,y)$ on the rectangle $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$ by tensoring (3.4) and (3.7). Let $\mathbf{w}^1_{\beta,\alpha}$ denote $\mathbf{q}_{ij}\left(x_i^\beta, \widehat{y}_j^\alpha\right)$, then

$$\begin{aligned}
\overline{\mathbf{w}}_{ij}^n &= \frac{1}{\Delta x \Delta y} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathbf{q}_{ij}(x,y) dx dy \\
&= \frac{1}{\Delta x \Delta y} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \left(\sum_{\beta=1}^L \mathbf{q}_{ij}(x_i^\beta, y) w_\beta \Delta x \right) dy
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{\beta=1}^{L} w_\beta \left(\frac{1}{\Delta y} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \mathbf{q}_{ij}(x_i^\beta, y) dy \right) \\
&= \sum_{\beta=1}^{L} w_\beta \left(\frac{1}{\Delta y} \sum_{\alpha=1}^{N} \mathbf{q}_{ij}(x_i^\beta, \widehat{y}_j^\alpha) \Delta y \widehat{w}_\alpha \right) \\
&= \sum_{\beta=1}^{L} \sum_{\alpha=1}^{N} w_\beta \widehat{w}_\alpha \mathbf{w}_{\beta,\alpha}^1 \qquad (3.9)
\end{aligned}
$$

Similarly, we can get another quadrature rule for $\mathbf{q}_{ij}(x,y)$ on the rectangle $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$ by tensoring (3.5) and (3.6). Let $\mathbf{w}_{\alpha,\beta}^2$ denote $\mathbf{q}_{ij}\left(\widehat{x}_i^\alpha, y_j^\beta\right)$, then

$$
\overline{\mathbf{w}}_{ij}^n = \sum_{\beta=1}^{L} \sum_{\alpha=1}^{N} w_\beta \widehat{w}_\alpha \mathbf{w}_{\alpha,\beta}^2 \qquad (3.10)
$$

Notice that $\mathbf{w}_{1,\beta}^2 = \mathbf{w}_{i-\frac{1}{2},\beta}^+$, $\mathbf{w}_{N,\beta}^2 = \mathbf{w}_{i+\frac{1}{2},\beta}^-$, $\mathbf{w}_{\beta,1}^1 = \mathbf{w}_{\beta,j-\frac{1}{2}}^+$ and $\mathbf{w}_{\beta,N}^1 = \mathbf{w}_{\beta,j+\frac{1}{2}}^-$. Let $\mu = a_1\lambda_1 + a_2\lambda_2$ and combine (3.9) and (3.10), we have

$$
\begin{aligned}
\overline{\mathbf{w}}_{ij}^n &= \frac{a_1\lambda_1}{\mu} \overline{\mathbf{w}}_{ij}^n + \frac{a_2\lambda_2}{\mu} \overline{\mathbf{w}}_{ij}^n \\
&= \frac{a_1\lambda_1}{\mu} \sum_{\beta=1}^{L} \sum_{\alpha=1}^{N} w_\beta \widehat{w}_\alpha \mathbf{w}_{\alpha,\beta}^2 + \frac{a_2\lambda_2}{\mu} \sum_{\beta=1}^{L} \sum_{\alpha=1}^{N} w_\beta \widehat{w}_\alpha \mathbf{w}_{\beta,\alpha}^1 \\
&= \frac{a_1\lambda_1}{\mu} \sum_{\alpha=2}^{N-1} \sum_{\beta=1}^{L} w_\beta \widehat{w}_\alpha \mathbf{w}_{\alpha,\beta}^2 + \frac{a_1\lambda_1}{\mu} \widehat{w}_1 \sum_{\beta=1}^{L} w_\beta \left(\mathbf{w}_{i-\frac{1}{2},\beta}^+ + \mathbf{w}_{i+\frac{1}{2},\beta}^- \right) \\
&\quad + \frac{a_2\lambda_2}{\mu} \sum_{\alpha=2}^{N-1} \sum_{\beta=1}^{L} w_\beta \widehat{w}_\alpha \mathbf{w}_{\beta,\alpha}^1 + \frac{a_2\lambda_2}{\mu} \widehat{w}_1 \sum_{\beta=1}^{L} w_\beta \left(\mathbf{w}_{\beta,j-\frac{1}{2},}^+ + \mathbf{w}_{\beta,j-\frac{1}{2}}^- \right), \qquad (3.11)
\end{aligned}
$$

where we have used the fact that $\widehat{w}_1 = \widehat{w}_N$.

Plugging (3.11) into (3.8), adding and subtracting $\mathbf{h}_1\left(\mathbf{w}_{i-\frac{1}{2},\beta}^+, \mathbf{w}_{i+\frac{1}{2},\beta,}^-\right)$ and $\mathbf{h}_2\left(\mathbf{w}_{\beta,j-\frac{1}{2}}^+, \mathbf{w}_{\beta,j+\frac{1}{2}}^-\right)$, then the scheme (3.8) can be written as

$$
\begin{aligned}
\overline{\mathbf{w}}_{ij}^{n+1} &= \frac{a_1\lambda_1}{\mu} \sum_{\alpha=2}^{N-1} \sum_{\beta=1}^{L} w_\beta \widehat{w}_\alpha \mathbf{w}_{\alpha,\beta}^2 + \frac{a_2\lambda_2}{\mu} \sum_{\alpha=2}^{N-1} \sum_{\beta=1}^{L} w_\beta \widehat{w}_\alpha \mathbf{w}_{\beta,\alpha}^1 \\
&\quad + \frac{a_1\lambda_1}{\mu} \widehat{w}_1 \sum_{\beta=1}^{L} w_\beta \left(\mathbf{w}_{i+\frac{1}{2},\beta}^- - \frac{\mu}{a_1\widehat{w}_1} \left[\mathbf{h}_1\left(\mathbf{w}_{i+\frac{1}{2},\beta}^-, \mathbf{w}_{i+\frac{1}{2},\beta}^+\right) - \mathbf{h}_1\left(\mathbf{w}_{i-\frac{1}{2},\beta}^+, \mathbf{w}_{i+\frac{1}{2},\beta}^-\right) \right] \right) \\
&\quad + \frac{a_1\lambda_1}{\mu} \widehat{w}_1 \sum_{\beta=1}^{L} w_\beta \left(\mathbf{w}_{i-\frac{1}{2},\beta}^+ - \frac{\mu}{a_1\widehat{w}_1} \left[\mathbf{h}_1\left(\mathbf{w}_{i-\frac{1}{2},\beta}^+, \mathbf{w}_{i+\frac{1}{2},\beta}^-\right) - \mathbf{h}_1\left(\mathbf{w}_{i-\frac{1}{2},\beta}^-, \mathbf{w}_{i-\frac{1}{2},\beta}^+\right) \right] \right)
\end{aligned}
$$

$$+\frac{a_2\lambda_2}{\mu}\widehat{w}_1\sum_{\beta=1}^{L} w_\beta\left(\mathbf{w}^-_{\beta,j+\frac{1}{2}}-\frac{\mu}{a_2\widehat{w}_1}\left[\mathbf{h}_2\left(\mathbf{w}^-_{\beta,j+\frac{1}{2}},\mathbf{w}^+_{\beta,j+\frac{1}{2}}\right)-\mathbf{h}_2\left(\mathbf{w}^+_{\beta,j-\frac{1}{2}},\mathbf{w}^-_{\beta,j+\frac{1}{2}}\right)\right]\right)$$
$$+\frac{a_2\lambda_2}{\mu}\widehat{w}_1\sum_{\beta=1}^{L} w_\beta\left(\mathbf{w}^+_{\beta,j-\frac{1}{2}}-\frac{\mu}{a_2\widehat{w}_1}\left[\mathbf{h}_2\left(\mathbf{w}^+_{\beta,j-\frac{1}{2}},\mathbf{w}^-_{\beta,j+\frac{1}{2}}\right)-\mathbf{h}_2\left(\mathbf{w}^-_{\beta,j+\frac{1}{2}},\mathbf{w}^+_{\beta,j-\frac{1}{2}}\right)\right]\right). \tag{3.12}$$

Starting from (3.12) and following the same lines as the proof of Theorem 2.1, we can easily prove

Theorem 3.1. For the scheme (3.8), if $\mathbf{w}^1_{\beta,\alpha}=\mathbf{q}_{ij}\left(x_i^\beta,\widehat{y}_j^\alpha\right)\in G$ and $\mathbf{w}^2_{\alpha,\beta}=\mathbf{q}_{ij}\left(\widehat{x}_i^\alpha,y_j^\beta\right)\in G$ for all i,j,α and β, then $\overline{\mathbf{w}}_j^{n+1}\in G$ under the CFL condition

$$\frac{\Delta t}{\Delta x}\parallel(|u|+c)\parallel_\infty+\frac{\Delta t}{\Delta y}\parallel(|v|+c)\parallel_\infty\leq\widehat{w}_1\alpha_0. \tag{3.13}$$

■

3.2 Limiter and implementation for the DG method

Given the vector of approximation polynomials $\mathbf{q}_{ij}(x,y)=(\rho_{ij}(x,y),m_{ij}(x,y),n_{ij}(x,y),E_{ij}(x,y))^T$ (either reconstruction polynomials or DG polynomials) with its cell average $\overline{\mathbf{w}}_{ij}^n=(\overline{\rho}_{ij}^n,\overline{m}_{ij}^n,\overline{n}_{ij}^n,\overline{E}_{ij}^n)^T\in G$, we would like to modify $\mathbf{q}_{ij}(x,y)$ into $\widetilde{\mathbf{q}}_{ij}(x,y)$ such that it satisfies:

- Accuracy: For smooth solutions, the limiter does not destroy accuracy

$$\parallel\widetilde{\mathbf{q}}_{ij}(x,y)-\mathbf{q}_{ij}(x,y)\parallel=O(\Delta x^{k+1}),\quad\forall(x,y)\in I_{ij},$$

  where $\|\cdot\|$ denotes the Euclidean norm.

- Positivity: $\widetilde{\mathbf{q}}_{ij}\left(x_i^\beta,\widehat{y}_j^\alpha\right)\in G$ and $\widetilde{\mathbf{q}}_{ij}\left(\widehat{x}_i^\alpha,y_j^\beta\right)\in G$ for $\alpha=1,2,\cdots,N$ and $\beta=1,2,\cdots,L$.

- Conservativity:

$$\frac{1}{\Delta x\Delta y}\iint\limits_{I_{ij}}\widetilde{\mathbf{q}}_{ij}(x,y)dxdy=\overline{\mathbf{w}}_{ij}^n.$$

Define $\overline{p}_{ij}^n = (\gamma-1)\left(\overline{E}_{ij}^n - \frac{1}{2}(\overline{m}_{ij}^n)^2/\overline{\rho}_{ij}^n - \frac{1}{2}(\overline{n}_{ij}^n)^2/\overline{\rho}_{ij}^n\right)$. Then $\overline{\rho}_{ij}^n > 0$ and $\overline{p}_{ij}^n > 0$ for all i, j. Assume there exists a small number $\varepsilon > 0$ such that $\overline{\rho}_{ij}^n \geq \varepsilon$ and $\overline{p}_{ij}^n \geq \varepsilon$ for all j.

The first step is to limit the density. Replace $\rho_{ij}(x, y)$ by

$$\widehat{\rho}_{ij}(x, y) = \theta_1(\rho_{ij}(x, y) - \overline{\rho}_{ij}^n) + \overline{\rho}_{ij}^n, \tag{3.14}$$

$$\theta_1 = \min\left\{\frac{\overline{\rho}_{ij}^n - \varepsilon}{\overline{\rho}_{ij}^n - \rho_{\min}}, 1\right\}, \quad \rho_{\min} = \min_{\alpha,\beta}\left\{\rho_{ij}\left(\widehat{x}_i^\alpha, y_j^\beta\right), \rho_{ij}\left(x_i^\beta, \widehat{y}_j^\alpha\right)\right\}. \tag{3.15}$$

The second step is to enforce the positivity of the pressure. Let

$$\widehat{\mathbf{q}}_{ij}(x, y) = (\widehat{\rho}_{ij}(x, y), m_{ij}(x, y), n_{ij}(x, y), E_{ij}(x, y))^T.$$

Denote $\widehat{\mathbf{q}}_{\beta,\alpha}^1 = \widehat{\mathbf{q}}_{ij}(x_i^\beta, \widehat{y}_j^\alpha)$ and $\widehat{\mathbf{q}}_{\alpha,\beta}^2 = \widehat{\mathbf{q}}_{ij}(\widehat{x}_i^\alpha, y_j^\beta)$. Define

$$\mathbf{s}_1^{\beta,\alpha}(t) = (1-t)\overline{\mathbf{w}}_{ij}^n + t\widehat{\mathbf{q}}_{\beta,\alpha}^1,$$

$$\mathbf{s}_2^{\alpha,\beta}(t) = (1-t)\overline{\mathbf{w}}_{ij}^n + t\widehat{\mathbf{q}}_{\alpha,\beta}^2.$$

Then calculate

$$t_{\varepsilon,1}^{\beta,\alpha} = \begin{cases} 1, & \text{if} \quad p\left(\widehat{\mathbf{q}}_{\beta,\alpha}^1\right) \geq \varepsilon \\ \text{the solution of} \quad p\left(\mathbf{s}_1^{\beta,\alpha}(t)\right) = \varepsilon, & \text{if} \quad p\left(\widehat{\mathbf{q}}_{\beta,\alpha}^1\right) < \varepsilon \end{cases} \tag{3.16}$$

$$t_{\varepsilon,2}^{\alpha,\beta} = \begin{cases} 1, & \text{if} \quad p\left(\widehat{\mathbf{q}}_{\alpha,\beta}^2\right) \geq \varepsilon \\ \text{the solution of} \quad p\left(\mathbf{s}_2^{\alpha,\beta}(t)\right) = \varepsilon, & \text{if} \quad p\left(\widehat{\mathbf{q}}_{\alpha,\beta}^2\right) < \varepsilon \end{cases} \tag{3.17}$$

We have the following new vector of polynomials

$$\widetilde{\mathbf{q}}_{ij}(x, y) = \theta_2\left(\widehat{\mathbf{q}}_{ij}(x) - \overline{\mathbf{w}}_{ij}^n\right) + \overline{\mathbf{w}}_{ij}^n, \tag{3.18}$$

$$\theta_2 = \min_{\alpha,\beta}\left\{t_{\varepsilon,1}^{\beta,\alpha}, t_{\varepsilon,2}^{\alpha,\beta}\right\} \tag{3.19}$$

For the same reason as in Section 2.2, we can show the limiting process (3.14), (3.15), (3.18) and (3.19) returns $\widetilde{\mathbf{q}}_{ij}(x, y)$ satisfying the accuracy, positivity and conservativity.

The algorithm flowchart of our algorithm for the Euler forward is

- Set up a small number $\varepsilon = \min_{i,j}\{10^{-13}, \overline{\rho}_{ij}^n, p\left(\overline{\mathbf{w}}_{ij}^n\right)\}$.

- In each cell, modify the density first: evaluate $\min_{\alpha,\beta}\left\{\rho_{ij}\left(\widehat{x}_i^{\alpha}, y_j^{\beta}\right), \rho_{ij}\left(x_i^{\beta}, \widehat{y}_j^{\alpha}\right)\right\}$ and get $\widehat{\rho}_{ij}(x,y)$ by (3.14) and (3.15), set $\widehat{\mathbf{q}}_{ij}(x,y) = \left(\widehat{\rho}_{ij}(x,y), m_{ij}(x,y), n_{ij}(x,y), E_{ij}(x,y)\right)^T$.

- Then modify the pressure: Solve $t_{\varepsilon,1}^{\beta,\alpha}$ and $t_{\varepsilon,2}^{\alpha,\beta}$ in (3.16) and (3.17). Get $\widetilde{\mathbf{q}}_{ij}(x,y)$ by (3.18) and (3.19).

- Use $\widetilde{\mathbf{q}}_{ij}(x,y)$ instead of $\mathbf{q}_{ij}(x,y)$ in the DG scheme with Euler forward in time under the CFL condition (2.1). ■

For SSP high order time discretizations, we need to use the limiter in each stage for a Runge-Kutta method or in each step for a multistep method.

4 Numerical tests for a third order DG scheme

In this section, we implement a third order ($k = 2$) DG scheme with our positivity preserving method. Time discretization is third order SSP Runge-Kutta method [25].

4.1 The accuracy test

Consider a two-dimensional low density problem for (3.1). The initial condition is

$$\rho_0(x,y) = 1 + 0.99\sin(x+y),\ u_0(x,y) = 1,\ v_0(x,y) = 1,\ p_0(x,y) = 1. \tag{4.1}$$

The domain is $[0, 2\pi] \times [0, 2\pi]$ and the boundary condition is periodic. The exact solution is

$$\rho(x,y,t) = 1 + 0.99\sin(x+y-2t),\ u(x,y,t) = 1,\ v(x,y,t) = 1,\ p(x,y,t) = 1.$$

The minimum density of the exact solution is 0.01. The accuracy result for a third order RKDG scheme with the method in previous sections, which is referred to as the positivity limiter, is listed in Table 4.1. We clearly observe the designed order of accuracy for this low density problem. The positivity limiter is actually of the same type as the linear scaling limiter in [26], for which extensive accuracy experiments have been performed for scalar problems, see [26] for more details. We have monitored the numerical evolution and have observed that the positivity limiter does get turned on for the coarser meshes.

Table 4.1: Third order RKDG scheme with the positivity limiter, for (3.1) with initial data (4.1), $\Delta x = \Delta y = \frac{2\pi}{N}$, t=0.1.

$N \times N$	L^1 error	order	L^∞ error	order
4×4	1.49e-1	-	4.22e-1	-
8×8	1.10e-2	3.76	6.83e-2	2.62
16×16	1.11e-3	3.31	9.49e-3	2.84
32×32	1.42e-4	2.97	1.73e-3	2.45
64×64	2.07e-5	2.78	2.83e-4	2.61
128×128	3.11e-6	2.74	3.28e-5	3.02

4.2 The Sedov blast waves

The Sedov point-blast wave is a typical low density problem involving shocks. The exact solution formula can be found in [22, 14].

If discontinuities emerge in the solution, then we should use the characteristic TVB limiter [4, 2] in the DG scheme. Although the positivity limiter can successfully preserve the positivity of density and pressure, the TVB limiter is still necessary for shocks. The DG scheme without the TVB limiter will produce blow-ups for the Sedov blast waves even if we use the positivity limiter. The TVB limiter is applied before the positivity limiter.

In the TVB limiter, there is a TVB corrected minmod function [23, 3] defined by

$$\overline{m}(a_1, \cdots, a_m) = \begin{cases} a_1, & \text{if} \quad |a_1| \le M\Delta x^2 \\ m(a_1, \cdots, a_m), & \text{otherwise,} \end{cases} \tag{4.2}$$

with the minmod function m defined by

$$m(a_1, \cdots, a_m) = \begin{cases} s \min_i |a_i|, & \text{if} \quad s = \text{sign}(a_1) = \cdots = \text{sign}(a_m), \\ 0, & \text{otherwise.} \end{cases}$$

In each characteristic field, we may need a different M in (4.2). We will use M_i to denote the M in (4.2) for the i-th characteristic field. For scalar problems, analysis in [23, 3] indicates that M is related to the size of the second derivative of the initial condition near smooth extrema. The estimates of M for systems are more complicated, especially when local characteristic decomposition is used. It is possible to estimate the correct range of M for the linearized system after characteristic decomposition using techniques similar to the scalar case in [23, 3], which could serve as guidelines for nonlinear systems.

Without the positivity limiter, the DG method with the TVB limiter may produce blow-ups for $M_i > 0$. $M_i = 0$ will stabilize the scheme but it reduces to a TVD limiter [3] which is a first order correction at extrema and the computational result is not satisfactory. The advantage of using the positivity limiter is, one can tune up M_i in a much larger range to get a much better computational result, without producing negative density or pressure.

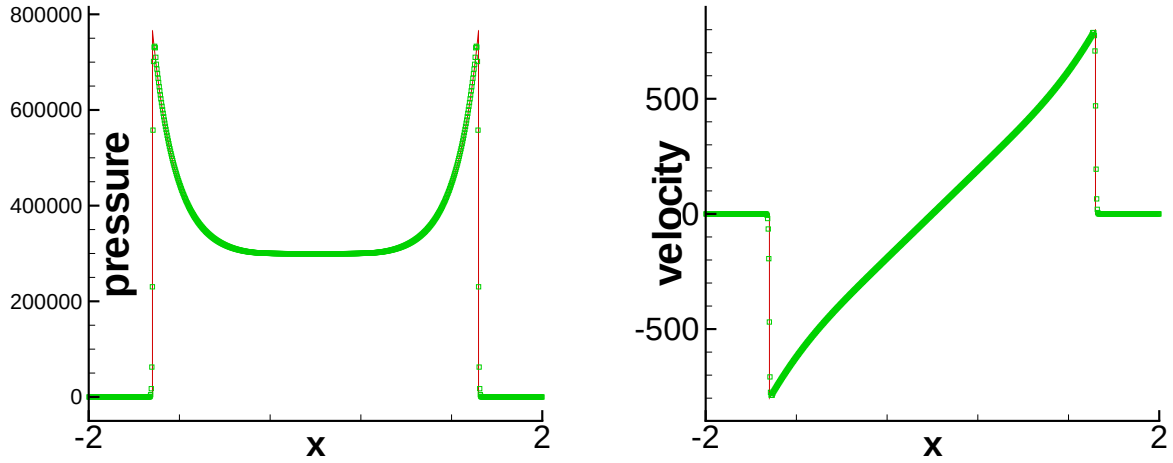
We test two blast waves here. The first one is one-dimensional. For the initial condition, the density is 1, velocity is zero, total energy is 10^{-12} everywhere except that the energy in the center cell is the constant $\frac{E_0}{\Delta x}$ with $E_0 = 3200000$ (emulating a δ-function at the center). $\gamma = 1.4$. The computational results are shown in Figure 4.1. We can see the shock is captured very well.

The second one is two-dimensional. The computational domain is a square. For the initial condition, the density is 1, velocity is zero, total energy is 10^{-12} everywhere except that the energy in the lower left corner cell is the constant $\frac{0.244816}{\Delta x \Delta y}$. $\gamma = 1.4$. The numerical boundary treatment is, extending ρ, v, E of the DG solutions as even functions and u as an odd function with respect to the left edge; extending ρ, u, E of the DG solutions as even functions and v as an odd function with respect to the bottom edge (symmetry). See Figure 4.2. The computational result is comparable to those in the literature, e.g. [16] (which uses a Lagrangian method to compute this problem).

4.3 Extreme Riemann problems

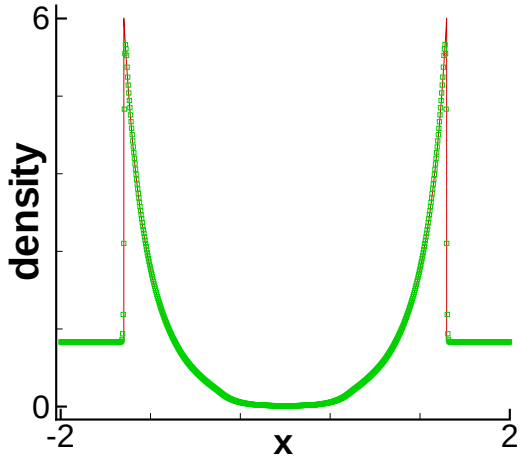
We consider two Riemann problems. The first one is a double rarefaction in [15]. We did two tests, one is a one-dimensional double rarefaction, for which the initial condition is $\rho_L = \rho_R = 7$, $u_L = -1$, $u_R = 1$, $p_L = p_R = 0.2$ and $\gamma = 1.4$. The other one is a two-dimensional double rarefaction with the initial condition $\rho_L = \rho_R = 7$, $u_L = -1$, $u_R = 1$, $v_L = v_R = 0$, $p_L = p_R = 0.2$.

The exact solution contains vacuum. Since there is no shock, we do not need the TVB limiter for this problem. See Figure 4.3 for the result of the DG scheme with the positivity



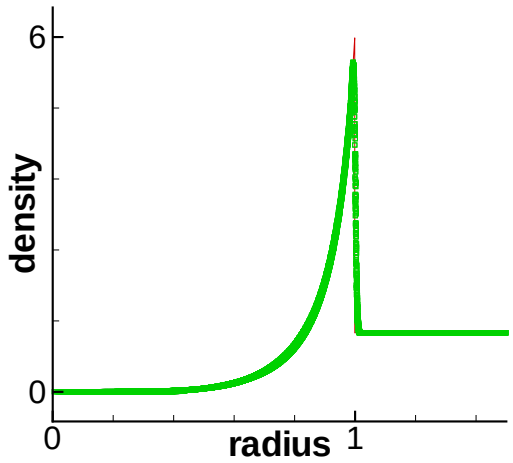
(a) pressure

(b) velocity

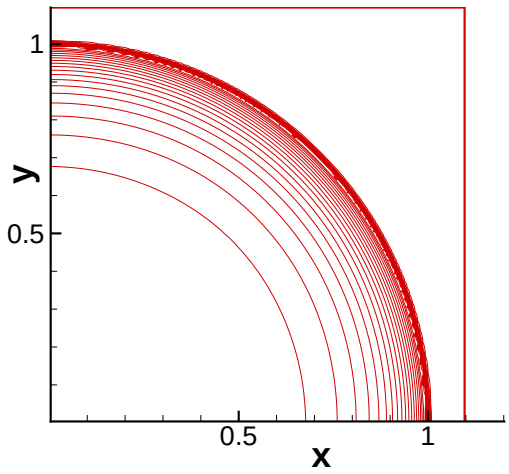


(c) density

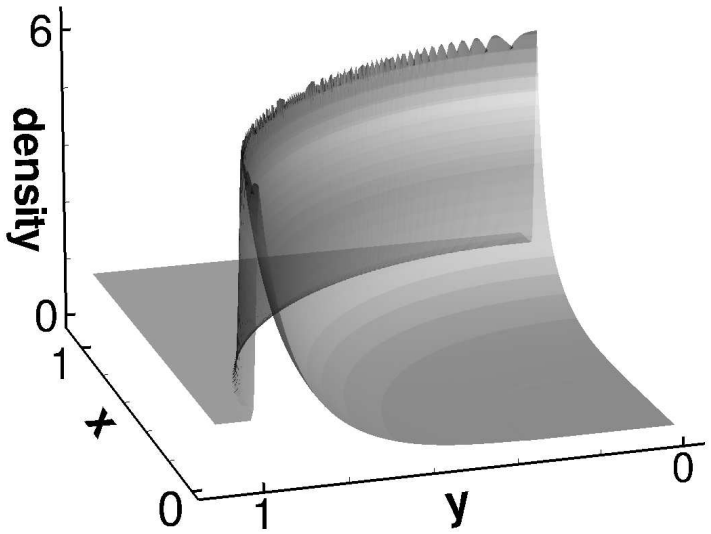
Figure 4.1: 1D Sedov blast. The solid line is the exact solution. Symbols are numerical solutions. $T = 0.001$. $N = 800$. $\Delta x = \frac{4}{N}$. TVB limiter parameters $(M_1, M_2, M_3) = (20000, 20000, 20000)$.



(a) Projection to radical coordinates. The solid line is the exact solution. Symbols are numerical solutions.



(b) 20 equally spaced contour lines from 0 to 5.5.



(c) Surface.

Figure 4.2: 2D Sedov blast, plot of density. $T = 1$. $N = 160$. $\Delta x = \Delta y = \frac{1.1}{N}$. TVB limiter parameters $(M_1, M_2, M_3, M_4) = (8000, 16000, 16000, 8000)$.

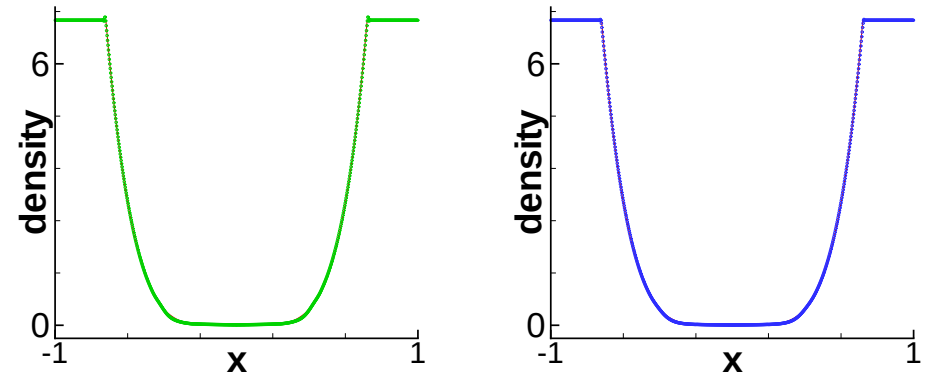
limiter. We can see that the low pressure and the low density are both captured very well. Without the positivity limiter, the DG scheme will blow up for this example. Even though the TVB limiter can make it stable, the result is poor compared to the one in Figure 4.3.

The second one is a 1D Leblanc shock tube problem. The initial condition is $\rho_L = 2$, $\rho_R = 0.001$, $u_L = u_R = 0$, $p_L = 10^9$, $p_R = 1$, and $\gamma = 1.4$. See Figure 4.4 for the results of 800 cells and 6400 cells. We tune up the TVB limiter parameter and find good results for $(M_1, M_2, M_3) = (10^{10}, 10, 10^{10})$ when $N = 800$ and $(M_1, M_2, M_3) = (10^{20}, 100, 10^{20})$ when $N = 6400$. There is a contact discontinuity in the exact solution, which is governed by the second characteristic field in the hyperbolic system. That explains why M_2 is different from M_1 and M_3. Without the positivity limiter, the DG scheme is unstable for such large values of M_i. The computational results for the DG scheme using only the TVB limiter with small M_i are much worse than those in Figure 4.4.

4.4 High Mach number astrophysical jets

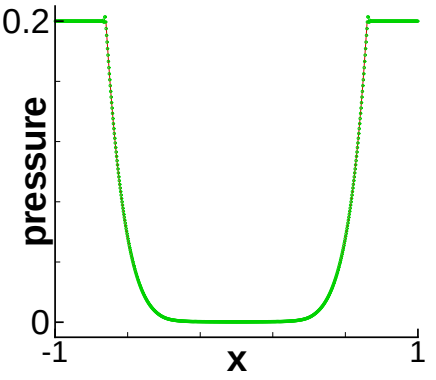
To simulate the gas flows and shock wave patterns which are revealed by the Hubble Space Telescope images, one can implement theoretical models in a gas dynamics simulator, see [9, 8, 7]. For example, the two-dimensional model without radiative cooling is governed by (3.1). The velocity of the gas flow is extremely high, and the Mach number could be hundreds or thousands. A big challenge for computation is, even for a state-of-the-art high order scheme solving (3.1), negative pressure could appear since the internal energy is very small compared to the huge kinetic energy. Therefore, we have a strong motivation to use the positivity limiter for this kind of problems.

First, we compute a Mach 80 (i.e. the Mach number of the jet inflow is 80 with respect to the soundspeed in the jet gas) problem without the radiative cooling in [9]. γ is set as 5/3. The computation domain is $[0, 2] \times [-0.5, 0.5]$, which is full of the ambient gas with $(\rho, u, v, p) = (5, 0, 0, 0.4127)$ initially. The boundary conditions for the right, top and bottom are outflow. For the left boundary, $(\rho, u, v, p) = (5, 30, 0, 0.4127)$ if $y \in [-0.05, 0.05]$

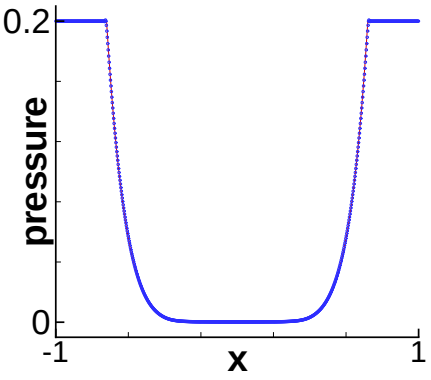


(a) density

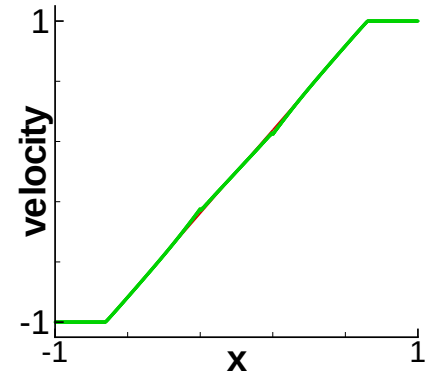
(b) density



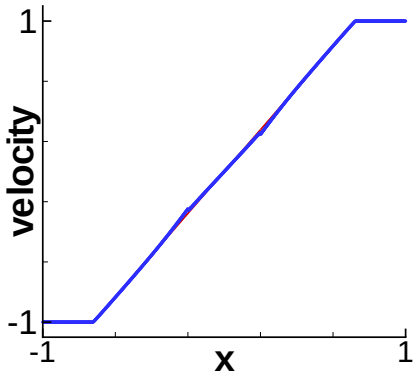
(c) pressure



(d) pressure

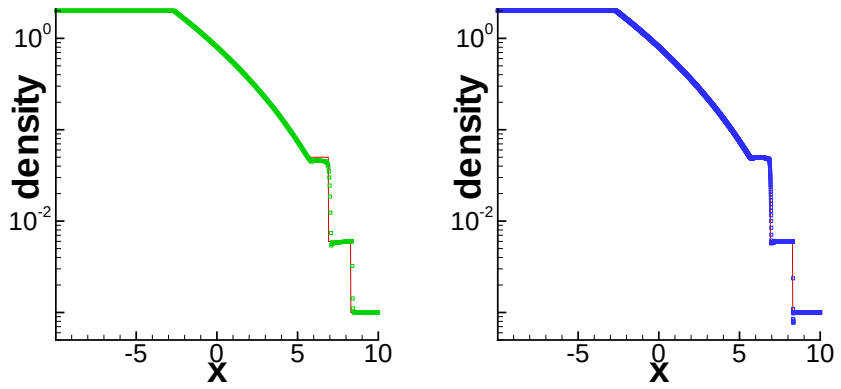


(e) velocity



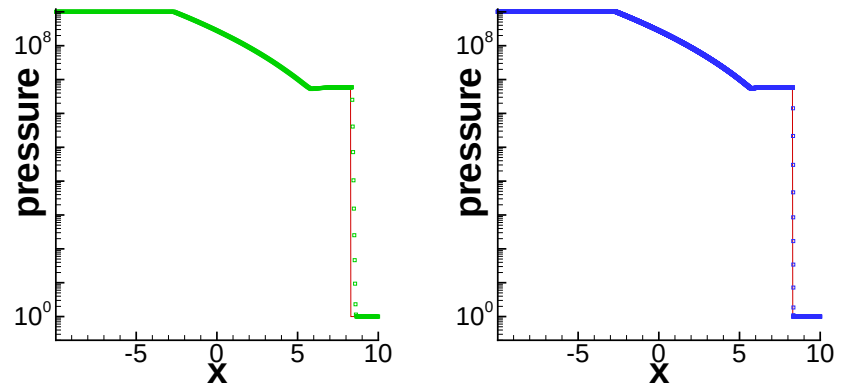
(f) velocity

Figure 4.3: Double rarefaction problem. T=0.6. Left: 1D problem. Right: Cut at $y = 0$ for the 2D problem. The solid line is the exact solution. Symbols are numerical solutions. $\Delta x = 0.01$, with the positivity limiter.



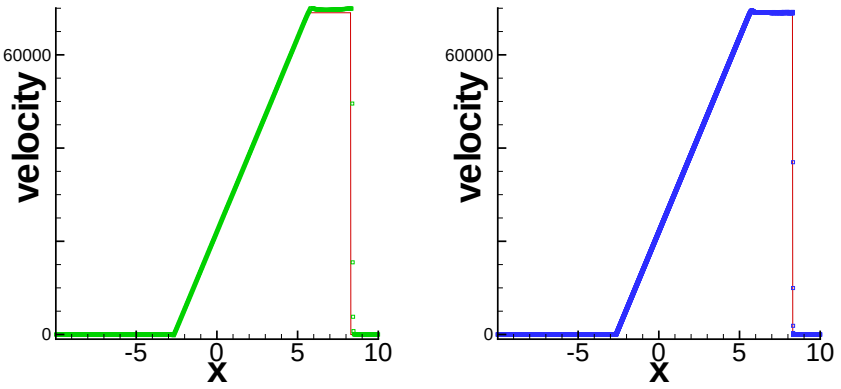
(a) log plot of density

(b) log plot of density



(c) log plot of pressure

(d) log plot of pressure



(e) velocity

(f) velocity

Figure 4.4: Leblanc problem. $T = 0.0001$. Left: $N = 800$. Right: $N = 6400$. The solid line is the exact solution. Symbols are numerical solutions. $\Delta x = \frac{20}{N}$ with the positivity limiter.

and $(\rho, u, v, p) = (5, 0, 0, 0.4127)$ otherwise. The terminal time is 0.07. The computation is performed on a 448×224 mesh. TVB limiter parameters are $M_1 = M_2 = M_3 = M_4 = 10000$. See Figure 4.5. The result is comparable to the one in [9].

Second, to demonstrate the robustness of our method, we compute a Mach 2000 problem. The basic scales of the computational units are the same as in [9]. The domain is $[0, 1] \times [-0.25, 0.25]$, initially full of the ambient gas with $(\rho, u, v, p) = (5, 0, 0, 0.4127)$. The boundary conditions for the right, top and bottom are outflow. For the left boundary, $(\rho, u, v, p) = (5, 800, 0, 0.4127)$ if $y \in [-0.05, 0.05]$ and $(\rho, u, v, p) = (5, 0, 0, 0.4127)$ otherwise. The terminal time is 0.001. The speed of the jet is 800, which is around Mach 2100 with respect to the soundspeed in the jet gas. The computation is performed on a 640×320 mesh. TVB limiter parameters are $M_1 = M_2 = M_3 = M_4 = 10000000$. See Figure 4.6.

4.5 Shock diffraction problem

Shock passing a backward facing corner (diffraction) has been used as a test problem for the DG method in [4]. It is easy to get negative density and/or pressure below and to the right of the corner. An ad hoc positivity correction procedure was used in [4] in order to avoid blow-ups of the DG code. Here we test our positivity preserving method. The setup is the following: the computational domain is the union of $[0, 1] \times [6, 11]$ and $[1, 13] \times [0, 11]$; the initial condition is a pure right-moving shock of $Mach = 5.09$, initially located at $x = 0.5$ and $6 \leq y \leq 11$, moving into undisturbed air ahead of the shock with a density of 1.4 and pressure of 1. The boundary conditions are inflow at $x = 0$, $6 \leq y \leq 11$, outflow at $x = 13$, $0 \leq y \leq 11$, $1 \leq x \leq 13$, $y = 0$ and $0 \leq x \leq 13$, $y = 11$, and reflective at the walls $0 \leq x \leq 1$, $y = 6$ and $x = 1$, $0 \leq y \leq 6$. $\gamma = 1.4$ and the TVB limiter parameters $M_i = 100$ for $i = 1, 2, 3, 4$. The density and pressure at $t = 2.3$ are presented in Figures 4.7 and 4.8. The result is comparable to the one in [4], with the advantage that negative pressure and density never appear.

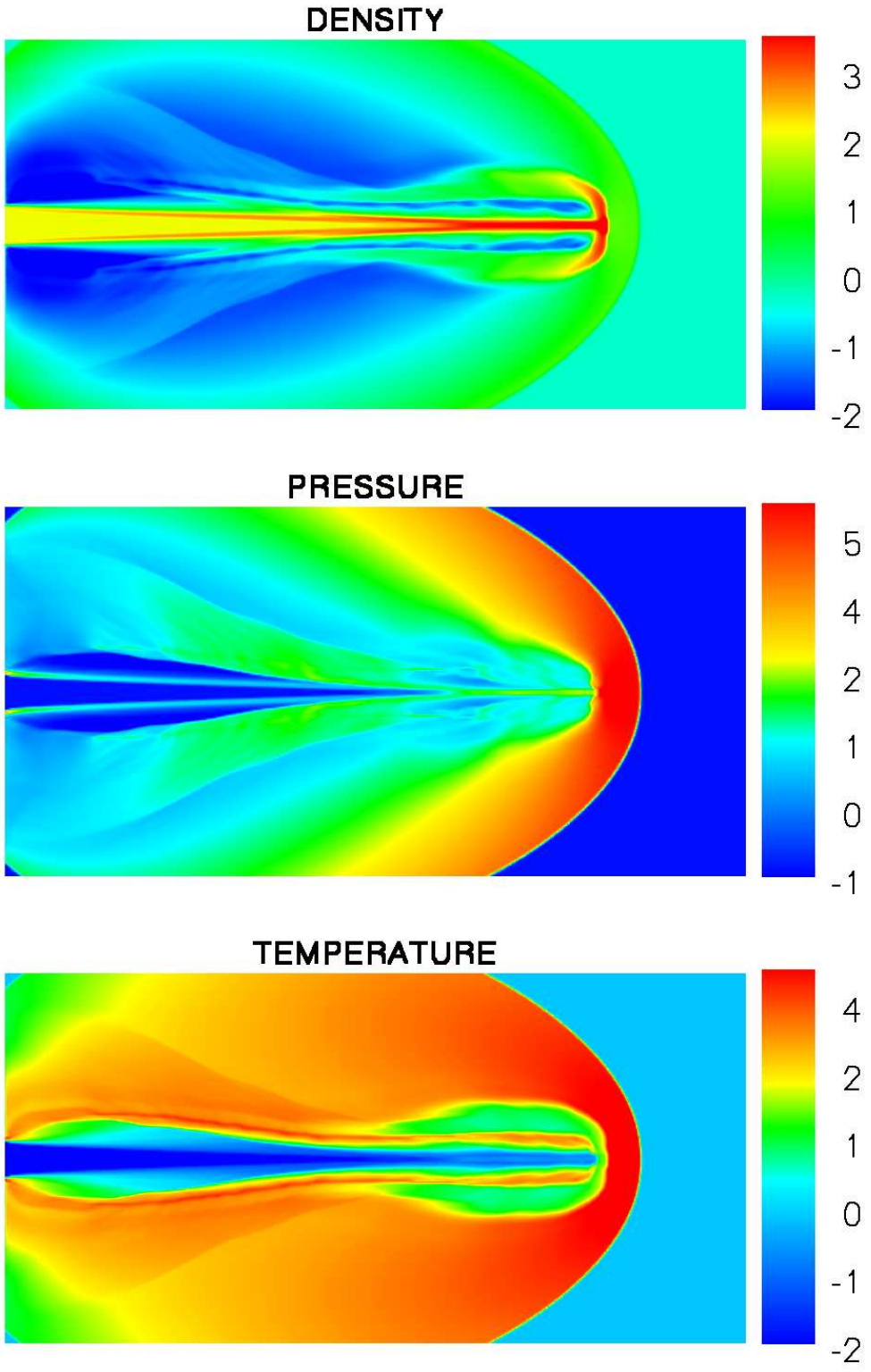


Figure 4.5: Simulation of Mach 80 jet without radiative cooling. Scales are logarithmic.

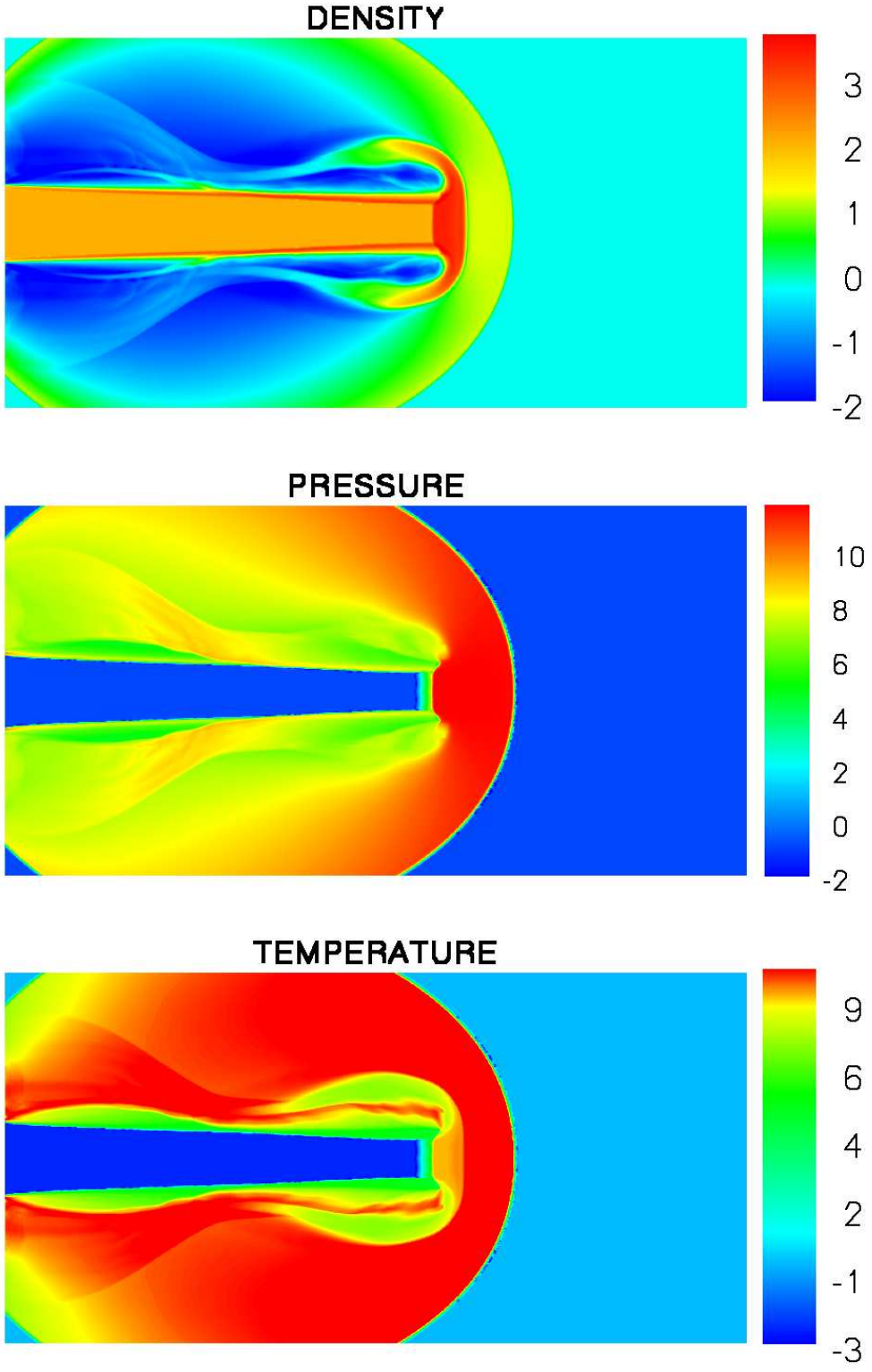
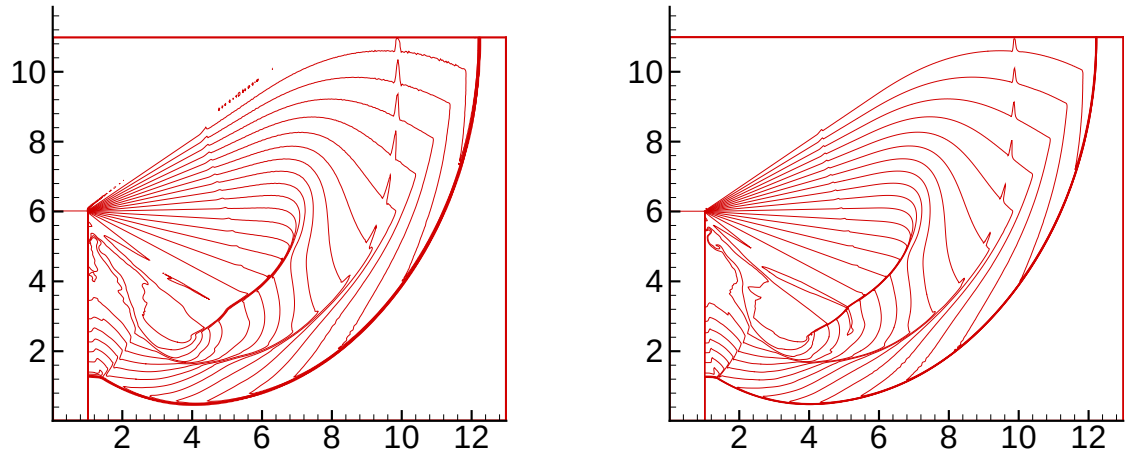
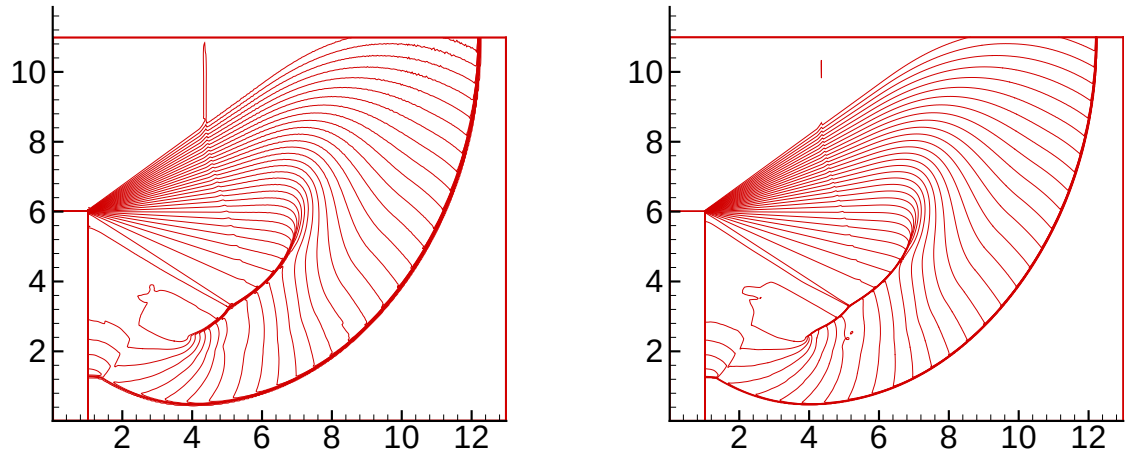


Figure 4.6: Simulation of Mach 2000 jet without radiative cooling. Scales are logarithmic.



(a) $\Delta x = \Delta y = 1/32$ (b) $\Delta x = \Delta y = 1/64$

Figure 4.7: Shock diffraction problem. Density: 20 equally spaced contour lines from $\rho = 0.066227$ to $\rho = 7.0668$.



(a) $\Delta x = \Delta y = 1/32$ (b) $\Delta x = \Delta y = 1/64$

Figure 4.8: Shock diffraction problem. Pressure: 40 equally spaced contour lines from $p = 0.091$ to $p = 37$.

5 Concluding remarks

In this paper, we improved the general framework in [20] to construct arbitrarily high order accurate finite volume and discontinuous Galerkin positivity preserving schemes for compressible Euler equations of gas dynamics in one dimension. An efficient and easy implementation is achieved by a simple linear scaling limiter, which can be proven to maintain high order accuracy and positivity of density and pressure. Then we extend the method to schemes on rectangular meshes for two dimensional Euler equations. Conceptually, extension of this technique to three dimensions is straightforward. With the addition of the positivity limiter in this paper, which involves small additional computational cost, to the DG scheme or the finite volume scheme (e.g. ENO and WENO), the numerical solutions will satisfy the positivity property in the sense that the density and pressure of the cell average are positive under suitable CFL condition.

We have tested the third order DG scheme with the positivity limiter on a variety of examples including two dimensional blast waves and high Mach astrophysical jets. Implementation details and numerical results for finite volume schemes using this positivity limiter will be reported elsewhere.

In this paper we have only shown a straightforward extension of the one dimensional algorithm to two dimensional finite volume or DG schemes on a rectangular mesh. For triangular meshes, the idea of rewriting the scheme as a convex combination of positive first order schemes is still plausible with the introduction of a special quadrature rule, see [28] for such an extension. For the astrophysical jets, the model with radiative cooling, i.e., (3.1) with a source term modeling the cooling effect, makes more sense physically, but it is more difficult to preserve the positivity. For radially-symmetric problems, e.g., the three-dimensional Sedov blast wave, we can consider the Euler systems in radially-symmetric form, which is in fact a one-dimensional system (1.1) with a source term. Generalizations to positivity preserving high order schemes for Euler systems with a source term also constitute our ongoing work.

References

[1] B. Cockburn, S. Hou and C.-W. Shu, *The Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws IV: the multidimensional case*, Mathematics of Computation, 54 (1990), 545-581.

[2] B. Cockburn, S.-Y. Lin and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws III: one dimensional systems*, Journal of Computational Physics, 84 (1989), 90-113.

[3] B. Cockburn and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws II: general framework*, Mathematics of Computation, 52 (1989), 411-435.

[4] B. Cockburn and C.-W. Shu, *The Runge-Kutta discontinuous Galerkin method for conservation laws V: multidimensional systems*, Journal of Computational Physics, 141 (1998), 199-224.

[5] B. Cockburn and C.-W. Shu, *Runge-Kutta discontinuous Galerkin methods for convection-dominated problems*, Journal of Scientific Computing, 16 (2001), 173-261.

[6] B. Einfeldt, C.D. Munz, P.L. Roe and B. Sjögreen, *On Godunov-Type methods near low densities*, Journal of Computational Physics, 92 (1991), 273-295.

[7] C. Gardner and S. Dwyer, *Numerical simulation of the XZ Tauri supersonic astrophysical jet*, Acta Mathemacita Scientia, 29B (2009), 1677-1683.

[8] Y. Ha and C. Gardner, *Positive scheme numerical simulation of high Mach number astrophysical jets*, Journal of Scientific Computing, 34 (2008), 247-259.

[9] Y. Ha, C. Gardner, A. Gelb and C.-W. Shu, *Numerical simulation of high Mach number astrophysical jets with radiative cooling*, Journal of Scientific Computing, 24 (2005), 597-612.

[10] A. Harten, *High resolution schemes for hyperbolic conservation laws*, Journal of Computational Physics, 49 (1983), 357-393.

[11] A. Harten, B. Engquist, S. Osher and S. Chakravarthy, *Uniformly high order essentially non-oscillatory schemes, III*, Journal of Computational Physics, 71 (1987), 231-303.

[12] A. Harten, P.D. Lax and B. van Leer, *On upstream differencing and Godunov type schemes for hyperbolic conservation laws*, SIAM Review, 25 (1983), 35-61.

[13] G.-S. Jiang and C.-W. Shu, *Efficient implementation of weighted ENO schemes*, Journal of Computational Physics, 126 (1996), 202-228.

[14] V.P. Korobeinikov, *Problems of Point-Blast Theory*, American Institute of Physics, 1991.

[15] T. Linde and P.L. Roe, *Robust Euler codes*, Thirteenth Computational Fluid Dynamics Conference, AIAA Paper-97-2098.

[16] W. Liu, J. Cheng and C.-W. Shu, *High order conservative Lagrangian schemes with Lax-Wendroff type time discretization for the compressible Euler equations*, Journal of Computational Physics, 228 (2009), 8872-8891.

[17] X.-D. Liu, S. Osher and T. Chan, *Weighted essentially non-oscillatory schemes*, Journal of Computational Physics, 115 (1994), 200-212.

[18] S. Osher and S. Chakravarthy, *High resolution schemes and the entropy condition*, SIAM Journal on Numerical Analysis, 21 (1984), 955-984.

[19] B. Perthame, *Second-order Boltzmann schemes for compressible Euler equations in one and two space dimensions*, SIAM Journal on Numerical Analysis, 29 (1992), 1-19.

[20] B. Perthame and C.-W. Shu, *On positivity preserving finite volume schemes for Euler equations*, Numerische Mathematik, 73 (1996), 119-130.

[21] R. Sanders, *A third-order accurate variation nonexpansive difference scheme for single nonlinear conservation law*, Mathematics of Computation, 51 (1988), 535-558.

[22] L.I. Sedov, *Similarity and Dimensional Methods in Mechanics*, Academic Press, New York, 1959.

[23] C.-W. Shu, *TVB uniformly high order schemes for conservation laws*, Mathematics of Computation, 49 (1987), 105-121.

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

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

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

[27] X. Zhang and C.-W. Shu, *A genuinely high order total variation diminishing scheme for one-dimensional scalar conservation laws*, SIAM Journal on Numerical Analysis, 48 (2010), 772-795.

[28] X. Zhang, Y. Xia and C.-W. Shu, *Maximum-principle-satisfying and positivity-preserving high order discontinuous Galerkin schemes for conservation laws on triangular meshes*, submitted to Journal of Scientific Computing.