On positivity-preserving high order discontinuous Galerkin schemes for compressible Navier-Stokes equations

Xiangxiong Zhang

Department of Mathematics, Purdue University,
150 N. University Street, West Lafayette, IN 47907-2067

Abstract

We construct a local Lax-Friedrichs type positivity-preserving flux for compressible Navier-Stokes equations, which can be easily extended to high dimensions for generic forms of equations of state, shear stress tensor and heat flux. With this positivity-preserving flux, any finite volume type schemes including discontinuous Galerkin (DG) schemes with strong stability preserving Runge-Kutta time discretizations satisfy a weak positivity property. With a simple and efficient positivity-preserving limiter, high order explicit Runge-Kutta DG schemes are rendered preserving the positivity of density and internal energy without losing local conservation or high order accuracy. Numerical tests suggest that the positivity-preserving flux and the positivity-preserving limiter do not induce excessive artificial viscosity, and the high order positivity-preserving DG schemes without other limiters can produce satisfying non-oscillatory solutions when the nonlinear diffusion in compressible Navier-Stokes equations is accurately resolved.

Keywords: discontinuous Galerkin method, high order accuracy, gas dynamics, compressible Navier-Stokes, positivity-preserving, high speed flows

1. Introduction

1.1. Gas dynamics equations in conservative form

The compressible viscous fluid dynamics equations without external forces in conservative form can be written as

$$\mathbf{U}_t + \nabla \cdot \mathbf{F}^a = \nabla \cdot \mathbf{F}^d. \tag{1}$$

The conservative variables are $\mathbf{U} = (\rho, \rho u, \rho v, \rho w, E)^t = (\rho, \rho \mathbf{u}^t, E)^t$ where ρ is the density, $\mathbf{u} = (u, v, w)^t$ denotes the velocity, E is the total energy and the

Email address: zhan1966@purdue.edu (Xiangxiong Zhang)

superscript t denotes transpose of a vector. The advection and diffusion fluxes are

$$\mathbf{F}^a = \begin{pmatrix} \rho\mathbf{u} \\ \rho\mathbf{u}\otimes\mathbf{u} + p\mathbb{I} \\ (E+p)\mathbf{u} \end{pmatrix}, \mathbf{F}^d = \begin{pmatrix} 0 \\ \boldsymbol{\tau} \\ \mathbf{u}\cdot\boldsymbol{\tau} - \mathbf{q} \end{pmatrix}$$

where p is pressure, $\mathbb{I}$ is the unit tensor, the shear stress tensor is

$$\boldsymbol{\tau} = \begin{pmatrix} \tau_{xx} & \tau_{xy} & \tau_{xz} \\ \tau_{yx} & \tau_{yy} & \tau_{yz} \\ \tau_{zx} & \tau_{zy} & \tau_{zz} \end{pmatrix},$$

and $\mathbf{q}$ denotes the heat diffusion flux. The total energy can be written as $E = \frac{1}{2}\rho\|\mathbf{u}\|^2 + \rho e$ where e denotes the internal energy. The relations between conserved variables $\mathbf{U}$ and temperature T and pressure p are given by equations of state (EOS).

With the Newtonian approximation, the shear stress tensor is given by $\boldsymbol{\tau} = \eta(\boldsymbol{\nabla}\mathbf{u} + (\boldsymbol{\nabla}\mathbf{u})^t) + (\eta_b - \frac{2}{3}\eta)(\nabla\cdot\mathbf{u})\mathbb{I}$ with coefficient of shear viscosity η and the coefficient of bulk viscosity η_b. The shear viscosity coefficient η strongly depends on the temperature T, e.g., Sutherland formula $\eta = \frac{C_1\sqrt{T}}{1+C_2/T}$ for a wide range of temperature. The dependence of η on pressure p cannot be neglected for high temperatures. The Stokes hypothesis states that $\eta_b = 0$, which can be used for monatomic gases however no longer holds for polyatomic gases [1].

With Fourier's heat conduction law, the heat flux is given by $\mathbf{q} = -\kappa\nabla T$ where the thermal conductivity coefficient κ is proportional to η in molecular theory. Assuming the specific heat at constant pressure c_p is a constant, the dimensionless quantity Prandtl number $\mathrm{Pr} \equiv \frac{c_p\eta}{\kappa}$ is a constant.

For simplicity, we will use the dimensionless form of equations for ideal gases as an example in this paper. Assuming $\eta_b = 0$, the dimensionless stress tensor is given by $\boldsymbol{\tau} = \frac{1}{\mathrm{Re}}(\boldsymbol{\nabla}\mathbf{u} + (\boldsymbol{\nabla}\mathbf{u})^t - \frac{2}{3}(\nabla\cdot\mathbf{u})\mathbb{I})$ where Re is the Reynolds number. For a calorically ideal gas one has $p = (\gamma-1)\rho e$ and $T = \frac{e}{c_v}$ where the specific heat capacity c_v and ratio of specific heats $\gamma = \frac{c_p}{c_v}$ are constants.

The following two-dimensional dimensionless compressible Navier-Stokes equations for ideal gases in [2] is considered here as an example even though the key discussions throughout this paper do not rely on the specific definitions of shear stress tensor, heat flux and pressure function:

$$\begin{pmatrix} \rho \\ \rho u \\ \rho v \\ E \end{pmatrix}_t + \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ (E+p)u \end{pmatrix}_x + \begin{pmatrix} \rho v \\ \rho vu \\ \rho v^2 + p \\ (E+p)v \end{pmatrix}_y = \frac{1}{\mathrm{Re}} \begin{pmatrix} 0 \\ \tau_{xx} \\ \tau_{yx} \\ \tau_{xx}u + \tau_{yx}v + \frac{\gamma e_x}{\mathrm{Pr}} \end{pmatrix}_x + \frac{1}{\mathrm{Re}} \begin{pmatrix} 0 \\ \tau_{xy} \\ \tau_{yy} \\ \tau_{xy}u + \tau_{yy}v + \frac{\gamma e_y}{\mathrm{Pr}} \end{pmatrix}_y, \tag{2a}$$

with

$$e = \frac{1}{\rho}\left(E - \frac{1}{2}\rho u^2 - \frac{1}{2}\rho v^2\right), \tag{2b}$$

$$p = (\gamma-1)\rho e, \tag{2c}$$

and

$$\begin{aligned}\tau_{xx} &= \frac{4}{3}u_x - \frac{2}{3}v_y,\\ \tau_{xy} &= \tau_{yx} = u_y + v_x,\\ \tau_{yy} &= \frac{4}{3}v_y - \frac{2}{3}u_x.\end{aligned} \tag{2d}$$

In this paper, we will use $\gamma = 1.4$ and $\mathrm{Pr} = 0.72$ for air.

1.2. Objective and motivation

For numerical schemes solving gas dynamics equations (1), it is imperative to preserve positivity of density and pressure (or internal energy). Not only is positivity-preserving needed for physically meaningful numerical solutions, but more importantly it is also well known to be critical for the sake of robustness of numerical computation.

For the ideal gas equation of state (2c), the eigen-values of $\frac{\partial \mathbf{F}^a}{\partial \mathbf{U}}$ contains the sound speed $\sqrt{\gamma p/\rho}$. With the presence of negative density or pressure, eigen-values of $\frac{\partial \mathbf{F}^a}{\partial \mathbf{U}}$ become imaginary thus the linearized compressible Euler system is no longer hyperbolic, which implies ill-posedness of the initial value problem of (1). In practice, emergence of negative density or pressure may easily result in blow-ups in numerical simulations by high order schemes. In some demanding problems involving low density or low pressure, e.g., high Mach number astrophysical jets, even popular robust high resolution and high order schemes including Monotonic Upstream-Centered Scheme for Conservation Laws (MUSCL) type schemes (without special positivity treatment) and classical Weighted Essentially Non-Oscillatory (WENO) schemes may blow up due to emergence of negative pressure [3].

The simplest ad-hoc approach of truncating negative values to zero or small positive ones may work in certain problems. But in a lot of other problems, especially high speed flows with presence of low pressure, the brutal truncation will eventually result in blow-ups because total mass or total energy is increased every time a negative density or pressure value is set to zero. In other words, both conservation and positivity must be satisfied for robustness. For example, a conservative finite volume scheme solving (1) satisfies the global conservation (up to proper boundary conditions) $\sum_i \overline{\rho}_i^n = \sum_i \overline{\rho}_i^{n+1}$, where $\overline{\rho}_i^n$ denotes the cell average of density on the i-th cell at time level n. If the scheme is positivity-preserving, then $|\overline{\rho}_i^n| = \overline{\rho}_i^n$ for any n and i. Thus global conservation and positivity imply L^1-stability: $\sum_i |\overline{\rho}_i^n| = \sum_i |\overline{\rho}_i^{n+1}|$. Similarly, we can have L^1-stability for total energy.

Towards robustness, it is desired to construct conservative schemes that are positivity-preserving of density and pressure (or internal energy). We will consider a general equation of state which satisfies $p > 0 \Longleftrightarrow e > 0$. Then we will focus on the positivity of internal energy instead of pressure because the internal energy has the same definition (2b) for any equation of state. We define

the set of admissible states as

$$G = \left\{ \mathbf{U} = \begin{pmatrix} \rho \\ \rho \mathbf{u} \\ E \end{pmatrix} : \rho > 0, \quad \rho e(\mathbf{U}) = E - \frac{1}{2}\rho \|\mathbf{u}\|^2 > 0. \right\}. \tag{3}$$

It is straightforward to check that the eigenvalues of the Hessian matrix $\frac{\partial^2}{\partial \mathbf{U}^2} \rho e$ are nonpositive if and only $\rho > 0$. Thus ρe is a concave function with respect to $\mathbf{U}$ and it satisfies a Jensen's inequality: $\forall \mathbf{U}_1, \mathbf{U}_2 \in G, \forall \lambda_1, \lambda_2 \geq 0, \lambda_1 + \lambda_2 = 1$,

$$\rho e(\lambda_1 \mathbf{U}_1 + \lambda_2 \mathbf{U}_2) \geq \lambda_1 \rho e(\mathbf{U}_1) + \lambda_2 \rho e(\mathbf{U}_2). \tag{4}$$

Therefore, G is a convex set.

The main objective in this paper is to construct conservative positivity-preserving high order accurate schemes solving (1), which is in general a difficult problem.

1.3. Positivity-preserving high order schemes for compressible Euler equations

Quite a few first order finite volume schemes are positivity-preserving for compressible Euler equations with EOS (2c), including the Godunov scheme, the Lax-Friedrichs scheme [4], and the HLLE [5, 6] schemes and kinetic schemes [7, 8]. Roughly speaking, to prove the positivity in Godunov and HLLE schemes, one must take advantage of the exact solution for Riemann problems, which is not available for a generic EOS. On the other hand, the positivity-preserving property of the Lax-Friedrichs scheme is an algebraic fact [9, 10]. Thus it is straightforward to show that the Lax-Friedrichs scheme is positivity-preserving for compressible Euler equations with a generic EOS [11].

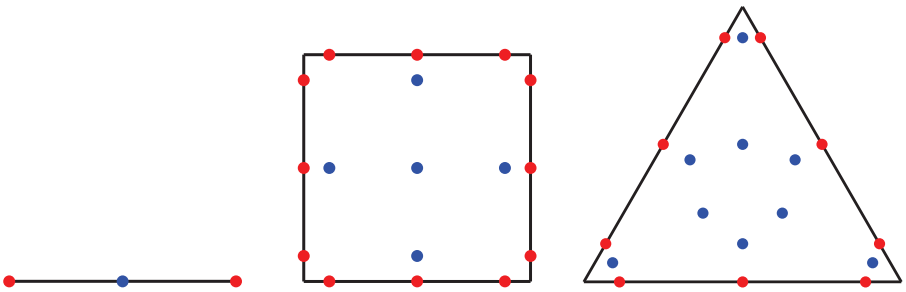


Figure 1: *An illustration of the weak monotonicity/positivity of third order finite volume and discontinuous Galerkin schemes using quadratic polynomials on 1D, 2D rectangular and triangular cells. Red points are quadrature points for computing the line integral of numerical fluxes along the cell boundary. Red points and blues points form a special quadrature and they are the points in the weak monotonicity/positivity.*

To construct positivity-preserving higher order accurate schemes, there are a handful of efforts in the literature, e.g., [4, 12, 13, 14, 15, 16]. One of the most successful approaches is the methodology proposed in [17, 10, 18, 19]. The details of this approach will be reviewed later. Here we first take a look at

the critical feature of this methodology, which is an intrinsic **weak positivity** property of arbitrarily high order finite volume and discontinuous Galerkin (DG) spatial discretizations. Let K be a polygonal cell with edges e_i $(i = 1, \cdots, E)$ in a two-dimensional mesh. Consider a high order finite volume scheme solving two-dimensional equations $\mathbf{U}_t + \nabla \cdot \mathbf{F}^a = \mathbf{0}$ on the cell K with forward Euler time discretization,

$$\overline{\mathbf{U}}_K^{n+1} = \overline{\mathbf{U}}_K^n - \frac{\Delta t}{|K|} \int_{\partial K} \widehat{\mathbf{F}^a \cdot \mathbf{n}}\, ds = \overline{\mathbf{U}}_K^n - \frac{\Delta t}{|K|} \sum_{i=1}^{E} \int_{e_i} \widehat{\mathbf{F}^a \cdot \mathbf{n}}\, ds, \tag{5}$$

where $\overline{\mathbf{U}}_K^n$ is the cell average on K at time step n, $\mathbf{n}$ is the unit outward normal vector to ∂K, $|K|$ denotes the area of K and $\widehat{\mathbf{F}^a \cdot \mathbf{n}}$ is a positivity-preserving flux. Positivity-preserving fluxes are those which make first order schemes positivity-preserving , e.g., Godunov, Lax-Friedrichs and HLLE fluxes, etc. For simplicity, we only consider a **local Lax-Friedrichs** flux for $\mathbf{F}^a$ in this paper,

$$\widehat{\mathbf{F}^a \cdot \mathbf{n}}(\mathbf{U}^-, \mathbf{U}^+)\Big|_{e_i} = \frac{1}{2}\left[\left(\mathbf{F}^a(\mathbf{U}^-) + \mathbf{F}^a(\mathbf{U}^+)\right) \cdot \mathbf{n} - \alpha_i(\mathbf{U}^+ - \mathbf{U}^-)\right], \tag{6}$$

where $\mathbf{U}^-$ and $\mathbf{U}^+$ denote the approximations to $\mathbf{U}$ on ∂K from interior of K and exterior of K respectively, and $\alpha_i = \max\left|\frac{\partial \mathbf{F}^a}{\partial \mathbf{U}}\right|$ with the maximum being taken over all $\mathbf{U}^-, \mathbf{U}^+$ along e_i. Here $\left|\frac{\partial \mathbf{F}^a}{\partial \mathbf{U}}\right|$ denotes the largest magnitude of the eigenvalues of the Jacobian matrix $\frac{\partial \mathbf{F}^a}{\partial \mathbf{U}}$, which is equal to the wave speed $|\mathbf{u} \cdot \mathbf{n}| + \sqrt{\gamma \frac{p}{\rho}}$ for the ideal gas EOS (2c).

The positivity property holds for the first order Lax-Friedrichs scheme under a CFL condition $\Delta t \frac{|\partial K|}{|K|} \max\left|\frac{\partial \mathbf{F}^a}{\partial \mathbf{U}}\right| \leq 2$ where $|\partial K|$ denotes the length of the boundary ∂K. In the high order scheme (5), $\overline{\rho}_K^{n+1}$ is not a monotone function of independent degree freedoms such as $\overline{\rho}_K^n$ and the boundary values of ρ_K^n along ∂K for any positive Δt, but under a suitable CFL condition $\overline{\rho}_K^{n+1}$ is a monotone function with respect to certain dependent degree of freedoms, e.g., some redundant point values of reconstruction or approximation polynomials illustrated in Figure 1. We call this property **weak monotonicity**, which was first explored in [17] for scalar conservation laws. Via Jensen's inequality (4), the weak monotonicity can be extended to weak positivity for pressure and internal energy: if the same set of point values as illustrated in Figure 1 for the conserved variable vector $\mathbf{U}$ belong to the set of admissible states G, then $\overline{\mathbf{U}}_K^{n+1} \in G$ in the scheme (5) using approximation polynomials of degree k under the CFL constraint

$$\Delta t \frac{|e_i|}{|K|} \alpha_i \leq a \frac{1}{N(N-1)}, \quad \forall i, \tag{7}$$

where $N = \lceil (k+3)/2 \rceil$, i.e., N is smallest integer satisfying $2N - 3 \geq k$, and $a = \frac{1}{2}$ for rectangular cells and $a = \frac{2}{3}$ for triangular cells in a two-dimensional case.

The weak positivity property implies that we only need to filter or limit negative point values as illustrated in Figure 1 to ensure the positivity of cell

averages in forward Euler, which makes construction of a conservative positivity-preserving high order scheme possible. A simple efficient scaling limiter can be used to modify negative point values to small positive ones without changing cell averages. For smooth solutions with a uniform positive lower bound on pressure $p \geq m > 0$, high order local truncation errors in spatial discretization of this limiter, can be rigorously justified. High order time accuracy can be achieved by Strong Stability Preserving (SSP) Runge-Kutta and multistep time discretizations [20], which are convex combinations of formal forward Euler schemes thus preserve the positivity if forward Euler is positivity-preserving.

In a nutshell, the key difference of the approach in [17, 10, 18] from all other positivity-preserving methods is the weak positivity, which allows not only rigorous justification of high order accuracy, but also easy constructions of explicit schemes with any order of accuracy, easy extensions to three dimensions [21] and general shapes of computational cells [22].

1.4. From Euler to Navier-Stokes: monotonicity in discrete Laplacian

To extend positivity-preserving schemes for Euler system to Navier-Stokes system, we only need to focus on the pressure or internal energy since the mass conservation equations in two systems are the same. However, it is much more difficult to guarantee positivity of the internal energy in compressible Navier-Stokes system. Positivity-preserving discretizations must be used for the nonlinear diffusion operator.

Consider a simple toy model $u_t = u_{xx}$. The simplest finite difference scheme solving it is $u_i^{n+1} = u_i^n + \frac{\Delta t}{\Delta x^2}(u_{i-1}^n - 2u_i^n + u_{i+1}^n)$. For that the right hand side of this scheme is a monotone function of u_i^n and $u_{i\pm1}^n$ if $\frac{\Delta t}{\Delta x^2} \leq \frac{1}{2}$, we call the second order accurate central difference $\frac{u_{i-1}-2u_i+u_{i+1}}{\Delta x^2}$ a *monotone* approximation to u_{xx}. By Taylor expansion, it is straightforward to prove that higher order accurate linear finite difference approximating the second order derivative cannot be *monotone*. In Appendix A, we will show that the second order central difference is positivity-preserving for the one-dimensional form of (2). However, it is difficult to extend positivity of this scheme to high dimensions since finite difference approximations for mixed second order derivatives are much more complicated.

In other words, it is already nontrivial to preserve the positivity of internal energy for second order schemes without losing conservation of total energy in high dimensions. There are few such results in the literature. In [23], an unconditionally stable staggered pressure correction second order accurate positivity-preserving implicit scheme was constructed for (2). The positivity in this scheme is heavily dependent upon the monotonicity of second order discrete Laplacian and the specific form of shear stress tensor (2d).

It is interesting to explore any weak monotonicity in diffusion discretizations. Unfortunately, weak monotonicity holds only up to second order accuracy for any linear finite volume scheme and most DG schemes [24], see Appendix D. Surprisingly, it is still possible to construct a third order linear DG scheme satisfying the weak monotonicity. With special parameters, the direct DG (DDG) method,

which is a generalized version of interior penalty DG method, indeed satisfies a weak monotonicity up to third order accuracy [25, 26, 27, 28]. However, if we use Taylor expansion to examine the local truncation error in the numerical flux of this scheme, only second order accuracy is obtained. Nonetheless, third order error estimate in the semi-discrete DDG scheme can be proven [29, 30, 31]. This phenomenon of inconsistency between the orders of local truncation error and actual error in DG methods is referred as *supraconvergence* [32], different from superconvergence. To fully understand how the second order errors are canceled out, we need the error estimate for the fully discretized scheme, which is not available. On the other hand, the supraconvergence in DDG method satisfying the weak monotonicity does not seem to reach beyond third order accuracy. One possible approach to construct positivity-preserving schemes for Navier-Stokes is to take advantage of the weak monotonicity in DDG. However, it is still quite difficult to extend the weak monotonicity to weak positivity of internal energy in high dimensions.

1.5. A different perspective: a positivity-preserving flux for nonlinear diffusion

In all approaches mentioned in the previous subsection, we regard $\nabla \cdot \mathbf{F}^d$ as a diffusion term when seeking monotonicity. Unfortunately, monotonicity and the weak monotonicity hold for finite difference, finite volume and most DG discretizations approximating Laplacian only up to second order accuracy. On the other hand, we can regard $\mathbf{F} = \mathbf{F}^a - \mathbf{F}^d$ as a single flux and formally treat $\nabla \cdot \mathbf{F}$ as a convection. This is perhaps a more natural perspective since the system (1) is derived from integral equations in the first place: $\iint_K \mathbf{U}_t \, dV = -\int_{\partial K} \mathbf{F} \cdot \mathbf{n} \, ds$. A finite volume scheme with forward Euler approximating this integral equation takes the form,

$$\overline{\mathbf{U}}_K^{n+1} = \overline{\mathbf{U}}_K^n - \frac{\Delta t}{|K|} \int_{\partial K} \widehat{\mathbf{F} \cdot \mathbf{n}} \, ds. \tag{8}$$

If $\widehat{\mathbf{F} \cdot \mathbf{n}}$ is a positivity-preserving flux, then (8) would satisfy the same weak positivity of pressure as for (5). The major challenge now boils down to construction of a positivity-preserving flux $\widehat{\mathbf{F} \cdot \mathbf{n}}$.

In this paper, we introduce a simple positivity-preserving flux for the nonlinear diffusion in the Navier-Stokes system, for which the design is inspired by the positivity-preserving property of the Lax-Friedrichs flux for Euler equations. Recall that the local Lax-Friedrichs flux (6) is positivity-preserving with $\alpha_i = \max\left(|\mathbf{u} \cdot \mathbf{n}| + \sqrt{\gamma \frac{p}{\rho}}\right)$ for the ideal gas EOS (2c). For a generic EOS, we can use $\alpha_i > \max\left(|\mathbf{u} \cdot \mathbf{n}| + \sqrt{\frac{p^2}{2\rho^2 e}}\right)$ to ensure positivity of internal energy, which will be reviewed.

Since the shear stress tensor and heat flux in the diffusion flux $\mathbf{F}^d$ are dependent on the derivatives of conserved variables $\mathbf{U}$, we introduce an auxiliary variable $\mathbf{S}$ as an approximation to $\nabla \mathbf{U}$. Now consider the following Lax-Friedrichs

type flux for the diffusion in Navier-Stokes equations,

$$\widehat{\mathbf{F}^d \cdot \mathbf{n}}(\mathbf{U}^-, \mathbf{S}^-, \mathbf{U}^+, \mathbf{S}^+)\Big|_{e_i} = \frac{1}{2}\left[\left(\mathbf{F}^d(\mathbf{U}^-, \mathbf{S}^-) + \mathbf{F}^d(\mathbf{U}^+, \mathbf{S}^+)\right) \cdot \mathbf{n} + \beta_i(\mathbf{U}^+ - \mathbf{U}^-)\right], \tag{9a}$$

where β_i is a nonnegative number. We will show that this flux is positivity-preserving if

$$\beta_i > \max \frac{1}{2\rho^2 e}\left(\sqrt{\rho^2|\mathbf{q}\cdot\mathbf{n}|^2 + 2\rho^2 e\|\boldsymbol{\tau}\cdot\mathbf{n}\|^2} + \rho|\mathbf{q}\cdot\mathbf{n}|\right), \tag{9b}$$

where the maximum is taken over $\mathbf{U}^-, \mathbf{S}^-, \mathbf{U}^+, \mathbf{S}^+$ along e_i. Here a positivity-preserving flux $\widehat{\mathbf{F}^d \cdot \mathbf{n}}$ means that a first order finite volume scheme with such a flux for solving the formal diffusion system $\mathbf{U}_t = \nabla \cdot \mathbf{F}^d(\mathbf{U}, \mathbf{S})$ is positivity-preserving.

Then we have a positivity-preserving flux in (8): $\widehat{\mathbf{F} \cdot \mathbf{n}} = \widehat{\mathbf{F}^a \cdot \mathbf{n}} - \widehat{\mathbf{F}^d \cdot \mathbf{n}}$ where $\widehat{\mathbf{F}^a \cdot \mathbf{n}}$ is any positivity-preserving flux for compressible Euler system and $\widehat{\mathbf{F}^d \cdot \mathbf{n}}$ is given in (9). Another slightly different positivity-preserving flux will be introduced in Section 2.4. Following the results in [10, 18], it becomes straightforward to show exactly the same weak positivity as illustrated in Figure 1 holds for (8) thus it is straightforward to construct positivity-preserving arbitrarily high order finite volume and DG schemes for (1).

1.6. Positivity-preserving DG schemes for compressible Navier-Stokes equations

In this paper, we discuss the construction of positivity-preserving DG schemes. For solving the compressible Navier-Stokes system (1), there are quite a few DG type schemes, e.g., the pioneering work by Bassi and Rebay [33, 34], the scheme by Baumann and Oden [35], Compact DG [36], correction procedure via reconstruction (CPR) [37, 38], Hybrid DG [39] and Embedded DG [40], etc. The major differences among these DG methods include how to approximate the derivatives of the solution and the choice of numerical fluxes, which are derived from various perspectives.

As a demonstration, we focus on one of the most popular approaches in [33]. For the derivatives of $\mathbf{U}$, an auxiliary variable $\mathbf{S} = \nabla\mathbf{U}$ is introduced. After multiplying the following system by test functions and integration by parts on a polygonal cell K,

$$\begin{cases} \mathbf{S} - \nabla\mathbf{U} = 0 \\ \mathbf{U}_t + \nabla \cdot \mathbf{F}^a(\mathbf{U}) - \nabla \cdot \mathbf{F}^d(\mathbf{U}, \mathbf{S}) = 0 \end{cases},$$

we obtain the following general form of a DG scheme,

$$\begin{cases} \displaystyle\iint_K \mathbf{S}_h \phi_h \, dV = -\iint_K \mathbf{U}_h \nabla\phi_h \, dV + \int_{\partial K} \widehat{\mathbf{U}\mathbf{n}}\phi_h \, ds \\ \displaystyle\iint_K \frac{d}{dt}\mathbf{U}_h \psi_h \, dV = \iint_K (\mathbf{F}^a - \mathbf{F}^d)\nabla\cdot\psi_h \, dV - \int_{\partial K} (\widehat{\mathbf{F}^a \cdot \mathbf{n}} - \widehat{\mathbf{F}^d \cdot \mathbf{n}})\psi_h \, ds \end{cases}, \tag{10}$$

where $\mathbf{U}_h$ and $\mathbf{S}_h$ are vectors of polynomials of degree k on K and ϕ_h and ψ_h are the polynomial of degree k test functions. The advection flux $\widehat{\mathbf{F}^a \cdot \mathbf{n}}$ is the same ones as mentioned above, e.g., the local Lax-Friedrichs flux. Central fluxes were used for the other two fluxes in [33], $\widehat{\mathbf{F}^d \cdot \mathbf{n}}(\mathbf{U}^-, \mathbf{S}^-, \mathbf{U}^+, \mathbf{S}^+) = \frac{1}{2}\left[\mathbf{F}^d(\mathbf{U}^-, \mathbf{S}^-) + \mathbf{F}^d(\mathbf{U}^+, \mathbf{S}^+)\right] \cdot \mathbf{n}$ and

$$\widehat{\mathbf{U}\mathbf{n}}(\mathbf{U}^-, \mathbf{U}^+) = \frac{1}{2}(\mathbf{U}^- + \mathbf{U}^+)\mathbf{n}, \tag{11}$$

where $-$ and $+$ denote the approximations on ∂K from interior of K and exterior of K respectively.

To render this high order DG scheme satisfying the weak positivity property, we can simply replace the central flux $\widehat{\mathbf{F}^d \cdot \mathbf{n}}$ by (9). Compared to the central flux, the extra term $\frac{1}{2}\beta(\mathbf{U}^+ - \mathbf{U}^-)$ in (9) contributes to the DG scheme (10) as an interior penalty term. In other words, we can add a nonlinear penalty defined by (9b) to the DG scheme in [33] so that it satisfies a weak positivity property under some CFL constraint. A slightly different positivity-preserving flux discussed in Section 2.4 can also be used to achieve the weak positivity property.

1.7. CFL constraints, implementation, and numerical performance

For the positivity-preserving flux (6) and (9) solving (1) with a generic EOS, the following time step constraint is a sufficient condition to ensure the weak positivity in a high order finite volume scheme (8),

$$\Delta t \frac{|e_i|}{|K|} \max\{\alpha_i, \beta_i\} \leq \frac{1}{2} a \frac{1}{N(N-1)}, \quad \forall i, \tag{12}$$

where $a = \frac{1}{2}$ for rectangular cells and $a = \frac{2}{3}$ for triangular cells.

To better understand (12), consider DG schemes solving a very smooth solution of one-dimensional form of (2). The linear stability on a simplified model $u_t = \frac{1}{\mathrm{Re}} u_{xx}$ would require $\Delta t = \mathcal{O}(\mathrm{Re}\,\Delta x^2)$ for an explicit scheme where Δx denotes the mesh size in the spatial discretization, whereas (12) roughly reduces to $\Delta t = \mathcal{O}(\Delta x)$. When shocks are present, (12) are roughly around $\Delta t = \mathcal{O}(\mathrm{Re}\,\Delta x^2)$. The inconsistency between the linear stability CFL and a nonlinear stability CFL (12) for smooth solutions implies that we also need to enforce the linear stability CFL beyond a positivity induced CFL. After all, the weak positivity property is a very weak stability, i.e., only cell averages are guaranteed to be positive in one time step in (8).

Numerical tests suggest that the linear stability CFL constraint $\Delta t = \mathcal{O}(\mathrm{Re}\,\Delta x^2)$ must be satisfied. Otherwise errors may grow exponentially even though the L^1 stability for density and total energy is still valid. There is no contradiction since stability itself does not guarantee convergence for nonlinear equations. In this paper, we only pursue nonlinear stability by enforcing positivity of density and internal energy in high order schemes for Navier-Stokes equations. Another approach towards nonlinear stability of high order schemes is to enforce entropy

bounds [41, 42, 43]. For convergence of high order schemes, entropy inequalities should be considered, which is in general a much harder problem than stability. Nonetheless, positivity is the first step towards entropy stability and entropy inequalities.

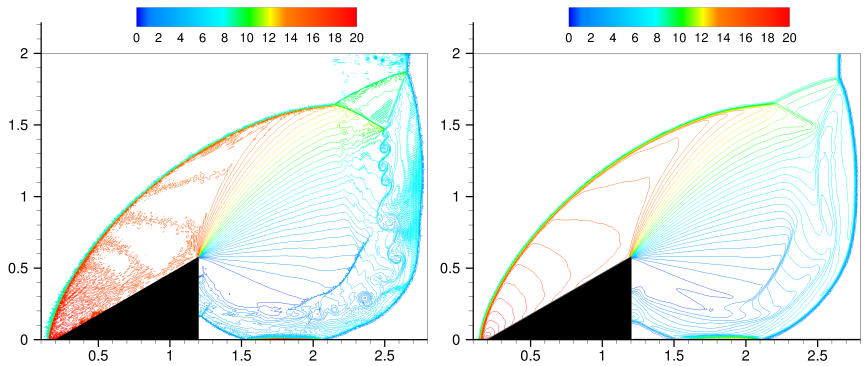
On the other hand, (7) and (12) should be not strictly enforced in Runge-Kutta time discretizations for several reasons:

1. The constraint (12) is hardly a necessary condition for $\overline{\mathbf{U}}_K^{n+1} \in G$ in (8) in practice.
2. It is very difficult to accurately predict wave speeds for all stages in a Runge-Kutta time discretization. For example, to enforce (7), we need to estimate $\max(|u| + \sqrt{\gamma p/\rho})$ in all inner stages before computing them in a high order Runge-Kutta time discretization. Similar difficulty arises in (12).
3. *Artificial stiffness* may arise in low density or low pressure problems. For instance, if density and internal energy are numerically close to zero, it is difficult to accurately evaluate the sound speed $\sqrt{\gamma \frac{p}{\rho}}$ due to the round-off errors. The wave speed computed in a low density region might be significantly larger than the actual one, which results in a much smaller time step numerically computed by (7) than a necessary time step for positivity. Similar difficulty arises in (12) due to the presence of density and internal energy in the denominator in (9b).

Instead, we can enforce the time step constraint (12) only when larger time steps give $\overline{\mathbf{U}}_K^{n+1} \notin G$. For each time step of Runge-Kutta discretization, we can start with a usually used time step for explicit schemes, e.g., $\Delta t = \mathcal{O}(\mathrm{Re}\,\Delta x^2)$. If density or internal energy of the cell average becomes negative, then we restart the computation with a smaller time step, e.g., a time step halved. The sufficiency of CFL (12) for the weak positivity excludes the possibility of endless loops. This ad hoc implementation was used in [44] for Euler equations with chemical reactions and works well in practice.

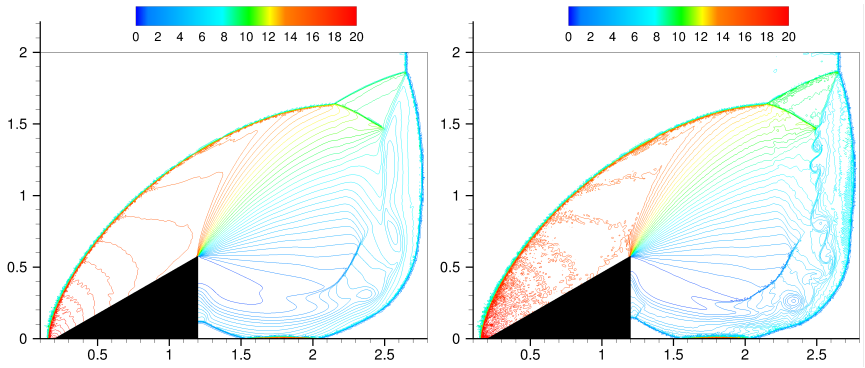
To implement an explicit positivity-preserving high order DG scheme for (1), we can do the following simple modifications to the scheme in [33],

1. Use SSP Runge-Kutta time discretizations.
2. Use a positivity-preserving flux for the advection, e.g., (6). Use the positivity-preserving flux (9) for the nonlinear diffusion. In other words, add an interior penalty term to the scheme in [33] with nonlinear penalty parameter defined by (9b).
3. Add a simple limiter to filter negative point values at quadrature points, e.g., Gauss quadrature, for computing all integrals in the DG scheme (10). We emphasize that we do not use the quadrature in Figure 1 to compute any integral in the actual implementation. The blue points in Figure 1 are highly redundant in high dimensions and are not used explicitly in the implementation.
4. Time stepping will be discussed in detail in Section 5.3.



(a) Without other limiters. $\mathrm{Re} = \infty$. (b) With TVB limiter. $\mathrm{Re} = \infty$.

Figure 2: Mach 10 shock passing a sharp corner. The DG scheme with positivity-preserving local Lax-Friedrichs flux (6), the positivity-preserving limiter, and the third order SSP Runge-Kutta on unstructured triangular meshes solving compressible Euler equations with ideal gas EOS. The mesh size is $\frac{1}{80}$. P^2 basis.



(a) Without other limiters. $\mathrm{Re} = 100$. (b) Without other limiters. $\mathrm{Re} = 1000$.

Figure 3: Mach 10 shock passing a sharp corner. The DG scheme with positivity-preserving fluxes (6) and (9), the positivity-preserving limiter, and the third order SSP Runge-Kutta on unstructured triangular meshes solving Navier-Stokes equations (2). The mesh size is $\frac{1}{80}$. P^2 basis.

To have a glance at the numerical performance of the high order DG scheme implemented as above, we consider the problem of Mach 10 shock diffracted at a sharp corner in [18]. This is quite a representative test problem since shocks, low density/pressure and fine structures from Kelvin-Helmholtz instability are all involved. See Figure 2 for the results of compressible Euler equations. The positivity-preserving RKDG scheme may produce oscillations, which is not a surprise because only positivity is guaranteed. To reduce oscillations, other type of limiters towards a non-oscillatory property should be used. For instance, a TVB limiter was used in [18]. We emphasize that other types of high order accurate limiters without the positivity-preserving limiter cannot stabilize the DG scheme for this low pressure problem. On the other hand, we can observe that the TVB limiter also smears out some interesting features such as roll-ups due to the Kelvin-Helmholtz instability, which is an indication of excessive numerical viscosity of the TVB limiter. For compressible Euler equations, the RKDG scheme with only the positivity-preserving limiter may produce excessive oscillations affecting the shock locations on finer mesh or with higher order basis for this problem. WENO type limiters [45, 46, 47] are less diffusive thus more suitable for reducing oscillations with a better resolution.

However, a very interesting observation is that **no other limiters are needed to reduce oscillations for positivity-preserving DG scheme solving Navier-Stokes equations**. See Figure 3 for results of DG with only the positivity-preserving limiter solving compressible Navier-Stokes equations (2). In Figure 2 (a), we can see that the local Lax-Friedrichs flux for advection and positivity-preserving limiter do not contribute excessive artificial viscosity. In Figure 3 (a), the main source of additional artificial viscosity is the extra term $\frac{1}{2}\beta(\mathbf{U}^+ - \mathbf{U}^-)$ in (9). On the other hand, the result for higher Reynolds number in Figure 3 (b) is a strong evidence that the flux (9) does not contribute excessive artificial viscosity either. With these observations and numerical evidence, we may conclude that the physical nonlinear diffusion starts to smooth out the numerical oscillations in high order positivity-preserving DG schemes when it is accurately resolved.

1.8. Contributions and organization of the paper

The main contributions of this paper include the construction of positivity-preserving flux (9) for the nonlinear diffusion in (1) and the construction and implementation of the very first high order schemes for compressible Navier-Stokes equations (1) in two dimensions, which can preserve positivity of internal energy without losing the conservation of total energy. For implementing positivity-preserving DG schemes, it is straightforward to add a nonlinear penalty term and the positivity-preserving limiter to the DG scheme in [33]. The interior penalty plays an essential role in stabilizing DG method solving diffusion problems. We have revealed how it may affect the nonlinear stability in terms of the positivity-preserving property for compressible Navier-Stokes equations.

An interesting and important observation is that only the positivity-preserving limiter is needed for high order DG schemes to produce non-oscillatory solutions

for compressible Navier-Stokes equations (2) even when strong shocks are involved, which is not the case for compressible Euler equations.

The high order positivity-preserving scheme in this paper has the following advantages and features:

- The construction of the positivity-preserving flux (9) depends on neither the specific form of the shear stress tensor $\boldsymbol{\tau}$, heat flux $\mathbf{q}$ and equations of state nor how the derivatives $\nabla \mathbf{U}$ are approximated in a scheme.
- Extensions to arbitrarily high order polynomial basis, multiple dimensions and unstructured meshes are straightforward.
- The full scheme is explicit with the time step constraint $\Delta t = \mathcal{O}(\mathrm{Re}\,\Delta x^2)$ thus it is more suitable for high Reynolds number flows.
- For compressible Navier-Stokes equations, only positivity-preserving fluxes and the positivity-preserving limiter are needed to stabilize the high order DG scheme producing non-oscillatory solutions. Numerical evidence suggests that the proposed high order positivity-preserving DG scheme does not produce excessive artificial viscosity.

The paper is organized as follows: we review the weak positivity property of high order finite volume schemes solving compressible Euler equations in one dimension then introduce two positivity-preserving fluxes for compressible Navier-Stokes equations in one dimension in Section 2. We discuss the weak positivity property of high order finite volume schemes with a positivity-preserving flux for compressible Navier-Stokes equations in two dimensions in Section 3. For compressible Euler equations, the CFL condition for triangular cells derived in Section 3 is slightly better than the one in [18]. Then we review the positivity-preserving limiter and an efficient implementation in Section 4. We emphasize that all the discussions in Sections 2, 3 and 4 apply to any finite volume type scheme including the scheme satisfied by cell averages in DG methods. Implementation details for DG schemes are discussed in Section 5. Numerical tests are given in Section 6. Section 7 consists of concluding remarks.

2. The weak positivity of high order schemes in one dimension

We will first review the positivity-preserving flux and the weak positivity for high order finite volume schemes solving compressible Euler equations. Then we construct two positivity-preserving fluxes for compressible Navier-Stokes equations and discuss the weak positivity for high order finite volume schemes.

2.1. The positivity of first order schemes for compressible Euler equations

For the one dimensional compressible Euler system $\mathbf{U}_t + \mathbf{F}^a(\mathbf{U})_x = \mathbf{0}$ with $\mathbf{U} = (\rho, \rho u, E)^t$ and the flux function in one dimension defined by $\mathbf{F}^a =$

$(\rho u, \rho u^2 + p, (E+p)u)^t$, we first consider a first order finite volume scheme on a cell $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$ with cell size Δx,

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - \frac{\Delta t}{\Delta x}\left[\widehat{\mathbf{F}^a}(\mathbf{U}_j^n, \mathbf{U}_{j+1}^n) - \widehat{\mathbf{F}^a}(\mathbf{U}_{j-1}^n, \mathbf{U}_j^n)\right],$$

with the local Lax-Friedrichs flux defined by

$$\widehat{\mathbf{F}^a}(\mathbf{U}_j^n, \mathbf{U}_{j+1}^n) = \frac{1}{2}\left[\mathbf{F}^a(\mathbf{U}_j^n) + \mathbf{F}^a(\mathbf{U}_{j+1}^n) - \alpha_{j+\frac{1}{2}}(\mathbf{U}_{j+1}^n - \mathbf{U}_j^n)\right), \tag{13}$$

where $\mathbf{U}_j^n$ is the approximation to the average of $\mathbf{U}$ on I_j at time level n and $\alpha_{j+\frac{1}{2}}$ is a positive number dependent upon $\mathbf{U}_j^n$ and $\mathbf{U}_{j+1}^n$. With the assumption $\mathbf{U}_j^n, \mathbf{U}_{j\pm1}^n \in G$, we want to find a proper $\alpha_{j+\frac{1}{2}}$ and a CFL condition so that $\mathbf{U}_j^{n+1} \in G$. The scheme can be rewritten as

$$\begin{aligned}\mathbf{U}_j^{n+1} = &\left(1 - \frac{1}{2}\alpha_{j-\frac{1}{2}}\frac{\Delta t}{\Delta x} - \frac{1}{2}\alpha_{j+\frac{1}{2}}\frac{\Delta t}{\Delta x}\right)\mathbf{U}_j^n + \frac{1}{2}\alpha_{j+\frac{1}{2}}\frac{\Delta t}{\Delta x}\left(\mathbf{U}_{j+1}^n - \alpha_{j+\frac{1}{2}}^{-1}\mathbf{F}^a(\mathbf{U}_{j+1}^n)\right)\\ &+ \frac{1}{2}\alpha_{j-\frac{1}{2}}\frac{\Delta t}{\Delta x}\left(\mathbf{U}_{j-1}^n + \alpha_{j-\frac{1}{2}}^{-1}\mathbf{F}^a(\mathbf{U}_{j-1}^n)\right).\end{aligned} \tag{14}$$

By Lemma 6 in Appendix B, if we set $\alpha_{j+\frac{1}{2}} > \max\limits_{\mathbf{U}_j^n, \mathbf{U}_{j+1}^n}\left(|u| + \sqrt{\frac{p^2}{2\rho^2 e}}\right)$, e.g., $\alpha_{j+\frac{1}{2}} = \max\limits_{\mathbf{U}_j^n, \mathbf{U}_{j+1}^n}\left(|u| + \sqrt{\gamma\frac{p}{\rho}}\right)$ for the ideal gas EOS (2c) with $\gamma > 0$, then we have $\mathbf{U}_{j+1}^n - \alpha_{j+\frac{1}{2}}^{-1}\mathbf{F}^a(\mathbf{U}_{j+1}^n) \in G$, and $\mathbf{U}_{j-1}^n + \alpha_{j-\frac{1}{2}}^{-1}\mathbf{F}^a(\mathbf{U}_{j-1}^n) \in G$. Under the CFL condition $\frac{\Delta t}{\Delta x}\max_j \alpha_{j+\frac{1}{2}} \le 1$, $\mathbf{U}_j^{n+1}$ in (14) is a convex combination of three vectors in G thus $\mathbf{U}_j^{n+1} \in G$.

2.2. The weak positivity of high order schemes for compressible Euler equations

Consider a $(k+1)$-th order finite volume type scheme with reconstruction polynomials or approximation polynomials of degree k. For one dimensional compressible Euler system with forward Euler time discretization on I_j, it takes the form,

$$\overline{\mathbf{U}}_j^{n+1} = \overline{\mathbf{U}}_j^n - \frac{\Delta t}{\Delta x}\left[\widehat{\mathbf{F}^a}(\mathbf{U}_{j+\frac{1}{2}}^-, \mathbf{U}_{j+\frac{1}{2}}^+) - \widehat{\mathbf{F}^a}(\mathbf{U}_{j-\frac{1}{2}}^-, \mathbf{U}_{j-\frac{1}{2}}^+)\right], \tag{15a}$$

where $\overline{\mathbf{U}}_j^n$ is the cell average on I_j at time level n, $\mathbf{U}_{j+\frac{1}{2}}^-$ and $\mathbf{U}_{j+\frac{1}{2}}^+$ are approximations to the point value of $\mathbf{U}$ at $x_{j+\frac{1}{2}}$ and time level n from the left and from the right respectively. The local Lax-Friedrichs flux is

$$\widehat{\mathbf{F}^a}(\mathbf{U}_{j+\frac{1}{2}}^-, \mathbf{U}_{j+\frac{1}{2}}^+) = \frac{1}{2}\left[\mathbf{F}^a(\mathbf{U}_{j+\frac{1}{2}}^-) + \mathbf{F}^a(\mathbf{U}_{j+\frac{1}{2}}^+) - \alpha_{j+\frac{1}{2}}(\mathbf{U}_{j+\frac{1}{2}}^+ - \mathbf{U}_{j+\frac{1}{2}}^-)\right], \tag{15b}$$

$$\alpha_{j+\frac{1}{2}} > \max_{\mathbf{U}_{j+\frac{1}{2}}^-, \mathbf{U}_{j+\frac{1}{2}}^+}\left(|u| + \sqrt{\frac{p^2}{2\rho^2 e}}\right). \tag{15c}$$

Let $N = \lceil (k+3)/2 \rceil$, i.e., N is smallest integer satisfying $2N-3 \geq k$. We consider an 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 integrals of polynomials of degree up to $2N-3$. Denote these quadrature points on I_j as

$$S_j = \{x_{j-\frac{1}{2}} = \widehat{x}_j^1, \widehat{x}_j^2, \cdots, \widehat{x}_j^{N-1}, \widehat{x}_j^N = x_{j+\frac{1}{2}}\}. \tag{16}$$

Let $\widehat{\omega}_\mu$ be the quadrature weights for the interval $[-\frac{1}{2}, \frac{1}{2}]$ such that $\sum_{\mu=1}^{N} \widehat{\omega}_\mu = 1$. Let $\mathbf{P}_j(x)$ be the reconstruction or approximation polynomials of degree k in the high order scheme (15a) on I_j with cell average $\overline{\mathbf{U}}_j^n$ and nodal values $\mathbf{U}_{j+\frac{1}{2}}^-$ and $\mathbf{U}_{j-\frac{1}{2}}^+$ at two cell ends, then

$$\overline{\mathbf{U}}_j^n = \frac{1}{\Delta x}\int_{I_j} \mathbf{P}_j(x)\,dx = \sum_{\mu=1}^{N} \widehat{\omega}_\mu \mathbf{P}_j(\widehat{x}_j^\mu) = \sum_{\mu=2}^{N-1} \widehat{\omega}_\mu \mathbf{P}_j(\widehat{x}_j^\mu) + \widehat{\omega}_1 \mathbf{U}_{j-\frac{1}{2}}^+ + \widehat{\omega}_N \mathbf{U}_{j+\frac{1}{2}}^-. \tag{17}$$

By plugging (17) into the scheme (15a), we obtain

$$\begin{aligned}
\overline{\mathbf{U}}_j^{n+1} =& (\widehat{\omega}_1 - \frac{1}{2}\frac{\Delta t}{\Delta x}\alpha_{j-\frac{1}{2}})\left(\mathbf{U}_{j-\frac{1}{2}}^+ + \frac{1}{2}\frac{\Delta t}{\Delta x}(\widehat{\omega}_1 - \frac{1}{2}\frac{\Delta t}{\Delta x}\alpha_{j-\frac{1}{2}})^{-1}\mathbf{F}^a(\mathbf{U}_{j-\frac{1}{2}}^+)\right)\\
&+ (\widehat{\omega}_N - \frac{1}{2}\frac{\Delta t}{\Delta x}\alpha_{j+\frac{1}{2}})\left(\mathbf{U}_{j+\frac{1}{2}}^- - \frac{1}{2}\frac{\Delta t}{\Delta x}(\widehat{\omega}_N - \frac{1}{2}\frac{\Delta t}{\Delta x}\alpha_{j+\frac{1}{2}})^{-1}\mathbf{F}^a(\mathbf{U}_{j+\frac{1}{2}}^-)\right)\\
&+ \frac{1}{2}\frac{\Delta t}{\Delta x}\alpha_{j-\frac{1}{2}}\left(\mathbf{U}_{j-\frac{1}{2}}^- + \alpha_{j-\frac{1}{2}}^{-1}\mathbf{F}^a(\mathbf{U}_{j-\frac{1}{2}}^-)\right)\\
&+ \frac{1}{2}\frac{\Delta t}{\Delta x}\alpha_{j+\frac{1}{2}}\left(\mathbf{U}_{j+\frac{1}{2}}^+ - \alpha_{j+\frac{1}{2}}^{-1}\mathbf{F}^a(\mathbf{U}_{j+\frac{1}{2}}^+)\right)\\
&+ \sum_{\mu=2}^{N-1} \widehat{\omega}_\mu \mathbf{P}_j(\widehat{x}_j^\mu).
\end{aligned} \tag{18}$$

Let $\widehat{\omega}$ denote the smallest weight in $\widehat{\omega}_\mu$, i.e., $\widehat{\omega} = \widehat{\omega}_1 = \widehat{\omega}_N$. For Gauss-Lobatto quadrature with N points, $\widehat{\omega} = \frac{1}{N(N-1)}$. Notice the fact that $\frac{\Delta t}{\Delta x}\alpha \leq \widehat{\omega}$ if and only if $0 \leq \frac{1}{2}\frac{\Delta t}{\Delta x}(\widehat{\omega} - \frac{1}{2}\frac{\Delta t}{\Delta x}\alpha)^{-1} \leq \alpha^{-1}$ for positive numbers $\frac{\Delta t}{\Delta x}, \alpha$ and $\widehat{\omega}$. By Lemma 6 in Appendix B, under the CFL condition $\frac{\Delta t}{\Delta x}\max_j \alpha_{j+\frac{1}{2}} \leq \widehat{\omega}$, we have

$$\mathbf{U}_{j-\frac{1}{2}}^-, \mathbf{U}_{j+\frac{1}{2}}^+ \in G \Rightarrow \mathbf{U}_{j-\frac{1}{2}}^- + \alpha_{j-\frac{1}{2}}^{-1}\mathbf{F}(\mathbf{U}_{j-\frac{1}{2}}^-), \mathbf{U}_{j+\frac{1}{2}}^+ - \alpha_{j+\frac{1}{2}}^{-1}\mathbf{F}(\mathbf{U}_{j+\frac{1}{2}}^+) \in G,$$

$$\mathbf{U}_{j-\frac{1}{2}}^+ \in G \Rightarrow \mathbf{U}_{j-\frac{1}{2}}^+ + \frac{1}{2}\frac{\Delta t}{\Delta x}(\widehat{\omega}_1 - \frac{1}{2}\frac{\Delta t}{\Delta x}\alpha_{j-\frac{1}{2}})^{-1}\mathbf{F}^a(\mathbf{U}_{j-\frac{1}{2}}^+) \in G,$$

$$\mathbf{U}_{j+\frac{1}{2}}^- \in G \Rightarrow \mathbf{U}_{j+\frac{1}{2}}^- - \frac{1}{2}\frac{\Delta t}{\Delta x}(\widehat{\omega}_N - \frac{1}{2}\frac{\Delta t}{\Delta x}\alpha_{j+\frac{1}{2}})^{-1}\mathbf{F}^a(\mathbf{U}_{j+\frac{1}{2}}^-) \in G.$$

Moreover, (18) is a convex combination under the same CFL condition. Thus we get the *weak positivity* for the high order scheme (15),

Theorem 1. *A sufficient condition for* $\overline{\mathbf{U}}_j^{n+1} \in G$ *in the scheme (15) with reconstruction or approximation polynomials of degree* k *is*

$$\mathbf{U}^{\pm}_{j\pm\frac{1}{2}} \in G \quad \textit{and} \quad \mathbf{P}_j(\widehat{x}_j^{\mu}) \in G \ (\mu = 2, \cdots, N-1), \quad \forall j, \tag{19}$$

under the CFL condition

$$\frac{\Delta t}{\Delta x} \max_j \alpha_{j+\frac{1}{2}} \le \widehat{\omega} = \frac{1}{N(N-1)}, \quad N = \lceil (k+3)/2 \rceil . \tag{20}$$

Remark 1. If using $\sum_{\mu=2}^{N-1} \widehat{\omega}_\mu \mathbf{P}_j(\widehat{x}_j^{\mu}) = (1-\widehat{\omega}_1-\widehat{\omega}_N) \sum_{\mu=2}^{N-1} \frac{\widehat{\omega}_\mu}{1-\widehat{\omega}_1-\widehat{\omega}_N} \mathbf{P}_j(\widehat{x}_j^{\mu}) = (1-2\widehat{\omega}) \sum_{\mu=2}^{N-1} \frac{\widehat{\omega}_\mu}{1-2\widehat{\omega}} \mathbf{P}_j(\widehat{x}_j^{\mu})$ in (18), we obtain a weaker sufficient condition than (19) for $\overline{\mathbf{U}}_j^{n+1} \in G$ in the scheme (15),

$$\mathbf{U}^{\pm}_{j\pm\frac{1}{2}} \in G \quad \text{and} \quad \sum_{\mu=2}^{N-1} \frac{\widehat{\omega}_\mu}{1-2\widehat{\omega}} \mathbf{P}_j(\widehat{x}_j^{\mu}) \in G, \quad \forall j, \tag{21}$$

under the same CFL constraint (20).

2.3. A positivity-preserving flux for the nonlinear diffusion

It is straightforward to construct a first order positivity-preserving scheme for the one-dimensional form of (2) since mixed second order derivatives do not appear in the one-dimensional equations. As a matter of fact, the simplest second order finite difference scheme is positivity-preserving for the one-dimensional compressible Navier-Stokes equations, see Appendix A. However, it is difficult to extend such a result to two dimensions due to the mixed second order derivatives. In this subsection, we introduce a simple positivity-preserving flux for the diffusion flux $\mathbf{F}^d$ which can be easily extended to high dimensions.

We first formally consider the diffusion system $\mathbf{U}_t - \mathbf{F}^d(\mathbf{U}, \mathbf{S})_x = \mathbf{0}$ with $\mathbf{S} = \mathbf{U}_x$ and the flux function in one dimension defined by $\mathbf{F}^d = (0, \tau, u\tau - q)^t$. Let $\mathbf{S}_j^n$ denote the approximation $\mathbf{U}_x$ on I_j at time step n. Consider a first order finite volume scheme on I_j,

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n + \frac{\Delta t}{\Delta x} \left[\widehat{\mathbf{F}^d}(\mathbf{U}_j^n, \mathbf{S}_j^n, \mathbf{U}_{j+1}^n, \mathbf{S}_{j+1}^n) - \widehat{\mathbf{F}^d}(\mathbf{U}_{j-1}^n, \mathbf{S}_{j-1}^n, \mathbf{U}_j^n, \mathbf{S}_j^n) \right].$$

Consider a Lax-Friedrichs type flux defined by

$$\widehat{\mathbf{F}^d}(\mathbf{U}_j^n, \mathbf{S}_j^n, \mathbf{U}_{j+1}^n, \mathbf{S}_{j+1}^n) = \frac{1}{2} \left[\mathbf{F}^d(\mathbf{U}_j^n, \mathbf{S}_j^n) + \mathbf{F}^d(\mathbf{U}_{j+1}^n, \mathbf{S}_{j+1}^n) + \beta_{j+\frac{1}{2}} (\mathbf{U}_{j+1}^n - \mathbf{U}_j^n) \right], \tag{22a}$$

where $\beta_{j+\frac{1}{2}}$ is positive number dependent upon $\mathbf{U}_j^n$, $\mathbf{S}_j^n$, $\mathbf{U}_{j+1}^n$ and $\mathbf{S}_{j+1}^n$. With the assumption $\mathbf{U}_j^n, \mathbf{U}_{j\pm1}^n \in G$, we want to find a proper $\beta_{j+\frac{1}{2}}$ and a CFL condition so that $\mathbf{U}_j^{n+1} \in G$. By Lemma 6 in Appendix B, if we set

$$\beta_{j+\frac{1}{2}} > \max_{\mathbf{U}_j^n, \mathbf{S}_j^n, \mathbf{U}_{j+1}^n, \mathbf{S}_{j+1}^n} \frac{1}{2\rho^2 e} \left(\sqrt{\rho^2 q^2 + 2\rho^2 e |\tau|^2} + \rho |q| \right), \tag{22b}$$

then we have $\mathbf{U}^n_{j+1}+\beta^{-1}_{j+\frac{1}{2}}\mathbf{F}^d(\mathbf{U}^n_{j+1},\mathbf{S}^n_{j+1})\in G$, and $\mathbf{U}^n_{j-1}-\beta^{-1}_{j-\frac{1}{2}}\mathbf{F}^d(\mathbf{U}^n_{j-1},\mathbf{S}^n_{j-1})\in G$. The scheme can be rewritten as

$$\begin{aligned}\mathbf{U}^{n+1}_j=&\left(1-\frac{1}{2}\beta_{j-\frac{1}{2}}\frac{\Delta t}{\Delta x}-\frac{1}{2}\beta_{j+\frac{1}{2}}\frac{\Delta t}{\Delta x}\right)\mathbf{U}^n_j+\frac{1}{2}\beta_{j+\frac{1}{2}}\frac{\Delta t}{\Delta x}\left(\mathbf{U}^n_{j+1}+\beta^{-1}_{j+\frac{1}{2}}\mathbf{F}^d(\mathbf{U}^n_{j+1},\mathbf{S}^n_{j+1})\right)\\&+\frac{1}{2}\beta_{j-\frac{1}{2}}\frac{\Delta t}{\Delta x}\left(\mathbf{U}^n_{j-1}-\beta^{-1}_{j-\frac{1}{2}}\mathbf{F}^d(\mathbf{U}^n_{j-1},\mathbf{S}^n_{j-1})\right).\end{aligned}$$

Under the CFL condition $\frac{\Delta t}{\Delta x}\max_j\beta_{j+\frac{1}{2}}\le 1$, $\mathbf{U}^{n+1}_j$ in the scheme above is a convex combination of three vectors in G thus $\mathbf{U}^{n+1}_j\in G$.

For the compressible Navier-Stokes equations $\mathbf{U}_t+\mathbf{F}^a(\mathbf{U})_x-\mathbf{F}^d(\mathbf{U},\mathbf{S})_x=\mathbf{0}$, it is straightforward to construct a first order positivity-preserving scheme,

$$\begin{aligned}\mathbf{U}^{n+1}_j=&\frac{1}{2}\left(\mathbf{U}^n_j-2\frac{\Delta t}{\Delta x}\left[\widehat{\mathbf{F}^a}(\mathbf{U}^n_j,\mathbf{U}^n_{j+1})-\widehat{\mathbf{F}^a}(\mathbf{U}^n_{j-1},\mathbf{U}^n_j)\right]\right)\\&+\frac{1}{2}\left(\mathbf{U}^n_j+2\frac{\Delta t}{\Delta x}\left[\widehat{\mathbf{F}^d}(\mathbf{U}^n_j,\mathbf{S}^n_j,\mathbf{U}^n_{j+1},\mathbf{S}^n_{j+1})-\widehat{\mathbf{F}^d}(\mathbf{U}^n_{j-1},\mathbf{S}^n_{j-1},\mathbf{U}^n_j,\mathbf{S}^n_j)\right]\right),\end{aligned}$$

where $\widehat{\mathbf{F}^a}$ is any positivity-preserving flux for Euler system, e..g, Godunov and HLLE fluxes, and $\widehat{\mathbf{F}^d}$ is the flux constructed above. The right hand side of this scheme is an average of the two formal schemes for $\mathbf{U}_t+\mathbf{F}^a(\mathbf{U})_x=\mathbf{0}$ and $\mathbf{U}_t-\mathbf{F}^d(\mathbf{U})_x=\mathbf{0}$. If $\widehat{\mathbf{F}^a}$ is (13) and $\widehat{\mathbf{F}^d}$ is (22), then the positivity-preserving CFL constraints are $\frac{\Delta t}{\Delta x}\max\left\{\alpha_{j+\frac{1}{2}},\beta_{j+\frac{1}{2}}\right\}\le\frac{1}{2}$ for all j.

2.4. Another positivity-preserving flux for compressible Navier-Stokes equations

By treating two fluxes $\mathbf{F}^a$ and $\mathbf{F}^d$ separately, we surely overlook the interaction of two fluxes, which may give more information thus possibly better CFL constraints. To this end, consider the following finite volume scheme for the compressible Navier-Stokes equations in the form $\mathbf{U}_t+\mathbf{F}(\mathbf{U},\mathbf{S})_x=\mathbf{0}$ with $\mathbf{F}=\mathbf{F}^a-\mathbf{F}^d$,

$$\mathbf{U}^{n+1}_j=\mathbf{U}^n_j-\frac{\Delta t}{\Delta x}\left[\widehat{\mathbf{F}}(\mathbf{U}^n_j,\mathbf{S}^n_j,\mathbf{U}^n_{j+1},\mathbf{S}^n_{j+1})-\widehat{\mathbf{F}}(\mathbf{U}^n_{j-1},\mathbf{S}^n_{j-1},\mathbf{U}^n_j,\mathbf{S}^n_j)\right].$$

Consider a Lax-Friedrichs type flux defined by

$$\widehat{\mathbf{F}}(\mathbf{U}^n_j,\mathbf{S}^n_j,\mathbf{U}^n_{j+1},\mathbf{S}^n_{j+1})=\frac{1}{2}\left[\mathbf{F}(\mathbf{U}^n_j,\mathbf{S}^n_j)+\mathbf{F}(\mathbf{U}^n_{j+1},\mathbf{S}^n_{j+1})-\beta_{j+\frac{1}{2}}(\mathbf{U}^n_{j+1}-\mathbf{U}^n_j)\right],\tag{23}$$

where $\beta_{j+\frac{1}{2}}$ is a positive number dependent upon $\mathbf{U}^n_j$,$\mathbf{S}^n_j$, $\mathbf{U}^n_{j+1}$ and $\mathbf{S}^n_{j+1}$. With the assumption $\mathbf{U}^n_j,\mathbf{U}^n_{j\pm1}\in G$, we want to find a proper $\beta_{j+\frac{1}{2}}$ and a CFL condition so that $\mathbf{U}^{n+1}_j\in G$. The scheme can be rewritten as

$$\begin{aligned}\mathbf{U}^{n+1}_j=&\left(1-\frac{1}{2}\beta_{j-\frac{1}{2}}\frac{\Delta t}{\Delta x}-\frac{1}{2}\beta_{j+\frac{1}{2}}\frac{\Delta t}{\Delta x}\right)\mathbf{U}^n_j+\frac{1}{2}\beta_{j+\frac{1}{2}}\frac{\Delta t}{\Delta x}\left(\mathbf{U}^n_{j+1}-\beta^{-1}_{j+\frac{1}{2}}\mathbf{F}(\mathbf{U}^n_{j+1},\mathbf{S}^n_{j+1})\right)\\&+\frac{1}{2}\beta_{j-\frac{1}{2}}\frac{\Delta t}{\Delta x}\left(\mathbf{U}^n_{j-1}+\beta^{-1}_{j-\frac{1}{2}}\mathbf{F}(\mathbf{U}^n_{j-1},\mathbf{S}^n_{j-1})\right).\end{aligned}\tag{24}$$

By Lemma 6 in Appendix B, if we set

$$\beta_{j+\frac{1}{2}} > \max_{\mathbf{U}_j^n,\mathbf{S}_j^n,\mathbf{U}_{j+1}^n,\mathbf{S}_{j+1}^n} \left[|u| + \frac{1}{2\rho^2 e}\left(\sqrt{\rho^2 q^2 + 2\rho^2 e|\tau - p|^2} + \rho|q|\right)\right],$$

then we have $\mathbf{U}_{j+1}^n - \beta_{j+\frac{1}{2}}^{-1}\mathbf{F}(\mathbf{U}_{j+1}^n, \mathbf{S}_{j+1}^n) \in G$, and $\mathbf{U}_{j-1}^n + \beta_{j-\frac{1}{2}}^{-1}\mathbf{F}(\mathbf{U}_{j-1}^n, \mathbf{S}_{j-1}^n) \in G$. Under the CFL condition $\frac{\Delta t}{\Delta x}\max_j \beta_{j+\frac{1}{2}} \le 1$, $\mathbf{U}_j^{n+1}$ in (24) is a convex combination of three vectors in G thus $\mathbf{U}_j^{n+1} \in G$.

2.5. The weak positivity of high order schemes for compressible Navier-Stokes equations

Consider a $(k+1)$-th order finite volume type scheme for one dimensional compressible Navier-Stokes system with forward Euler time discretization on I_j,

$$\overline{\mathbf{U}}_j^{n+1} = \overline{\mathbf{U}}_j^n - \frac{\Delta t}{\Delta x}\left[\widehat{\mathbf{F}}(\mathbf{U}_{j+\frac{1}{2}}^-, \mathbf{S}_{j+\frac{1}{2}}^-, \mathbf{U}_{j+\frac{1}{2}}^+, \mathbf{S}_{j+\frac{1}{2}}^+) - \widehat{\mathbf{F}}(\mathbf{U}_{j-\frac{1}{2}}^-, \mathbf{S}_{j-\frac{1}{2}}^-, \mathbf{U}_{j-\frac{1}{2}}^+, \mathbf{S}_{j-\frac{1}{2}}^+)\right], \tag{25a}$$

where $\widehat{\mathbf{F}}$ is a positivity-preserving flux. We can use either $\widehat{\mathbf{F}} = \widehat{\mathbf{F}^a} - \widehat{\mathbf{F}^d}$ with any positivity-preserving flux $\widehat{\mathbf{F}^a}$ for the Euler system and $\widehat{\mathbf{F}^d}$ as defined in (22a), or the flux $\widehat{\mathbf{F}}$ as as defined in (23). For simplicity, we only discuss the latter one,

$$\widehat{\mathbf{F}}(\mathbf{U}_{j+\frac{1}{2}}^-, \mathbf{S}_{j+\frac{1}{2}}^-, \mathbf{U}_{j+\frac{1}{2}}^+, \mathbf{S}_{j+\frac{1}{2}}^+) = \frac{1}{2}\left[\mathbf{F}(\mathbf{U}_{j+\frac{1}{2}}^-, \mathbf{S}_{j+\frac{1}{2}}^-) + \mathbf{F}(\mathbf{U}_{j+\frac{1}{2}}^+, \mathbf{S}_{j+\frac{1}{2}}^+) - \beta_{j+\frac{1}{2}}(\mathbf{U}_{j+\frac{1}{2}}^+ - \mathbf{U}_{j+\frac{1}{2}}^-)\right], \tag{25b}$$

$$\beta_{j+\frac{1}{2}} > \max_{\mathbf{U}_{j+\frac{1}{2}}^-,\mathbf{S}_{j+\frac{1}{2}}^-,\mathbf{U}_{j+\frac{1}{2}}^+,\mathbf{S}_{j+\frac{1}{2}}^+} \left[|u| + \frac{1}{2\rho^2 e}\left(\sqrt{\rho^2 q^2 + 2\rho^2 e|\tau - p|^2} + \rho|q|\right)\right]. \tag{25c}$$

Plugging the cell average decomposition (17) into the scheme (25a), we obtain

$$\begin{aligned}
\overline{\mathbf{U}}_j^{n+1} = {} & (\widehat{\omega}_1 - \frac{1}{2}\frac{\Delta t}{\Delta x}\beta_{j-\frac{1}{2}})\left(\mathbf{U}_{j-\frac{1}{2}}^+ + \frac{1}{2}\frac{\Delta t}{\Delta x}(\widehat{\omega}_1 - \frac{1}{2}\frac{\Delta t}{\Delta x}\beta_{j-\frac{1}{2}})^{-1}\mathbf{F}(\mathbf{U}_{j-\frac{1}{2}}^+, \mathbf{S}_{j-\frac{1}{2}}^+)\right) \\
& + (\widehat{\omega}_N - \frac{1}{2}\frac{\Delta t}{\Delta x}\beta_{j+\frac{1}{2}})\left(\mathbf{U}_{j+\frac{1}{2}}^- - \frac{1}{2}\frac{\Delta t}{\Delta x}(\widehat{\omega}_N - \frac{1}{2}\frac{\Delta t}{\Delta x}\beta_{j+\frac{1}{2}})^{-1}\mathbf{F}(\mathbf{U}_{j+\frac{1}{2}}^-, \mathbf{S}_{j+\frac{1}{2}}^-)\right) \\
& + \frac{1}{2}\frac{\Delta t}{\Delta x}\beta_{j-\frac{1}{2}}(\mathbf{U}_{j-\frac{1}{2}}^- + \beta_{j-\frac{1}{2}}^{-1}\mathbf{F}(\mathbf{U}_{j-\frac{1}{2}}^-, \mathbf{S}_{j-\frac{1}{2}}^-)) \\
& + \frac{1}{2}\frac{\Delta t}{\Delta x}\beta_{j+\frac{1}{2}}(\mathbf{U}_{j+\frac{1}{2}}^+ - \beta_{j+\frac{1}{2}}^{-1}\mathbf{F}(\mathbf{U}_{j+\frac{1}{2}}^+, \mathbf{S}_{j+\frac{1}{2}}^+)) \\
& + \sum_{\mu=2}^{N-1}\widehat{\omega}_\mu \mathbf{P}_j(\widehat{x}_j^\mu).
\end{aligned} \tag{26}$$

By Lemma 6 in Appendix B, under the CFL condition $\frac{\Delta t}{\Delta x}\max_j \beta_{j+\frac{1}{2}} \le \widehat{\omega}$, we have

$$\mathbf{U}_{j-\frac{1}{2}}^-, \mathbf{U}_{j+\frac{1}{2}}^+ \in G \Rightarrow \mathbf{U}_{j-\frac{1}{2}}^- + \beta_{j-\frac{1}{2}}^{-1}\mathbf{F}(\mathbf{U}_{j-\frac{1}{2}}^-, \mathbf{S}_{j-\frac{1}{2}}^-), \mathbf{U}_{j+\frac{1}{2}}^+ - \beta_{j+\frac{1}{2}}^{-1}\mathbf{F}(\mathbf{U}_{j+\frac{1}{2}}^+, \mathbf{S}_{j+\frac{1}{2}}^+) \in G,$$

$$\mathbf{U}^+_{j-\frac{1}{2}} \in G \Rightarrow \mathbf{U}^+_{j-\frac{1}{2}} + \frac{1}{2}\frac{\Delta t}{\Delta x}(\widehat{\omega}_1 - \frac{1}{2}\frac{\Delta t}{\Delta x}\beta_{j-\frac{1}{2}})^{-1}\mathbf{F}(\mathbf{U}^+_{j-\frac{1}{2}}, \mathbf{S}^+_{j-\frac{1}{2}}) \in G,$$

$$\mathbf{U}^-_{j+\frac{1}{2}} \in G \Rightarrow \mathbf{U}^-_{j+\frac{1}{2}} - \frac{1}{2}\frac{\Delta t}{\Delta x}(\widehat{\omega}_N - \frac{1}{2}\frac{\Delta t}{\Delta x}\beta_{j+\frac{1}{2}})^{-1}\mathbf{F}(\mathbf{U}^-_{j+\frac{1}{2}}, \mathbf{S}^-_{j+\frac{1}{2}}) \in G.$$

Moreover, (26) is a convex combination under the same CFL condition. Thus we get the *weak positivity* for the high order scheme (25),

Theorem 2. *A sufficient condition for* $\overline{\mathbf{U}}_j^{n+1} \in G$ *in the scheme (25) with reconstruction or approximation polynomials of degree* k *is*

$$\mathbf{U}^{\pm}_{j\pm\frac{1}{2}} \in G \quad \textit{and} \quad \mathbf{P}_j(\widehat{x}_j^{\mu}) \in G \ (\mu = 2, \cdots, N-1), \quad \forall j, \tag{27}$$

under the CFL condition

$$\frac{\Delta t}{\Delta x}\max_j \beta_{j+\frac{1}{2}} \le \widehat{\omega} = \frac{1}{N(N-1)}, \quad N = \lceil (k+3)/2 \rceil . \tag{28}$$

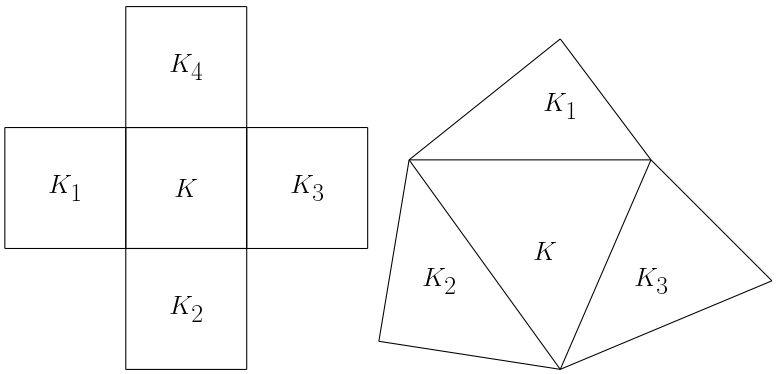
Following Remark 1, a weaker condition to replace (27) is

$$\mathbf{U}^{\pm}_{j\pm\frac{1}{2}} \in G \quad \textit{and} \quad \sum_{\mu=2}^{N-1}\frac{\widehat{\omega}_\mu}{1-2\widehat{\omega}}\mathbf{P}_j(\widehat{x}_j^{\mu}) \in G, \quad \forall j, \tag{29}$$

Remark 2. For a very smooth solution, $|u| + \frac{1}{2\rho^2 e}\left(\sqrt{\rho^2 q^2 + 2\rho^2 e|\tau - p|^2} + \rho|q|\right) = \mathcal{O}(1)$ thus the CFL constraint (28) implies $\Delta t = \mathcal{O}(\Delta x)$, which is inconsistent with the time step constraint suggested by a linear stability analysis. For explicit schemes solving (2), we should take $\Delta t = \mathcal{O}(\mathrm{Re}\,\Delta x^2)$ as implied by the linear stability analysis on a toy model $u_t = \frac{1}{\mathrm{Re}}u_{xx}$. The weak positivity property is a very weak nonlinear stability, which does not necessarily imply the linear stability. In numerical tests, if (28) is satisfied but $\Delta t = \mathcal{O}(\mathrm{Re}\,\Delta x^2)$ is violated, we observe that errors may grow exponentially in time for smooth solutions of (2), even though the weak positivity still holds.

3. The weak positivity of high order schemes in two dimensions

As we have seen in the previous section, the discussion of positivity of first order schemes and weak positivity of high order schemes for Euler equations is similar to the ones for the diffusion operator and the Navier-Stokes equations. For simplicity, we only discuss the positivity of the Lax-Friedrichs type flux (23) for the Navier-Stokes equations $\mathbf{U}_t + \nabla\cdot\mathbf{F}(\mathbf{U},\mathbf{S}) = \mathbf{0}$ with $\mathbf{S} = \nabla\mathbf{U}$ in two dimensions in this section. All discussions in this section also apply to the flux (13) for Euler equations $\mathbf{U}_t + \nabla\cdot\mathbf{F}^a(\mathbf{U}) = \mathbf{0}$ and the flux (22a) for the diffusion equations $\mathbf{U}_t - \nabla\cdot\mathbf{F}^d(\mathbf{U},\mathbf{S}) = \mathbf{0}$.



(a) A rectangular mesh. (b) An unstructured triangular mesh.

Figure 4: An illustration of a computational cell K in two dimensions.

3.1. The positivity of first order schemes

Let K be a polygonal cell with edges e_i $(i = 1, \cdots, E)$ in a mesh and K_i be the adjacent cell which shares the edge e_i with K. For simplicity, we only consider a rectangular mesh or an unstructured triangular mesh, e.g., as illustrated in Figure 4. Let $|e_i|$ denote the length of the edge e_i. For the two dimensional compressible Navier-Stokes system $\mathbf{U}_t + \nabla \cdot \mathbf{F}(\mathbf{U}, \mathbf{S}) = \mathbf{0}$, consider a first order finite volume scheme on the cell K,

$$\mathbf{U}_K^{n+1} = \mathbf{U}_K^n - \frac{\Delta t}{|K|} \sum_{i=1}^{E} |e_i| \widehat{\mathbf{F} \cdot \mathbf{n}}(\mathbf{U}_K^n, \mathbf{S}_K^n, \mathbf{U}_{K_i}^n, \mathbf{S}_{K_i}^n),$$

with the numerical flux defined by

$$\widehat{\mathbf{F} \cdot \mathbf{n}}(\mathbf{U}_K^n, \mathbf{S}_K^n, \mathbf{U}_{K_i}^n, \mathbf{S}_{K_i}^n) = \frac{1}{2} \left[\mathbf{F}(\mathbf{U}_K^n, \mathbf{S}_K^n) \cdot \mathbf{n}_i + \mathbf{F}(\mathbf{U}_{K_i}^n, \mathbf{S}_{K_i}^n) \cdot \mathbf{n}_i - \beta_i (\mathbf{U}_{K_i}^n - \mathbf{U}_K^n) \right],$$

where $\mathbf{U}_K^n, \mathbf{S}_K^n$ are the approximation to the average of $\mathbf{U}, \mathbf{S}$ on K at time level n, $\mathbf{n}_i$ is the unit vector normal to e_i pointing outward of K, and β_i is a positive number dependent upon $\mathbf{U}_K^n$ and $\mathbf{U}_{K_i}^n$. With the assumption $\mathbf{U}_K^n, \mathbf{U}_{K_i}^n \in G$, we want to find a proper β_i and a CFL condition so that $\mathbf{U}_K^{n+1} \in G$.

We have $\sum_{i=1}^{E} \mathbf{F}(\mathbf{U}_K^n, \mathbf{S}_K^n) \cdot \mathbf{n}_i |e_i| = 0$ due to the fact that $\sum_{i=1}^{E} \mathbf{f} \cdot \mathbf{n}_i |e_i| = \int_{\partial K} \mathbf{f} \cdot \mathbf{n} \, ds = \iint_K \nabla \cdot \mathbf{f} dV = 0$ for any constant vector $\mathbf{f}$. Thus the scheme can be rewritten as

$$\mathbf{U}_K^{n+1} = \left(1 - \frac{1}{2} \frac{\Delta t}{|K|} \sum_{i=1}^{E} |e_i| \beta_i \right) \mathbf{U}_K^n + \frac{1}{2} \frac{\Delta t}{|K|} \sum_{i=1}^{E} |e_i| \beta_i \left[\mathbf{U}_{K_i}^n - \beta_i^{-1} \mathbf{F}(\mathbf{U}_{K_i}^n, \mathbf{S}_{K_i}^n) \cdot \mathbf{n}_i \right],$$

which is a convex combination of $\mathbf{U}_K^n$ and $\mathbf{U}_{K_i}^n - \beta_i^{-1} \mathbf{F}(\mathbf{U}_{K_i}^n, \mathbf{S}_{K_i}^n) \cdot \mathbf{n}_i$ under the CFL constraint $\Delta t \frac{|\partial K|}{|K|} \max_i \beta_i \leq 2$ with $|\partial K| = \sum_{i=1}^{E} |e_i|$. We emphasize

that the CFL condition $\Delta t \frac{|\partial K|}{|K|} \max_i \beta_i \le 2$ derived here is better than the one derived following discussions in [4, 18], which is $\Delta t \frac{|\partial K|}{|K|} \max_i \beta_i \le 1$.

By Lemma 6 in Appendix B, we have $\mathbf{U}^n_{K_i} - \beta_i^{-1}\mathbf{F}(\mathbf{U}^n_{K_i}, \mathbf{S}^n_{K_i}) \cdot \mathbf{n}_i \in G$ if we take

$$\beta_i > \max \left[|\mathbf{u} \cdot \mathbf{n}_i| + \frac{1}{2\rho^2 e} \left(\sqrt{\rho^2 |\mathbf{q} \cdot \mathbf{n}_i|^2 + 2\rho^2 e \|\boldsymbol{\tau} \cdot \mathbf{n}_i - p\mathbf{n}_i^t\|^2} + \rho |\mathbf{q} \cdot \mathbf{n}_i| \right) \right],$$

where the maximum is taken over $\mathbf{U}^n_K, \mathbf{S}^n_K, \mathbf{U}^n_{K_i}, \mathbf{S}^n_{K_i}$.

3.2. High order schemes for the compressible Navier-Stokes equations

Consider a $(k+1)$-th order finite volume type scheme with reconstruction polynomials or approximation polynomials of degree k. For the two dimensional compressible Euler system, with forward Euler time discretization on K, it takes the form (5). Assume we use the L-point Gauss quadrature for integrals along each edge e_i. Let w_ν $(\nu = 1, \cdots, L)$ denotes the L-point Gauss quadrature weights on interval $[-\frac{1}{2}, \frac{1}{2}]$, so that $\sum_{\nu=1}^L w_\nu = 1$. Let $\mathbf{x}_{\nu,i}$ denote the ν-th Gauss quadrature point on the i-th edge. Replacing the integrals by the Gauss quadrature, we obtain

$$\overline{\mathbf{U}}^{n+1}_K = \overline{\mathbf{U}}^n_K - \frac{\Delta t}{|K|} \sum_{i=1}^E |e_i| \sum_{\nu=1}^L w_\nu \widehat{\mathbf{F} \cdot \mathbf{n}}(\mathbf{U}^{\nu,i}_K, \mathbf{S}^{\nu,i}_K, \mathbf{U}^{\nu,i}_{K_i}, \mathbf{S}^{\nu,i}_{K_i}), \tag{30a}$$

where $\overline{\mathbf{U}}^n_K$ is the cell average on K at time level n, $\mathbf{U}^{\nu,i}_K$ and $\mathbf{U}^{\nu,i}_{K_i}$ are approximations to $\mathbf{U}(\mathbf{x}_{\nu,i}, t^n)$ from K and from K_i respectively. The local Lax-Friedrichs flux is

$$\widehat{\mathbf{F} \cdot \mathbf{n}}(\mathbf{U}^{\nu,i}_K, \mathbf{S}^{\nu,i}_K, \mathbf{U}^{\nu,i}_{K_i}, \mathbf{S}^{\nu,i}_{K_i}) = \frac{1}{2} \left[\left(\mathbf{F}(\mathbf{U}^{\nu,i}_K, \mathbf{S}^{\nu,i}_K) + \mathbf{F}(\mathbf{U}^{\nu,i}_{K_i}, \mathbf{S}^{\nu,i}_{K_i}) \right) \cdot \mathbf{n}_i - \beta_i (\mathbf{U}^{\nu,i}_{K_i} - \mathbf{U}^{\nu,i}_K) \right], \tag{30b}$$

$$\beta_i > \max \left[|\mathbf{u} \cdot \mathbf{n}_i| + \frac{1}{2\rho^2 e} \left(\sqrt{\rho^2 |\mathbf{q} \cdot \mathbf{n}_i|^2 + 2\rho^2 e \|\boldsymbol{\tau} \cdot \mathbf{n}_i - p\mathbf{n}_i^t\|^2} + \rho |\mathbf{q} \cdot \mathbf{n}_i| \right) \right], \tag{30c}$$

where the maximum is taken over $\mathbf{U}^{\nu,i}_K, \mathbf{S}^{\nu,i}_K, \mathbf{U}^{\nu,i}_{K_i}, \mathbf{S}^{\nu,i}_{K_i}$ for $\nu = 1, \cdots, L$.

Next, assume there exists a quadrature on K satisfying the following properties:

1. It is exact for integrals of polynomials of degree k on K.
2. The quadrature points include $\mathbf{x}_{\nu,i}$ for all ν and i.
3. The smallest quadrature weight must be strictly positive.

This quadrature will play the same role in two dimensions as the Gauss-Lobatto quadrature in Section 2.2 so that we can decompose the cell average as a convex combination of certain point values. The existence of such a quadrature rule is not obvious due to the third constraint above. For rectangular cells, we can use

the tensor product of Gauss quadrature and Gauss-Lobatto quadrature [17, 10]. For triangular cells, we can construct it by a Dubinar transform from rectangles to triangles [18]. For more general polygonal cells, see [22]. In [21], a more straightforward construction was discussed, but the quadrature weights for $\mathbf{x}_{\nu,i}$ on triangles are smaller.

Assume we have a M-point quadrature rule satisfying the three constraints above. Let $\mathbf{x}_{\nu,i}$ ($\nu = 1, \cdots, L$, $i = 1, \cdots, E$) and $\mathbf{x}_\lambda$ ($\lambda = EL+1, \cdots, M$) denote the M quadrature points on K. See Figure 1 for an illustration, where the red points are $\mathbf{x}_{\nu,i}$ and blue points are $\mathbf{x}_\lambda$. Let $\omega_{\nu,i}$ and ω_λ denote the corresponding normalized quadrature weights so that $\sum_{\nu=1}^{L}\sum_{i=1}^{E}\omega_{\nu,i} + \sum_{\lambda=EL+1}^{M}\omega_\lambda = 1$.

Consider a rectangular cell K with sides parallel to x-axis and y-axis. Assume lengths of sides are Δx and Δy. To construct a suitable quadrature, we use $\overline{\mathbf{U}}_K^n = \frac{\Delta x}{\Delta x+\Delta y}\overline{\mathbf{U}}_K^n + \frac{\Delta y}{\Delta x+\Delta y}\overline{\mathbf{U}}_K^n$. Following [17, 10], for the integral in $\frac{\Delta x}{\Delta x+\Delta y}\overline{\mathbf{U}}_K^n = \frac{\Delta x}{\Delta x+\Delta y}\frac{1}{|K|}\iint_K \mathbf{P}_K(\mathbf{x})\,dxdy$, we can use the tensor product of L-point Gauss quadrature for x-variable and N-point Gauss-Lobatto quadrature for y-variable. For the integral in $\frac{\Delta y}{\Delta x+\Delta y}\overline{\mathbf{U}}_K^n = \frac{\Delta y}{\Delta x+\Delta y}\frac{1}{|K|}\iint_K \mathbf{P}_K(\mathbf{x})\,dxdy$, we can use the tensor product of L-point Gauss quadrature for y-variable and N-point Gauss-Lobatto quadrature for x-variable. Take the union of these two tensor products, we get a quadrature rule needed and we have $\omega_{\nu,i}/w_\nu = \frac{|e_i|}{\Delta x+\Delta y}\widehat{\omega} = \frac{|e_i|}{\Delta x+\Delta y}\frac{1}{N(N-1)}$ for any $\nu = 1, \cdots, L$. For a triangular cell, we can use the quadrature constructed in [18], which is also based on N-point Gauss-Lobatto quadrature and $\omega_{\nu,i}/w_\nu = \frac{2}{3}\widehat{\omega} = \frac{2}{3}\frac{1}{N(N-1)}$ for any $\nu = 1, \cdots, L$. For both cases, we have $\sum_{i=1}^{E}\sum_{\nu=1}^{L}\omega_{\nu,i} = 2\widehat{\omega}$.

Let $\mathbf{P}_K(\mathbf{x})$ be the reconstruction or approximation polynomial of degree k in the $(k+1)$-th order finite volume type scheme (30a), i.e., the cell average of $\mathbf{P}_K(\mathbf{x})$ is $\overline{\mathbf{U}}_K^n$ and $\mathbf{U}_K^{\nu,i} = \mathbf{P}_K(\mathbf{x}_{\nu,i})$. Then we have

$$\overline{\mathbf{U}}_K^n = \frac{1}{|K|}\iint_K \mathbf{P}_K(\mathbf{x})\,dV = \sum_{j=1}^{E}\sum_{\nu=1}^{L}\omega_{\nu,j}\mathbf{U}_K^{\nu,j} + \sum_{\lambda=EL+1}^{M}\omega_\lambda \mathbf{P}_K(\mathbf{x}_\lambda).$$

Plugging the cell average decomposition above into (30a), we obtain,

$$\begin{aligned}
\overline{\mathbf{U}}_K^{n+1} &= \sum_{\lambda=EL+1}^{M} \omega_\lambda \mathbf{P}_K(\mathbf{x}_\lambda) + \sum_{i=1}^{E}\sum_{\nu=1}^{L}\left[\omega_{\nu,i}\mathbf{U}_K^{\nu,i} - \Delta t\frac{|e_i|}{|K|}w_\nu \widehat{\mathbf{F}\cdot\mathbf{n}}(\mathbf{U}_K^{\nu,i},\mathbf{S}_K^{\nu,i},\mathbf{U}_{K_i}^{\nu,i},\mathbf{S}_{K_i}^{\nu,i})\right] \\
&= \sum_{\lambda=EL+1}^{M} \omega_\lambda \mathbf{P}_K(\mathbf{x}_\lambda) + \frac{1}{2}\sum_{i=1}^{E}\sum_{\nu=1}^{L}\Delta t\frac{|e_i|}{|K|}w_\nu\beta_i\left(\mathbf{U}_{K_i}^{\nu,i} - \beta_i^{-1}\mathbf{F}(\mathbf{U}_{K_i}^{\nu,i},\mathbf{S}_{K_i}^{\nu,i})\cdot\mathbf{n}_i\right) \\
&\quad + \sum_{i=1}^{E}\sum_{\nu=1}^{L}\left[\omega_{\nu,i}\mathbf{U}_K^{\nu,i} - \frac{1}{2}\Delta t\frac{|e_i|}{|K|}w_\nu\left(\beta_i\mathbf{U}_K^{\nu,i} + \mathbf{F}(\mathbf{U}_K^{\nu,i},\mathbf{S}_K^{\nu,i})\cdot\mathbf{n}_i\right)\right] \\
&= \sum_{\lambda=EL+1}^{M} \omega_\lambda \mathbf{P}_K(\mathbf{x}_\lambda) + \sum_{i=1}^{E}\sum_{\nu=1}^{L}\Lambda_{\nu,i}\beta_i\left(\mathbf{U}_{K_i}^{\nu,i} - \beta_i^{-1}\mathbf{F}(\mathbf{U}_{K_i}^{\nu,i},\mathbf{S}_{K_i}^{\nu,i})\cdot\mathbf{n}_i\right) \\
&\quad + \sum_{i=1}^{E}\sum_{\nu=1}^{L}\left(\omega_{\nu,i} - \Lambda_{\nu,i}\beta_i\right)\left[\mathbf{U}_K^{\nu,i} - \Lambda_{\nu,i}\left(\omega_{\nu,i} - \Lambda_{\nu,i}\beta_i\right)^{-1}\mathbf{F}(\mathbf{U}_K^{\nu,i},\mathbf{S}_K^{\nu,i})\cdot\mathbf{n}_i\right],
\end{aligned} \tag{31}$$

where $\Lambda_{\nu,i} = \frac{1}{2}\Delta t\frac{|e_i|}{|K|}w_\nu$.

Notice the fact that $\Lambda_{\nu,i}(\omega_{\nu,i} - \Lambda_{\nu,i}\beta_i)^{-1} \le \beta_i^{-1}$ if and only if $0 \le \Delta t\frac{|e_i|}{|K|}\beta_i \le \frac{\omega_{\nu,i}}{w_\nu}$. By Lemma 6 in Appendix B, under the CFL constraint $\Delta t\frac{|e_i|}{|K|}\max_i\beta_i \le \min_{i,\nu}\frac{\omega_{\nu,i}}{w_\nu}$, we have

$$\mathbf{U}_K^{\nu,i} \in G \Rightarrow \mathbf{U}_K^{\nu,i} - \Lambda_{\nu,i}\left(\omega_{\nu,i} - \Lambda_{\nu,i}\beta_i\right)^{-1}\mathbf{F}(\mathbf{U}_K^{\nu,i},\mathbf{S}_K^{\nu,i})\cdot\mathbf{n}_i \in G,$$

$$\mathbf{U}_{K_i}^{\nu,i} \in G \Rightarrow \mathbf{U}_{K_i}^{\nu,i} - \beta_i^{-1}\mathbf{F}(\mathbf{U}_{K_i}^{\nu,i},\mathbf{S}_{K_i}^{\nu,i})\cdot\mathbf{n}_i \in G.$$

Thus we obtain the weak positivity for the high order scheme (30).

Theorem 3. *Consider the scheme (30) with reconstruction or approximation polynomials of degree k. Let $N = \lceil (k+3)/2 \rceil$. A sufficient condition for $\overline{\mathbf{U}}_j^{n+1} \in G$ in is*

$$\mathbf{U}_K^{\nu,i}, \mathbf{U}_{K_i}^{\nu,i} \in G \quad \textit{and} \quad \mathbf{P}_K(\mathbf{x}_\lambda) \in G \ (\lambda = EL+1, \cdots, M), \tag{32}$$

under the CFL condition

$$\begin{aligned}
&\Delta t\left(\frac{1}{\Delta x} + \frac{1}{\Delta y}\right)\beta_i \le \widehat{\omega} = \frac{1}{N(N-1)}, \quad \textit{on a rectangular cell } K, \\
&\Delta t\frac{|e_i|}{|K|}\beta_i \le \frac{2}{3}\widehat{\omega} = \frac{2}{3}\frac{1}{N(N-1)}, \quad \textit{on a triangular cell } K.
\end{aligned} \tag{33}$$

Following Remark 1, with the fact

$$\sum_{\lambda=EL+1}^{M}\frac{\omega_\lambda}{1 - \sum_{i=1}^{E}\sum_{\nu=1}^{L}\omega_{\nu,i}}\mathbf{P}_K(\mathbf{x}_\lambda) = \sum_{\lambda=EL+1}^{M}\frac{\omega_\lambda}{1-2\widehat{\omega}}\mathbf{P}_K(\mathbf{x}_\lambda),$$

we get a weaker sufficient condition to replace (32) as,

$$\mathbf{U}_K^{\nu,i}, \mathbf{U}_{K_i}^{\nu,i} \in G \quad \textit{and} \quad \sum_{\lambda=EL+1}^{M} \frac{\omega_\lambda}{1-2\widehat{\omega}} \mathbf{P}_K(\mathbf{x}_\lambda) \in G. \tag{34}$$

Remark 3. The discussion of weak positivity in this section, i.e., (31), is different from the one in [18] for triangular cells. Following the discussion in [18], we would get a constraint $\Delta t \frac{\sum_{i=1}^{E} |e_i|}{|K|} \beta_i \le \frac{2}{3}\widehat{\omega}$ for a triangular cell. The CFL condition (33) is certainly a better one, even though these CFL constraints are sufficient rather than necessary conditions for the weak positivity.

4. Review of the positivity-preserving limiter

To enforce the sufficient conditions for the weak positivity in Theorem 1, Theorem 2 and Theorem 3, we can use the simple positivity-preserving limiter in [17, 10, 18, 44]. In this section, we review this limiter in one dimension but all discussions can be extended to higher dimensions in a straightforward way. We will describe the limiter for DG schemes in two dimensions in Section 5.

4.1. A simple limiter to enforce bounds

We first discuss a simpler case of enforcing bounds of a piecewise polynomial approximation to a scalar function. Consider piecewise polynomials $p_j(x)$ on each interval $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$ approximating a smooth function $u(x)$. Let m and M be the minimum and the maximum of $u(x)$, i.e., $u(x) \in [m, M]$ for any x. Assume the cell averages $\overline{p}_j = \frac{1}{\Delta x_j}\int_{I_j} p_j(x)\,dx \in [m, M]$. If $p_j(x) \notin [m, M]$ for some $x \in I_j$, then we seek a modified approximation polynomial $\widetilde{p}_j(x)$ satisfying $\widetilde{p}_j(x) \in [m, M]$ for any $x \in I_j$, with the same cell average $\frac{1}{\Delta x_j}\int_{I_j} \widetilde{p}_j(x)\,dx = \frac{1}{\Delta x_j}\int_{I_j} p_j(x)\,dx$. For instance, if $p_j(x)$ is the L^2 projection of $u(x)$ onto the vector space of polynomials of degree k on the I_j, then we have $\overline{p}_j \in [m, M]$ but not necessarily $p_j(x) \in [m, M]$ for any $x \in I_j$.

The following limiter was first discussed in [48]:

$$\widetilde{p}_j(x) = \theta\left[p_j(x) - \overline{p}_j\right] + \overline{p}_j, \quad \theta = \min\left\{1, \left|\frac{M-\overline{p}_j}{M_j-\overline{p}_j}\right|, \left|\frac{m-\overline{p}_j}{m_j-\overline{p}_j}\right|\right\}, \tag{35a}$$

$$M_j = \max_{x\in I_j} p_j(x), m_j = \min_{x\in I_j} p_j(x). \tag{35b}$$

It is straightforward to see that $\widetilde{p}_j(x) \in [m, M]$ for any $x \in I_j$ and the cell average of $\widetilde{p}_j(x)$ is still $\overline{p}_j$. On the other hand, this simple limiter does not destroy the approximation accuracy of $p_j(x)$.

Theorem 4. *For the modified polynomial of degree k in the limiter (35), we have $|p_j(x) - \widetilde{p}_j(x)| \le C_k \max_{x\in I_j} |p_j(x) - u(x)|$, where C_k is a constant depending only on the polynomial degree k.*

PROOF. We only need to discuss the case that $p_j(x)$ is not a constant and $\theta = \left|\frac{M-\overline{p}_j}{M_j-\overline{p}_j}\right|$. The other cases are similar. Since $\overline{p}_j \le M$ and $\overline{p}_j \le M_j$, we have $\theta = (M-\overline{p}_j)/(M_j-\overline{p}_j)$. Therefore,

$$\begin{aligned}\widetilde{p}_j(x) - p_j(x) &= \theta[p_j(x)-\overline{p}_j] + \overline{p}_j - p_j(x)\\ &= (\theta-1)[p_j(x)-\overline{p}_j]\\ &= \frac{M-M_j}{M_j-\overline{p}_j}[p_j(x)-\overline{p}_j]\\ &= (M-M_j)\frac{p_j(x)-\overline{p}_j}{M_j-\overline{p}_j}.\end{aligned}$$

Thus $|\widetilde{p}_j(x) - p_j(x)| \le |M-M_j|\left|\frac{p_j(x)-\overline{p}_j}{M_j-\overline{p}_j}\right|$. The assumption $\theta = \left|\frac{M-\overline{p}_j}{M_j-\overline{p}_j}\right|$ implies the overshoot $M_j > M$. Suppose $p_j(x^*) = M_j$ for some $x^* \in I_j$, then $u(x^*) \le M < M_j = p_j(x^*)$. Thus we have $|M-M_j| \le |u(x^*)-p_j(x^*)| \le \max_{x\in I_j}|p_j(x)-u(x)|$. We need to show $\left|\frac{p_j(x)-\overline{p}_j}{M_j-\overline{p}_j}\right| \le C_k$.

Consider a new polynomial $q(x) = p_j\left(x\Delta x + x_{j-\frac{1}{2}}\right) - \overline{p}_j$. Then $\overline{q} = \int_0^1 q(x)\,dx = 0$, $\max\limits_{x\in[0,1]} q(x) = \max\limits_{x\in I_j} p_j(x) - \overline{p}_j$ and $\min\limits_{x\in[0,1]} q(x) = \min\limits_{x\in I_j} p_j(x) - \overline{p}_j$. We have

$$\left|\frac{p_j(x)-\overline{p}_j}{M_j-\overline{p}_j}\right| = \frac{|q(x)|}{\max\limits_{x\in[0,1]} q(x)} \le \frac{\max\limits_{x\in[0,1]}|q(x)|}{\max\limits_{x\in[0,1]} q(x)} = \max\left\{\frac{\max\limits_{x\in[0,1]} q(x)}{\max\limits_{x\in[0,1]} q(x)}, \frac{-\min\limits_{x\in[0,1]} q(x)}{\max\limits_{x\in[0,1]} q(x)}\right\}.$$

Thus we only need to prove $\frac{\max\limits_{x\in[0,1]}|q(x)|}{\max\limits_{x\in[0,1]} q(x)} \le C_k$ or $\left|\frac{\min\limits_{x\in[0,1]} q(x)}{\max\limits_{x\in[0,1]} q(x)}\right| \le C_k$. For quadratic polynomials $k=2$ in one dimension, $C_k = 3$ was proven by explicit calculations in [48]. For general k and higher dimensions, see Appendix C.

Remark 4. The proof above can be easily extended to any kind of cells in any dimensions.

In practice, the limiter (35) is not very interesting since evaluating the maximum and the minimum of a high order polynomial in (35b) is computationally demanding in high dimensions. A practical limiter is (35) with M_j and m_j redefined as

$$\widetilde{p}_j(x) = \theta\left[p_j(x)-\overline{p}_j\right] + \overline{p}_j, \quad \theta = \min\left\{1, \left|\frac{M-\overline{p}_j}{M_j-\overline{p}_j}\right|, \left|\frac{m-\overline{p}_j}{m_j-\overline{p}_j}\right|\right\}, \tag{36a}$$

$$M_j = \max_{x\in S_j} p_j(x), m_j = \min_{x\in S_j} p_j(x), \tag{36b}$$

where S_j are Gauss-Lobatto quadrature points (16). The simplified limiter (36) was first used in [17] to enforce the sufficient conditions for the weak monotonicity in any high order finite volume schemes with monotone fluxes solving scalar conservation laws. Since (35) is a more stringent limiter than (36), Theorem 4 also applies to the simplified limiter (36).

4.2. A simple limiter to enforce positivity

To enforce the condition (19), we can apply the limiter (36) to the density and extend such a limiter to enforce the positivity of internal energy. Let $\mathbf{P}_j(x) = (\rho_j(x), \rho u_j(x), E_j(x))^t$ be a vector of polynomials of degree k on I_j with the cell average $\overline{\mathbf{P}}_j = (\overline{\rho}_j, \overline{\rho u}_j, \overline{E}_j)^t$. Define $\overline{\rho e}_j = \overline{E}_j - \frac{1}{2}\overline{\rho u}_j^2/\overline{\rho}_j$. Assume $\overline{\mathbf{P}}_j$ has positive density and internal energy, i.e., $\overline{\rho}_j > 0, \overline{\rho e}_j > 0$. Let ε be a small positive number as the desired lower bound for density and internal energy, e.g., $\varepsilon = 10^{-8}$ or $\varepsilon = 10^{-13}$.

Numerically the set of admissible states is

$$G^\varepsilon = \left\{ \mathbf{U} = \begin{pmatrix} \rho \\ \rho \mathbf{u} \\ E \end{pmatrix} : \rho \geq \varepsilon, \quad \rho e(\mathbf{U}) \geq \varepsilon. \right\}.$$

Assume $\overline{\mathbf{P}}_j \in G^\varepsilon$. If $\mathbf{P}_j(\widehat{x}_j^\mu) \notin G^\varepsilon$ for some $\mu = 1, \cdots, N$, then we seek a modified polynomial $\widetilde{\mathbf{P}}_j(x)$ with the same cell average so that $\widetilde{\mathbf{P}}_j(\widehat{x}_j^\mu) \in G^\varepsilon$ for all μ.

We first modify density to enforce the positivity by,

$$\widehat{\rho}_j(x) = \theta_\rho(\rho_j(x) - \overline{\rho}_j) + \overline{\rho}_j, \quad \theta_\rho = \min\left\{1, \frac{\overline{\rho}_j - \varepsilon}{\overline{\rho}_j - \min\limits_{\mu=1,\cdots,N} \rho_j(\widehat{x}_j^\mu)}\right\}. \tag{37a}$$

Then let $\widehat{\mathbf{P}}_j(x) = (\widehat{\rho}_j(x), \rho u_j(x), E_j(x))^t$. Let $\widehat{\rho e}_j(x) = E_j(x) - \frac{1}{2}\rho u_j(x)^2/\widehat{\rho}_j(x)$. To enforce the positivity of internal energy, we can use the simplified limiter in [44],

$$\widetilde{\mathbf{P}}_j(x) = \theta_e(\widehat{\mathbf{P}}_j(x) - \overline{\mathbf{P}}_j) + \overline{\mathbf{P}}_j, \quad \theta_e = \min\left\{1, \frac{\overline{\rho e}_j - \varepsilon}{\overline{\rho e}_j - \min\limits_{\mu=1,\cdots,N} \widehat{\rho e}_j(\widehat{x}_j^\mu)}\right\}. \tag{37b}$$

By the convex combination $\overline{\mathbf{P}}_j = \sum_{\mu=1}^N \widehat{\omega}_\mu \mathbf{P}_j(\widehat{x}_j^\mu)$ and the Jensen's inequality (4), we have $\min\limits_{\mu=1,\cdots,N} \rho_j(\widehat{x}_j^\mu) \leq \overline{\rho}_j$ and $\min\limits_{\mu=1,\cdots,N} \widehat{\rho e}_j(\widehat{x}_j^\mu) \leq \overline{\rho e}_j$, which implies $0 \leq \theta_\rho, \theta_e \leq 1$. By the Jensen's inequality (4), it is straightforward to check that $\widetilde{\mathbf{P}}_j(\widehat{x}_j^\mu) \in G^\varepsilon$ for $\mu = 1, \cdots, N$.

Now consider the polynomial $\mathbf{P}_j(x)$ approximating a smooth solution $\mathbf{U}(x) = (\rho(x), \rho u(x), E(x))^t \in G^\varepsilon$ for all x. If $\rho(x) \geq \varepsilon$ for any x, then the accuracy of the limiter on density (37a) can be understood in the sense of Theorem 4.

Suppose the smooth solution $\mathbf{U}(x)$ satisfies $\rho e(x) = E(x) - \frac{1}{2}\rho u(x)^2/\rho(x) = \varepsilon$ for some x, then the limiter (37b) may induce at least a second order error $\mathcal{O}(\Delta x^2)$ around the minimum of $\rho e(x)$, see [43] for a discussion on a similar issue when enforcing entropy bounds. In other words, (37b) is a crude limiter if the smooth solution $\mathbf{U}(x)$ lies on the boundary of the convex set G^ε for some x.

If the internal energy of $\mathbf{U}(x)$ is uniformly bounded away from ε, then (37b) is still a high order accurate limiter as Δx goes to zero. Assume $\rho e(x) \geq m$ for any x and $m > 2\varepsilon$. Assume $\|\mathbf{P}_j(x) - \mathbf{U}(x)\| = \mathcal{O}(\Delta x^{k+1})$ for any x, then

$\|\widehat{\mathbf{P}}_j(x)-\mathbf{U}(x)\| = \mathcal{O}(\Delta x^{k+1})$ for any x by Theorem 4. Since $\|\mathbf{P}_j(x)-\mathbf{U}(x)\| = \mathcal{O}(\Delta x^{k+1})$, we have $\overline{\rho e}_j = \frac{1}{\Delta x}\int_{I_j} \rho e(x)\,dx + \mathcal{O}(\Delta x^{k+1})$. Thus $\overline{\rho e}_j \geq \frac{1}{2}m > \varepsilon$ if Δx is small enough. Without loss of generality, assume $\theta_e = \frac{\overline{\rho e}_j-\varepsilon}{\overline{\rho e}_j-\widehat{\rho e}_j(\widehat{x}_j^1)}$, then $\widehat{\rho e}_j(\widehat{x}_j^1) < \varepsilon$ thus $\overline{\rho e}_j - \widehat{\rho e}_j(\widehat{x}_j^1) > m/2-\varepsilon > 0$. Therefore, $1-\theta_e = \frac{\varepsilon-\widehat{\rho e}_j(\widehat{x}_j^1)}{\overline{\rho e}_j-\widehat{\rho e}_j(\widehat{x}_j^1)} \leq \frac{\widehat{\rho e}_j(\widehat{x}_j^1)-\varepsilon}{m/2-\varepsilon}$. On the other hand, $\|\widehat{\mathbf{P}}_j(x)-\mathbf{U}(x)\| = \mathcal{O}(\Delta x^{k+1})$ implies the undershoot $\widehat{\rho e}_j(\widehat{x}_j^1)-\varepsilon = \mathcal{O}(\Delta x^{k+1})$ thus $1-\theta_e = \mathcal{O}(\Delta x^{k+1})$ if Δx is small enough. So we get $\|\widetilde{\mathbf{P}}_j(x)-\mathbf{P}_j(x)\| = \|(1-\theta_e)[\mathbf{P}_j(x)-\overline{\mathbf{P}}_j]\| = \mathcal{O}(\Delta x^{k+1})$. In a nutshell, for smooth solutions without vacuum, i.e., the internal energy or pressure is uniformly bounded away from zero, the limiter on internal energy (37b) is a high order accurate limiter.

We remark that it is straightforward to define an optimal limiter in terms of accuracy as an optimization problem, i.e., finding a polynomial $\widetilde{\mathbf{P}}_j(x)$ to minimize $\|\widetilde{\mathbf{P}}_j(x)-\mathbf{P}_j(x)\|$ under the constraints $\int_{I_j}\widetilde{\mathbf{P}}_j(x)\,dx = \int_{I_j}\mathbf{P}_j(x)\,dx$ and $\widetilde{\mathbf{P}}_j(\widehat{x}_j^\mu) \in G^\varepsilon$. But accurately solving such a convex optimization problem is much more computationally demanding.

4.3. An efficient implementation of the positivity-preserving limiter

In the limiter (37), we have to evaluate $\mathbf{P}_j(\widehat{x}_j^\mu)$ for $\mu = 2,\cdots,N-2$, which are the blue point values as illustrated in Figure 1. In two and higher dimensions, these blue point values are not needed in any standard finite volume and DG schemes. It will be a more efficient implementation if we can avoid evaluating these redundant point values in the limiter. To this end, we can enforce (29) instead of (27).

Given polynomials $\mathbf{P}_j(x)$ with cell averages $\overline{\mathbf{P}}_j \in G^\varepsilon$, we seek polynomials $\widetilde{\mathbf{P}}_j(x)$ with the same cell averages so that $\widetilde{\mathbf{P}}_j(\widehat{x}_j^1), \widetilde{\mathbf{P}}_j(\widehat{x}_j^N), \sum_{\mu=2}^{N-1}\frac{\widehat{\omega}_\mu}{1-2\widehat{\omega}}\widetilde{\mathbf{P}}_j(\widehat{x}_j^\mu) \in G^\varepsilon$.

By the Mean Value Theorem, the convex combination $\sum_{\mu=2}^{N-1}\frac{\widehat{\omega}_\mu}{1-2\widehat{\omega}}\rho_j(\widehat{x}_j^\mu)$ is equal to $\rho_j(x_j^*)$ where x_j^* is some point in the cell I_j. The convex combination $\overline{\rho}_j = \sum_{\mu=1}^{N}\widehat{\omega}_\mu\rho_j(\widehat{x}_j^\mu)$ implies $\overline{\rho}_j = \widehat{\omega}\rho_j(\widehat{x}_j^1) + \widehat{\omega}\rho_j(\widehat{x}_j^N) + (1-2\widehat{\omega})\rho_j(x_j^*)$. Even though the value of x_j^* is unknown, we can compute the point value of $\rho_j(x)$ at x_j^* as $\rho_j(x_j^*) = \frac{1}{1-2\widehat{\omega}}(\overline{\rho}_j - \widehat{\omega}\rho_j(\widehat{x}_j^1) - \widehat{\omega}\rho_j(\widehat{x}_j^N))$. We first modify density by,

$$\widehat{\rho}_j(x) = \theta_\rho(\rho_j(x)-\overline{\rho}_j)+\overline{\rho}_j, \quad \theta_\rho = \min\left\{1, \frac{\overline{\rho}_j-\varepsilon}{\overline{\rho}_j - \min\limits_{x\in\{\widehat{x}_j^1,\widehat{x}_j^N,x_j^*\}}\rho_j(x)}\right\}. \tag{38a}$$

Notice that we have the convex combination $\overline{\rho}_j = \widehat{\omega}\rho_j(\widehat{x}_j^1) + \widehat{\omega}\rho_j(\widehat{x}_j^N) + (1-2\widehat{\omega})\rho_j(x_j^*)$, so $\overline{\rho}_j \geq \min\{\rho_j(\widehat{x}_j^1), \rho_j(\widehat{x}_j^N), \rho_j(x_j^*)\}$ thus $\theta_\rho \in [0,1]$. It is straightforward to see that the limiter (38a) is a more relaxed one than (37a), thus Theorem

4 also applies to the limiter (38a). Moreover, we have $\widehat{\rho}_j(\widehat{x}_j^1), \widehat{\rho}_j(\widehat{x}_j^N) \geq \varepsilon$ and

$$\begin{aligned}\sum_{\mu=2}^{N-1} \frac{\widehat{\omega}_\mu}{1-2\widehat{\omega}} \widehat{\rho}_j(\widehat{x}_j^\mu) &= \sum_{\mu=2}^{N-1} \frac{\widehat{\omega}_\mu}{1-2\widehat{\omega}} [\theta_\rho(\rho_j(\widehat{x}_j^\mu) - \overline{\rho}_j) + \overline{\rho}_j] \\ &= \theta_\rho \left(\sum_{\mu=2}^{N-1} \frac{\widehat{\omega}_\mu}{1-2\widehat{\omega}} \rho_j(\widehat{x}_j^\mu) - \overline{\rho}_j \right) + \overline{\rho}_j \\ &= \theta_\rho \left(\rho_j(x_j^*) - \overline{\rho}_j \right) + \overline{\rho}_j \geq \varepsilon.\end{aligned}$$

Let $\widehat{\mathbf{P}}_j(x) = (\widehat{\rho}_j(x), \rho u_j(x), E_j(x))^t$ and $\widehat{\rho e}_j(x) = E_j(x) - \frac{1}{2}\rho u_j(x)^2/\widehat{\rho}_j(x)$. By the Mean Value Theorem, the convex combination $\sum_{\mu=2}^{N-1} \frac{\widehat{\omega}_\mu}{1-2\widehat{\omega}} \widehat{\mathbf{P}}_j(\widehat{x}_j^\mu)$ is equal to $(\widehat{\rho}_j(x_j^{*,1}), \rho u_j(x_j^{*,2}), E_j(x_j^{*,3}))^t$, where $x_j^{*,1}, x_j^{*,2}, x_j^{*,3}$ are three different points in the cell I_j. We abuse the notation by using $\widehat{\mathbf{P}}_j(x_j^{**})$ to denote the vector $(\widehat{\rho}_j(x_j^{*,1}), \rho u_j(x_j^{*,2}), E_j(x_j^{*,3}))^t$. Then $\widehat{\mathbf{P}}_j(x_j^{**}) = \sum_{\mu=2}^{N-1} \frac{\widehat{\omega}_\mu}{1-2\widehat{\omega}} \widehat{\mathbf{P}}_j(\widehat{x}_j^\mu)$ and we compute its value by $\widehat{\mathbf{P}}_j(x_j^{**}) = \frac{1}{1-2\widehat{\omega}}(\overline{\mathbf{P}}_j - \widehat{\omega}\mathbf{P}_j(\widehat{x}_j^1) - \widehat{\omega}\mathbf{P}_j(\widehat{x}_j^N))$. For any vector $\mathbf{U} = (\rho, \rho u, E)^t$, define the internal energy function as,

$$\Psi(\mathbf{U}) = E - \frac{1}{2}\frac{|\rho u|^2}{\rho}.$$

The internal energy of the vector $\widehat{\mathbf{P}}_j(x_j^{**})$ is denoted by $\widehat{\rho e}_j(x_j^{**}) = \Psi\left(\widehat{\mathbf{P}}_j(x_j^{**})\right)$. To enforce the positivity of internal energy, we define a limiter as,

$$\widetilde{\mathbf{P}}_j(x) = \theta_e(\widehat{\mathbf{P}}_j(x) - \overline{\mathbf{P}}_j) + \overline{\mathbf{P}}_j, \quad \theta_e = \min\left\{1, \frac{\overline{\rho e}_j - \varepsilon}{\overline{\rho e}_j - \min\limits_{x\in\{\widehat{x}_j^1, \widehat{x}_j^N, x_j^{**}\}} \widehat{\rho e}_j(x)}\right\}. \tag{38b}$$

Since the cell average of $\widehat{\mathbf{P}}_j(x)$ is still $\overline{\mathbf{P}}_j$, we have the convex combination $\overline{\mathbf{P}}_j = \sum_{\mu=1}^{N} \widehat{\omega}_\mu \widehat{\mathbf{P}}_j(\widehat{x}_j^\mu) = \widehat{\omega}\widehat{\mathbf{P}}_j(\widehat{x}_j^1) + \widehat{\omega}\widehat{\mathbf{P}}_j(\widehat{x}_j^N) + (1-2\widehat{\omega})\widehat{\mathbf{P}}_j(x_j^{**})$. The Jensen's inequality (4) implies,

$$\overline{\rho e}_j \geq \widehat{\omega}\widehat{\rho e}_j(\widehat{x}_j^1) + \widehat{\omega}\widehat{\rho e}_j(\widehat{x}_j^N) + (1-2\widehat{\omega})\widehat{\rho e}_j(x_j^{**}).$$

So $\overline{\rho e}_j \geq \min\{\widehat{\rho e}_j(\widehat{x}_j^1), \widehat{\rho e}_j(\widehat{x}_j^N), \widehat{\rho e}_j(x_j^{**})\}$ thus $\theta_e \in [0,1]$. It is straightforward to see that the limiter (38b) is a more relaxed one than (37b) because of $\widehat{\rho e}_j(x_j^{**}) \geq \sum_{\mu=2}^{N-1} \frac{\widehat{\omega}_\mu}{1-2\widehat{\omega}} \widehat{\rho e}_j(\widehat{x}_j^\mu)$ implied by the Jensen's inequality (4). Therefore, (38b) is also a high order accurate limiter for approximating a smooth solution $\mathbf{U}(x)$ with $\rho e(x)$ having a positive uniform lower bound.

Moreover, it is straightforward to check $\widetilde{\mathbf{P}}_j(\widehat{x}_j^1), \widetilde{\mathbf{P}}_j(\widehat{x}_j^N) \in G^\varepsilon$. Let $\widetilde{\mathbf{P}}_j(x) =$

$(\widetilde{\rho}_j(x), \widetilde{\rho u}_j(x), \widetilde{E}_j(x))^t$ and $\widetilde{\rho e}_j(x) = \widetilde{E}_j(x) - \frac{1}{2}\widetilde{\rho u}_j(x)^2/\widetilde{\rho}_j(x)$. Then we have

$$\begin{aligned}
\Psi\left(\sum_{\mu=2}^{N-1}\frac{\widehat{\omega}_\mu}{1-2\widehat{\omega}}\widetilde{\mathbf{P}}_j(\widehat{x}_j^\mu)\right) &= \Psi\left(\sum_{\mu=2}^{N-1}\frac{\widehat{\omega}_\mu}{1-2\widehat{\omega}}[\theta_e(\widehat{\mathbf{P}}_j(\widehat{x}_j^\mu)-\overline{\mathbf{P}}_j)+\overline{\mathbf{P}}_j]\right)\\
&= \Psi\left(\theta_e\left[\sum_{\mu=2}^{N-1}\frac{\widehat{\omega}_\mu}{1-2\widehat{\omega}}\widehat{\mathbf{P}}_j(\widehat{x}_j^\mu)-\overline{\mathbf{P}}_j\right]+\overline{\mathbf{P}}_j\right)\\
&= \Psi\left(\theta_e\left[\widehat{\mathbf{P}}_j(x_j^{**})-\overline{\mathbf{P}}_j\right]+\overline{\mathbf{P}}_j\right)\\
&= \Psi\left(\theta_e\widehat{\mathbf{P}}_j(x_j^{**})+(1-\theta_e)\overline{\mathbf{P}}_j\right)\\
&\geq \theta_e\Psi\left(\widehat{\mathbf{P}}_j(x_j^{**})\right)+(1-\theta_e)\Psi\left(\overline{\mathbf{P}}_j\right)\\
&= \theta_e\widehat{\rho e}_j(x_j^{**})+(1-\theta_e)\overline{\rho e}_j \geq \varepsilon.
\end{aligned}$$

Therefore, the limiter (38) returns a $\widetilde{\mathbf{P}}_j(x)$ satisfying the condition (29) without evaluating the point values at $\widehat{x}_j^\mu$ for $\mu = 2, \cdots, N-2$.

5. Implementation of positivity-preserving high order DG schemes

5.1. DG schemes

For solving the compressible Navier-Stokes system (1), there are quite a few very different DG type schemes, e.g., the pioneering work by Bassi and Rebay [33, 34], the scheme by Baumann and Oden [35], Compact DG [36], correction procedure via reconstruction (CPR) [37, 38], Hybrid DG [39] and Embedded DG [40], etc. In this paper, we focus on the mixed finite element formulation (10) used in [33, 34]. Let $\mathbf{F} = \mathbf{F}^a - \mathbf{F}^d$ and $\widehat{\mathbf{F}\cdot\mathbf{n}} = \widehat{\mathbf{F}^a\cdot\mathbf{n}} - \widehat{\mathbf{F}^d\cdot\mathbf{n}}$, then (10) becomes

$$\begin{cases}
\displaystyle\iint_K \mathbf{S}_h\phi_h\,dV = -\iint_K \mathbf{U}_h\nabla\phi_h\,dV + \int_{\partial K}\widehat{\mathbf{U}\mathbf{n}}\phi_h\,ds\\
\displaystyle\iint_K \frac{d}{dt}\mathbf{U}_h\psi_h\,dV = \iint_K \mathbf{F}\nabla\cdot\psi_h\,dV - \int_{\partial K}\widehat{\mathbf{F}\cdot\mathbf{n}}\psi_h\,ds
\end{cases},$$

where $\mathbf{U}_h$ and $\mathbf{S}_h$ are vectors of polynomials of degree k and ϕ_h and ψ_h are the polynomial of degree k test functions defined on K, which is a two-dimensional cell as illustrated in Figure 4.

For evaluating the flux $\mathbf{F}^d(\mathbf{U}, \mathbf{S})$ in (2), we need to compute the derivatives of velocity and internal energy based on the derivatives of the conserved variables, which can be done by the following formulas obtained from applying product and quotient rules to ρu and $e = \frac{1}{\rho}\left(E - \frac{1}{2}\rho u^2 - \frac{1}{2}\rho v^2\right) = \frac{E}{\rho} - \frac{1}{2}u^2 - \frac{1}{2}v^2$:

$$u_x = \frac{1}{\rho}[(\rho u)_x - \rho_x u], \quad e_x = \frac{1}{\rho^2}[E_x\rho - \rho_x E] - uu_x - vv_x. \tag{39}$$

Assume we use the L-point Gauss quadrature for each edge e_i in integrals along ∂K. Let w_ν ($\nu = 1, \cdots, L$) denotes the L-point Gauss quadrature weights

on interval $[-\frac{1}{2},\frac{1}{2}]$, so that $\sum_{\nu=1}^{L} w_\nu = 1$. Let $\mathbf{x}_{\nu,i}$ denote the ν-th Gauss quadrature point on the i-th edge. Let $\mathbf{U}_K(\mathbf{x})$ and $\mathbf{U}_{K_i}(\mathbf{x})$ denote the solution polynomials restricted on the elements K and K_i respectively, i.e., $\mathbf{U}_K(\mathbf{x}) = \mathbf{U}_h(\mathbf{x})|_K$ and $\mathbf{U}_{K_i}(\mathbf{x}) = \mathbf{U}_h(\mathbf{x})|_{K_i}$. Let $\mathbf{U}_K^{\nu,i} = \mathbf{U}_K(\mathbf{x}_{\nu,i})$ and $\mathbf{U}_{K_i}^{\nu,i} = \mathbf{U}_{K_i}(\mathbf{x}_{\nu,i})$. Replacing the line integrals by Gauss quadrature, we obtain

$$\begin{cases} \iint_K \mathbf{S}_h\phi_h\,dV = -\iint_K \mathbf{U}_h\nabla\phi_h\,dV + \sum_{i=1}^{E}|e_i|\sum_{\nu=1}^{L} w_\nu \widehat{\mathbf{U}\mathbf{n}}(\mathbf{U}_K^{\nu,i},\mathbf{U}_{K_i}^{\nu,i})\phi_h(\mathbf{x}_{\nu,i}) \\ \iint_K \frac{d}{dt}\mathbf{U}_h\psi_h\,dV = \iint_K \mathbf{F}\nabla\cdot\psi_h\,dV - \sum_{i=1}^{E}|e_i|\sum_{\nu=1}^{L} w_\nu \widehat{\mathbf{F}\cdot\mathbf{n}}(\mathbf{U}_K^{\nu,i},\mathbf{S}_K^{\nu,i},\mathbf{U}_{K_i}^{\nu,i},\mathbf{S}_{K_i}^{\nu,i})\psi_h(\mathbf{x}_{\nu,i}) \end{cases} \tag{40}$$

Following [33], the numerical flux $\widehat{\mathbf{U}\mathbf{n}}$ in (40) can be taken as a centered one

$$\widehat{\mathbf{U}\mathbf{n}}(\mathbf{U}_K^{\nu,i},\mathbf{U}_{K_i}^{\nu,i}) = \frac{1}{2}(\mathbf{U}_K^{\nu,i}+\mathbf{U}_{K_i}^{\nu,i})\mathbf{n}.$$

To render the high order DG scheme satisfying the weak positivity property, the numerical flux $\widehat{\mathbf{F}\cdot\mathbf{n}} = \widehat{\mathbf{F}^a\cdot\mathbf{n}} - \widehat{\mathbf{F}^d\cdot\mathbf{n}}$ should be positivity-preserving. For instance, $\widehat{\mathbf{F}^a\cdot\mathbf{n}}$ can be any positivity-preserving flux for compressible Euler equations and $\widehat{\mathbf{F}^d\cdot\mathbf{n}}$ can be taken as one similar to (22). For simplicity, we simply use the local Lax-Friedrichs type positivity-preserving flux (30b) and (30c).

By setting the test function $\psi_h \equiv 1$ in the scheme (40), we obtain the scheme satisfied by the cell averages, the forward Euler discretization of which is precisely (30a). Thus Theorem 3 also applies to the high order DG scheme (40) with the flux (30b) and (30c).

5.2. The positivity-preserving limiter in two dimensions

To enforce the condition in (34) in Theorem 3, we can use the limiter as described in Section 4.3 to avoid evaluating point values at highly redundant blues points as illustrated in Figure 1. Let $\mathbf{P}_K(\mathbf{x}) = (\rho_K(\mathbf{x}),\rho\mathbf{u}_K(\mathbf{x}),E_K(\mathbf{x}))^t$ denote the DG polynomial on the element K. Assume its cell average $\overline{\mathbf{P}}_K = (\overline{\rho}_K,\overline{\rho\mathbf{u}}_K,\overline{E}_K)^t \in G^\varepsilon$. Let $\mathbf{x}^m$ $(m=1,\cdots,P)$ be quadrature points used for computing integrals $\iint_K \mathbf{F}\nabla\cdot\psi_h\,dV$ in the scheme (40). In practice, it is also preferred to have $\mathbf{P}_K(\mathbf{x}^m)\in G^\varepsilon$. To this end, we can include the points $\mathbf{x}^m$ $(m=1,\cdots,P)$ in the positivity-preserving limiter.

By the Mean Value Theorem, there exists $\mathbf{x}^*\in K$ such that

$$\rho_K(\mathbf{x}^*) = \sum_{\lambda=EM+1}^{M}\frac{\widehat{\omega}_\lambda}{1-2\widehat{\omega}}\rho_K(\mathbf{x}_\lambda).$$

Let

$$S_\rho = \{\mathbf{x}_{\nu,i}:\nu=1,\cdots,L;i=1,\cdots,E\}\cup\{\mathbf{x}^m:m=1,\cdots,P\}\cup\{\mathbf{x}^*\}.$$

In other words, the set S_ρ contains $\mathbf{x}^*$ and all quadrature points for computing line integrals along ∂K and double integrals over K in the DG scheme (40). Given only point values of $\rho_K(\mathbf{x})$ at $\mathbf{x}_{\nu,i}(\nu = 1, \cdots, L; i = 1, \cdots, E)$ and $\mathbf{x}^m(m = 1, \cdots, P)$, we can compute

$$\rho_K(\mathbf{x}^*) = \frac{1}{1-2\widehat{\omega}}\left(\overline{\rho}_K - \sum_{i=1}^{E}\sum_{\nu=1}^{L}\omega_{\nu,i}\rho_K(\mathbf{x}_{\nu,i})\right),$$

and

$$\widehat{\rho}_K(\mathbf{x}) = \theta_\rho(\rho_K(\mathbf{x}) - \overline{\rho}_K) + \overline{\rho}_K, \quad \theta_\rho = \min\left\{1, \frac{\overline{\rho}_K - \varepsilon}{\overline{\rho}_K - \min\limits_{\mathbf{x}\in S_\rho}\rho_K(\mathbf{x})}\right\}. \tag{41a}$$

Let $\widehat{\mathbf{P}}_K(\mathbf{x}) = (\widehat{\rho}_K(\mathbf{x}), \rho\mathbf{u}_K(\mathbf{x}), E_K(\mathbf{x}))^t$. Let $\widehat{\rho e}_K(\mathbf{x}) = E_K(\mathbf{x}) - \frac{1}{2}\|\rho\mathbf{u}_K(\mathbf{x})\|^2/\widehat{\rho}_K(\mathbf{x})$ where $\|\cdot\|$ denotes the standard l^2 norm of a vector. By abusing the notation, we define

$$\widehat{\mathbf{P}}_K(\mathbf{x}^{**}) = \sum_{\lambda=EM+1}^{M}\frac{\widehat{\omega}_\lambda}{1-2\widehat{\omega}}\widehat{\mathbf{P}}_K(\mathbf{x}_\lambda),$$

even though the point $\mathbf{x}^{**}$ may not exist. Let

$$S_e = \{\mathbf{x}_{\nu,i} : \nu = 1, \cdots, L; i = 1, \cdots, E\} \cup \{\mathbf{x}^m : m = 1, \cdots, P\} \cup \{\mathbf{x}^{**}\}.$$

Define $\overline{\rho e}_K = \overline{E}_K - \frac{1}{2}\|\overline{\rho\mathbf{u}}_K\|^2/\overline{\rho}_K$. Given only point values of $\widehat{\rho e}_K(\mathbf{x})$ at $\mathbf{x}_{\nu,i}(\nu = 1, \cdots, L; i = 1, \cdots, E)$ and $\mathbf{x}^m(m = 1, \cdots, P)$, we can compute

$$\widehat{\mathbf{P}}(\mathbf{x}^{**}) = \frac{1}{1-2\widehat{\omega}}\left(\overline{\mathbf{P}}_K - \sum_{i=1}^{E}\sum_{\nu=1}^{L}\omega_{\nu,i}\widehat{\mathbf{P}}_K(\mathbf{x}_{\nu,i})\right),$$

and

$$\widetilde{\mathbf{P}}_K(\mathbf{x}) = \theta_e(\widehat{\mathbf{P}}_K(\mathbf{x}) - \overline{\mathbf{P}}_K) + \overline{\mathbf{P}}_K, \quad \theta_e = \min\left\{1, \frac{\overline{\rho e}_K - \varepsilon}{\overline{\rho e}_K - \min\limits_{\mathbf{x}\in S_e}\widehat{\rho e}_K(\mathbf{x})}\right\}. \tag{41b}$$

5.3. Time discretizations

We discretize the semi-discrete DG scheme (40) in time by a SSP Runge-Kutta method. Let $\frac{d}{dt}\mathbf{U}_h = \mathcal{L}(\mathbf{U}_h)$ denote (40), then the third order SSP Runge-Kutta method used in this paper is given by,

$$\begin{aligned}
&\mathbf{U}_h^{(1)} = \mathbf{U}_h^n + \Delta t\mathcal{L}(\mathbf{U}_h^n),\\
&\mathbf{U}_h^{(2)} = \tfrac{3}{4}\mathbf{U}_h^n + \tfrac{1}{4}(\mathbf{U}_h^{(1)} + \Delta t\mathcal{L}(\mathbf{U}_h^{(1)})),\\
&\mathbf{U}_h^{n+1} = \tfrac{1}{3}\mathbf{U}_h^n + \tfrac{2}{3}(\mathbf{U}_h^{(2)} + \Delta t\mathcal{L}(\mathbf{U}_h^{(2)})).
\end{aligned} \tag{42}$$

By Theorem 3, the DG scheme with forward Euler is positivity-preserving under the suitable CFL condition (33) with β_i defined in (30c). But we should not simply use a time step suggested by (33) for compressible Navier-Stokes equations due to the following reasons:

1. The constraint (33) may not be a necessary condition for $\overline{\mathbf{U}}_K^{n+1} \in G$ in practice.
2. To enforce (33) for three stages in (42), we need to estimate $\max_i \beta_i$ for each stage. Given DG solutions $\mathbf{U}_h^n$ at time step n, it is hard to accurately estimate $\max_i \beta_i$ with β_i defined in (30c) for the two inner time stages $\mathbf{U}_h^{(1)}$ and $\mathbf{U}_h^{(2)}$ in a third order Runge-Kutta method.
3. *Artificial stiffness* may result in unnecessarily small time steps. The wave speed $\sqrt{\gamma \frac{p}{\rho}}$ for the ideal gas EOS may not be accurately evaluated numerically for low density or low pressure problems as explained in the introduction. Another type of artificial stiffness may emerge near a very strong shock. Notice that $\boldsymbol{\tau}$ and $\mathbf{q}$ defined in (2) contains the derivatives of $\mathbf{U}$, which are not well defined near discontinuities of $\mathbf{U}$. Numerically $\boldsymbol{\tau}$ and $\mathbf{q}$ could contain huge numbers near strong shocks, thus β_i computed in (30c) might be a huge number, which is still necessary in the preserving-positivity flux though. But the time step computed by (33) could be unnecessarily small for preserving positivity.
4. For smooth solutions, (33) is inconsistent with a time step implied by the linear stability analysis. See Remark 2.

Instead, we can use the following simple time-stepping strategy: for each Runge-Kutta step, start with some desired time step, then restart the computation with a time step halved when negative cell averages emerge in any stage of Runge-Kutta. This ad hoc approach can be applied to any scheme, but Theorem 3 ensures that there will be no endless loops for such a positivity-preserving scheme. In other words, the recomputation will be ended at least when Δt is small enough to satisfy (33) for each of the three time stages. For high order schemes, it is nontrivial to find the largest time step for positivity to hold, see [49] for such an effort.

We implement the positivity-preserving high order DG scheme (40) using the flux (30b) and (30c) with the third order SSP Runge-Kutta (42) for equations (2) as follows:

1. At time level n, in each cell K, we are given DG polynomials $\mathbf{U}_K^n(\mathbf{x})$ with the cell average $\overline{\mathbf{U}}_K^n \in G^\varepsilon$, where ε be a parameter of desired lower bound for density and internal energy, e.g., $\varepsilon = 10^{-13}$ or $\varepsilon = \min\{10^{-13}, \overline{\rho}_K, \overline{\rho e}_K\}$. Apply the limiter (41) to $\mathbf{U}_K^n(\mathbf{x})$ and we obtain $\widetilde{\mathbf{U}}_K^n(\mathbf{x})$.
2. Compute the wave speed $\alpha_i = \max\limits_{\nu=1,\cdots,L} \max\limits_{\mathbf{U}_K^{\nu,i}, \mathbf{U}_{K_i}^{\nu,i}} \left(|\mathbf{u} \cdot \mathbf{n}_i| + \sqrt{\gamma \frac{p}{\rho}} \right)$ for the ideal gas EOS (2c), and $\alpha_i > \max\limits_{\nu=1,\cdots,L} \max\limits_{\mathbf{U}_K^{\nu,i}, \mathbf{U}_{K_i}^{\nu,i}} \left[|\mathbf{u} \cdot \mathbf{n}_i| + \sqrt{\frac{p^2}{2\rho^2 e}} \right]$ for a generic EOS based on $\widetilde{\mathbf{U}}_K^n(\mathbf{x})$ for each edge in the mesh. Let $\alpha^* = \max_i \alpha_i$ where the maximum is taken over all edges in the mesh. For each cell K, let e_K denote its longest edge. By abusing notation, let $\Delta x = \min_K \frac{|K|}{|e_K|}$

in a triangular mesh. Set the time step

$$\Delta t = \min\left\{ a\frac{1}{\alpha^*}\Delta x, b\,\mathrm{Re}\,\Delta x^2 \right\}, \tag{43}$$

where a and b are two parameters. For instance, (33) implies we can set $a = \frac{2}{3}\widehat{\omega}$ for a triangular cell K. Since we do not enforce (33) by using (43), we may use larger a, e.g., $a = 2\widehat{\omega}$. We choose not to use a time step dependent on β_i defined in (30c) mainly because of the artificial stiffness for strong shocks.

3. Compute the first stage based on $\widetilde{\mathbf{U}}_K^n(\mathbf{x})$, denoted by $\mathbf{U}_K^{(1)}$.
   - If the cell averages $\overline{\mathbf{U}}_K^{(1)} \in G^\varepsilon$, then proceed to next step.
   - Otherwise, then recompute the first stage with either a halved time step or the stringent CFL (33), e.g., $\Delta t = \frac{2}{3}\frac{\widehat{\omega}}{\beta^*}\Delta x$ for a triangular mesh where $\beta^* = \max_i \beta_i$ with β_i defined in (30c). Notice that Theorem 3 guarantees that the cell averages $\overline{\mathbf{U}}_K^{(1)} \in G^\varepsilon$ if (33) is used.
4. Given the DG polynomials $\mathbf{U}_K^{(1)}(\mathbf{x})$ with cell averages $\overline{\mathbf{U}}_K^{(1)} \in G^\varepsilon$, apply the limiter (41) to $\mathbf{U}_K^{(1)}(\mathbf{x})$ and we obtain $\widetilde{\mathbf{U}}_K^{(1)}(\mathbf{x})$. Compute the second stage $\mathbf{U}_K^{(2)}$ based on $\widetilde{\mathbf{U}}_K^{(1)}(\mathbf{x})$.
   - If the cell averages $\overline{\mathbf{U}}_K^{(2)} \in G^\varepsilon$, then proceed to next step.
   - Otherwise, then return to Step 3 and restart the computation with a time step halved. Notice that even if the time step (33) is used in Step 3, there is no guarantee that $\overline{\mathbf{U}}_K^{(2)} \in G^\varepsilon$ because (33) is based on $\widetilde{\mathbf{U}}_K^n$ rather than $\widetilde{\mathbf{U}}_K^{(1)}$.
5. Given the DG polynomials $\mathbf{U}_K^{(2)}(\mathbf{x})$ with cell averages $\overline{\mathbf{U}}_K^{(2)} \in G^\varepsilon$, apply the limiter (41) to $\mathbf{U}_K^{(2)}(\mathbf{x})$ and we obtain $\widetilde{\mathbf{U}}_K^{(2)}(\mathbf{x})$. Compute $\mathbf{U}_K^{n+1}(\mathbf{x})$.
   - If the cell averages $\overline{\mathbf{U}}_K^{n+1} \in G^\varepsilon$, then computation to time step $n+1$ is done.
   - Otherwise, then return to Step 3 and restart the computation with a time step halved.

6. Numerical tests

6.1. Preliminaries

We test the high order DG schemes (40) with the positivity-preserving local Lax-Friedrichs type flux (30b) and (30c) and the third order SSP Runge-Kutta (42) for the two-dimensional compressible Navier-Stokes equations with the ideal gas EOS (2) and its one-dimensional version (A.1). We use P^k or Q^k basis on rectangular cells and P^k basis on triangular cells, where P^k denotes polynomials

of degree k and Q^k denotes tensor products of one-dimensional polynomials of degree k.

In the equations (2) and (A.1), we set the parameters as $\gamma = 1.4$, $\eta = \frac{4}{3}$ and $\mathrm{Pr} = 0.72$.

In the evaluation of the flux function $\mathbf{F}$, we need to compute $u = \frac{\rho u}{\rho}$ and (39) with density in the denominator, we may encounter numerical problems if density is close to zero. To this end, if $\rho < \varepsilon^*$ where ε^* is a small positive number, we need to modify the definition of velocity and (39). In the following numerical tests, we use $\varepsilon^* = 10^{-8}$ and an ad hoc modification by setting velocity, internal energy and their derivatives as zero when $\rho < \varepsilon^*$.

In all numerical tests in this section, only the positivity-preserving limiter (41) is used. For the compressible Euler equations, i.e, (2) with $\mathrm{Re} = \infty$, high order DG schemes with only the positivity-preserving limiter in general may produce highly oscillatory numerical solutions, thus other type of limiters [45, 46, 47] should also be used to reduce oscillations. For the compressible Navier-Stokes equations, as we will see in the following examples, high order DG schemes with only the positivity-preserving limiter can produce satisfying non-oscillatory solutions when the mesh size is small enough or polynomial basis order is high enough so that the nonlinear diffusion is accurately resolved.

6.2. One-dimensional case

Example 1. Accuracy test of the positivity-preserving flux for compressible Navier-Stokes equations. *We test the accuracy of one-dimensional DG scheme with the positivity-preserving flux* (25b) *and* (25c) *for the equations* (A.1) *with* $Re = 100$. *We compare it with the central flux for the diffusion* $\mathbf{F}^d$ *used in [33], i.e., we can use the local Lax-Friedrichs flux for convection* (15b) *and* (15c), *and the central flux for diffusion* $\widehat{\mathbf{F}}^d(\mathbf{U}^-_{j+\frac{1}{2}}, \mathbf{S}^-_{j+\frac{1}{2}}, \mathbf{U}^+_{j+\frac{1}{2}}, \mathbf{S}^+_{j+\frac{1}{2}}) = \frac{1}{2}\left[\mathbf{F}^d(\mathbf{U}^-_{j+\frac{1}{2}}, \mathbf{S}^-_{j+\frac{1}{2}}) + \mathbf{F}^d(\mathbf{U}^+_{j+\frac{1}{2}}, \mathbf{S}^+_{j+\frac{1}{2}})\right]$. *The initial data is* $\rho = 1$, $u = 0$, $E = \frac{12}{\gamma-1} + \frac{1}{2}\exp(-4\cos x^2)$. *Boundary conditions are periodic on the interval* $[0, 2\pi]$. *Reference solution was generated by a Fourier collocation spectral method using* 1024 *points.* L^∞ *error over* $k+1$ *Gauss-Lobatto quadrature points in each cell at time* $T = 0.1$ *is listed in Table 1. The difference in* L^2 *errors between the two schemes is less than* 0.1%.

Table 1: Example 1. L^∞ error in total energy for DG schemes with P^k basis. The mesh size $h = \frac{2\pi}{10}$.

	central fluxes for diffusion					positivity-preserving flux for diffusion				
k	h	h/2	order	h/4	order	h	h/2	order	h/4	order
2	1.92E-2	3.43E-3	2.49	4.45E-4	2.94	1.92E-2	3.43E-3	2.49	4.45E-4	2.94
3	2.10E-3	4.78E-4	2.13	3.67E-5	3.70	2.10E-3	4.78E-4	2.13	3.67E-5	3.70
4	1.51E-3	5.29E-5	4.84	1.73E-6	4.93	1.51E-3	5.29E-5	4.84	1.73E-6	4.93
5	2.49E-4	5.53E-6	5.49	9.39E-8	5.88	2.49E-4	5.53E-6	5.49	9.39E-8	5.88

Example 2. The Lax shock tube problem. *We test the performance of the positivity-preserving DG scheme for the Lax shock tube problem. The initial condition can be found in [50]. See Figure 5 for the numerical solutions for compressible Euler equations, where oscillations can be observed. See Figure 6 for the numerical solutions for compressible Navier-Stokes equations, where the reference solution was generated by the second order finite difference scheme discussed in Appendix A using a fifth order positivity-preserving WENO flux for* $\mathbf{F}^a$ *in [51] with the second order approximation for diffusion on a mesh of* 64000 *grid points. For compressible Euler equations, the time step* (43) *for this example is replaced by* $\Delta t = \frac{1}{4}\frac{1}{N(N-1)}\frac{1}{\alpha^*}\Delta x$. *For compressible Navier-Stokes equations, the two parameters in the time step* (43) *for this example are set as* $a = \frac{1}{4}\frac{1}{N(N-1)}$ *and* $b = 0.001$.

Example 3. Double rarefaction. *This is a Riemann problem with the initial condition as*

$$\begin{pmatrix}\rho\\u\\p\end{pmatrix} = \begin{pmatrix}7\\-1\\0.2\end{pmatrix} \quad \text{if} \quad x \le 0; \qquad \begin{pmatrix}\rho\\u\\p\end{pmatrix} = \begin{pmatrix}7\\1\\0.2\end{pmatrix} \quad \text{if} \quad x > 0.$$

The exact solution for the compressible Euler equations contains zero density and zero pressure, see [10, 44]. See Figure 7 for the numerical solutions for compressible Navier-Stokes equations with $\mathrm{Re} = 1000$, *which contains low density and low pressure. The reference solution was generated by the second order finite difference scheme discussed in the Appendix Appendix A on a mesh of* 32000 *points. The two parameters in the time step* (43) *for this example are set as* $a = \frac{1}{4}\frac{1}{N(N-1)}$ *and* $b = 0.001$.

6.3. Two-dimensional case

Example 4. An accuracy test of the positivity-preserving limiter. *Consider an isentropic vortex evolution problem for 2D Euler equations, i.e.,* (2) *with* $\mathrm{Re} = \infty$, *and the following exact solution:* $\rho^{(\gamma-1)} = 1 - \frac{(\gamma-1)\epsilon^2}{8\gamma\pi^2}\exp(1-r^2), p = \rho^\gamma, u = 1 - \frac{\epsilon}{2\pi}\exp 0.5(1-r^2)(y-5-t), v = 1 + \frac{\epsilon}{2\pi}\exp 0.5(1-r^2)(x-5-t)$, *where* $r^2 = (x-5-t)^2 + (y-5-t)^2$ *and the vortex strength* ϵ *is a constant. We test the accuracy of the positivity-preserving limiter* (41) *on DG schemes at time* $T = 0.1$ *on uniform rectangular meshes and unstructured triangular meshes for a square* $[0,10]\times[0,10]$. *The vortex strength is taken as* $\epsilon = 9.5$ *and the lowest density and pressure of the exact solution is* 4.22×10^{-3} *and* 4.74×10^{-4} *respectively. We check the error for the* α*-optimized nodal values [52] in each triangular cell and* $(k+1)^2$ *uniform grid point values in each rectangular cell in the region* $[2,8]\times[2,8]$ *for polynomials of degree* k. *See Table 2 and Table 3, where* L^1 *error is defined as the average of the magnitude of errors over these points.*

Example 5. An accuracy test of the positivity-preserving flux for compressible Navier-Stokes equations. *Consider the compressible Navier-Stokes*

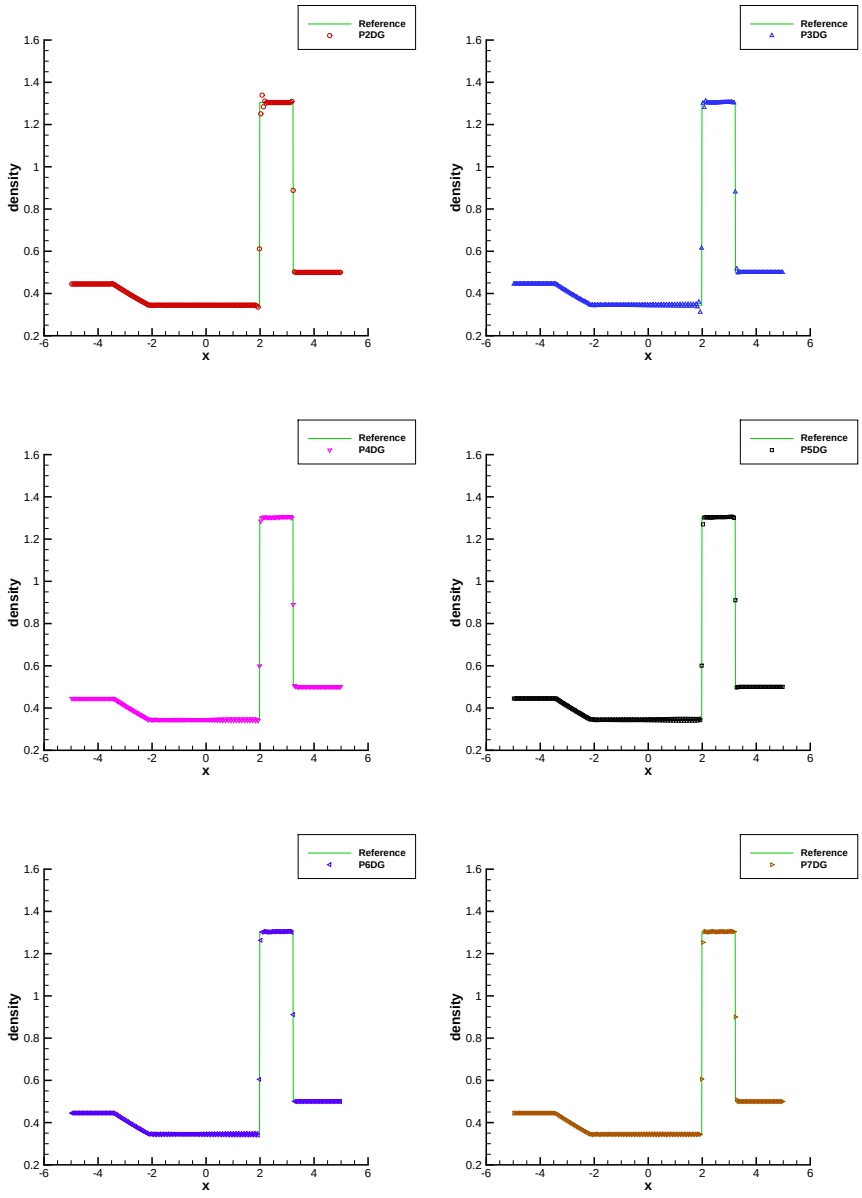
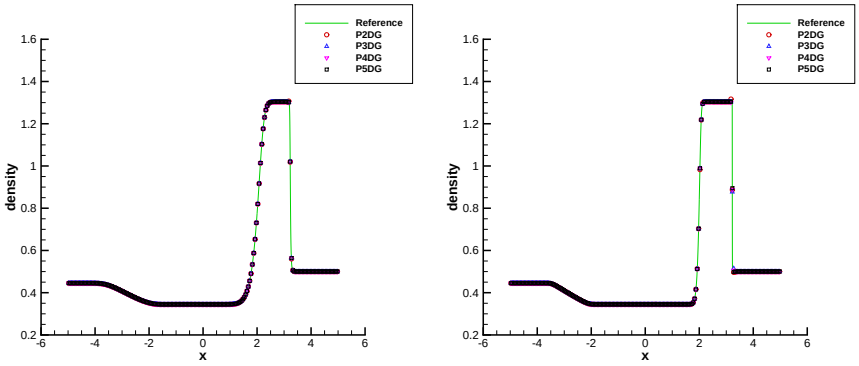
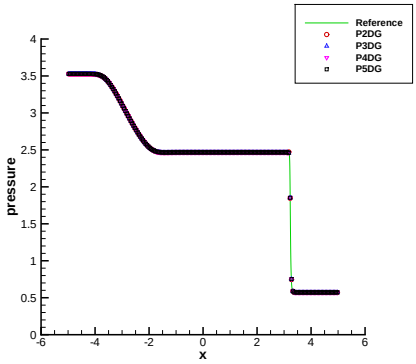


Figure 5: Example 2. DG using P^k basis with only the positivity-preserving limiter on 200 uniform cells for compressible Euler equations. Only cell averages are plotted.

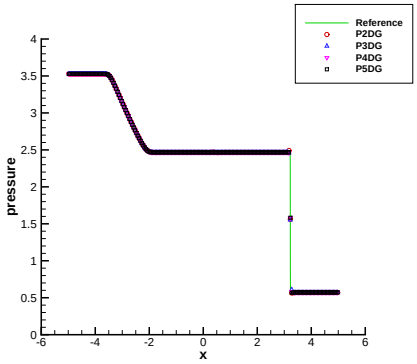


(a) Re = 100.

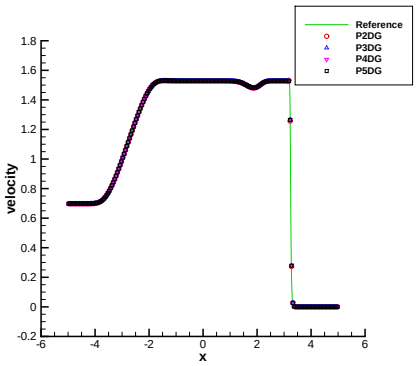
(b) Re = 1000.



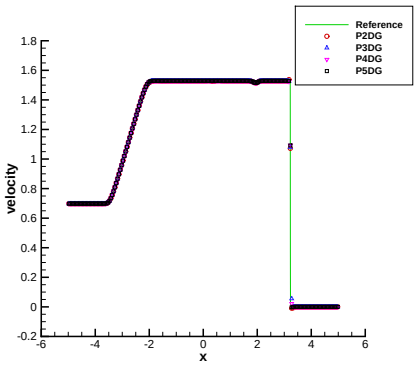
(c) Re = 100.



(d) Re = 1000.

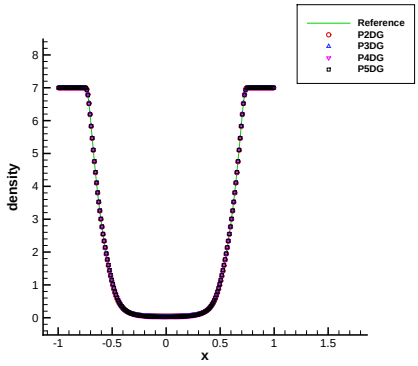


(e) Re = 100.

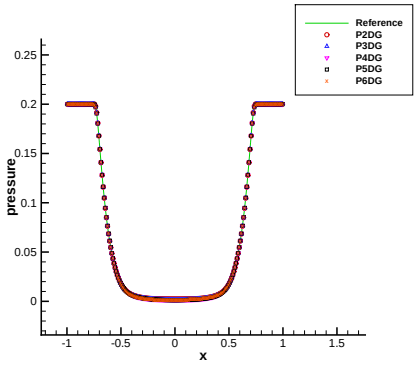


(f) Re = 1000.

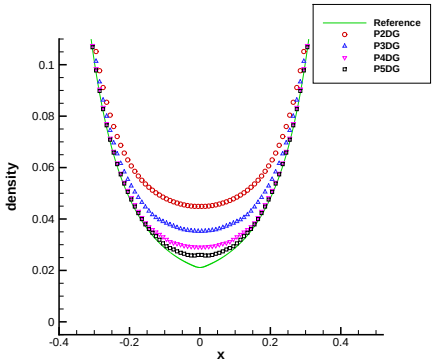
Figure 6: Example 2. DG using P^k basis with only the positivity-preserving limiter on 200 uniform cells for compressible Navier-Stokes equations. Only cell averages are plotted.



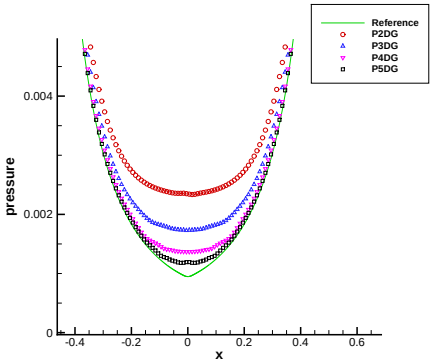
(a) Density on 200 uniform cells.



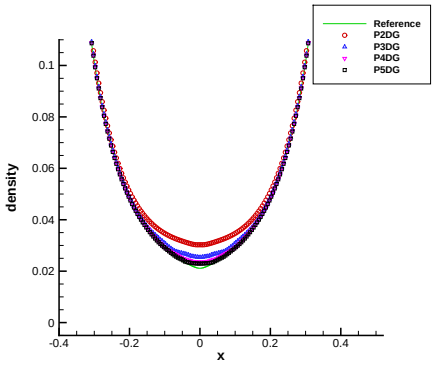
(b) Pressure on 200 uniform cells.



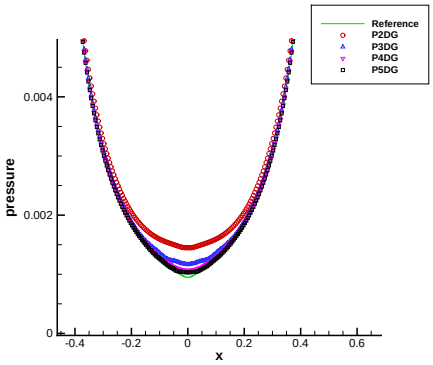
(c) Density on 200 uniform cells.



(d) Pressure on 200 uniform cells.



(e) Density on 400 uniform cells.



(f) Pressure on 400 uniform cells.

Figure 7: Example 3. DG with P^k basis for 1D double rarefaction wave, Re = 1000. Only cell averages are plotted.

Table 2: Example 4. L^1 error in total energy for positivity-preserving DG schemes with P^k basis on unstructured triangular meshes. The mesh sizes of three unstructured triangular meshes are around h, $h/2$ and $h/4$ respectively, where $h = \frac{10}{16}$.

k	Mesh1	Mesh2	order	Mesh3	order
2	2.86E-3	3.70E-4	2.95	5.11E-5	2.86
3	3.87E-4	2.30E-5	4.08	1.49E-6	3.94
4	5.29E-5	1.68E-6	4.98	5.92E-8	4.83
5	7.68E-6	1.20E-7	6.00	1.84E-9	6.03

Table 3: Example 4. L^1 error in total energy for positivity-preserving DG schemes with P^k basis on uniform rectangular meshes. The mesh size $h = \frac{10}{16}$.

k	h	h/2	order	h/4	order
2	3.35E-3	4.35E-4	2.95	5.51E-5	2.98
3	1.61E-3	5.19E-5	4.96	2.79E-6	4.22
4	5.11E-4	4.88E-6	6.71	1.64E-7	4.90
5	2.50E-5	4.76E-7	5.71	8.58E-9	5.79

equations (2) *with the initial condition:* $\rho = 1$, $u = v = 0$, $E = \frac{12}{\gamma-1} + \frac{1}{2}\exp(-4\cos(\frac{x}{2})^2 - 4\cos(\frac{y}{2})^2)$ *and periodic boundary conditions. on the domain* $[0, 2\pi] \times [0, 2\pi]$. *The reference solution was generated by a Fourier collocation spectral method on a* 512×512 *grid.*

For compressible Euler equations, the time step (43) *for this example is replaced by* $\Delta t = \frac{1}{2}\frac{1}{N(N-1)}\frac{1}{\alpha^*}\Delta x$ *on uniform rectangular meshes with* $\Delta x = \Delta y$, *and by* $\Delta t = \frac{2}{3}\frac{1}{N(N-1)}\frac{1}{\alpha^*}\Delta x$ *on unstructured triangular meshes* ($\Delta x = \min_K \frac{|K|}{e_K}$). *For compressible Navier-Stokes equations, the two parameters in the time step* (43) *for this example are set as* $a = \frac{1}{2}\frac{1}{N(N-1)}$ *and* $b = 0.001$ *on uniform rectangular meshes, and* $a = \frac{2}{3}\frac{1}{N(N-1)}$ *and* $b = 1$ *on unstructured triangular meshes.*

For P^k *basis and* Q^k *basis on rectangles,* L^∞ *error is defined as the maximum pointwise error for uniform* $(k+1)^2$ *points in each cell. For* P^k *basis on triangles,* L^∞ *error is defined as the maximum pointwise error for the* α*-optimized nodal values [52] in each triangle. See the errors at time* $T = 0.1$ *in Table (4).*

Example 6. Sedov blast wave. *We use a uniform* 160×160 *rectangular mesh for the domain* $[0, 1.1] \times [0, 1.1]$. *The initial condition is set up as piecewise constants: density is constant* 1 *and velocity is zero everywhere; the pressure is* 10^{-5} *on all cells except the one in the lower left corner; the total energy in the cell of lower left corner is* $0.244816/\Delta x^2$ *where the mesh size* $\Delta x = \frac{1.1}{160}$. *The physical meaning of this initial condition for compressible Navier-Stokes system is very limited. Nonetheless, this is a good low density and low pressure test for a positivity-preserving scheme.*

The boundary conditions for the left and bottom edges of the domain is re-

Table 4: Example 5. L^∞ error in total energy for DG schemes with the positivity-preserving flux (30b) and (30c) on uniform rectangular meshes and unstructured triangular meshes. The mesh size for the coarsest uniform rectangular mesh is $h = \frac{2\pi}{8}$. The maximum edge length in three unstructured triangular meshes are around h, $h/2$ and $h/4$ respectively.

	P^k basis on rectangular meshes									
	Euler equations, i.e., Re $= \infty$					Navier-Stokes equations, Re $= 100$				
k	h	h/2	order	h/4	order	h	h/2	order	h/4	order
2	1.19E-2	2.72E-3	2.13	3.57E-4	2.93	1.17E-2	2.58E-3	2.18	3.42E-4	2.92
3	5.14E-3	3.01E-4	4.09	2.49E-5	3.60	5.01E-3	2.91E-4	4.11	2.34E-5	3.64
4	6.34E-4	4.17E-5	3.93	1.29E-6	5.01	6.01E-4	3.87E-5	3.96	1.23E-6	4.98
5	1.91E-4	3.62E-6	5.72	7.48E-8	5.60	1.90E-4	3.41E-6	5.80	6.50E-8	5.71
	Q^k basis on rectangular meshes									
	Euler equations, i.e., Re $= \infty$					Navier-Stokes equations, Re $= 100$				
k	h	h/2	order	h/4	order	h	h/2	order	h/4	order
2	6.52E-3	1.49E-3	2.13	2.22E-4	2.75	6.61E-3	1.48E-3	2.16	2.21E-4	2.74
3	2.76E-3	1.23E-4	4.48	9.46E-6	3.70	2.59E-3	1.17E-4	4.46	8.91E-6	3.72
4	1.04E-4	1.09E-5	3.24	3.89E-7	4.82	1.07E-4	1.10E-5	3.28	3.88E-7	4.83
5	4.66E-5	5.54E-7	6.39	1.05E-8	5.72	4.47E-5	4.84E-7	6.53	8.78E-9	5.78
	P^k basis on unstructured triangular meshes									
	Euler equations, i.e., Re $= \infty$					Navier-Stokes equations, Re $= 100$				
k	Mesh1	Mesh2	order	Mesh2	order	Mesh1	Mesh2	order	Mesh3	order
2	1.29E-2	1.54E-3	3.06	1.32E-4	3.54	1.25E-2	1.44E-3	3.12	1.33E-4	3.43
3	1.52E-3	1.30E-4	3.55	8.32E-6	3.96	1.54E-3	1.21E-4	3.67	7.26E-6	4.05
4	3.28E-4	1.10E-5	4.90	2.33E-7	5.56	3.15E-4	1.03E-5	4.93	2.95E-7	5.12
5	3.90E-5	8.56E-7	5.51	1.09E-8	6.30	4.06E-5	8.40E-7	5.60	9.99E-9	6.39

flective, defined as follows: we extend the density, total energy and tangential velocity of $\mathbf{U}_h$ *in* (40) *as an even function across the boundary and extend the normal velocity of* $\mathbf{U}_h$ *as an odd function across the boundary. For the auxiliary variable* $\mathbf{S}_h$ *in* (40) *approximating derivatives of the conserved variables, we also need to specify the boundary conditions. We extend the tangential derivatives of conserved variables as an even function across the boundary. Then we extend the normal derivatives of density, total energy and tangential component of momentum as an odd function and extend the normal derivatives of the normal component of momentum as an even function. The reflective boundary conditions mimic the symmetry of the solution across the left and bottom edges of the domain in this example.*

We test the DG scheme with only the positivity-preserving limiter for solving compressible Euler equations and compressible Navier-Stokes equations with $\mathrm{Re} = 200$. *For compressible Navier-Stokes equations, the two parameters in the time step* (43) *for this example are set as* $a = \frac{1}{2}\frac{1}{N(N-1)}$ *and* $b = 0.002$. *For compressible Euler equations in this example, we replace* (43) *by* $\Delta t = \max\{\frac{1}{2}\frac{1}{N(N-1)}\frac{1}{\alpha^*}\Delta x, 2\Delta x^2\}$. *The exact solution of Sedov blast wave for compressible Euler equations contains zero density, thus we do encounter the issue of artificial stiffness due to low density/pressure since the wave speed* $\sqrt{\gamma p/\rho}$ *may be a huge number numerically, which results in a much smaller time step than a necessary one for positivity. The max with* $2\Delta x^2$ *in the time step above is used to avoid this artificial stiffness.*

See Figure 8 for the plots of density for DG schemes with P^2, P^3 and P^4 bases.

Example 7. Shock diffraction. *The initial condition is a pure right-moving shock of Mach number* 5.09*, initially located at* $x = 0.5$ *and* $6 \le y \le 11$*, moving into undisturbed air ahead of the shock with a density of* 1.4 *and a pressure of* 1*. See [53] for a more detailed description of the set up. The boundary conditions are inflow at* $x = 0$*,* $6 \le y \le 11$*, outflow at* $x = 13$*,* $6 \le y \le 11$*, reflective at* $0 \le x \le 1$*,* $y = 6$ *and at* $x = 1$*,* $0 \le y \le 6$*. See Example 6 for the description of the reflective boundary condition, which mimics a weakly imposed no-penetration boundary condition in this example.*

We use uniform rectangular meshes with $\Delta x = \Delta y$*. For compressible Euler equations, the time step (43) for this example is replaced by* $\Delta t = \frac{1}{2}\frac{1}{N(N-1)}\frac{1}{\alpha^*}\Delta x$*. For compressible Navier-Stokes equations, the two parameters in the time step (43) for this example are set as* $a = \frac{1}{2}\frac{1}{N(N-1)}$ *and* $b = 0.005$*.*

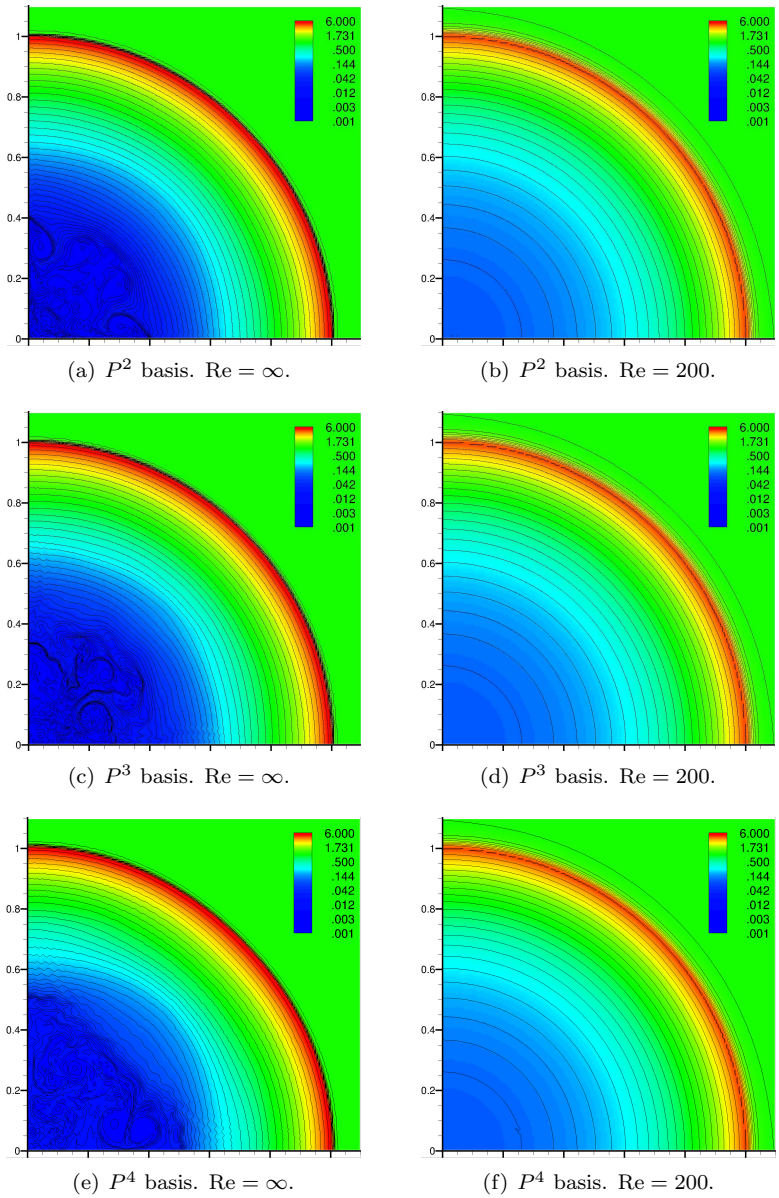
No special treatment is done at the corner which is a singularity of the solution. It is well known that the diffraction of high speed shocks at a sharp corner may result in low density and low pressure. See the plot of density at $T = 2.3$ *in Figure 9 for DG using* P^k *and* Q^k *bases solving compressible Euler equations and compressible Navier-Stokes equations with* $\mathrm{Re} = 200$*. For Euler equations, We can observe that high order DG schemes with only the positivity-preserving limiter may produce highly oscillatory solutions and such oscillations may affect the shock location. For instance, see DG with* P^3 *basis for compressible Euler equations in Figure 9.*

Example 8. Double Mach reflection of a Mach 10 **shock.** *The set up of the initial condition and boundary conditions are exactly the same as those in [53], i.e., a Mach* 10 *shock initially making a sixty degree angle with a reflecting wall. The same reflective boundary conditions as described in Example 6 are used for the reflecting wall.*

We use uniform rectangular meshes with $\Delta x = \Delta y$*. For compressible Euler equations, the time step (43) for this example is replaced by* $\Delta t = \frac{1}{2}\frac{1}{N(N-1)}\frac{1}{\alpha^*}\Delta x$*. For compressible Navier-Stokes equations, the two parameters in the time step (43) for this example are set as* $a = \frac{1}{2}\frac{1}{N(N-1)}$ *and* $b = 0.0001$*.*

For compressible Euler equations, see Figure 10 for DG schemes with only the positivity-preserving limiter. Compared to the numerical results of DG schemes with high order TVB limiter in [53], DG schemes with only the positivity-preserving limiter can capture more detailed structure in the blown-up region shown in Figure 10 (e), which suggests the low numerical dissipation of the positivity-preserving limiter. However, the solutions for DG schemes with only the positivity-preserving limiter solving compressible Euler equations are more oscillatory on a finer mesh or with a higher order basis. **Figure 10 (c) suggests that other type limiters must also be used to reduce oscillations for compressible Euler equations.**

On the other hand, the performance of the DG schemes with only the positivity-preserving limiter for compressible Navier-Stokes equations is somehow the op-



(a) P^2 basis. $\mathrm{Re} = \infty$. (b) P^2 basis. $\mathrm{Re} = 200$.

(c) P^3 basis. $\mathrm{Re} = \infty$. (d) P^3 basis. $\mathrm{Re} = 200$.

(e) P^4 basis. $\mathrm{Re} = \infty$. (f) P^4 basis. $\mathrm{Re} = 200$.

Figure 8: Sedov blast wave in Example 6. DG schemes using P^k basis **with only the positivity-preserving limiter** on a 160×160 rectangular mesh. 50 exponentially distributed contour lines of density from 0.001 to 6. Left column are results for compressible Euler equations. Right column are results for compressible Navier-Stokes equations with $\mathrm{Re} = 200$.

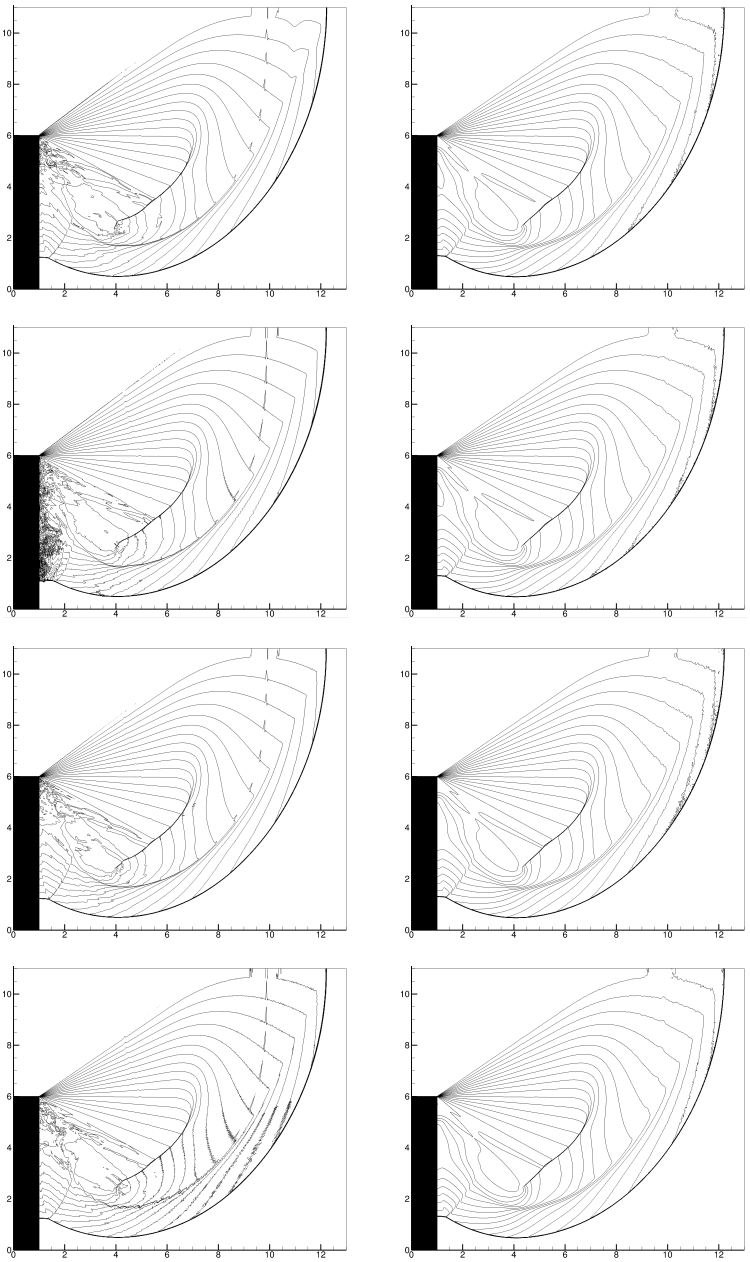


Figure 9: Example 7. DG schemes using P^k and Q^k bases **with only the positivity-preserving limiter** on a uniform rectangular mesh with mesh size $\frac{1}{64}$. Plot of density: 20 equally spaced contour lines from 0.066227 to 7.0668. Left column are results for compressible Euler equations. Right column are results for compressible Navier-Stokes equations with $\mathrm{Re} = 200$. From top to bottom: P^2 basis, P^3 basis, Q^2 basis and Q^3 basis.

posite: solutions are less oscillatory on a finer mesh or with a higher order basis. This is a strong numerical evidence that the physical diffusion starts to smooth the numerical solutions when it is marginally resolved on a fine enough mesh or by an accurate enough scheme. See similar observations in [54]. It also indicates that no excessive artificial viscosity is added to high order DG schemes by the positivity-preserving local Lax-Friedrichs type flux (30b) *and* (30c) *and the positivity-preserving limiter. See Figure 11 for numerical solutions for* $\mathrm{Re} = 100$*, and Figure 12 and Figure 13 for numerical solutions for* $\mathrm{Re} = 500$*.*

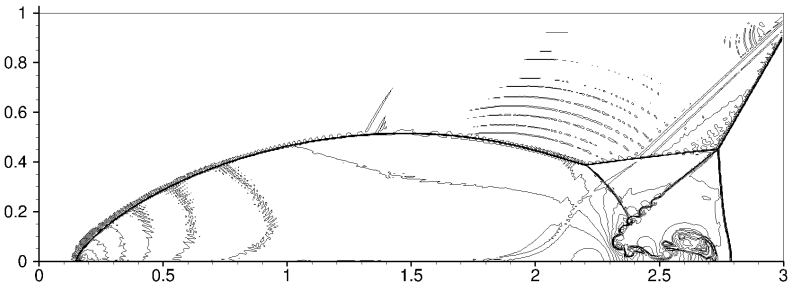
Example 9. Mach 10 Shock reflection and diffraction. *The domain is a wedge is bounded by segments connecting the points* $(0.1, 0), (0.2, 0), (1.2, \frac{\sqrt{3}}{3})$, $(1.2, 0), (2.8, 0), (2.8, 2.0), (0.1, 2)$*. See Figure 14 for an illustration of the domain. The initial condition is a right-moving Mach* 10 *shock located at the line* $x = 0.2$*. For the area where* $x > 0.2$*,* $(\rho, u, p) = (1.4, 0, 1)$*. The boundary conditions are set up as follows: the exact solution of a right-moving Mach* 10 *shock for compressible Euler equations is used for the top edge; inflow and outflow boundary conditions are used for the left and right edges respectively; reflective boundary conditions as described in Example 6 are used to weakly impose a no-penetration boundary condition for the rest of boundary.*

The right-moving shock will be first reflected by the wall making sixty degree to the shock, which is exactly the same setup as in Example 8. After the shock passing the sharp corner, diffraction happens, which is similar to the set up in Example 7. In a nutshell, this test case is a combination of Example 8 and Example 7 involving not only shocks but also low density, low pressure and complicated fine structure due to the Kelvin Helmholtz instability generated in the shock reflection. These features make this test quite a representative numerical test for a positivity-preserving high order scheme.

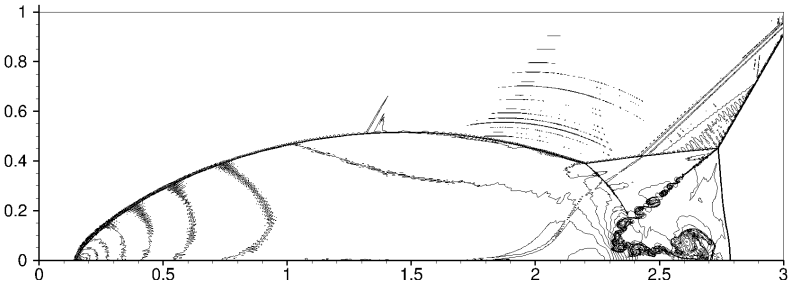
For compressible Euler equations, the time step (43) *is replaced by* $\Delta t = \max\{\frac{1}{N(N-1)}\frac{1}{\alpha^*}\Delta x, 10\Delta x^2\}$ *where the max with* $10\Delta x^2$ *is to avoid artificial stiffness induced by the low density/pressure. For compressible Navier-Stokes equations, the two parameters in the time step* (43) *for this example are set as* $a = \frac{1}{N(N-1)}$ *and* $b = 0.01$*.*

See Figure 2 first for the effect of the TVB limiter. The TVB limiter successfully reduces the oscillations, even though DG schemes with only the TVB limiter are not stable for this example due to emergence of negative density or negative pressure. On the other hand, the TVB limiter induces more artificial viscosity than the positivity-preserving limiter, which smears interesting fine features generated by the Kelvin-Helmholtz instability on the relatively coarse mesh in Figure 2. See Figure 15 for more results of DG schemes with only the positivity-preserving limiter on an unstructured triangular mesh for compressible Euler equations. As we have seen in previous examples, the solutions are more oscillatory on a finer mesh or with a higher order basis.

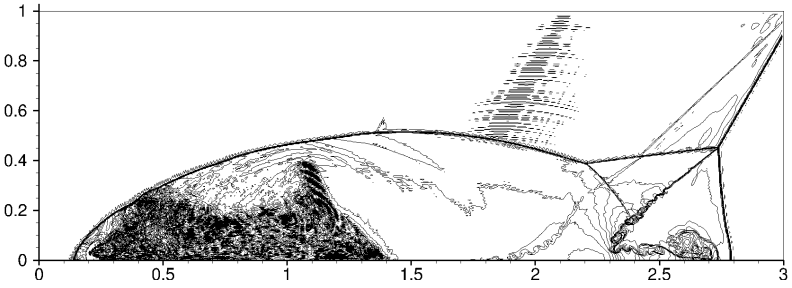
See Figure 16 for the results for compressible Navier-Stokes equations with $\mathrm{Re} = 100$*. We observe no Kelvin-Helmholtz instability and consistent numerical*



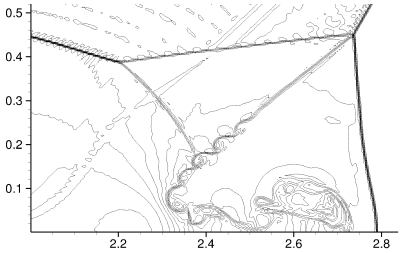
(a) P^2 basis. $\Delta x = \frac{1}{240}$.



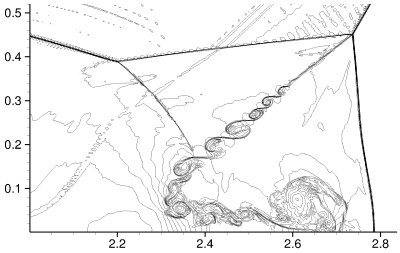
(b) P^2 basis. $\Delta x = \frac{1}{480}$.



(c) P^3 basis. $\Delta x = \frac{1}{240}$.

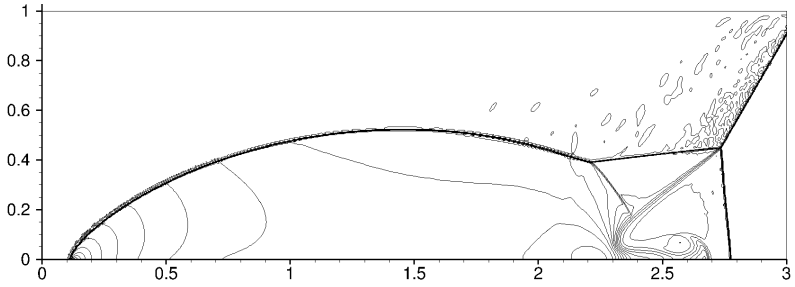


(d) P^2 basis. $\Delta x = \frac{1}{240}$.

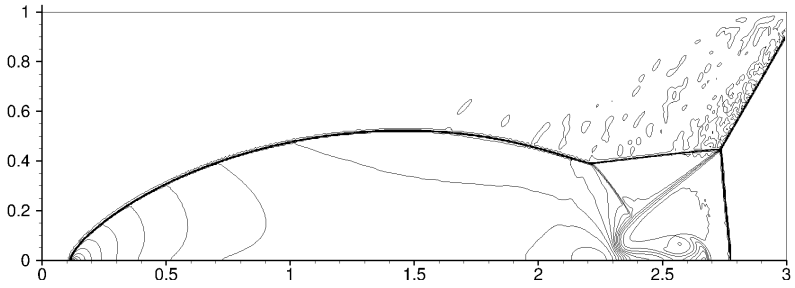


(e) P^2 basis. $\Delta x = \frac{1}{480}$.

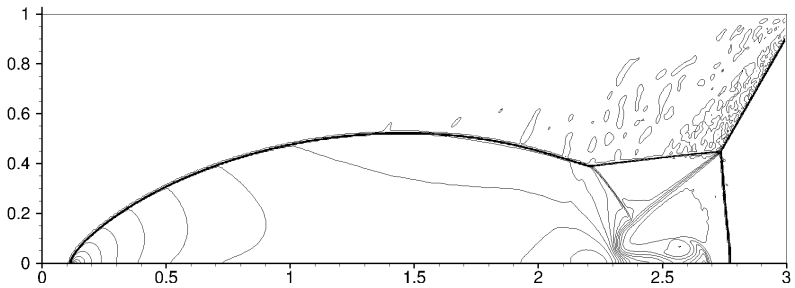
Figure 10: Double Mach Reflection in Example 8. DG schemes using P^k basis **with only the positivity-preserving limiter** on a uniform rectangular mesh with $\Delta x = \Delta y$ for compressible Euler equations. Plot of density: 30 equally spaced contour lines from 1.3965 to 22.682. The solutions are more oscillatory on a finer mesh or with a higher order basis.



(a) P^2 basis. Mesh size is $\Delta x = \frac{1}{240}$.

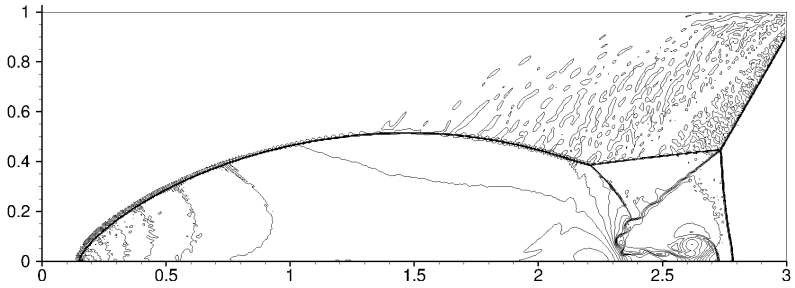


(b) P^3 basis. Mesh size is $\Delta x = \frac{1}{192}$.

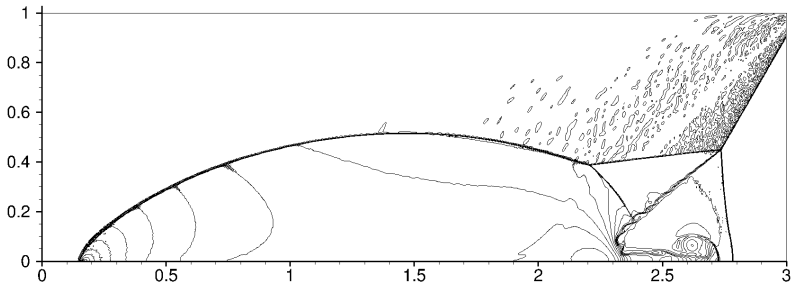


(c) P^4 basis. Mesh size is $\Delta x = \frac{1}{192}$.

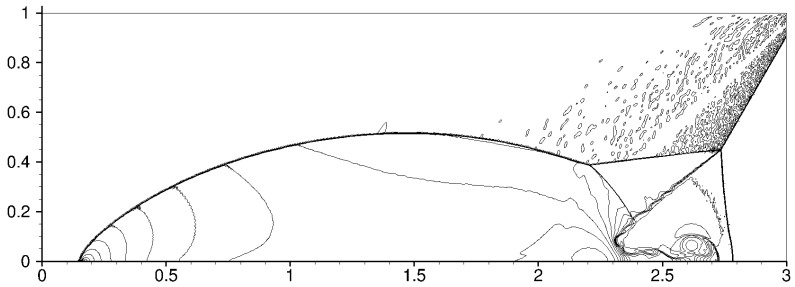
Figure 11: Double Mach Reflection in Example 8. DG schemes using P^k basis **with only the positivity-preserving limiter** on a uniform rectangular mesh with $\Delta x = \Delta y$ for compressible Navier-Stokes equations with Re = 100. Plot of density: 30 equally spaced contour lines from 1.3965 to 22.682.



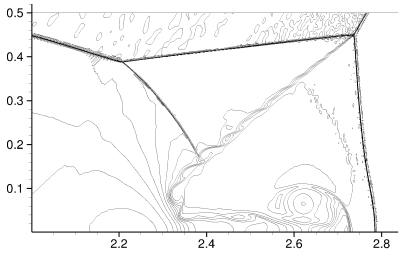
(a) P^2 basis. Mesh size is $\Delta x = \frac{1}{240}$.



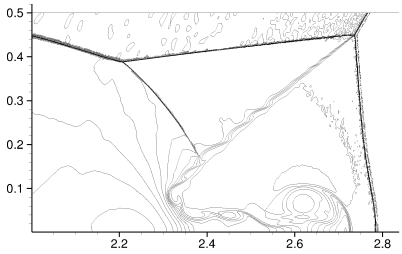
(b) P^2 basis. Mesh size is $\Delta x = \frac{1}{480}$.



(c) P^3 basis. Mesh size is $\Delta x = \frac{1}{480}$.

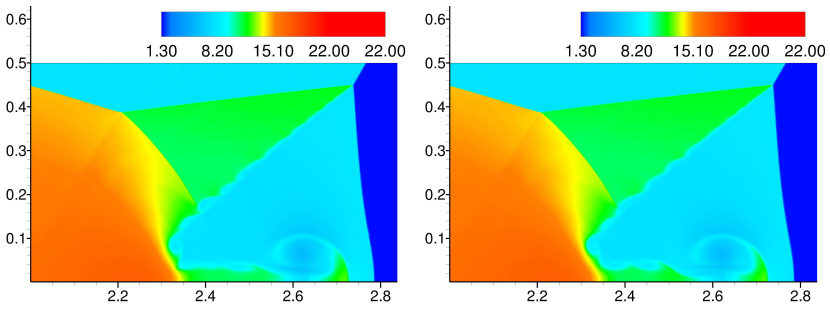


(d) P^2 basis. Mesh size is $\Delta x = \frac{1}{480}$.



(e) P^3 basis. Mesh size is $\Delta x = \frac{1}{480}$.

Figure 12: Double Mach Reflection in Example 8. DG schemes using P^k basis **with only the positivity-preserving limiter** on a uniform rectangular mesh with $\Delta x = \Delta y$ for compressible Navier-Stokes equations with Re = 500. Plot of density: 30 equally spaced contour lines from 1.3965 to 22.682.



(a) P^2 basis. Mesh size is $\frac{1}{480}$. (b) P^3 basis. Mesh size is $\frac{1}{480}$.

Figure 13: Double Mach Reflection in Example 8. DG schemes using P^k basis **with only the positivity-preserving limiter** on a uniform rectangular mesh with $\Delta x = \Delta y$ for compressible Navier-Stokes equations with Re = 500. Color contour of density for the blown-up region.

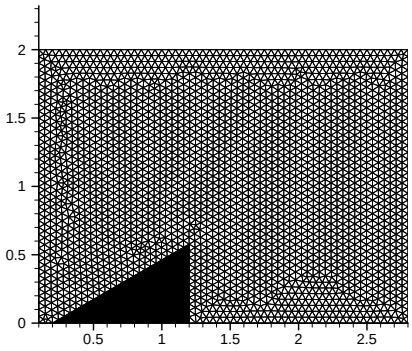
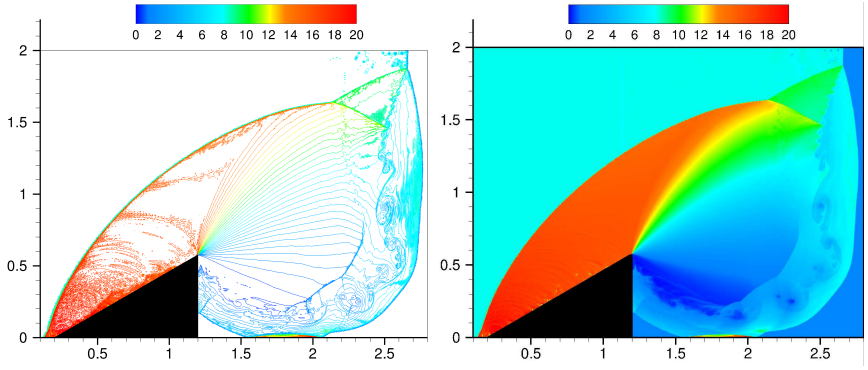


Figure 14: Example 9. An illustration of the domain and an unstructured triangular mesh with mesh size (the maximum edge length) equal to $\frac{1}{20}$.

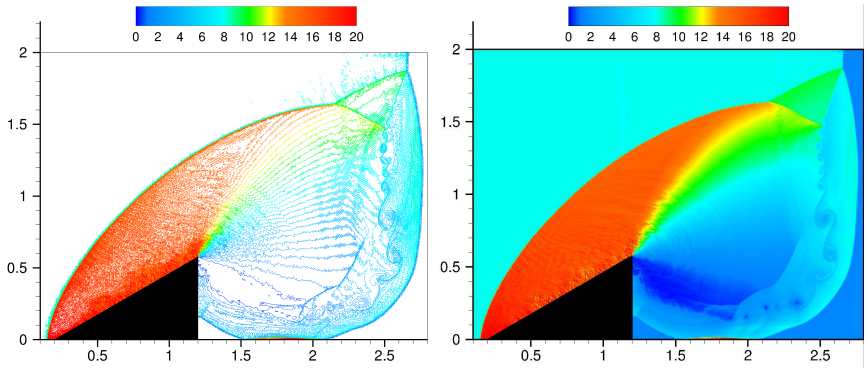
results from DG schemes using different polynomial bases on different meshes. A higher order scheme on a finer mesh produces less oscillatory solutions.

See Figure 17 for the results for compressible Navier-Stokes equations with $\mathrm{Re} = 1000$. *The numerical results from different DG schemes are at least qualitatively comparable. A higher order scheme on the same mesh also produces less oscillatory solutions.*



(a) P^2 basis. Mesh size is $\frac{1}{160}$.

(b) P^2 basis. Mesh size is $\frac{1}{160}$.



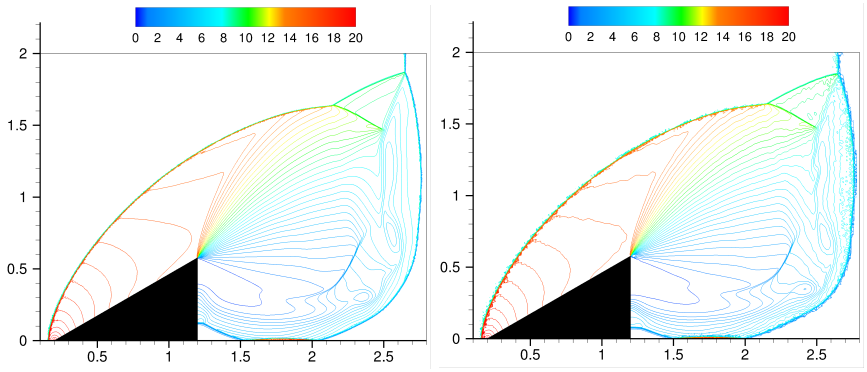
(c) P^4 basis. Mesh size is $\frac{1}{80}$.

(d) P^4 basis. Mesh size is $\frac{1}{80}$.

Figure 15: Example 9. DG schemes using P^k basis **with only the positivity-preserving limiter** on an unstructured triangular mesh for compressible Euler equations. Plot of density: 50 equally spaced contour lines from 0.05 to 25. Compared to Figure 2, the solutions are more oscillatory on a finer mesh or with a higher order basis.

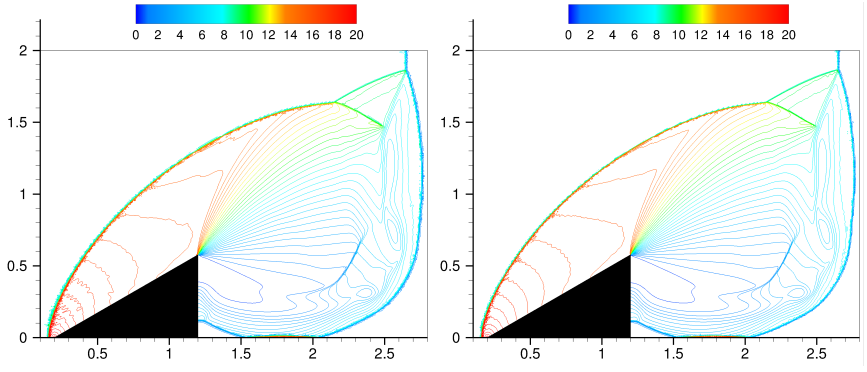
7. Concluding remarks

In this paper, we have constructed a positivity-preserving local Lax-Friedrichs type flux for compressible Navier-Stokes equations. Finite volume and DG schemes with this flux satisfy the same weak positivity property as in [10, 18, 19],



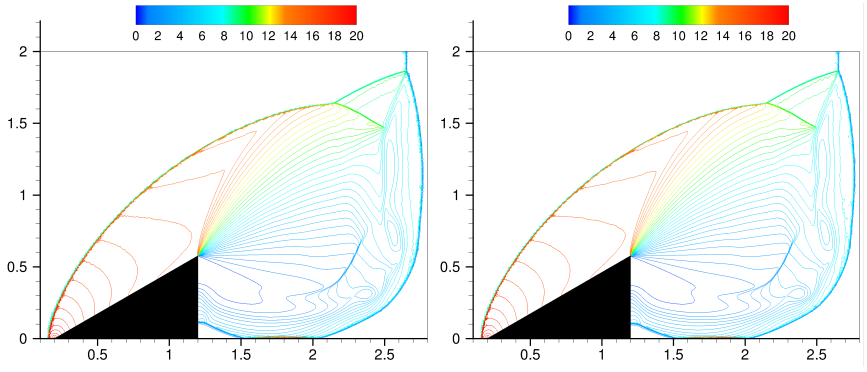
(a) P^2 basis. Mesh size is $\frac{1}{160}$.

(b) P^8 basis. Mesh size is $\frac{1}{40}$.



(c) P^2 basis. Mesh size is $\frac{1}{80}$.

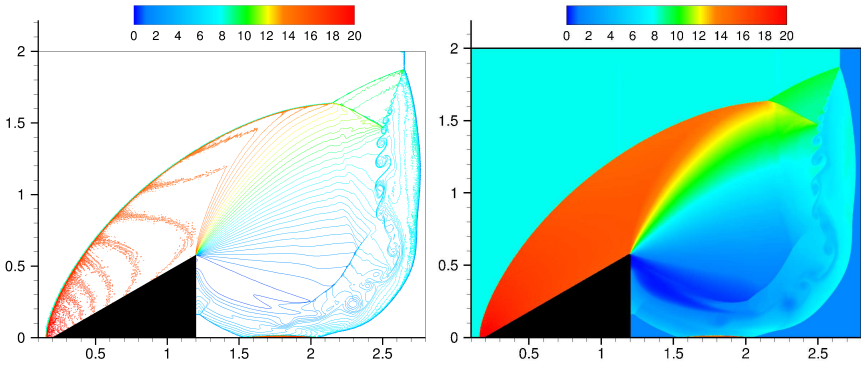
(d) P^3 basis. Mesh size is $\frac{1}{80}$.



(e) P^4 basis. Mesh size is $\frac{1}{80}$.

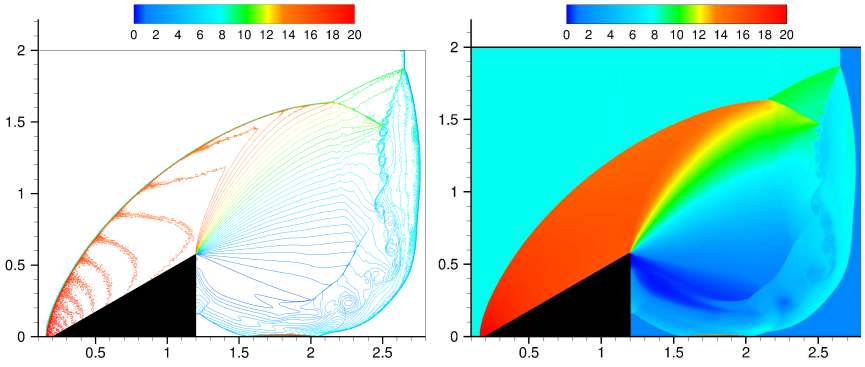
(f) P^5 basis. Mesh size is $\frac{1}{80}$.

Figure 16: Example 9. DG schemes using P^k basis **with only the positivity-preserving limiter** on an unstructured triangular mesh for compressible Navier-Stokes equations with $\mathrm{Re} = 100$. Plot of density: 50 equally spaced contour lines from 0.05 to 25.



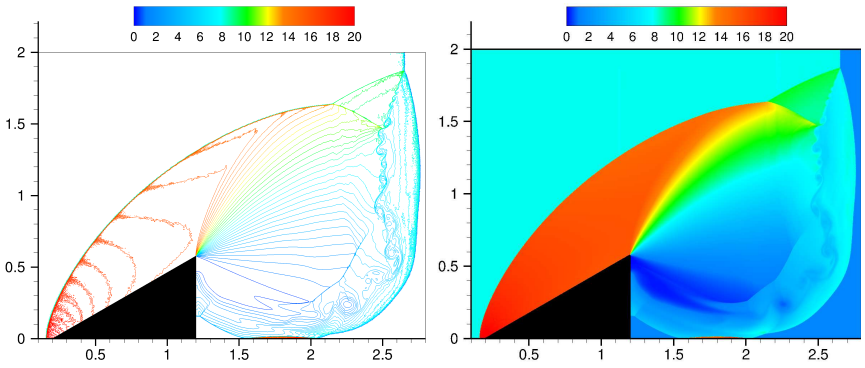
(a) P^4 basis. Mesh size is $\frac{1}{160}$.

(b) P^4 basis. Mesh size is $\frac{1}{160}$.



(c) P^5 basis. Mesh size is $\frac{1}{160}$.

(d) P^5 basis. Mesh size is $\frac{1}{160}$.



(e) P^7 basis. Mesh size is $\frac{1}{160}$.

(f) P^7 basis. Mesh size is $\frac{1}{160}$.

Figure 17: Example 9. DG schemes using P^k basis **with only the positivity-preserving limiter** on an unstructured triangular mesh for compressible Navier-Stokes equations with Re = 1000. Plot of density: 50 equally spaced contour lines from 0.05 to 25. Solutions of higher order schemes are less oscillatory.

i.e., the cell averages in a forward Euler time discretization will have positive density and positive internal energy if certain initial point values have positive density and internal energy under a suitable time step constraint. Higher order time discretizations are achieved by SSP Runge-Kutta methods. The weak positivity property makes it possible to construct a conservative positivity-preserving high order scheme.

We demonstrate that a high order DG scheme with a simple positivity-preserving flux and an efficient positivity-preserving limiter is conservative, positivity-preserving and high order accurate. This approach of constructing positivity-preserving high order schemes has the following features:

- The scheme is fully explicit and in practice time step must be not larger than $\mathcal{O}(\mathrm{Re}\,\Delta x^2)$, which makes the scheme more suitable for high Reynolds number flows.
- It applies to arbitrarily high order polynomial basis on cells of general shapes.
- The construction of the positivity-preserving flux does not depend on how derivatives of solutions are approximated in numerical schemes or specific forms of the equations of state, the stress tensor and the heat flux.

Numerical tests suggest that the proposed DG scheme does not induce excessive artificial viscosity even if strong shocks are present. In particular, for compressible Navier-Stokes equations, a higher order positivity-preserving DG scheme is less oscillatory, which is an indication that the physical diffusion may properly smooth numerical solutions.

Appendix A.

Consider the one-dimensional compressible Navier-Stokes equations in the nondimensional form:

$$\begin{pmatrix}\rho\\ \rho u\\ E\end{pmatrix}_t + \begin{pmatrix}\rho u\\ \rho u^2+p\\ (E+p)u\end{pmatrix}_x = \frac{1}{\mathrm{Re}}\begin{pmatrix}0\\ \tau\\ \tau u+\frac{\gamma}{\mathrm{Pr}}e_x\end{pmatrix}_x,$$

with

$$e=\frac{1}{\rho}\left(E-\frac{1}{2}\rho u^2-\frac{1}{2}\rho v^2\right),\quad p=(\gamma-1)\rho e,\qquad \text{and}\quad \tau=\eta u_x,$$

and $\gamma, \mathrm{Pr}, \eta$ are positive constants. It can be rewritten as

$$\begin{pmatrix}\rho\\ \rho u\\ E\end{pmatrix}_t + \begin{pmatrix}\rho u\\ \rho u^2+p\\ (E+p)u\end{pmatrix}_x = \frac{\eta}{\mathrm{Re}}\begin{pmatrix}0\\ u\\ \frac{u^2}{2}+\frac{\gamma}{\mathrm{Pr}\,\eta}e\end{pmatrix}_{xx}. \tag{A.1}$$

We will show that the standard second order central difference for the diffusion term in the right hand side can preserve the positivity under a suitable time step constraint.

Consider the equation

$$\mathbf{U}_t = \frac{\eta}{\mathrm{Re}}\mathbf{r}_{xx} \tag{A.2}$$

where $\mathbf{U} = (\rho, \rho u, E)^t$ and $\mathbf{r} = (0, u, \frac{u^2}{2} + \frac{\gamma}{\mathrm{Pr}\,\eta}e)^t$. With forward Euler in time and central difference for the spatial derivative, we obtain a finite difference scheme,

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n + \frac{\Delta t}{\Delta x^2}\frac{\eta}{\mathrm{Re}}(\mathbf{r}_{i-1}^n - 2\mathbf{r}_i^n + \mathbf{r}_{i+1}^n), \tag{A.3}$$

where the superscripts denote the time step and the subscripts denote the spatial index. For convenience, we drop the superscript n in the right hand side. Let $\mu = 2\frac{\Delta t}{\Delta x^2}\frac{\eta}{\mathrm{Re}}$, then we have

$$\begin{aligned}
\mathbf{U}_i^{n+1} = & \begin{pmatrix}\rho\\ \rho u\\ E\end{pmatrix}_i + \frac{\mu}{2}\left[\begin{pmatrix}0\\ u\\ \frac{u^2}{2}+\frac{\gamma}{\mathrm{Pr}\,\eta}e\end{pmatrix}_{i-1} - 2\begin{pmatrix}0\\ u\\ \frac{u^2}{2}+\frac{\gamma}{\mathrm{Pr}\,\eta}e\end{pmatrix}_i + \begin{pmatrix}0\\ u\\ \frac{u^2}{2}+\frac{\gamma}{\mathrm{Pr}\,\eta}e\end{pmatrix}_{i+1}\right] \\
= & \begin{pmatrix}\rho\\ \rho u\\ E\end{pmatrix}_i + \frac{\mu}{2}\left[\begin{pmatrix}1\\ u\\ \frac{u^2}{2}+\frac{\gamma}{\mathrm{Pr}\,\eta}e\end{pmatrix}_{i-1} - 2\begin{pmatrix}1\\ u\\ \frac{u^2}{2}+\frac{\gamma}{\mathrm{Pr}\,\eta}e\end{pmatrix}_i + \begin{pmatrix}1\\ u\\ \frac{u^2}{2}+\frac{\gamma}{\mathrm{Pr}\,\eta}e\end{pmatrix}_{i+1}\right] \\
= & \frac{\mu}{2}\begin{pmatrix}1\\ u\\ \frac{u^2}{2}+\frac{\gamma}{\mathrm{Pr}\,\eta}e\end{pmatrix}_{i-1} + \frac{\mu}{2}\begin{pmatrix}1\\ u\\ \frac{u^2}{2}+\frac{\gamma}{\mathrm{Pr}\,\eta}e\end{pmatrix}_{i+1} + \left[\begin{pmatrix}\rho\\ \rho u\\ E\end{pmatrix}_i - \mu\begin{pmatrix}1\\ u\\ \frac{u^2}{2}+\frac{\gamma}{\mathrm{Pr}\,\eta}e\end{pmatrix}_i\right]
\end{aligned}$$

For any vector $\mathbf{U} = (\rho, \rho u, E)^t$, define the function

$$\chi(\mathbf{U}) = \rho E - \frac{1}{2}|\rho u|^2.$$

Thus a vector $\mathbf{U} \in G$ if and only if its first component and $\chi(\mathbf{U})$ are positive. Assuming $\mathbf{U}_i \in G$ for all i then it is obvious that $\begin{pmatrix}1\\ u\\ \frac{u^2}{2}+\frac{\gamma}{\mathrm{Pr}\,\eta}e\end{pmatrix}_{i\pm1} \in G$. Consider $\mathbf{V} = \begin{pmatrix}\rho\\ \rho u\\ E\end{pmatrix} - \mu\begin{pmatrix}1\\ u\\ \frac{u^2}{2}+\frac{\gamma}{\mathrm{Pr}\,\eta}e\end{pmatrix}$ with $\begin{pmatrix}\rho\\ \rho u\\ E\end{pmatrix} \in G$, then the first component of $\mathbf{V}$ is $\rho - \mu \geq 0$ if

$$2\frac{\Delta t}{\Delta x^2}\frac{\eta}{\mathrm{Re}} \leq \min_i \rho_i. \tag{A.4}$$

For the internal energy, we have

$$\begin{aligned}\chi(\mathbf{V}) &= (\rho-\mu)\left(E-\mu(\frac{u^2}{2}+\frac{\gamma}{\Pr\eta}e)\right)-\frac{1}{2}(\rho u-\mu u)^2\\ &= \rho E-\mu\rho(\frac{u^2}{2}+\frac{\gamma}{\Pr\eta}e)-\mu E+\mu^2(\frac{u^2}{2}+\frac{\gamma}{\Pr\eta}e)-\frac{1}{2}\rho^2u^2-\frac{1}{2}\mu^2u^2+\rho\mu u^2\\ &= \frac{\gamma}{\Pr\eta}e\mu^2-(1+\frac{\gamma}{\Pr\eta})\rho e\mu+\rho^2 e\\ &= e(\mu-\rho)(\frac{\gamma}{\Pr\eta}\mu-\rho).\end{aligned}$$

With (A.4), we now have $\chi(\mathbf{V}_i)\geq 0$ if $\mathbf{U}_i\in G$ and $\frac{\gamma}{\Pr\eta}\mu-\rho_i\leq 0$, i.e.,

$$2\frac{\Delta t}{\Delta x^2}\frac{\gamma}{\Pr\mathrm{Re}}\leq\min_i\rho_i. \tag{A.5}$$

We have proved the following fact,

Lemma 5. *For the second order finite difference scheme (A.3) solving (A.2), if $\mathbf{U}_i^n\in G$ for all i, then $\mathbf{U}_i^{n+1}\in G$ under the CFL constraint,*

$$\Delta t\leq\frac{1}{2}\min\left\{\frac{1}{\eta},\frac{\Pr}{\gamma}\right\}\min_i\rho_i\,\mathrm{Re}\,\Delta x^2.$$

Unfortunately, the proof of Lemma 5 heavily relies on the special structure of the one-dimensional equations and second order finite difference operator. It is highly nontrivial to extend it directly to higher dimensions or higher order accuracy if not impossible. Nonetheless, it allows us to easily construct a positivity-preserving scheme for (A.1). Let $\mathbf{F}^a=(\rho u,\rho u^2+p,(E+p)u)^t$ and consider any finite difference scheme in the form of

$$\mathbf{U}_i^{n+1}=\mathbf{U}_i^n-\frac{\Delta t}{\Delta x}(\widehat{\mathbf{F}^a}_{i+\frac{1}{2}}-\widehat{\mathbf{F}^a}_{i-\frac{1}{2}})+\frac{\Delta t}{\Delta x^2}(\mathbf{r}_{i-1}^n-2\mathbf{r}_i^n+\mathbf{r}_{i+1}^n), \tag{A.6}$$

where $\widehat{\mathbf{F}^a}$ is a positivity-preserving numerical flux, e.g., (13). See [51] for how to construct a fifth order accurate positivity-preserving numerical flux $\widehat{\mathbf{F}^a}$ by high order WENO reconstruction. With third order SSP Runge-Kutta in time (42), we can easily construct a positivity-preserving finite difference scheme that is fifth order accurate for the advection and second order accurate for the diffusion. We have used this scheme to generate some reference solutions in Section 6.2.

Appendix B.

Lemma 6. *Consider any $\mathbf{U}=(\rho,\rho u,\rho v,\rho w,E)^t=(\rho,\rho\mathbf{u}^t,E)^t\in G$, and*

$$\mathbf{F}^a(\mathbf{U})=\begin{pmatrix}\rho\mathbf{u}\\ \rho\mathbf{u}\otimes\mathbf{u}+p\mathbb{I}\\ (E+p)\mathbf{u}\end{pmatrix},\mathbf{F}^d(\mathbf{U})=\begin{pmatrix}0\\ \boldsymbol{\tau}\\ \mathbf{u}\cdot\boldsymbol{\tau}-\mathbf{q}\end{pmatrix},$$

where p, $\boldsymbol{\tau}$ and $\mathbf{q}$ are not necessarily dependent on $\mathbf{U}$. Let $e = \frac{1}{\rho}(E - \frac{1}{2}\rho\|\mathbf{u}\|^2)$. For any unit vector $\mathbf{n} = (n_1, n_2, n_3)^t$, let $\mathbb{v} = \mathbf{u}\cdot\mathbf{n}$, $\mathbb{q} = \mathbf{q}\cdot\mathbf{n}$ and $\vec{\tau} = \mathbf{n}\cdot\boldsymbol{\tau}$. Then we have the following

(a) $\alpha\mathbf{U} \pm \mathbf{F}^a(\mathbf{U})\cdot\mathbf{n} \in G$ *if and only if* $\alpha > |\mathbb{v}| + \sqrt{\frac{p^2}{2\rho^2 e}}$,

(b) $\beta\mathbf{U} \pm \mathbf{F}^d(\mathbf{U})\cdot\mathbf{n} \in G$ *if and only if* $\beta > \frac{1}{2\rho^2 e}\left(\sqrt{\rho^2\mathbb{q}^2 + 2\rho^2 e\|\vec{\tau}\|^2} + \rho|\mathbb{q}|\right)$.

(c) $\beta\mathbf{U} \pm (\mathbf{F}^a(\mathbf{U}) - \mathbf{F}^d(\mathbf{U}))\cdot\mathbf{n} \in G$ *if and only if*

$$\beta > |\mathbb{v}| + \frac{1}{2\rho^2 e}\left(\sqrt{\rho^2\mathbb{q}^2 + 2\rho^2 e\|\vec{\tau} - p\mathbf{n}\|^2} + \rho|\mathbb{q}|\right).$$

Proof. For any $\mathbf{U} = (\rho, \rho\mathbf{u}^t, E)^t$, define the function

$$\chi(\mathbf{U}) = \rho E - \frac{1}{2}\|\rho\mathbf{u}\|^2.$$

Thus a vector $\mathbf{U} \in G$ if and only if its first component and $\chi(\mathbf{U})$ are positive.

(a) First we have

$$\mathbf{F}^a(\mathbf{U})\cdot\mathbf{n} = \begin{pmatrix} \rho\mathbf{u} \\ \rho\mathbf{u}\otimes\mathbf{u} + p\mathbb{I} \\ (E+p)\mathbf{u} \end{pmatrix}\cdot\mathbf{n} = \begin{pmatrix} \rho\mathbb{v} \\ \rho\mathbb{v}\mathbf{u} + p\mathbf{n} \\ (E+p)\mathbb{v} \end{pmatrix},$$

thus

$$\alpha\mathbf{U} \pm \mathbf{F}^a(\mathbf{U})\cdot\mathbf{n} = (\alpha \pm \mathbb{v})\begin{pmatrix} \rho \\ \rho\mathbf{u} \\ E \end{pmatrix} \pm \begin{pmatrix} 0 \\ p\mathbf{n} \\ p\mathbb{v} \end{pmatrix} = \overline{\alpha}\begin{pmatrix} \rho \\ \rho\mathbf{u} \\ E \end{pmatrix} \pm \begin{pmatrix} 0 \\ p\mathbf{n} \\ p\mathbb{v} \end{pmatrix} = \begin{pmatrix} \overline{\alpha}\rho \\ \overline{\alpha}\rho\mathbf{u} \pm p\mathbf{n} \\ \overline{\alpha}E \pm p\mathbb{v} \end{pmatrix}.$$

where $\overline{\alpha}$ denotes $\alpha \pm \mathbb{v}$ for convenience. Finally we have

$$\begin{aligned}\chi(\alpha\mathbf{U} \pm \mathbf{F}^a(\mathbf{U})\cdot\mathbf{n}) &= \overline{\alpha}^2\rho E \pm \overline{\alpha}\rho p\mathbb{v} - \frac{1}{2}\|\overline{\alpha}\rho\mathbf{u} \pm p\mathbf{n}\|^2 \\ &= \overline{\alpha}^2\rho(E - \frac{1}{2}\rho\|\mathbf{u}\|^2) - \frac{1}{2}p^2 = \overline{\alpha}^2\rho^2 e - \frac{1}{2}p^2.\end{aligned}$$

(b) First we have

$$\mathbf{F}^d(\mathbf{U})\cdot\mathbf{n} = \begin{pmatrix} 0 \\ \boldsymbol{\tau} \\ \mathbf{u}\cdot\boldsymbol{\tau} - \mathbf{q} \end{pmatrix}\cdot\mathbf{n} = \begin{pmatrix} 0 \\ \vec{\tau} \\ \mathbf{u}\cdot\vec{\tau} - \mathbb{q} \end{pmatrix},$$

thus

$$\beta\mathbf{U} \pm \mathbf{F}^d(\mathbf{U})\cdot\mathbf{n} = \beta\begin{pmatrix} \rho \\ \rho\mathbf{u} \\ E \end{pmatrix} \pm \begin{pmatrix} 0 \\ \vec{\tau} \\ \mathbf{u}\cdot\vec{\tau} - \mathbb{q} \end{pmatrix} = \begin{pmatrix} \beta\rho \\ \beta\rho\mathbf{u} \pm \vec{\tau} \\ \beta E \pm \mathbf{u}\cdot\vec{\tau} \mp \mathbb{q} \end{pmatrix}.$$

Then we get

$$\begin{aligned}\chi(\beta\mathbf{U} \pm \mathbf{F}^d(\mathbf{U})\cdot\mathbf{n}) &= \beta^2\rho E \pm \beta\rho\mathbf{u}\cdot\vec{\tau} \mp \beta\rho\mathbb{q} - \frac{1}{2}\|\beta\rho\mathbf{u} \pm \vec{\tau}\|^2 \\ &= \beta^2\rho(E - \frac{1}{2}\rho\|\mathbf{u}\|^2) \mp \beta\rho\mathbb{q} - \frac{1}{2}\|\vec{\tau}\|^2 \\ &= \rho^2 e\beta^2 \mp \rho\mathbb{q}\beta - \frac{1}{2}\|\vec{\tau}\|^2,\end{aligned}$$

which are two quadratic forms of β. Since $\rho^2 e$ is assumed to be positive and $\|\vec{\tau}\|^2 \geq 0$, either quadratic equation has at least one nonnegative root. Let β_0 be the largest root among the four roots for two quadratic equations, then $\rho^2 e\beta^2 \mp \rho\mathbb{q}\beta - \frac{1}{2}\|\vec{\tau}\|^2 > 0$ if $\beta > \beta_0$. And $\beta_0 = \frac{1}{2\rho^2 e}\left(\sqrt{\rho^2\mathbb{q}^2 + 2\rho^2 e\|\vec{\tau}\|^2} + \rho|\mathbb{q}|\right)$.

(c) First we have

$$\begin{aligned}\beta\mathbf{U} \pm (\mathbf{F}^a(\mathbf{U}) - \mathbf{F}^d(\mathbf{U}))\cdot\mathbf{n} &= (\beta \pm \mathbb{v})\begin{pmatrix}\rho\\ \rho\mathbf{u}\\ E\end{pmatrix} \pm \begin{pmatrix}0\\ p\mathbf{n} - \vec{\tau}\\ p\mathbb{v} - \mathbf{u}\cdot\vec{\tau} + \mathbb{q}\end{pmatrix} \\ &= \begin{pmatrix}\overline{\beta}\rho\\ \overline{\beta}\rho\mathbf{u} \pm (p\mathbf{n} - \vec{\tau})\\ \overline{\beta}E \pm (p\mathbb{v} - \mathbf{u}\cdot\vec{\tau} + \mathbb{q})\end{pmatrix}.\end{aligned}$$

where $\overline{\beta}$ denotes $\beta \pm \mathbb{v}$ for convenience. Then we get

$$\begin{aligned}&\chi(\beta\mathbf{U} \pm (\mathbf{F}^a(\mathbf{U}) - \mathbf{F}^d(\mathbf{U}))\cdot\mathbf{n}) \\ &= \overline{\beta}^2\rho E \pm \overline{\beta}\rho p\mathbb{v} \mp \overline{\beta}\rho\mathbf{u}\cdot\vec{\tau} \pm \overline{\beta}\rho\mathbb{q} - \frac{1}{2}\|\overline{\beta}\rho\mathbf{u} \pm (p\mathbf{n} - \vec{\tau})\|^2 \\ &= \overline{\beta}^2\rho(E - \frac{1}{2}\rho\|\mathbf{u}\|^2) \pm \overline{\beta}\rho\mathbb{q} - \frac{1}{2}\|p\mathbf{n} - \vec{\tau}\|^2 \\ &= \rho^2 e\overline{\beta}^2 \pm \rho\mathbb{q}\overline{\beta} - \frac{1}{2}\|p\mathbf{n} - \vec{\tau}\|^2.\end{aligned}$$

Following the same arguments as in (b), we obtain the conditions on β such that $\chi(\beta\mathbf{U} \pm (\mathbf{F}^a(\mathbf{U}) - \mathbf{F}^d(\mathbf{U}))\cdot\mathbf{n}) > 0$.

Appendix C.

Lemma 7. *Let $q(x)$ be a non-constant polynomial of degree k with $\int_0^1 q(x)\,dx = 0$, then*

$$\left|\frac{\min\limits_{x\in[0,1]} q(x)}{\max\limits_{x\in[0,1]} q(x)}\right| \leq (k^2 + k - 1)\Lambda_{k+1}[0,1],$$

where $\Lambda_{k+1}[0,1]$ is the Lebesgue constant for the $(k+1)$-point Gauss-Lobatto quadrature points on the interval $[0,1]$.

PROOF. Let $M' = \max\limits_{x\in[0,1]} q(x)$ and $m' = \min\limits_{x\in[0,1]} q(x)$, then $M' > 0$ and $m' < 0$. If $M' \le -m'$, then $\left|\frac{M'}{m'}\right| \le 1$. Next we consider the case $M' > -m'$.

Let x_j $(j = 1, \cdots, k+1)$ denote the $(k+1)$-point Gauss-Lobatto quadrature points for the interval $[0,1]$ and $\widehat{\omega}_j$ $(j = 1, \cdots, k+1)$ denote the corresponding weights. Let $l_j(x)$ $(j = 1, \cdots, k+1)$ denote the Lagrange interpolation polynomials at points x_j $(j = 1, \cdots, k+1)$. Then

$$q(x) = \sum_{j=1}^{k+1} q(x_j) l_j(x).$$

Let $M'' = \max_j q(x_j)$ and $m'' = \min_j q(x_j)$. If $q(x_j) = 0$ for all j, then $q(x) = \sum\limits_{j=1}^{k+1} q(x_j) l_j(x) = 0$, which is impossible for a non-constant polynomial $q(x)$. On the other hand, $\sum_{j=1}^{k+1} \widehat{\omega}_j q(x_j) = \int_0^1 q(x)dx = \overline{q} = 0$. Thus we have $m'' < 0 < M''$.

Then

$$q(x) \le \sum_{j=1}^{k+1} |q(x_j)||l_j(x)| < \max\{M'', -m''\} \sum_{j=1}^{k+1} |l_j(x)|.$$

Thus $M' \le \max\{M'', -m''\} \max\limits_{x\in[0,1]} \sum\limits_{j=1}^{k+1} |l_j(x)| = \max\{M'', -m''\}\Lambda_{k+1}[0,1]$ where $\Lambda_{k+1}[0,1] = \max\limits_{x\in[0,1]} \sum\limits_{j=1}^{k+1} |l_j(x)|$ is the Lebesgue constant. So we have

$$m' \le m'' < 0 < M' \le \max\{M'', -m''\}\Lambda_{k+1}[0,1].$$

Without loss of generality, assume $q(x_1) = \max_j q(x_j) = M''$. Since $\sum\limits_{j=1}^{k+1} \widehat{\omega}_j q(x_j) = 0$, we get $\widehat{\omega}_1 M'' = \widehat{\omega}_1 q(x_1) = -\sum\limits_{j=2}^{k+1} \widehat{\omega}_j q(x_j) \le -\sum\limits_{j=2}^{k+1} \widehat{\omega}_j m'' = -m'' \sum\limits_{j=2}^{k+1} \widehat{\omega}_j$, thus

$$\frac{M''}{-m''} \le \frac{1}{\widehat{\omega}_1} \sum_{j=2}^{k+1} \widehat{\omega}_j \le \frac{1}{\min\limits_j \widehat{\omega}_j} \sum_{j=2}^{k+1} \widehat{\omega}_j \le \frac{1 - \min\limits_j \widehat{\omega}_j}{\min\limits_j \widehat{\omega}_j}.$$

Therefore,

$$0 < \frac{M'}{-m'} \le \frac{\max\{M'', -m''\}\Lambda_{k+1}[0,1]}{-m''} \le \max\left\{\frac{M''}{-m''}, 1\right\}\Lambda_{k+1}[0,1] \le \frac{1 - \min\limits_j \widehat{\omega}_j}{\min\limits_j \widehat{\omega}_j}\Lambda_{k+1}[0,1].$$

Moreover,

$$\frac{1 - \min\limits_j \widehat{\omega}_j}{\min\limits_j \widehat{\omega}_j}\Lambda_{k+1}[0,1] = \frac{1 - \frac{1}{(k+1)k}}{\frac{1}{(k+1)k}}\Lambda_{k+1}[0,1] = (k^2 + k - 1)\Lambda_{k+1}[0,1].$$

Thus we have proved $\left|\frac{M'}{m'}\right| \leq (k^2+k-1)\Lambda_{k+1}[0,1]$. By replacing $q(x)$ by $-q(x)$ in the proof above, we can get $\left|\frac{m'}{M'}\right| \leq (k^2+k-1)\Lambda_{k+1}[0,1]$.

Remark 5. To extend Lemma 7 to multiple dimensions, one would need a quadrature rule with positive weights for Lagrangian interpolation points on a multidimensional cell, which is in general nontrivial.

The following result can be easily extended to any cells in multiple dimensions, the proof of which is similar to the proof of the equivalence of any two norms of a finite dimensional vector space.

Lemma 8. *Let $q(x)$ be a non-constant polynomial of degree k with $\int_0^1 q(x)\,dx = 0$, then*

$$\frac{\max\limits_{x\in[0,1]} |q(x)|}{\max\limits_{x\in[0,1]} q(x)} \leq C_k,$$

where C_k is a constant depending only on k.

Proof. Let V denote the $k-$dimensional vector space consisting of all polynomials of degree k whose averages on the interval $[0,1]$ are zero. For any $q(x) \in V$, define three functionals on V by $f_1[q] = \left|\max\limits_{x\in[0,1]} q(x)\right| = \max\limits_{x\in[0,1]} q(x)$, $f_2[q] = \left|\min\limits_{x\in[0,1]} q(x)\right| = -\min\limits_{x\in[0,1]} q(x)$ and $f_0[q] = \max\limits_{x\in[0,1]} |q(x)| = \max\{f_1[q], f_2[q]\}$. Let $\mathbf{e}_i$ $(i = 1, \cdots, k)$ be a basis of V. For any vector $c = \begin{bmatrix} c_1 & \cdots & c_k \end{bmatrix}^T \in \mathbb{R}^k$, define $f^j(c) = f_j\left[\sum\limits_i c_i \mathbf{e}_i\right]$ for $j = 0, 1, 2$. Notice that $f_0[\cdot]$ is a norm of V and can be denoted as $f_0[q] = \|q\|_\infty$ on the interval $[0,1]$.

For any $p(x), q(x) \in V$, f_1 satisfies the following properties (similar ones hold for f_2):

1. $\forall a > 0$, $f_1[aq(x)] = \max\limits_{x\in[0,1]} aq(x) = af_1[q(x)]$.
2. $f_1[-q] = \left|\max\limits_{x\in[0,1]} -q(x)\right| = \max\limits_{x\in[0,1]} -q(x) = -\min\limits_{x\in[0,1]} q(x) = f_2[q]$.
3. $f_1[p+q] = \max\limits_{x\in[0,1]} (p+q) \leq \max\limits_{x\in[0,1]} p + \max\limits_{x\in[0,1]} q = f_1[p] + f_1[q]$.
4. $f_1[q] = 0 \Rightarrow q \equiv 0$.

Thus, for any $c, d \in \mathbb{R}^k$, we have

$$f^1(c) \leq f^1(d) + f^1(c-d) \leq f^1(d) + f^0(c-d),$$

and

$$f^1(c) \geq f^1(d) - f^1(d-c) = f^1(d) - f^2(c-d) \geq f^1(d) - f^0(c-d),$$

which implies

$$|f^1(c)-f^1(d)| \le f^0(c-d) = f_0\left[\sum_i (c_i-d_i)\mathbf{e}_i\right] \le \sum_i |c_i-d_i|\|\mathbf{e}_i\|_\infty \le \sqrt{\sum_i |c_i-d_i|^2}\sqrt{\|\mathbf{e}_i\|_\infty^2}.$$

Therefore, $f^1(c)$ is uniformly continuous w.r.t. the variable c. Notice that the unit sphere $S^1 = \{c \in \mathbf{R}^k : \|c\| = 1\}$ is a compact set, so f^1 attains its maximum and minimum values on S^1:

$$D_1 \le f^1(d) \le D_2, \quad \forall d \in S^1,$$

where D_1 and D_2 are constants. If there exists $d \in S^1$ such that $f^1(d) = 0$, then $d = \mathbf{0}$ by Property 4 above, which is a contradiction to $d \in S^1$. So we have $D_1 > 0$. By Property 1, we get $f^1(c/\|c\|) = f^1(c)/\|c\|$, thus we have

$$0 < D_1\|c\| \le f^1(c) \le D_2\|c\|, \quad \forall c \in \mathbb{R}^k, c \ne \mathbf{0}.$$

Notice that $f^0(c)$ is a norm of $\mathbb{R}^k$, thus by the equivalence of any two norms of $\mathbb{R}^k$, we get

$$0 < D_3\|c\| \le f^0(c) \le D_4\|c\|, \quad \forall c \in \mathbb{R}^k, c \ne \mathbf{0}.$$

Therefore, for $q = \sum_i c_i\mathbf{e}_i$, we have

$$\frac{\max\limits_{x\in[0,1]}|q(x)|}{\max\limits_{x\in[0,1]}q(x)} = \frac{f_0[q]}{f_1[q]} = \frac{f^0(c)}{f^1(c)} \le \frac{D_4}{D_1}.$$

Appendix D.

We briefly explain why the weak monotonicity holds only up to second order accuracy in a local truncation error analysis for explicit linear finite volume type schemes solving the heat equation. Consider a uniform mesh with grid points x_j and a finite volume type scheme with forward Euler in time for $u_t = u_{xx}$ on an interval $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$,

$$\bar{u}_j^{n+1} = \bar{u}_j^n + \frac{\Delta t}{\Delta x}(\widehat{q}_{j+\frac{1}{2}} - \widehat{q}_{j-\frac{1}{2}}), \tag{D.1}$$

where $\bar{u}_j^n$ is the cell average on I_j and $\widehat{q}_{j+\frac{1}{2}}$ approximates u_x at $x_{j+\frac{1}{2}}$. For a linear scheme, without loss of generality, consider $\widehat{q}_{j+\frac{1}{2}}$ as a linear function

$$\widehat{q}_{j+\frac{1}{2}} = a_l u_{j-l} + a_{l-1}u_{j-l+1} + \cdots + a_1 u_{j-1} + a_0 u_j + b_1 u_{j+1} + \cdots + b_m u_{j+m}, \tag{D.2}$$

where u_j is the approximation to the solution at x_j at time step n and the coefficients a_i and b_i are constants. Then

$$\widehat{q}_{j-\frac{1}{2}} = a_l u_{j-l-1} + a_{l-1}u_{j-l+1} + \cdots + a_1 u_{j-2} + a_0 u_{j-1} + b_1 u_j + \cdots + b_m u_{j+m-1}, \tag{D.3}$$

We rewrite the right hand side of (D.1) by plugging (D.2) and (D.3) in and rewriting $\bar{u}_j^n$ as a linear combination of point values of u on the interval I_j, e.g., Gauss-Lobatto quadrature points on I_j. The scheme (D.1) is said to be weakly monotone if the rewritten right hand side of (D.1) is a monotonically increasing function of all point values involved. By requiring the right hand side of (D.1) to have nonnegative partial derivatives with respect to all point values involved, we get

$$b_1 \geq b_2 \geq \cdots \geq b_{m-1} \geq b_m \geq 0, \quad a_1 \leq a_2 \leq \cdots \leq a_{l-1} \leq a_l \leq 0.$$

With the constraint above, it is straightforward to check that $\widehat{q}_{j+\frac{1}{2}}$ in (D.2) can be at best a second order approximation to u_x at $x_{j+\frac{1}{2}}$ by Taylor expansion.

Acknowledgments

The author is grateful to Prof. Mark Ainsworth for inspirations on the proof of Lemma 7 in Appendix C. The research was supported by the NSF grant DMS-1522593.

References

[1] G. Emanuel, Bulk viscosity of a dilute polyatomic gas, Physics of Fluids A: Fluid Dynamics (1989-1993) 2 (12) (1990) 2252–2254. doi:10.1063/1.857813.

[2] F. Bassi, F. Grasso, A. Jameson, L. Martinelli, M. Savini, Solution of the Compressible Navier-Stokes Equations for a Double Throat Nozzle, in: M. O. Bristeau, R. Glowinski, J. Periaux, H. Viviand (Eds.), Numerical Simulation of Compressible Navier-Stokes Flows: A GAMM-Workshop, Vieweg+Teubner Verlag, Wiesbaden, 1987, pp. 237–254. doi:10.1007/978-3-322-87873-1_14.

[3] Y. Ha, C. L. Gardner, A. Gelb, C.-W. Shu, Numerical simulation of high mach number astrophysical jets with radiative cooling, Journal of Scientific Computing 24 (1) (2005) 29–44.

[4] B. Perthame, C.-W. Shu, On positivity preserving finite volume schemes for Euler equations, Numerische Mathematik 73 (1) (1996) 119–130.

[5] B. Einfeldt, C.-D. Munz, P. L. Roe, B. Sjögreen, On Godunov-type methods near low densities, Journal of Computational Physics 92 (2) (1991) 273–295.

[6] P. Batten, N. Clarke, C. Lambert, D. Causon, On the choice of wavespeeds for the HLLC Riemann solver, SIAM Journal on Scientific Computing 18 (6) (1997) 1553–1570.

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

[8] T. Tao, K. Xu, Gas-kinetic schemes for the compressible Euler equations: positivity-preserving analysis, Zeitschrift für angewandte Mathematik und Physik ZAMP 50 (2) (1999) 258–281.

[9] H.-Z. Tang, K. Xu, Positivity-preserving analysis of explicit and implicit Lax–Friedrichs schemes for compressible Euler equations, Journal of Scientific Computing 15 (1) (2000) 19–28.

[10] X. Zhang, C.-W. Shu, On positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations on rectangular meshes, Journal of Computational Physics 229 (23) (2010) 8918–8934. doi:10.1016/j.jcp.2010.08.016.

[11] X. Zhang, C.-W. Shu, Positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations with source terms, Journal of Computational Physics 230 (4) (2011) 1238–1248.

[12] J. Estivalezes, P. Villedieu, High-order positivity-preserving kinetic schemes for the compressible Euler equations, SIAM Journal on Numerical Analysis 33 (5) (1996) 2050–2067.

[13] T. Linde, P. L. Roe, Robust Euler codes, in: Thirteenth Computational Fluid Dynamics Conference, AIAA Paper-97-2098, 1997.

[14] J. Gressier, P. Villedieu, J.-M. Moschetta, Positivity of flux vector splitting schemes, Journal of Computational Physics 155 (1) (1999) 199–220.

[15] X. Y. Hu, N. A. Adams, C.-W. Shu, Positivity-preserving method for high-order conservative schemes solving compressible Euler equations, Journal of Computational Physics 242 (2013) 169–180.

[16] T. Xiong, J.-M. Qiu, Z. Xu, Parametrized positivity preserving flux limiters for the high order finite difference WENO scheme solving compressible Euler equations, Journal of Scientific Computing (2014) 1–23.

[17] X. Zhang, C.-W. Shu, On maximum-principle-satisfying high order schemes for scalar conservation laws, Journal of Computational Physics 229 (9) (2010) 3091–3120. doi:10.1016/j.jcp.2009.12.030.

[18] X. Zhang, Y. Xia, C.-W. Shu, Maximum-principle-satisfying and positivity-preserving high order discontinuous Galerkin schemes for conservation laws on triangular meshes, Journal of Scientific Computing 50 (1) (2012) 29–62. doi:10.1007/s10915-011-9472-8.

[19] X. Zhang, C.-W. Shu, Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: survey and new developments, in: Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, Vol. 467, The Royal Society, 2011, pp. 2752–2776.

[20] S. Gottlieb, D. I. Ketcheson, C.-W. Shu, Strong stability preserving Runge-Kutta and multistep time discretizations, World Scientific, 2011.

[21] Y. Lv, M. Ihme, Entropy-bounded discontinuous Galerkin scheme for Euler equations, Journal of Computational Physics 295 (2015) 715–739.

[22] F. Vilar, C.-W. Shu, P.-H. Maire, Positivity-preserving cell-centered Lagrangian schemes for multi-material compressible flows: From first-order to high-orders, Journal of Computational Physics 312 (2016) 385–415.

[23] D. Grapsas, R. Herbin, W. Kheriji, J.-C. Latché, An unconditionally stable staggered pressure correction scheme for the compressible Navier-Stokes equations, preprint (Feb. 2015).
URL https://hal.archives-ouvertes.fr/hal-01115250

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

[25] H. Liu, H. Yu, Maximum-Principle-Satisfying Third Order Discontinuous Galerkin Schemes for Fokker–Planck Equations, SIAM Journal on Scientific Computing 36 (5) (2014) A2296–A2325.

[26] J. Yan, Maximum Principle Satisfying Direct discontinuous Galerkin method and its vairation for convection diffusion equations, submitted for publication (2015).

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

[28] H. Liu, Z.-M. Wang, An entropy satisfying discontinuous Galerkin method for nonlinear Fokker-Planck equations, Journal of Scientific Computing 68 (3) (2016) 1217–1240.

[29] H. Liu, J. Yan, The direct discontinuous Galerkin (DDG) method for diffusion with interface corrections, Communications in Computational Physics 8 (3) (2010) 541.

[30] M. Zhang, J. Yan, Fourier type error analysis of the direct discontinuous Galerkin method and its variations for diffusion equations, Journal of Scientific Computing 52 (3) (2012) 638–655.

[31] H. Liu, Optimal error estimates of the direct discontinuous Galerkin method for convection-diffusion equations, Mathematics of Computation 84 (295) (2015) 2263–2295.

[32] M. Zhang, C.-W. Shu, An analysis of three different formulations of the discontinuous Galerkin method for diffusion equations, Mathematical Models and Methods in Applied Sciences 13 (03) (2003) 395–413.

[33] F. Bassi, S. Rebay, A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations, Journal of Computational Physics 131 (2) (1997) 267–279.

[34] F. Bassi, S. Rebay, Numerical evaluation of two discontinuous Galerkin methods for the compressible Navier–Stokes equations, International Journal for Numerical Methods in Fluids 40 (1-2) (2002) 197–207.

[35] C. E. Baumann, J. T. Oden, A discontinuous hp finite element method for the Euler and Navier–Stokes equations, International Journal for Numerical Methods in Fluids 31 (1) (1999) 79–95.

[36] A. Uranga, P.-O. Persson, M. Drela, J. Peraire, Implicit large eddy simulation of transitional flows over airfoils and wings, Proceedings of the 19th AIAA Computational Fluid Dynamics, AIAA 4131 (7) (2009) 67.

[37] Z. Wang, H. Gao, A unifying lifting collocation penalty formulation including the discontinuous Galerkin, spectral volume/difference methods for conservation laws on mixed grids, Journal of Computational Physics 228 (21) (2009) 8161–8186.

[38] T. Haga, H. Gao, Z. J. Wang, A high-order unifying discontinuous formulation for the Navier-Stokes equations on 3D mixed grids, Mathematical Modelling of Natural Phenomena 6 (03) (2011) 28–56.

[39] J. Peraire, N. C. Nguyen, B. Cockburn, A Hybridizable Discontinuous Galerkin Method for the Compressible Euler and Navier-Stokes Equations, in: Proceedings of 48th AIAA Aerospace Sciences Meeting and Exhibit, Orlando, Florida (AIAA Paper 2010-363), 2010.

[40] J. Peraire, N. Nguyen, B. Cockburn, An embedded discontinuous Galerkin method for the compressible Euler and Navier-Stokes equations, in: Proceedings of the 20th AIAA Computational Fluid Dynamics Conference, Honolulu, Hawaii,(AIAA Paper 2011-3228), 2011.

[41] M. H. Carpenter, T. C. Fisher, E. J. Nielsen, S. H. Frankel, Entropy Stable Spectral Collocation Schemes for the Navier–Stokes Equations: Discontinuous Interfaces, SIAM Journal on Scientific Computing 36 (5) (2014) B835–B867.

[42] M. Parsani, M. H. Carpenter, E. J. Nielsen, Entropy stable wall boundary conditions for the three-dimensional compressible Navier–Stokes equations, Journal of Computational Physics 292 (2015) 88–113.

[43] X. Zhang, C.-W. Shu, A minimum entropy principle of high order schemes for gas dynamics equations, Numerische Mathematik 121 (3) (2012) 545–563.

[44] C. Wang, X. Zhang, C.-W. Shu, J. Ning, Robust high order discontinuous Galerkin schemes for two-dimensional gaseous detonations, Journal of Computational Physics 231 (2) (2012) 653–665.

[45] J. Qiu, C.-W. Shu, Runge–Kutta Discontinuous Galerkin Method Using WENO Limiters, SIAM Journal on Scientific Computing 26 (3) (2005) 907–929. doi:10.1137/S1064827503425298.

[46] X. Zhong, C.-W. Shu, A simple weighted essentially nonoscillatory limiter for Runge–Kutta discontinuous Galerkin methods, Journal of Computational Physics 232 (1) (2013) 397–415. doi:10.1016/j.jcp.2012.08.028.

[47] J. Zhu, X. Zhong, C.-W. Shu, J. Qiu, Runge–Kutta discontinuous Galerkin method using a new type of WENO limiters on unstructured meshes, Journal of Computational Physics 248 (2013) 200–220.

[48] X.-D. Liu, S. Osher, Nonoscillatory high order accurate self-similar maximum principle satisfying shock capturing schemes I, SIAM Journal on Numerical Analysis 33 (2) (1996) 760–779.

[49] S. Patkar, M. Aanjaneya, W. Lu, M. Lentine, R. Fedkiw, Towards positivity preservation for monolithic two-way solid–fluid coupling, Journal of Computational Physics 312 (2016) 82–114.

[50] B. Cockburn, S.-Y. Lin, 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 (1) (1989) 90–113.

[51] X. Zhang, C.-W. Shu, Positivity-preserving high order finite difference WENO schemes for compressible Euler equations, Journal of Computational Physics 231 (5) (2012) 2245–2258.

[52] J. Hesthaven, T. Warburton, Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications, Texts in Applied Mathematics, Springer, 2007.

[53] B. Cockburn, C.-W. Shu, The Runge–Kutta discontinuous Galerkin method for conservation laws V: multidimensional systems, Journal of Computational Physics 141 (2) (1998) 199–224.

[54] J. Chan, L. Demkowicz, R. Moser, A DPG method for steady viscous compressible flow, Computers & Fluids 98 (2014) 69–90.