A positivity-preserving implicit-explicit scheme with high order polynomial basis for compressible Navier–Stokes equations

Chen Liu[a], Xiangxiong Zhang[a,]

[a]*Department of Mathematics, Purdue University, 150 North University Street, West Lafayette, Indiana 47907.*

Abstract

In this paper, we are interested in constructing a scheme solving compressible Navier–Stokes equations, with desired properties including high order spatial accuracy, conservation, and positivity-preserving of density and internal energy under a standard hyperbolic type CFL constraint on the time step size, e.g., $\Delta t = \mathcal{O}(\Delta x)$. Strang splitting is used to approximate convection and diffusion operators separately. For the convection part, i.e., the compressible Euler equation, the high order accurate postivity-preserving Runge–Kutta discontinuous Galerkin method can be used. For the diffusion part, the equation of internal energy instead of the total energy is considered, and a first order semi-implicit time discretization is used for the ease of achieving positivity. A suitable interior penalty discontinuous Galerkin method for the stress tensor can ensure the conservation of momentum and total energy for any high order polynomial basis. In particular, positivity can be proven with $\Delta t = \mathcal{O}(\Delta x)$ if the Laplacian operator of internal energy is approximated by the $\mathbb{Q}^k$ spectral element method with $k = 1, 2, 3$. So the full scheme with $\mathbb{Q}^k$ ($k = 1, 2, 3$) basis is conservative and positivity-preserving with $\Delta t = \mathcal{O}(\Delta x)$, which is robust for demanding problems such as solutions with low density and low pressure induced by high-speed shock diffraction. Even though the full scheme is only first order accurate in time, numerical tests indicate that higher order polynomial basis produces much better numerical solutions, e.g., better resolution for capturing the roll-ups during shock reflection.

Keywords: compressible Navier–Stokes, discontinuous Galerkin, spectral element, implicit-explicit, high-order accuracy, conservation, positivity-preserving

2000 MSC: 35L65, 65M12, 65M60, 65N30

1. Introduction

1.1. Motivation of positivity

The compressible Navier–Stokes (NS) equations are one of the most popular and important models in gas dynamics as well as computational fluid dynamics applications. The equations in dimensionless form on a bounded spatial domain $\Omega \subset \mathbb{R}^d$ over the time interval $[0, T]$ are given by:

$$\partial_t \rho + \nabla \cdot (\rho \boldsymbol{u}) = 0 \qquad \text{in } [0,T] \times \Omega, \tag{1a}$$

$$\partial_t (\rho \boldsymbol{u}) + \nabla \cdot (\rho \boldsymbol{u} \otimes \boldsymbol{u}) + \nabla p - \tfrac{1}{\mathrm{Re}} \nabla \cdot \boldsymbol{\tau}(\boldsymbol{u}) = \boldsymbol{0} \qquad \text{in } [0,T] \times \Omega, \tag{1b}$$

$$\partial_t E + \nabla \cdot ((E + p)\boldsymbol{u}) + \tfrac{1}{\mathrm{Re}} \nabla \cdot \boldsymbol{q} - \tfrac{1}{\mathrm{Re}} \nabla \cdot (\boldsymbol{\tau}(\boldsymbol{u})\boldsymbol{u}) = 0 \qquad \text{in } [0,T] \times \Omega, \tag{1c}$$

where ρ, $\boldsymbol{u}$, p, and E are the density, velocity, pressure, and total energy respectively, and Re denotes the Reynolds number. Let $\boldsymbol{m} = \rho \boldsymbol{u}$ denote the momentum, then the conservative variables are $\boldsymbol{U} = [\rho, \boldsymbol{m}, E]^{\mathrm{T}}$. Assume the fluid is Newtonian, as well as the Stokes hypothesis, which states that the bulk viscosity equals to zero. Then the shear stress tensor is given by $\boldsymbol{\tau}(\boldsymbol{u}) = 2\boldsymbol{\varepsilon}(\boldsymbol{u}) - \frac{2}{3}(\nabla \cdot \boldsymbol{u})\mathbf{I}$, where $\boldsymbol{\varepsilon}(\boldsymbol{u}) = \frac{1}{2}(\nabla \boldsymbol{u} + (\nabla \boldsymbol{u})^{\mathrm{T}})$ and

Email addresses: liu3373@purdue.edu (Chen Liu), zhan1966@purdue.edu (Xiangxiong Zhang)

$\mathbf{I} \in \mathbb{R}^{d\times d}$ is an identity matrix. The total energy can be expressed as $E = \rho e + \frac{1}{2}\rho\|\boldsymbol{u}\|^2$, where e denotes the internal energy. For simplicity, we consider the ideal gas equation of state $p = (\gamma - 1)\rho e$, with parameter $\gamma > 0$ where $\gamma = 1.4$ for air. With the Fourier's heat conduction law, the heat flux $\boldsymbol{q}$ is defined by $\boldsymbol{q} = -\lambda\nabla e$, where parameter $\lambda = \frac{\gamma}{\mathrm{Pr}} > 0$ and Pr denotes the Prandtl number.

Physically meaningful solutions $\boldsymbol{U}$ should have positive density and positive internal energy. Define the set of admissible states as:

$$G = \{\boldsymbol{U} = [\rho, \boldsymbol{m}, E]^{\mathrm{T}} : \ \rho > 0, \ \rho e(\boldsymbol{U}) = E - \frac{\|\boldsymbol{m}\|^2}{2\rho} > 0\}.$$

The set G is convex and the ρe is a concave function with respect to $\boldsymbol{U}$, see [1]. With an initial condition $\boldsymbol{U}_0 = [\rho_0, \boldsymbol{m}_0, E_0]^{\mathrm{T}} \in G$, it is a wide open question whether the solution of compressible NS equations (1) should have positive density and internal energy for a given positive initial data, though it is partially justified for special systems, e.g., see [2, 3] and the references therein. On the other hand, empirically we would expect a reasonable numerical solution to this initial value problem should belong to the set G for any time $t > 0$.

In general, classical numerical methods for a convection-diffusion system like (1) are not positivity-preserving without any limiters. In practice, one often observes blow-ups once negative density or negative pressure (corresponding to negative internal energy) is generated during numerical simulations. The linearized compressible Euler equations with negative density or negative internal energy will no longer be hyperbolic thus its initial value problem becomes ill-posed [1]. When negative values emerge, the simple ad-hoc approach of truncating negative values to zero destroys conservation, which is equivalent to adding mass or internal energy into a conservative system, thus the computation will eventually still blow up. Therefore for the sake of robustness, it is desired to construct a numerical scheme which is both conservative and positivity-preserving.

1.2. Existing positivity-preserving schemes for compressible Navier–Stokes equations

In the literature there are many different methods to construct positivity-preserving schemes for compressible Euler equations. However, it is much more difficult to construct a conservative and positive scheme for the compressible NS equations in multiple dimensions due to the mixed second order derivatives in the diffusion operator. In the past decade, significant progress of practical conservative and positive schemes has been made for the fully nonlinear compressible NS equations (1). Notable efforts include at least the following three different kinds of schemes.

The first approach proposed by Grapas et al. in [4] is to solve the internal energy equation directly instead of solving the total energy equation (1c). By solving the internal energy equation, preserving positivity of internal energy becomes simpler but conservation of total energy becomes difficult. The fully implicit pressure correction scheme on staggered grids in [4] can be proven unconditionally stable, positivity-preserving and conservative. Nonlinear equations must be solved in the implementation. The spatial accuracy of this approach is at most second order accurate and it seems difficult to extend it to higher order spatial accuracy especially for a fully implicit scheme on a staggered grid.

The second approach is a fully explicit scheme proposed by the second author in [5]. By solving the conservative system (1), conservation is straightforward to achieve but positivity of internal energy is difficult to enforce. With a simple nonlinear diffusion numerical flux, it was proven in [5] that arbitrarily high order Runge–Kutta discontinuous Galerkin (DG) schemes solving (1) can be rendered positivity-preserving without losing conservation and accuracy by a simple limiter, which can be regarded as an easy extension of the Zhang–Shu method for conservation laws in [6, 1, 7] to the compressible NS equations. The advantages of such a fully explicit approach include easy extensions to general shear stress models and heat fluxes, and possible extensions to other type of schemes such as high order accurate finite volume schemes [8] and the high order accurate finite difference WENO (weighted essentially nonoscillatory) scheme [9]. However, the major drawback of any fully explicit scheme for the convection diffusion system (1) in [5, 8, 9] is a time step constraint like $\Delta t = \mathcal{O}(\mathrm{Re}\,\Delta x^2)$, which is suitable and practical only for high Reynolds number problems.

The third approach proposed by Guermond et al. in [10] introduces a semi-implicit continuous finite element scheme with positivity-preserving property under standard hyperbolic CFL condition like $\mathcal{O}(\Delta x)$. By applying the Strang splitting to the compressible NS model [11], the equations (1) are splitted into a hyperbolic subproblem (H) and a parabolic subproblem (P), which represent two asymptotic regimes, namely the vanishing viscosity limit, i.e., the compressible Euler equations, and the dominant of diffusive terms. The definition of these subproblems is as follows:

$$\text{(H)}\begin{cases}\partial_t\rho+\nabla\cdot(\rho\boldsymbol{u})=0\\ \partial_t(\rho\boldsymbol{u})+\nabla\cdot(\rho\boldsymbol{u}\otimes\boldsymbol{u}+p\mathbf{I})=\mathbf{0}\\ \partial_t E+\nabla\cdot((E+p)\boldsymbol{u})=0\end{cases},\qquad \text{(P)}\begin{cases}\partial_t\rho=0\\ \partial_t(\rho\boldsymbol{u})-\frac{1}{\mathrm{Re}}\nabla\cdot\boldsymbol{\tau}(\boldsymbol{u})=\mathbf{0}\\ \partial_t E+\frac{1}{\mathrm{Re}}\nabla\cdot(\boldsymbol{q}-\boldsymbol{\tau}(\boldsymbol{u})\boldsymbol{u})=0\end{cases}. \tag{2}$$

The first equation in (P) implies variable ρ in parabolic subproblem is time independent. Multiply the second equation in (P) by $\boldsymbol{u}$, use the heat flux $\boldsymbol{q}=-\lambda\nabla e$ and the identity $\nabla\cdot(\boldsymbol{\tau}(\boldsymbol{u})\boldsymbol{u})=(\nabla\cdot\boldsymbol{\tau}(\boldsymbol{u}))\cdot\boldsymbol{u}+\boldsymbol{\tau}(\boldsymbol{u}):\nabla\boldsymbol{u}$, we obtain the equivalent non-conservative form of equations for (P):

$$\text{(P)}\begin{cases}\partial_t\rho=0, & \text{(3a)}\\ \rho\partial_t\boldsymbol{u}-\frac{1}{\mathrm{Re}}\nabla\cdot\boldsymbol{\tau}(\boldsymbol{u})=\mathbf{0}, & \text{(3b)}\\ \rho\partial_t e-\frac{\lambda}{\mathrm{Re}}\Delta e=\frac{1}{\mathrm{Re}}\boldsymbol{\tau}(\boldsymbol{u}):\nabla\boldsymbol{u}. & \text{(3c)}\end{cases}$$

In [10], a semi-implicit time discretization is used for the internal energy equation (3c) such that only a linear system needs to be solved for implementing the scheme, without affecting the conservation of momentum and total energy. The positivity of internal energy in piecewise linear finite element method can also be easily proven due to the well-known fact that piecewise linear methods can form an M-matrix for the Laplacian operator.

1.3. Motivation and difficulty of high order spatial accuracy in implicit schemes

Even though schemes constructed from high order polynomials are high order accurate on a uniform or quasi-uniform mesh only for smooth solutions, they produce less artificial viscosity thus resolve small scale structures better than first order and second order schemes even for the gas dynamics problems involving with strong shocks, see examples in [5, 9]. In other words, less artificial viscosity is the main motivation of pursuing a high order scheme, e.g., DG methods with polynomial basis of degree at least two.

To see the key challenge in constructing a positivity-preserving high order scheme for compressible NS equations, we consider the heat equation $\partial_t e=\partial_{xx}e$ with homogeneious Dirichlet boundary conditions as a simplification of equation (3c). The simple second order centered difference $\partial_{xx}e\approx\frac{e_{i-1}-2e_i+e_{i+1}}{\Delta x^2}$ is monotone with both explicit and implicit time stepping. With forward Euler time stepping, the scheme

$$e_i^{n+1}=e_i^n+\Delta t\frac{e_{i-1}^n-2e_i^n+e_{i+1}^n}{\Delta x^2}=\frac{\Delta t}{\Delta x^2}e_{i-1}^n+\left(1-2\frac{\Delta t}{\Delta x^2}\right)e_i^n+\frac{\Delta t}{\Delta x^2}e_{i+1}^n$$

is monotone in the sense that e_i^{n+1} is a convex combination of e_i^n and $e_{i\pm1}^n$ if $\frac{\Delta t}{\Delta x^2}\le\frac{1}{2}$. Such monotonicity is in general not true for high order schemes, but some explicit high order schemes in [12, 13, 14, 15, 16, 17] were shown to have weak monotonicity for the parabolic equations, which means that the cell averages can still be a monotone function. In principle, all these explicit schemes can be applied to (3c) for constructing a positivity-preserving scheme for (1) but under a small time step constraint $\Delta t=\mathcal{O}(\mathrm{Re}\,\Delta x^2)$.

With backward Euler time stepping, the scheme

$$e_i^{n+1}=e_i^n+\Delta t\frac{e_{i-1}^{n+1}-2e_i^{n+1}+e_{i+1}^{n+1}}{\Delta x^2}$$

gives a linear system $\mathbf{A}\mathbf{e}^{n+1} = \mathbf{e}^n$, where $\mathbf{A}$ is a tridiagonal matrix with $\lambda = \frac{\Delta t}{\Delta x^2}$,

$$\mathbf{A} = \begin{pmatrix} 1+2\lambda & -\lambda & & & \\ -\lambda & 1+2\lambda & -\lambda & & \\ & \ddots & \ddots & \ddots & \\ & & -\lambda & 1+2\lambda & -\lambda \\ & & & -\lambda & 1+2\lambda \end{pmatrix}.$$

This implicit scheme is monotone because $\mathbf{A}^{-1}$ has nonnegative entries thus one can also show e_i^{n+1} is a convex combination of e_j^n for all j without any time step constraint. The matrix $\mathbf{A}$ is diagonally dominant with non-positive off diagonal entries, so $\mathbf{A}$ is an M-matrix [18] thus $\mathbf{A}^{-1} \geq 0$. It is well-known that the monotonicity in implicit schemes holds in piecewise linear finite element method, e.g., [10]. In general, the monotonicity is not true in implicit high order schemes, e.g., the continuous finite element method with quadratic polynomials cannot be monotone on unstructured meshes [19]. However, it is possible to show that continuous finite element method with quadratic and cubic polynomial basis can still be monotone on a uniform rectangular mesh under practical time step and mesh constraints [20, 21].

1.4. The main results

In this paper, we are interested in constructing a conservative and positivity-preserving scheme which is high order accurate for spatial variables, without a restrictive time step constraint such as $\Delta t = \mathcal{O}(\mathrm{Re}\,\Delta x^2)$. For problems involved with low density and low pressure, loss of positivity is the main source of instabilities of high order schemes. In order to avoid small time steps like $\Delta t = \mathcal{O}(\mathrm{Re}\,\Delta x^2)$, we follow the third approach in Section 1.2 by solving the non-conservative form of diffusion equations (3).

We will mainly consider the high order DG methods, which have a lot of advantages and have been successful in many scientific and industrial applications. In particular, high order DG methods have been quite popular for the compressible NS equations since the pioneering work in [22]. For the sake of easy extensions to arbitrarily high order polynomial basis, we use the positivity-preserving Runge–Kutta DG method for the compressible Euler equations [1, 7, 5] for solving the hyperbolic subproblem (H) in (2).

For shear stress tensor terms $\nabla \cdot \boldsymbol{\tau}(\boldsymbol{u})$ and $\boldsymbol{\tau}(\boldsymbol{u}) : \nabla \boldsymbol{u}$ in the parabolic subproblem (P) in (3), we will also use a DG method. In the literature, many different types of DG methods have been developed for solving diffusion equations, including local DG [23, 24], compact DG [25, 26], direct DG [27, 28, 29], hybridizable DG [30, 31, 32], interior penalty DG (IPDG) [33, 34, 35, 36], weak Galerkin methods [37, 38], and many others [39, 40]. In particular, we will use the IPDG method since the global conservation of momentum and total energy can be easily achieved via a proper choice of IPDG discretizations for approximating $\nabla \cdot \boldsymbol{\tau}(\boldsymbol{u})$ and $\boldsymbol{\tau}(\boldsymbol{u}) : \nabla \boldsymbol{u}$.

In order to achieve positivity of internal energy for solving equation (3c), we can utilize either IPDG with $\mathbb{Q}^1$ element or spectral element method with $\mathbb{Q}^2$ or $\mathbb{Q}^3$ element on uniform rectangular meshes for the Laplace operator $-\Delta e$. The monotonicity of spectral element method with $\mathbb{Q}^2$ and $\mathbb{Q}^3$ element for Laplacian has recently been proven in [20, 21].

To summarize, our numerical scheme for solving (1) consists of the following main ingredients:

1. With Strang splitting, the compressible Euler equations, i.e., the hyperbolic subproblem in (2) and parabolic subproblem (3) are solved separately. The compressible Euler equations are solved by the positivity-preserving Runge–Kutta DG method with $\mathbb{Q}^k$ element on rectangular meshes [1].

2. The time stepping for the parabolic subproblem consists of Crank–Nicolson method to (3b) and a first order semi-implicit time discretization to (3c). When a proper IPDG method is used for $\nabla \cdot \boldsymbol{\tau}(\boldsymbol{u})$ and $\boldsymbol{\tau}(\boldsymbol{u}) : \nabla \boldsymbol{u}$, global conservation of momentum and total energy is ensured.

3. The diffusion term $-\Delta e$ is treated implicitly. We will prove positivity of IPDG method with $\mathbb{Q}^1$ element. For positivity of higher order elements, we use the spectral element method with $\mathbb{Q}^2$ and $\mathbb{Q}^3$ element (i.e., continuous finite element method with Gauss–Labotto quadrature), for which monotonicity has

been proven in [20, 21]. We emphasize that no limiters are used at all in the fully discretized scheme for solving the parabolic subproblem.

So the overall scheme is at most first order accurate in time for the system (1) but fourth order accurate in space when $\mathbb{Q}^3$ element is used. At first glance, the high order spatial accuracy may not look necessary since the order of time accuracy is low. However, empirically the spatial resolution is more important than the temporal for many fluid dynamics problems. In particular, computational evidence often suggests that a spatially higher order accurate scheme can produce better solutions even if the temporal order of accuracy is low. For instance, see Figure 1 for results of our schemes solving a Mach 10 shock reflection-diffraction problem, which involves strong shock, very low density and pressure, as well as Kelvin–Helmholtz instability. In Figure 1, the $\mathbb{Q}^3$ scheme with less degrees of freedom can better capture the instability roll-ups than the $\mathbb{Q}^1$ scheme, even though both schemes are first order accurate in time for the internal energy equation (3c). See also the numerical examples for the superiority of $\mathbb{Q}^2$ element over $\mathbb{Q}^1$ element for scalar convection-diffusion problems in [41, 42, 43].

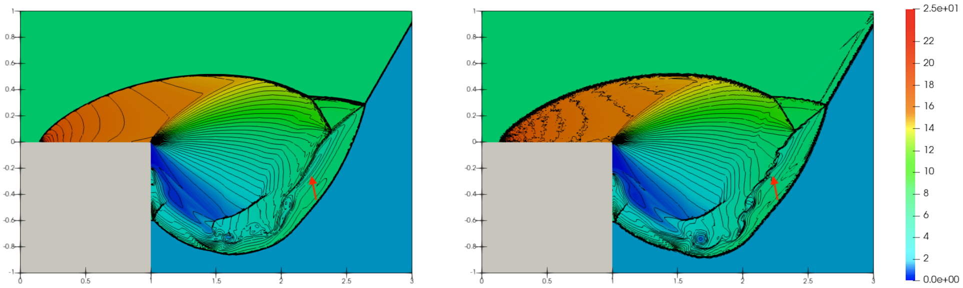

Figure 1: Mach 10 shock reflection and diffraction with Reynolds number 1000. Plot of density: 50 equally space contour lines from 0 to 25. Left snapshot from $\mathbb{Q}^1$ scheme in this paper on a uniform mesh with mesh resolution 1/480. Right snapshot from $\mathbb{Q}^3$ scheme in this paper on a uniform mesh with mesh resolution 1/120.

1.5. Contributions and organization of this paper

To the best of our knowledge, this is the first time that an implicit conservative positivity-preserving scheme with high order elements like $\mathbb{Q}^2$ and $\mathbb{Q}^3$ elements is constructed for the compressible NS equations. Morever, numerical tests suggest that the $\mathbb{Q}^3$ scheme is indeed robust with much better resolutions.

It is in general nontrivial to achieve global conservation when solving equations of the non-conservative form (3). Even though we only consider rectangular meshes in this paper, the global conservation of IPDG methods for the parabolic subproblem (3) can be easily extended to unstructured meshes. There are many variants of IPDG methods, including the symmetric version (SIPG), the nonsymmetric version (NIPG), and the incomplete version (IIPG). In particular, we prove that the global conservation can be achieved if the shear stress tensor terms are discretized by the NIPG method.

We also prove that the second order accurate IIPG method with $\mathbb{Q}^1$ element for the Laplacian term $-\Delta e$ forms an M-matrix. Even though it is well known that it is possible to achieve an M-matrix structure when using piecewise linear finite element method, to the best of our knowledge this is the first time that such an M-matrix structure is proven among the family of IPDG methods beyond one dimension.

The rest of this paper is organized as follows. In Section 2, we introduce the fully discrete numerical scheme and discuss the conservation property. In Section 3, we discuss the positivity-preserving property. In particular we prove that the IPDG method with $\mathbb{Q}^1$ element forms an M-matrix thus is monotone in Appendix A. Numerical tests are shown in Section 4. Concluding remarks are given in Section 5.

2. The full numerical scheme

In this section, we describe the fully discretized numerical scheme for solving the compressible NS equations (1) that utilizes DG discretization in space within the Strang splitting framework. Then we show that our method preserves the global conservation.

2.1. Time discretization

Given the conserved variables $\boldsymbol{U}^n$ at time step t_n $(n \geq 0)$, the Strang splitting for evolving to time step $t_{n+1} = t_n + \Delta t$ for the system (1) is to solve (P) and (H) in (2) separately:

$$\boldsymbol{U}^n \xrightarrow[\text{step size } \frac{\Delta t}{2}]{\text{solve (H)}} \boldsymbol{U}^{\mathrm{H}} \xrightarrow[\text{step size } \Delta t]{\text{solve (P)}} \boldsymbol{U}^{\mathrm{P}} \xrightarrow[\text{step size } \frac{\Delta t}{2}]{\text{solve (H)}} \boldsymbol{U}^{n+1}. \tag{4}$$

Define the advection flux as

$$\boldsymbol{F}^{\mathrm{a}} = [\rho \boldsymbol{u}, \rho \boldsymbol{u} \otimes \boldsymbol{u} + p\mathbf{I}, (E+p)\boldsymbol{u}]^{\mathrm{T}}.$$

For any $n \geq 0$, the time discretization methods in one time step of Strang splitting consists of the following steps:

Step 1. Given $\boldsymbol{U}^n = [\rho^n, \boldsymbol{m}^n, E^n]^{\mathrm{T}}$, we use the third order strong stability preserving (SSP) Runge–Kutta method [44] to obtain $\boldsymbol{U}^{\mathrm{H}} = [\rho^{\mathrm{H}}, \boldsymbol{m}^{\mathrm{H}}, E^{\mathrm{H}}]^{\mathrm{T}}$ in the first step in Strang splitting (4),

$$\boldsymbol{U}^{(1)} = \boldsymbol{U}^n - \frac{\Delta t}{2} \nabla \cdot \boldsymbol{F}^{\mathrm{a}}(\boldsymbol{U}^n), \tag{5a}$$

$$\boldsymbol{U}^{(2)} = \frac{3}{4}\boldsymbol{U}^n + \frac{1}{4}\Big[\boldsymbol{U}^{(1)} - \frac{\Delta t}{2} \nabla \cdot \boldsymbol{F}^{\mathrm{a}}(\boldsymbol{U}^{(1)})\Big], \tag{5b}$$

$$\boldsymbol{U}^{\mathrm{H}} = \frac{1}{3}\boldsymbol{U}^n + \frac{2}{3}\Big[\boldsymbol{U}^{(2)} - \frac{\Delta t}{2} \nabla \cdot \boldsymbol{F}^{\mathrm{a}}(\boldsymbol{U}^{(2)})\Big]. \tag{5c}$$

Step 2. Given $\boldsymbol{U}^{\mathrm{H}} = [\rho^{\mathrm{H}}, \boldsymbol{m}^{\mathrm{H}}, E^{\mathrm{H}}]^{\mathrm{T}}$, compute $(\boldsymbol{u}^{\mathrm{H}}, e^{\mathrm{H}})$ by solving

$$\boldsymbol{m}^{\mathrm{H}} = \rho^{\mathrm{H}} \boldsymbol{u}^{\mathrm{H}} \quad \text{and} \quad E^{\mathrm{H}} = \rho^{\mathrm{H}} e^{\mathrm{H}} + \frac{1}{2}\rho^{\mathrm{H}} \|\boldsymbol{u}^{\mathrm{H}}\|^2.$$

Step 3. Notice that equation (3a) implies that $\rho^{\mathrm{P}} = \rho^{\mathrm{H}}$ in the second step in Strang splitting (4). Apply the second order Crank–Nicolson method to (3b) and a first order semi-implicit time discretization to (3c),

$$\begin{aligned}
&\boldsymbol{u}^* = \frac{1}{2}\boldsymbol{u}^{\mathrm{P}} + \frac{1}{2}\boldsymbol{u}^{\mathrm{H}}, \\
&\rho^{\mathrm{P}} \frac{\boldsymbol{u}^{\mathrm{P}} - \boldsymbol{u}^{\mathrm{H}}}{\Delta t} - \frac{1}{\mathrm{Re}} \nabla \cdot \boldsymbol{\tau}(\boldsymbol{u}^*) = \mathbf{0}, \\
&\rho^{\mathrm{P}} \frac{e^{\mathrm{P}} - e^{\mathrm{H}}}{\Delta t} - \frac{1}{\mathrm{Re}} \boldsymbol{\tau}(\boldsymbol{u}^*) : \nabla \boldsymbol{u}^* = \frac{\lambda}{\mathrm{Re}} \Delta e^{\mathrm{P}},
\end{aligned}$$

which can be implemented as first solving two decoupled linear systems for $\boldsymbol{u}^*$ and e^{P}

$$\rho^{\mathrm{P}} \boldsymbol{u}^* - \frac{\Delta t}{2\mathrm{Re}} \nabla \cdot \boldsymbol{\tau}(\boldsymbol{u}^*) = \rho^{\mathrm{H}} \boldsymbol{u}^{\mathrm{H}}, \tag{6a}$$

$$\rho^{\mathrm{P}} e^{\mathrm{P}} - \frac{\Delta t\, \lambda}{\mathrm{Re}} \Delta e^{\mathrm{P}} = \rho^{\mathrm{H}} e^{\mathrm{H}} + \frac{\Delta t}{\mathrm{Re}} \boldsymbol{\tau}(\boldsymbol{u}^*) : \nabla \boldsymbol{u}^*, \tag{6b}$$

then setting $\boldsymbol{u}^{\mathrm{P}} = 2\boldsymbol{u}^* - \boldsymbol{u}^{\mathrm{H}}$.

Step 4. Given $(\rho^{\mathrm{P}}, \boldsymbol{u}^{\mathrm{P}}, e^{\mathrm{P}})$, compute $(\boldsymbol{m}^{\mathrm{P}}, E^{\mathrm{P}})$ by

$$\boldsymbol{m}^{\mathrm{P}} = \rho^{\mathrm{P}} \boldsymbol{u}^{\mathrm{P}} \quad \text{and} \quad E^{\mathrm{P}} = \rho^{\mathrm{P}} e^{\mathrm{P}} + \frac{1}{2}\rho^{\mathrm{P}} \|\boldsymbol{u}^{\mathrm{P}}\|^2.$$

Step 5. Given $\boldsymbol{U}^{\mathrm{P}} = [\rho^{\mathrm{P}}, \boldsymbol{m}^{\mathrm{P}}, E^{\mathrm{P}}]^{\mathrm{T}}$, to obtain $\boldsymbol{U}^{n+1} = [\rho^{n+1}, \boldsymbol{m}^{n+1}, E^{n+1}]^{\mathrm{T}}$ in (4), solve (H) for another $\frac{1}{2}\Delta t$ by the third order SSP Runge–Kutta.

2.2. Space discretization

Let $\mathcal{T}_h$ be a polygonal mesh of the computational domain Ω, where each element K is a square in two dimension and degenerates to an interval in one dimension. Let h denote the mesh size, namely the diagonal length of a square element in two dimension and the interval length in one dimension.

Let $\mathbb{Q}^k(K)$ be the space of tensor product of one-dimensional polynomials of degree k on an element K. Define the following discontinuous polynomial spaces:

$$M_h^k = \left\{\chi_h \in L^2(\Omega):\ \forall K \in \mathcal{T}_h,\ \chi_h|_K \in \mathbb{Q}^k(K)\right\},$$
$$\mathbf{X}_h^k = \left\{\boldsymbol{\theta}_h \in L^2(\Omega)^d:\ \forall K \in \mathcal{T}_h,\ \boldsymbol{\theta}_h|_K \in \mathbb{Q}^k(K)^d\right\}.$$

We first briefly review the Runge–Kutta DG scheme for Euler equations, then we describe the IPDG scheme for the parabolic subproblem.

Hyperbolic subproblem. For solving (H), we utilize the same scheme as described in [1], in which a simple limiter can preserve positivity without destroying conservation and accuracy in high order DG methods. The positivity-preserving property will be reviewed in Section 3. Here we briefly review the scheme.

The semi-discrete DG scheme on an element K for the compressible Euler equations $\partial_t \boldsymbol{U} + \nabla\cdot\boldsymbol{F}^{\mathrm{a}}(\boldsymbol{U}) = \boldsymbol{0}$ is defined by finding the piecewise polynomial solution $\boldsymbol{U}_h$ satisfying

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_K \boldsymbol{U}_h\Psi_h = \int_K \boldsymbol{F}^{\mathrm{a}}(\boldsymbol{U}_h)\cdot\nabla\Psi_h - \int_{\partial K}\widehat{\boldsymbol{F}^{\mathrm{a}}\cdot\boldsymbol{n}_K}(\boldsymbol{U}_h^-,\boldsymbol{U}_h^+)\Psi_h, \tag{7}$$

for any piecewise polynomial test function Ψ_h on any element K, where $\boldsymbol{n}_K$ is the unit outward normal of K and the $\widehat{\boldsymbol{F}^{\mathrm{a}}\cdot\boldsymbol{n}_K}$ is a Lax–Friedrichs flux for $\boldsymbol{F}^{\mathrm{a}}$. On a face or an edge $e \subset \partial K$, the local Lax–Friedrichs flux is defined by

$$\widehat{\boldsymbol{F}^{\mathrm{a}}\cdot\boldsymbol{n}_K}(\boldsymbol{U}_h^-,\boldsymbol{U}_h^+) = \frac{\boldsymbol{F}^{\mathrm{a}}(\boldsymbol{U}_h^-)+\boldsymbol{F}^{\mathrm{a}}(\boldsymbol{U}_h^+)}{2}\cdot\boldsymbol{n}_K - \frac{\alpha_e}{2}(\boldsymbol{U}_h^+ - \boldsymbol{U}_h^-),$$

where the $\boldsymbol{U}_h^-$ (resp. $\boldsymbol{U}_h^+$) denotes the trace of a function $\boldsymbol{U}_h$ on the face ∂K coming from the interior (resp. exterior) of K. Here, α_e denotes the maximum wave speed with maximum taken over all $\boldsymbol{U}_h^-$ and $\boldsymbol{U}_h^+$ along the face or edge e, i.e., the largest magnitude of the eigenvalues of the Jacobian matrix $\frac{\partial \boldsymbol{F}^{\mathrm{a}}}{\partial \boldsymbol{U}}$, which equals to the wave speed $|\boldsymbol{u}\cdot\boldsymbol{n}_K| + \sqrt{\gamma\frac{p}{\rho}}$ for ideal gas equation of state.

By convention, we replace $\boldsymbol{U}_h^+$ by an appropriate boundary function which realizes the boundary conditions when $\partial K \cap \partial\Omega \neq \emptyset$. For instance, if purely inflow condition $\boldsymbol{U} = \boldsymbol{U}_{\mathrm{D}}$ is imposed on ∂K, then $\boldsymbol{U}_h^+$ is replaced by $\boldsymbol{U}_{\mathrm{D}}$; if purely outflow condition is imposed on ∂K, then set $\boldsymbol{U}_h^+ = \boldsymbol{U}_h^-$; and if reflective boundary condition for fluid–solid interfaces is imposed on ∂K, then set $\boldsymbol{U}_h^+ = [\rho_h^-, \boldsymbol{m}_h^- - 2(\boldsymbol{m}_h^-\cdot\boldsymbol{n}_K)\boldsymbol{n}_K, E_h^-]^{\mathrm{T}}$.

Parabolic subproblem. We use the IPDG method for discretizing (P). For convenience of introducing discrete forms in parabolic subproblem, we partition the boundary of the domain Ω into the union of two disjoint sets, namely $\partial\Omega = \partial\Omega_{\mathrm{D}} \cup \partial\Omega_{\mathrm{N}}$, where the Dirichlet boundary conditions ($\boldsymbol{u} = \boldsymbol{u}_{\mathrm{D}}$ and $e = e_{\mathrm{D}}$) are applied on $\partial\Omega_{\mathrm{D}}$ and the Neumann-type boundary conditions ($\boldsymbol{\tau}(\boldsymbol{u})\cdot\boldsymbol{n} = \boldsymbol{0}$ and $\nabla e\cdot\boldsymbol{n} = 0$) are applied on $\partial\Omega_{\mathrm{N}}$. Here, $\boldsymbol{n}$ denotes the unit outer normal of domain Ω.

Let Γ_h denote the set of interior faces. For each interior face $e \in \Gamma_h$ shared by elements K_{i^-} and K_{i^+}, with $i^- < i^+$, we define a unit normal vector $\boldsymbol{n}_e$ that points from K_{i^-} into K_{i^+}. For a boundary face e, i. e., $e = \partial K_{i^-} \cap \partial\Omega$, the normal $\boldsymbol{n}_e$ is taken to be the unit outward vector to $\partial\Omega$. We define the broken Sobolev spaces, for any $r \geq 1$,

$$H^r(\mathcal{T}_h) = \{\omega \in L^2(\Omega):\ \forall K \in \mathcal{T}_h,\ \omega|_K \in H^r(K)\}.$$

The average and jump operators of any scalar quantity $\omega \in H^1(\mathcal{T}_h)$ are defined for each interior face $e \in \Gamma_h$ by

$$\{\!\{\omega\}\!\}|_e = \frac{1}{2}\,\omega|_{K_{i^-}} + \frac{1}{2}\,\omega|_{K_{i^+}}, \quad [\![\omega]\!]|_e = \omega|_{K_{i^-}} - \omega|_{K_{i^+}}, \quad e = \partial K_{i^-} \cap \partial K_{i^+}.$$

If a face e belongs to the boundary $\partial\Omega$, the jump and average of ω coincide with its trace on face e. The related definitions of any vector quantity are similar. For more details see [33].

The main focus here is the conservation of momentum and total energy, since we solve the non-conservative form of the parabolic subproblem (3). The fluxes across the element interfaces should be designed such that no extra discrete momentum or discrete total energy is created or eliminated over the whole domain. We utilize the NIPG method to discretize (6a). The bilinear forms $a_{\boldsymbol{\varepsilon}} : H^2(\mathcal{T}_h)^d \times H^2(\mathcal{T}_h)^d \to \mathbb{R}$ and $a_\lambda : H^2(\mathcal{T}_h)^d \times H^2(\mathcal{T}_h)^d \to \mathbb{R}$ associated with terms $-2\nabla \cdot \boldsymbol{\varepsilon}(\boldsymbol{u})$ and $\nabla \cdot ((\nabla \cdot \boldsymbol{u})\mathbf{I})$ are defined as follows:

$$\begin{aligned}
a_{\boldsymbol{\varepsilon}}(\boldsymbol{u}, \boldsymbol{\theta}) = {} & 2 \sum_{K\in\mathcal{T}_h} \int_K \boldsymbol{\varepsilon}(\boldsymbol{u}) : \boldsymbol{\varepsilon}(\boldsymbol{\theta}) - 2 \sum_{e\in\Gamma_h\cup\partial\Omega_{\mathrm{D}}} \int_e \{\!\!\{\boldsymbol{\varepsilon}(\boldsymbol{u})\, \boldsymbol{n}_e\}\!\!\} \cdot [\![\boldsymbol{\theta}]\!] \\
& + 2 \sum_{e\in\Gamma_h\cup\partial\Omega_{\mathrm{D}}} \int_e \{\!\!\{\boldsymbol{\varepsilon}(\boldsymbol{\theta})\, \boldsymbol{n}_e\}\!\!\} \cdot [\![\boldsymbol{u}]\!] + \frac{\sigma}{h} \sum_{e\in\Gamma_h\cup\partial\Omega_{\mathrm{D}}} \int_e [\![\boldsymbol{u}]\!] \cdot [\![\boldsymbol{\theta}]\!], \\
a_\lambda(\boldsymbol{u}, \boldsymbol{\theta}) = {} & - \sum_{K\in\mathcal{T}_h} \int_K (\nabla \cdot \boldsymbol{u})(\nabla \cdot \boldsymbol{\theta}) + \sum_{e\in\Gamma_h\cup\partial\Omega_{\mathrm{D}}} \int_e \{\!\!\{\nabla \cdot \boldsymbol{u}\}\!\!\}\, [\![\boldsymbol{\theta} \cdot \boldsymbol{n}_e]\!] - \sum_{e\in\Gamma_h\cup\partial\Omega_{\mathrm{D}}} \int_e \{\!\!\{\nabla \cdot \boldsymbol{\theta}\}\!\!\}\, [\![\boldsymbol{u} \cdot \boldsymbol{n}_e]\!].
\end{aligned}$$

And the linear form $b_{\boldsymbol{\tau}} : H^2(\mathcal{T}_h)^d \to \mathbb{R}$ associated with the term $-\nabla \cdot \boldsymbol{\tau}(\boldsymbol{u})$ for the Dirichlet boundary $\partial\Omega_{\mathrm{D}}$ in (6a) is defined by

$$b_{\boldsymbol{\tau}}(\boldsymbol{\theta}) = 2 \sum_{e\in\partial\Omega_{\mathrm{D}}} \int_e (\boldsymbol{\varepsilon}(\boldsymbol{\theta})\, \boldsymbol{n}) \cdot \boldsymbol{u}_{\mathrm{D}} + \frac{\sigma}{h} \sum_{e\in\partial\Omega_{\mathrm{D}}} \int_e \boldsymbol{u}_{\mathrm{D}} \cdot \boldsymbol{\theta} - \frac{2}{3} \sum_{e\in\partial\Omega_{\mathrm{D}}} \int_e \nabla \cdot \boldsymbol{\theta}\, (\boldsymbol{u}_{\mathrm{D}} \cdot \boldsymbol{n}).$$

In order to achieve monotonicity for at least $\mathbb{Q}^1$ element, we employ the IIPG method to discretize the term $-\Delta e$ in (6b). In Appendix A, we will prove that the $\mathbb{Q}^1$ IIPG discretization enjoys an M-matrix structure unconditionally. For the IIPG discretization, we define the bilinear form $a_{\mathcal{D}} : H^2(\mathcal{T}_h) \times H^2(\mathcal{T}_h) \to \mathbb{R}$ and the linear form $b_{\mathcal{D}} : H^2(\mathcal{T}_h) \to \mathbb{R}$ for term $-\Delta e$ as follows:

$$\begin{aligned}
a_{\mathcal{D}}(e, \chi) &= \sum_{K\in\mathcal{T}_h} \int_K \nabla e \cdot \nabla \chi - \sum_{e\in\Gamma_h\cup\partial\Omega_{\mathrm{D}}} \int_e \{\!\!\{\nabla e \cdot \boldsymbol{n}_e\}\!\!\}\, [\![\chi]\!] + \frac{\tilde{\sigma}}{h} \sum_{e\in\Gamma_h\cup\partial\Omega_{\mathrm{D}}} \int_e [\![e]\!]\, [\![\chi]\!], \\
b_{\mathcal{D}}(\chi) &= \frac{\tilde{\sigma}}{h} \sum_{e\in\partial\Omega_{\mathrm{D}}} \int_e e_{\mathrm{D}} \chi.
\end{aligned}$$

For the sake of global conservation of total energy, to discrete term $\boldsymbol{\tau}(\boldsymbol{u}) : \nabla \boldsymbol{u} = 2\boldsymbol{\varepsilon}(\boldsymbol{u}) : \nabla \boldsymbol{u} - \frac{2}{3}((\nabla \cdot \boldsymbol{u})\mathbf{I}) : \nabla \boldsymbol{u}$ in (6b), by using the tensor identity $\boldsymbol{\varepsilon}(\boldsymbol{u}) : \nabla \boldsymbol{u} = \boldsymbol{\varepsilon}(\boldsymbol{u}) : \boldsymbol{\varepsilon}(\boldsymbol{u})$, the DG forms $b_{\varepsilon} : H^2(\mathcal{T}_h)^d \times H^2(\mathcal{T}_h) \to \mathbb{R}$ and $b_\lambda : H^2(\mathcal{T}_h)^d \times H^2(\mathcal{T}_h) \to \mathbb{R}$ are designed for terms $2\boldsymbol{\varepsilon}(\boldsymbol{u}) : \nabla \boldsymbol{u}$ and $-((\nabla \cdot \boldsymbol{u})\mathbf{I}) : \nabla \boldsymbol{u}$, respectively.

$$\begin{aligned}
b_{\varepsilon}(\boldsymbol{u}, \chi) &= 2 \sum_{K\in\mathcal{T}_h} \int_K \boldsymbol{\varepsilon}(\boldsymbol{u}) : \boldsymbol{\varepsilon}(\boldsymbol{u}) \chi + \frac{\sigma}{h} \sum_{e\in\Gamma_h} \int_e [\![\boldsymbol{u}]\!] \cdot [\![\boldsymbol{u}]\!]\, \{\!\!\{\chi\}\!\!\} + \frac{\sigma}{h} \sum_{e\in\partial\Omega_{\mathrm{D}}} \int_e (\boldsymbol{u} - \boldsymbol{u}_{\mathrm{D}}) \cdot (\boldsymbol{u} - \boldsymbol{u}_{\mathrm{D}}) \chi, \\
b_\lambda(\boldsymbol{u}, \chi) &= - \sum_{K\in\mathcal{T}_h} \int_K (\nabla \cdot \boldsymbol{u})(\nabla \cdot \boldsymbol{u}) \chi.
\end{aligned}$$

We note that the DG forms above employ penalty parameters σ and $\tilde{\sigma}$. For any $\sigma \geq 0$, the bilinear form of the NIPG method is coercive. In particular, NIPG0 refers to the choice $\sigma = 0$, namely the penalty term is removed. The NIPG0 method is convergent for polynomial degrees greater than or equal to two in two dimension [33]. For IIPG method, the penalty $\tilde{\sigma}$ needs to be large enough for coercivity. The penalty parameters used in our numerical tests will be given in Section 4. Next we summarize our fully discrete scheme.

The fully discrete scheme. Let $(\cdot, \cdot)$ and $\langle \cdot, \cdot \rangle$ denote the L^2 inner products associated with the quadrature rules which are employed in hyperbolic and parabolic subproblems, respectively. The quadrature rules should be accurate enough for $\mathbb{Q}^k$ polynomial basis. On a rectangular cell, the $(k+1)^d$-point Gauss quadrature and $(k+1)^d$-point Gauss–Lobatto quadrature are accurate for integrating $(2k+1)^{\text{th}}$-order polynomials and $(2k-1)^{\text{th}}$-order polynomials, respectively. The quadrature rule for solving (H) can be the same as in [6, 1, 5].

Our fully discrete scheme for solving (1) can be stated as follows:

Step 1. Given $\boldsymbol{U}_h^n \in M_h^k \times \mathbf{X}_h^k \times M_h^k$, compute $\boldsymbol{U}_h^{\mathrm{H}} \in M_h^k \times \mathbf{X}_h^k \times M_h^k$ by the DG method (7) with the positivity-preserving SSP Runge–Kutta (5) [1, 5] using step size $\frac{\Delta t}{2}$.

Step 2. Given $\boldsymbol{U}_h^{\mathrm{H}} \in M_h^k \times \mathbf{X}_h^k \times M_h^k$, compute $(\boldsymbol{u}_h^{\mathrm{H}}, e_h^{\mathrm{H}}) \in \mathbf{X}_h^k \times M_h^k$ by L^2 projection

$$\langle \boldsymbol{m}_h^{\mathrm{H}}, \boldsymbol{\theta}_h\rangle = \langle \rho_h^{\mathrm{H}} \boldsymbol{u}_h^{\mathrm{H}}, \boldsymbol{\theta}_h\rangle, \quad \forall \boldsymbol{\theta}_h \in \mathbf{X}_h^k \quad \text{and} \quad \langle E_h^{\mathrm{H}}, \chi_h\rangle = \langle \rho_h^{\mathrm{H}} e_h^{\mathrm{H}}, \chi_h\rangle + \frac{1}{2}\langle \rho_h^{\mathrm{H}} \boldsymbol{u}_h^{\mathrm{H}}, \boldsymbol{u}_h^{\mathrm{H}} \chi_h\rangle, \quad \forall \chi_h \in M_h^k. \tag{8}$$

Step 3. First given $\rho_h^{\mathrm{H}} \in M_h^k$, and set $\rho_h^{\mathrm{P}} = \rho_h^{\mathrm{H}}$. Given $(\rho_h^{\mathrm{H}}, \rho_h^{\mathrm{P}}, \boldsymbol{u}_h^{\mathrm{H}}) \in M_h^k \times M_h^k \times \mathbf{X}_h^k$, solve for $(\boldsymbol{u}_h^*, \boldsymbol{u}_h^{\mathrm{P}}) \in \mathbf{X}_h^k \times \mathbf{X}_h^k$, such that for all $\boldsymbol{\theta}_h \in \mathbf{X}_h^k$

$$\langle \rho_h^{\mathrm{P}} \boldsymbol{u}_h^*, \boldsymbol{\theta}_h\rangle + \frac{\Delta t}{2\mathrm{Re}} a_{\varepsilon}(\boldsymbol{u}_h^*, \boldsymbol{\theta}_h) + \frac{\Delta t}{3\mathrm{Re}} a_{\lambda}(\boldsymbol{u}_h^*, \boldsymbol{\theta}_h) = \langle \rho_h^{\mathrm{H}} \boldsymbol{u}_h^{\mathrm{H}}, \boldsymbol{\theta}_h\rangle + \frac{\Delta t}{2\mathrm{Re}} b_{\boldsymbol{\tau}}(\boldsymbol{\theta}_h), \tag{9a}$$

$$\boldsymbol{u}_h^{\mathrm{P}} = 2\boldsymbol{u}_h^* - \boldsymbol{u}_h^{\mathrm{H}}. \tag{9b}$$

Then given $(\rho_h^{\mathrm{H}}, \rho_h^{\mathrm{P}}, \boldsymbol{u}_h^*, e_h^{\mathrm{H}}) \in M_h^k \times M_h^k \times \mathbf{X}_h^k \times M_h^k$, solve for $e_h^{\mathrm{P}} \in M_h^k$, such that for all $\chi_h \in M_h^k$

$$\langle \rho_h^{\mathrm{P}} e_h^{\mathrm{P}}, \chi_h\rangle + \frac{\Delta t \lambda}{\mathrm{Re}} a_{\mathcal{D}}(e_h^{\mathrm{P}}, \chi_h) = \langle \rho_h^{\mathrm{H}} e_h^{\mathrm{H}}, \chi_h\rangle + \frac{\Delta t}{\mathrm{Re}} b_{\varepsilon}(\boldsymbol{u}_h^*, \chi_h) + \frac{2\Delta t}{3\mathrm{Re}} b_{\lambda}(\boldsymbol{u}_h^*, \chi_h) + \frac{\Delta t \lambda}{\mathrm{Re}} b_{\mathcal{D}}(\chi_h). \tag{9c}$$

Step 4. Given $(\rho_h^{\mathrm{P}}, \boldsymbol{u}_h^{\mathrm{P}}, e_h^{\mathrm{P}}) \in M_h^k \times \mathbf{X}_h^k \times M_h^k$, compute $(\boldsymbol{m}_h^{\mathrm{P}}, E_h^{\mathrm{P}}) \in \mathbf{X}_h^k \times M_h^k$ by L^2 projection

$$\langle \boldsymbol{m}_h^{\mathrm{P}}, \boldsymbol{\theta}_h\rangle = \langle \rho_h^{\mathrm{P}} \boldsymbol{u}_h^{\mathrm{P}}, \boldsymbol{\theta}_h\rangle, \quad \forall \boldsymbol{\theta}_h \in \mathbf{X}_h^k \quad \text{and} \quad \langle E_h^{\mathrm{P}}, \chi_h\rangle = \langle \rho_h^{\mathrm{P}} e_h^{\mathrm{P}}, \chi_h\rangle + \frac{1}{2}\langle \rho_h^{\mathrm{P}} \boldsymbol{u}_h^{\mathrm{P}}, \boldsymbol{u}_h^{\mathrm{P}} \chi_h\rangle, \quad \forall \chi_h \in M_h^k. \tag{10}$$

Postprocess $\boldsymbol{U}_h^{\mathrm{P}}$ by the positivity-preserving limiter in [1].

Step 5. Given $\boldsymbol{U}_h^{\mathrm{P}} \in M_h^k \times \mathbf{X}_h^k \times M_h^k$, compute $\boldsymbol{U}_h^{n+1} \in M_h^k \times \mathbf{X}_h^k \times M_h^k$ by (7) with step size $\frac{\Delta t}{2}$. Postprocess $\boldsymbol{U}_h^{\mathrm{n+1}}$ by the positivity-preserving limiter in [1].

The initial value $\boldsymbol{U}_h^0$ is obtained via postprocessing the L^2 projection of $\boldsymbol{U}_0$ by the positivity-preserving limiter [1]. The positivity-preserving limiter will be briefly reviewed in Section 3.

In Step 3, the two linear systems (9a) and (9c) are solved sequentially. The unique solvability of the linear systems is a straightforward conclusion due to the coercivity, see Chapter 2 and Chapter 5 in [33].

2.3. The global conservation

Next we show that the fully discrete scheme preserves the global conservation of conserved variables. For simplicity, we discuss the conservation only for periodic boundary conditions. It is straightforward to extend the discussion to many other types of boundary conditions, such as the ones implemented in the numerical tests in this paper.

Both the explicit Runge–Kutta DG scheme for the compressible Euler equations and the positivity-preserving limiter conserve mass, momentum, and total energy [1, 5]. Thus we have

$$(\rho_h^n, 1) = (\rho_h^{\mathrm{H}}, 1), \quad (\boldsymbol{m}_h^n, \mathbf{1}) = (\boldsymbol{m}_h^{\mathrm{H}}, \mathbf{1}), \quad (E_h^n, 1) = (E_h^{\mathrm{H}}, 1),$$

and $(\rho_h^{n+1}, 1) = (\rho_h^{\mathrm{P}}, 1)$. Therefore, $(\rho_h^n, 1) = (\rho_h^{n+1}, 1)$ holds, since in Step 3, we set $\rho_h^{\mathrm{H}} = \rho_h^{\mathrm{P}}$.

Notice that we have $(\boldsymbol{m}_h^n, \mathbf{1}) = (\boldsymbol{m}_h^{\mathrm{H}}, \mathbf{1})$ and $(\boldsymbol{m}_h^{n+1}, \mathbf{1}) = (\boldsymbol{m}_h^{\mathrm{P}}, \mathbf{1})$. Assume that $(\cdot,\cdot)$ and $\langle\cdot,\cdot\rangle$ are accurate enough quadratures, we also have $(\boldsymbol{m}_h^{\mathrm{H}}, \mathbf{1}) = \langle \boldsymbol{m}_h^{\mathrm{H}}, \mathbf{1}\rangle$ and $(\boldsymbol{m}_h^{\mathrm{P}}, \mathbf{1}) = \langle \boldsymbol{m}_h^{\mathrm{P}}, \mathbf{1}\rangle$. Take $\boldsymbol{\theta}_h = \mathbf{1}$ in (8) and (10), we get $\langle \boldsymbol{m}_h^{\mathrm{H}}, \mathbf{1}\rangle = \langle \rho_h^{\mathrm{H}}\boldsymbol{u}_h^{\mathrm{H}}, \mathbf{1}\rangle$ and $\langle \boldsymbol{m}_h^{\mathrm{P}}, \mathbf{1}\rangle = \langle \rho_h^{\mathrm{P}}\boldsymbol{u}_h^{\mathrm{P}}, \mathbf{1}\rangle$. Thus, above identities indicate $(\boldsymbol{m}_h^n, \mathbf{1}) = \langle \rho_h^{\mathrm{H}}\boldsymbol{u}_h^{\mathrm{H}}, \mathbf{1}\rangle$ and $(\boldsymbol{m}_h^{n+1}, \mathbf{1}) = \langle \rho_h^{\mathrm{P}}\boldsymbol{u}_h^{\mathrm{P}}, \mathbf{1}\rangle$. By selecting $\boldsymbol{\theta}_h = \mathbf{1}$ in (9a), we obtain $\langle \rho_h^{\mathrm{H}}\boldsymbol{u}_h^{\mathrm{H}}, \mathbf{1}\rangle = \langle \rho_h^{\mathrm{P}}\boldsymbol{u}_h^{\mathrm{P}}, \mathbf{1}\rangle$, namely the discrete momentum conservation holds.

Similarly, we have $(E_h^n, 1) = \langle \rho_h^{\mathrm{H}} e_h^{\mathrm{H}}, 1\rangle + \frac{1}{2}\langle \rho_h^{\mathrm{H}}\boldsymbol{u}_h^{\mathrm{H}}, \boldsymbol{u}_h^{\mathrm{H}}\rangle$ and $(E_h^{n+1}, 1) = \langle \rho_h^{\mathrm{P}} e_h^{\mathrm{P}}, 1\rangle + \frac{1}{2}\langle \rho_h^{\mathrm{P}}\boldsymbol{u}_h^{\mathrm{P}}, \boldsymbol{u}_h^{\mathrm{P}}\rangle$. Recall that $b_{\boldsymbol{\tau}}(\boldsymbol{\theta}) = 0$ and $b_{\mathfrak{D}}(\chi) = 0$ for periodic boundary conditions, thus by (9b) and $\rho_h^{\mathrm{H}} = \rho_h^{\mathrm{P}}$, the step (9a) can be written as

$$\langle \rho_h^{\mathrm{P}}\boldsymbol{u}_h^{\mathrm{P}}, \boldsymbol{\theta}_h\rangle + \frac{\Delta t}{\mathrm{Re}} a_{\varepsilon}(\boldsymbol{u}_h^*, \boldsymbol{\theta}_h) + \frac{2\Delta t}{3\mathrm{Re}} a_{\lambda}(\boldsymbol{u}_h^*, \boldsymbol{\theta}_h) = \langle \rho_h^{\mathrm{H}}\boldsymbol{u}_h^{\mathrm{H}}, \boldsymbol{\theta}_h\rangle.$$

Plugging in $\boldsymbol{\theta}_h = (\boldsymbol{u}_h^{\mathrm{P}} + \boldsymbol{u}_h^{\mathrm{H}})/2 = \boldsymbol{u}_h^*$, we have

$$\frac{1}{2}\langle \rho_h^{\mathrm{P}}\boldsymbol{u}_h^{\mathrm{P}}, \boldsymbol{u}_h^{\mathrm{P}}\rangle + \frac{\Delta t}{\mathrm{Re}} a_{\varepsilon}(\boldsymbol{u}_h^*, \boldsymbol{u}_h^*) + \frac{2\Delta t}{3\mathrm{Re}} a_{\lambda}(\boldsymbol{u}_h^*, \boldsymbol{u}_h^*) = \frac{1}{2}\langle \rho_h^{\mathrm{H}}\boldsymbol{u}_h^{\mathrm{H}}, \boldsymbol{u}_h^{\mathrm{H}}\rangle. \tag{11}$$

With $\chi_h = 1$ in (9c), we have

$$\langle \rho_h^{\mathrm{P}} e_h^{\mathrm{P}}, 1\rangle + \frac{\Delta t\lambda}{\mathrm{Re}} a_{\mathfrak{D}}(e_h^{\mathrm{P}}, 1) = \langle \rho_h^{\mathrm{H}} e_h^{\mathrm{H}}, 1\rangle + \frac{\Delta t}{\mathrm{Re}} b_{\varepsilon}(\boldsymbol{u}_h^*, 1) + \frac{2\Delta t}{3\mathrm{Re}} b_{\lambda}(\boldsymbol{u}_h^*, 1). \tag{12}$$

By adding two equations above, with the fact that $a_{\mathfrak{D}}(e_h^{\mathrm{P}}, 1) = 0$ and the identities $a_{\varepsilon}(\boldsymbol{u}_h^*, \boldsymbol{u}_h^*) = b_{\varepsilon}(\boldsymbol{u}_h^*, 1)$ and $a_{\lambda}(\boldsymbol{u}_h^*, \boldsymbol{u}_h^*) = b_{\lambda}(\boldsymbol{u}_h^*, 1)$, we obtain

$$\langle \rho_h^{\mathrm{H}} e_h^{\mathrm{H}}, 1\rangle + \frac{1}{2}\langle \rho_h^{\mathrm{H}}\boldsymbol{u}_h^{\mathrm{H}}, \boldsymbol{u}_h^{\mathrm{H}}\rangle = \langle \rho_h^{\mathrm{P}} e_h^{\mathrm{P}}, 1\rangle + \frac{1}{2}\langle \rho_h^{\mathrm{P}}\boldsymbol{u}_h^{\mathrm{P}}, \boldsymbol{u}_h^{\mathrm{P}}\rangle.$$

Theorem 1. *For $\mathbb{Q}^k$ scheme, assume the quadrature rules in hyperbolic and parabolic subproblems are both exact for integrating polynomials of degree k, then the fully discrete scheme conserves density, momentum, and total energy,*

$$(\rho_h^n, 1) = (\rho_h^{n+1}, 1), \quad (\boldsymbol{m}_h^n, \mathbf{1}) = (\boldsymbol{m}_h^{n+1}, \mathbf{1}), \quad (E_h^n, 1) = (E_h^{n+1}, 1).$$

3. The positivity-preserving property

From Section 2, a schematic flowchart of our fully discrete scheme at step $n \geq 0$ is as follows:

$$\boldsymbol{U}_h^n \xrightarrow[\text{step size } \frac{\Delta t}{2}]{\text{solve (H)}} \boldsymbol{U}_h^{\mathrm{H}} \xrightarrow{L^2 \text{ proj.}} (\boldsymbol{u}_h^{\mathrm{H}}, e_h^{\mathrm{H}}) \xrightarrow[\text{step size } \Delta t]{\text{solve (P)}} (\boldsymbol{u}_h^{\mathrm{P}}, e_h^{\mathrm{P}}) \xrightarrow{L^2 \text{ proj.}} \boldsymbol{U}_h^{\mathrm{P}} \xrightarrow[\text{step size } \frac{\Delta t}{2}]{\text{solve (H)}} \boldsymbol{U}_h^{n+1}.$$

At a given time step n, the numerical solution $\boldsymbol{U}_h^n$ is a piecewise polynomial. Usually it is impractical to have $\boldsymbol{U}_h^n(\boldsymbol{x}) \in G$, for all $\boldsymbol{x} \in \Omega$, i.e, positivity holds everywhere. On the other hand, notice that the scheme is implemented with quadrature, thus it suffices to enforce positivity only at quadrature points.

Quadratures and basis. We utilize different quadrature rules for different integral terms such as volume integrals and surface integrals. For $\mathbb{Q}^k$ scheme, the quadrature rules employed in hyperbolic and parabolic subproblems are defined as follows:

1. For face integrals in (H), we use the $(k+1)$-point Gauss quadrature. Denote the set of associated quadrature points here by $S_K^{\mathrm{H,face}}$ on a cell K.
2. For volume integrals in (H), we use a quadrature rule constructed by the tensor product of Gauss quadrature and request this quadrature is accurate for at least $(2k+1)$-order polynomials. Denote the set of associated quadrature points here by $S_K^{\mathrm{H,vol}}$ on a cell K.
3. For all (face and volume) integrals in (P), we use a quadrature rule constructed by the tensor product of $(k+1)$-point Gauss–Lobatto quadrature. Denote the set of associated quadrature points here by S_K^{P} on a cell K.

In addition, we consider the points for weak positivity of the compressible Euler equations [5], which are constructed by $(k+1)$-point Gauss quadrature tensor product with N-point Gauss–Lobatto quadrature in both x and y directions and we request $2N-3 \geq k$. Let $S_K^{\mathrm{H,aux}}$ denote a collection from these points, where each point in $S_K^{\mathrm{H,aux}}$ is located on the interior of a cell K.

As an example, we illustrate the quadrature points in $\mathbb{Q}^2$ scheme. The red points on the left of Figure 2 are used for computing the face integrals of numerical fluxes along the cell boundary when solving the hyperbolic subproblem. The black points together with the red points form a special quadrature for weak positivity. Notice, the black points are not used in computing any numerical integrals [5]. The red points in the middle of Figure 2 are used for computing the volume integrals of numerical fluxes when solving the hyperbolic subproblem. The blue points on the right of Figure 2 are used for computing all of the integrals when solving the parabolic subproblem.

Let $\hat{K} = [-\frac{1}{2}, \frac{1}{2}]^d$ be the reference element. For $\mathbb{Q}^k$ scheme, we use $(k+1)^d$ Gauss–Lobatto points to construct Lagrange interpolation polynomials, which serve as basis functions. For example, the blue points in Figure 2 are used for constructing the bases of our $\mathbb{Q}^2$ scheme. The total number of bases on $\hat{K}$, namely the number of local degrees of freedom is $N_{\mathrm{loc}} = (k+1)^d$. Let $\hat{\boldsymbol{q}}_\nu$ denote the ν^{th} Gauss–Lobatto point on $\hat{K}$, where $\nu = 0, \cdots, N_{\mathrm{loc}}-1$. We assign a basis with an index j, if it equals to 1 when evaluated it at $\hat{\boldsymbol{q}}_j$. From this construction, we have $\hat{\varphi}_j(\hat{\boldsymbol{q}}_\nu) = \delta_{j\nu}$, where δ denotes the Kronecker delta. Let $\boldsymbol{F}_i : \hat{K} \longrightarrow K_i$ denote the invertible mapping from the reference element $\hat{K}$ to $K_i \in \mathcal{T}_h$, then the basis functions on element K_i are defined by $\varphi_{ij} = \hat{\varphi}_j \circ \boldsymbol{F}_i^{-1}$. Thus, we have $\varphi_{i_1 j}(\boldsymbol{q}_{i_2\nu}) = \delta_{i_1 i_2}\delta_{j\nu}$, which indicates the points $\boldsymbol{q}_{i\nu} = \boldsymbol{F}_i(\hat{\boldsymbol{q}}_\nu)$ are not only quadrature nodes but also representing all degrees of freedom on cell K_i. It is obvious that these bases are numerically orthogonal with respect to $(k+1)^d$-point Gauss–Lobatto rule.

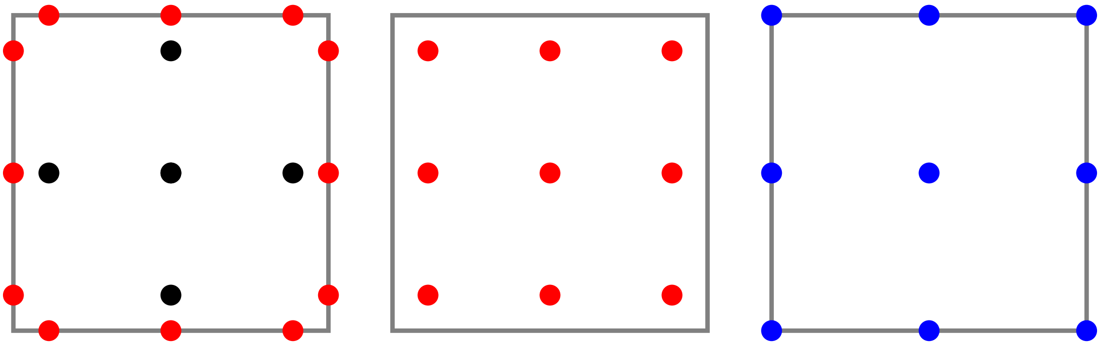

Figure 2: An illustration of quadratures used in $\mathbb{Q}^2$ schemes. Left: Gauss quadrature tensor product with Gauss–Lobatto quadrature in both x and y directions. The points along the boundary are exactly $S_K^{\mathrm{H,face}}$, which are marked red. The other points in $S_K^{\mathrm{H,aux}}$ are marked black. Middle: Gauss quadrature tensor product with Gauss quadrature. The points in $S_K^{\mathrm{H,vol}}$ are marked red. Right: Gauss–Lobatto quadrature tensor product with Gauss–Lobatto quadrature. The points in S_K^{P} are marked blue.

The outline of proving positivity. Let $S_K^{\mathrm{H}} = S_K^{\mathrm{H,face}} \cup S_K^{\mathrm{H,aux}} \cup S_K^{\mathrm{H,vol}}$ and let S_h be the union of set S_K^{H}, for all $K \in \mathcal{T}_h$. On each time iteration of the fully discrete scheme, we apply the positivity-preserving limiter on the following quadrature points on each cell K.

1. In Step 1 and Step 5, on each stage of SSP Runge–Kutta method, all points in set S_K^{H} need to be limited. As an example, for $\mathbb{Q}^2$ scheme, all of the red and black points in Figure 2.

2. In Step 1, on the last stage of SSP Runge–Kutta method, all points in set $S_K^{\mathrm{H}} \cup S_K^{\mathrm{P}}$ need to be limited. As an example, for $\mathbb{Q}^2$ scheme, all of the red, black, and blue points in Figure 2.

3. In Step 4, all points in set S_K^{H} need to be limited. As an example, for $\mathbb{Q}^2$ scheme, all of the red and black points in Figure 2.

To prove our fully discrete scheme is positivity-preserving, we need to show

$$\boldsymbol{U}_h^n(\boldsymbol{x}) \in G, \forall \boldsymbol{x} \in S_h \quad \Rightarrow \quad \boldsymbol{U}_h^{n+1}(\boldsymbol{x}) \in G, \forall \boldsymbol{x} \in S_h$$

by the following steps:

1. The positivity-preserving property of Runge–Kutta DG scheme for compressible Euler equations will be briefly reviewed in Section 3.1.
2. In Section 3.2, we will show that the simple positivity-preserving limiter can ensure positivity in the L^2 projection steps.
3. In Section 3.3 and Appendix A, we will show that the system matrix of (9c) in parabolic subproblem is monotone. Thus, the scheme preserves positivity of internal energy.

We emphasize that the first two steps above can be easily extended to unstructured meshes. But in the third step, the monotonicity of high order schemes only holds on uniform rectangular meshes. For the rest of this section, we only consider a uniform rectangular mesh for a computational domain $\Omega \subset \mathbb{R}^d$.

3.1. Positivity of hyperbolic subproblem and the positivity-preserving limiter

One of the most popular approaches of constructing a positivity-preserving high order DG method for conservation laws was introduced by Zhang and Shu in [6, 1], see also [45, 46, 47, 48, 49, 5]. A high-order SSP Runge–Kutta method (5) is a convex combination of several forward Euler steps, thus the positivity of forward Euler time discretization of (7) also carries over to Runge–Kutta method (5) due to the convex combination.

Define the numerical admissible state set G^ϵ as

$$G^\epsilon = \{\boldsymbol{U} = [\rho, \boldsymbol{m}, E]^{\mathrm{T}} : \ \rho \geq \epsilon, \ \rho e(\boldsymbol{U}) = E - \frac{\|\boldsymbol{m}\|^2}{2\rho} \geq \epsilon\},$$

where ϵ is a small positive number. Let $\boldsymbol{U}_K(\boldsymbol{x})$ denote the DG solution polynomial on a cell K and $\overline{\boldsymbol{U}}_K$ be its cell average on K. The main results in [6, 1] include a sufficient condition for positivity of cell averages $\overline{\boldsymbol{U}}_K^{n+1} \in G^\epsilon$ in the forward Euler discretization of high order DG schemes (7) and a simple positivity-preserving limiter to enforce the sufficient condition without destroying conservation and high order accuracy. To be specific, the sufficient condition for $\overline{\boldsymbol{U}}_K^{n+1} \in G^\epsilon$ is to have certain special quadrature point values of $\boldsymbol{U}_K^n$ to be in G^ϵ, as well as a typical hyperbolic type CFL condition. We emphasize that this special quadrature merely serves as a sufficient condition for positivity of $\overline{\boldsymbol{U}}_K^{n+1} \in G^\epsilon$ and it should not be used for computing any integrals. We refer to [5] for a review of these conditions.

The positivity-preserving limiter modifies the DG polynomial solution $\boldsymbol{U}_h(\boldsymbol{x}) = [\rho_h(\boldsymbol{x}), \boldsymbol{m}_h(\boldsymbol{x}), E_h(\boldsymbol{x})]^{\mathrm{T}}$ with the following two steps under the assumption that the cell average is positive $\overline{\boldsymbol{U}}_K \in G^\epsilon$.

1. First enforce positivity of density by

$$\hat{\rho}_K(\boldsymbol{x}) = \theta_\rho(\rho_K(\boldsymbol{x}) - \overline{\rho}_K) + \overline{\rho}_K, \quad \text{where } \theta_\rho = \min\left\{1, \frac{\overline{\rho}_K - \epsilon}{\overline{\rho}_K - \min\limits_{\boldsymbol{x} \in S_K} \rho_K(\boldsymbol{x})}\right\}.$$

In above, $\overline{\rho}_K$ denotes the cell average of ρ_K on K. Notice that $\hat{\rho}_K$ and ρ_K have the same cell average, and $\hat{\rho}_K(\boldsymbol{x}) = \rho_K(\boldsymbol{x})$ if $\min\limits_{\boldsymbol{x} \in S_K} \rho_K(\boldsymbol{x}) \geq \epsilon$.

2. Define $\widehat{\boldsymbol{U}}(\boldsymbol{x}) = [\hat{\rho}(\boldsymbol{x}), \boldsymbol{m}(\boldsymbol{x}), E(\boldsymbol{x})]^{\mathrm{T}}$ and enforce positivity of internal energy by

$$\widetilde{\boldsymbol{U}}_K(\boldsymbol{x}) = \theta_e(\widehat{\boldsymbol{U}}(\boldsymbol{x}) - \overline{\boldsymbol{U}}_K) + \overline{\boldsymbol{U}}_K, \quad \text{where } \theta_e = \min\left\{1, \frac{\overline{\rho e}_K - \epsilon}{\overline{\rho e}_K - \min\limits_{\boldsymbol{x} \in S_K} \rho e_K(\boldsymbol{x})}\right\}.$$

In above, $\overline{\rho e}_K = \overline{E}_K - \frac{1}{2}\frac{\|\overline{\boldsymbol{m}}_K\|^2}{\overline{\rho}_K}$ and $\rho e(\boldsymbol{x}) = E(\boldsymbol{x}) - \frac{1}{2}\frac{\|\boldsymbol{m}(\boldsymbol{x})\|^2}{\rho(\boldsymbol{x})}$.

We refer to [5] for details of the sufficient condition of positivity of cell averages, the CFL condition, and the rigorous justification of the high order accuracy of such a simple limiter.

3.2. The positivity of the L^2 projection steps

For the quadrature rule in the projection steps (8) and (10), we simply use the tensor product of $(k+1)$-point Gauss–Lobatto quadrature. As an example, for $\mathbb{Q}^2$ scheme, we use the blue points in Figure 2. It is straightforward to verify that this quadrature satisfy the condition for preserving conservation in Section 2, since it is exact for integrating $\mathbb{Q}^k$ polynomials.

Next we show the L^2 projections in (8) and (10) preserve positivity. Since the L^2 projection is local, we only need to consider a cell K_i. Recall that the basis functions are constructed by using Lagrange interpolation polynomials at $(k+1)^d$ Gauss–Lobatto points and they are numerically orthogonal with respect to the employed Gauss–Lobatto rule. Thus, the coefficients of basis functions also represent the values of DG solution polynomials at associated Gauss–Lobatto points. We use subscript ij to denote the point value on cell K_i at the j^{th} Gauss–Lobatto node. We have the following results.

Lemma 1. *If $\boldsymbol{U}_h^{\mathrm{H}}(\boldsymbol{x}) \in G^\epsilon$, for all $\boldsymbol{x} \in S_{K_i}^{\mathrm{H}}$, then after applying the positivity-preserving limiter to $\boldsymbol{U}_h^{\mathrm{H}}$ on all points in $S_{K_i}^{\mathrm{P}}$ and taking the L^2 projection, we have $\rho_h^{\mathrm{H}}(\boldsymbol{q}_{ij}) \geq \epsilon$ and $\rho_h^{\mathrm{H}}(\boldsymbol{q}_{ij})e_h^{\mathrm{H}}(\boldsymbol{q}_{ij}) \geq \epsilon$, for all Gauss–Lobatto points $\boldsymbol{q}_{ij} \in S_{K_i}^{\mathrm{P}}$.*

Proof. The condition $\boldsymbol{U}_h^{\mathrm{H}}(\boldsymbol{x}) \in G^\epsilon$, for all $\boldsymbol{x} \in S_{K_i}^{\mathrm{H}}$, implies $\overline{\boldsymbol{U}_h^{\mathrm{H}}}_{K_i} \in G^\epsilon$. Applying the positivity-preserving limiter on all Gauss–Lobatto points $\boldsymbol{q}_{ij} \in S_{K_i}^{\mathrm{P}}$, we obtain $\rho_h^{\mathrm{H}}(\boldsymbol{q}_{ij}) \geq \epsilon$ and $\rho e(\boldsymbol{U}_h^{\mathrm{H}})\big|_{\boldsymbol{q}_{ij}} \geq \epsilon$. By taking test functions $\boldsymbol{\theta}_h = \boldsymbol{e}_\ell \varphi_{ij}$ and $\chi_h = \varphi_{ij}$ in (8), due to the numerical orthogonality of the Lagrange bases, we get $\boldsymbol{m}_{ij}^{\mathrm{H}} = \rho_{ij}^{\mathrm{H}}\boldsymbol{u}_{ij}^{\mathrm{H}}$ and $E_{ij}^{\mathrm{H}} = \rho_{ij}^{\mathrm{H}}e_{ij}^{\mathrm{H}} + \frac{1}{2}\rho_{ij}^{\mathrm{H}}\|\boldsymbol{u}_{ij}^{\mathrm{H}}\|^2$. Therefore, we have

$$\rho_{ij}^{\mathrm{H}}e_{ij}^{\mathrm{H}} = E_{ij}^{\mathrm{H}} - \frac{1}{2}\rho_{ij}^{\mathrm{H}}\|\boldsymbol{u}_{ij}^{\mathrm{H}}\|^2 = E_{ij}^{\mathrm{H}} - \frac{\|\boldsymbol{m}_{ij}^{\mathrm{H}}\|^2}{2\rho_{ij}^{\mathrm{H}}} = \rho e(\boldsymbol{U}_h^{\mathrm{H}})\big|_{\boldsymbol{q}_{ij}} \geq \epsilon.$$

□

Lemma 2. *If $\rho_h^{\mathrm{P}}(\boldsymbol{q}_{ij}) \geq \epsilon$ and $\rho_h^{\mathrm{P}}(\boldsymbol{q}_{ij})e_h^{\mathrm{P}}(\boldsymbol{q}_{ij}) \geq \epsilon$, for all Gauss–Lobatto points $\boldsymbol{q}_{ij} \in S_{K_i}^{\mathrm{P}}$, then after taking the L^2 projection and applying the positivity-preserving limiter to $\boldsymbol{U}_h^{\mathrm{P}}$ on all points in $S_{K_i}^{\mathrm{H}}$ we have $\boldsymbol{U}_h^{\mathrm{P}}(\boldsymbol{x}) \in G^\epsilon$, for all $\boldsymbol{x} \in S_{K_i}^{\mathrm{H}}$.*

Proof. The density ρ_h^{P} equals to ρ_h^{H}. Thus, we only need to show the positivity of internal energy. By taking test functions $\boldsymbol{\theta}_h = \boldsymbol{e}_\ell \varphi_{ij}$ and $\chi_h = \varphi_{ij}$ in (10), due to the numerical orthogonality of the Lagrange bases, we have $\boldsymbol{m}_{ij}^{\mathrm{P}} = \rho_{ij}^{\mathrm{P}}\boldsymbol{u}_{ij}^{\mathrm{P}}$ and $E_{ij}^{\mathrm{P}} = \rho_{ij}^{\mathrm{P}}e_{ij}^{\mathrm{P}} + \frac{1}{2}\rho_{ij}^{\mathrm{P}}\|\boldsymbol{u}_{ij}^{\mathrm{P}}\|^2$. By $\rho_{ij}^{\mathrm{P}} = \rho_h^{\mathrm{P}}(\boldsymbol{q}_{ij})$ and $e_{ij}^{\mathrm{P}} = e_h^{\mathrm{P}}(\boldsymbol{q}_{ij})$, we have

$$\rho e(\boldsymbol{U}_h^{\mathrm{P}})\big|_{\boldsymbol{q}_{ij}} = E_{ij}^{\mathrm{P}} - \frac{\|\boldsymbol{m}_{ij}^{\mathrm{P}}\|^2}{2\rho_{ij}^{\mathrm{P}}} = E_{ij}^{\mathrm{P}} - \frac{1}{2}\rho_{ij}^{\mathrm{P}}\|\boldsymbol{u}_{ij}^{\mathrm{P}}\|^2 = \rho_{ij}^{\mathrm{P}}e_{ij}^{\mathrm{P}} \geq \epsilon.$$

With the ideal gas equation of state, the ρe is concave with respect to $\boldsymbol{U}$, see [5]. By Jensen's inequality, we have

$$\overline{\rho e}_{K_i} = \rho e(\overline{\boldsymbol{U}_h^{\mathrm{P}}}_{K_i}) = \rho e\Big(\sum_{j=0}^{N_{\mathrm{loc}}-1} \hat{\omega}_j \boldsymbol{U}_{ij}^{\mathrm{P}}\Big) \geq \sum_{j=0}^{N_{\mathrm{loc}}-1} \hat{\omega}_j\ \rho e(\boldsymbol{U}_h^{\mathrm{P}})\big|_{\boldsymbol{q}_{ij}} \geq \epsilon,$$

where the $\hat{\omega}_j$ denotes the j^{th} Gauss–Lobatto quadrature weights on the reference element. Thus, the cell average $\overline{\boldsymbol{U}_h^{\mathrm{P}}}_{K_i} \in G^\epsilon$. Applying the positivity-preserving limiter on points in $S_{K_i}^{\mathrm{H}}$ gives $\boldsymbol{U}_h^{\mathrm{P}}(\boldsymbol{x}) \in G^\epsilon$, for all $\boldsymbol{x} \in S_{K_i}^{\mathrm{H}}$. □

The Lemma 1 implies: if the positivity-preserving limiter is applied on all Lagrange node points in the last stage of SSP Runge–Kutta method on Step 1, then after taking the L^2 projection on Step 2 the internal energy is positive at each Lagrange node point. The Lemma 2 implies: if the solution of (9c) is positive on all Lagrange node points, then applying the positivity-preserving limiter on Step 4 guarantees the input of Step 5 is positive on set S_h.

3.3. Positivity of high-order scheme for parabolic subproblem

A matrix $\mathbf{A}$ is monotone if all entries of its inverse are nonnegative, namely $\mathbf{A}^{-1} \geq 0$. In the rest of this paper, all inequalities related with matrices are entry-wise inequalities. A matrix $\mathbf{A}$ is called an M-matrix if it can be expressed in the form $\mathbf{A} = s\mathbf{I} - \mathbf{B}$, where $\mathbf{B} \geq 0$ and s is greater than or equal to the spectral radius of $\mathbf{B}$. A non-singular M-matrix is inverse-positive, thus is monotone [18].

A convenient way to obtain a sufficient condition on the positivity of internal energy is by proving the monotonicity of a system matrix. To be precise, consider a linear system $(\mathbf{M}+\Delta t\mathbf{L})\boldsymbol{x} = \boldsymbol{b}$, where the matrix $\mathbf{M}$ is diagonal with strictly positive diagonal entries; the matrix $\mathbf{L}$ is an approximation of the Laplace operator such that $\mathbf{L}\mathbf{1} = \mathbf{0}$. Assume the right-hand side vector satisfies $\mathbf{M}^{-1}\boldsymbol{b} \geq \epsilon$, then $(\mathbf{I} + \Delta t\mathbf{M}^{-1}\mathbf{L})\boldsymbol{x} = \mathbf{M}^{-1}\boldsymbol{b} \geq \epsilon$.

Since $(\mathbf{I}+\Delta t\mathbf{M}^{-1}\mathbf{L})\mathbf{1} = \mathbf{1}$, each row of $(\mathbf{I}+\Delta t\mathbf{M}^{-1}\mathbf{L})^{-1}$ sums to one. Notice that if $\mathbf{I}+\Delta t\mathbf{M}^{-1}\mathbf{L}$ is monotone, then each row of $(\mathbf{I} + \Delta t\mathbf{M}^{-1}\mathbf{L})^{-1}$ has nonnegative entries thus forms a convex combination coefficients, thus $\boldsymbol{x} \geq \epsilon$.

Since $\mathbf{M}^{-1} > 0$, $(\mathbf{M} + \Delta t\mathbf{L})^{-1} \geq 0 \Leftrightarrow (\mathbf{I} + \Delta t\mathbf{M}^{-1}\mathbf{L})^{-1} \geq 0$. Thus, the monotonicity of $\mathbf{M} + \Delta t\mathbf{L}$ is sufficient for positivity of $\boldsymbol{x}$.

In order to obtain a monotone system matrix, we use either the IIPG method with $\mathbb{Q}^1$ element or the spectral element method with $\mathbb{Q}^k$ $(k = 2, 3)$ element to discretize the Laplace operator $-\Delta e$ in (6b).

3.3.1. Preserve positivity through the IIPG method

Consider (9c) in a matrix formulation. The entry of a matrix with row index $N_{\text{loc}}i' + j'$ and column index $N_{\text{loc}}i + j$ is denoted by $[\cdot]_{i'j';ij}$. The entry of a vector with index $N_{\text{loc}}i' + j'$ is denoted by $[\cdot]_{i'j'}$. Given ρ_h^{H}, ρ_h^{P}, $\boldsymbol{u}_h^*$, and e_h^{H}, we define the following matrices and vectors:

$$
\begin{aligned}
&[\mathbf{A}_{\mathcal{M}}]_{i'j';ij} = \langle \rho_h^{\text{P}}\varphi_{ij}, \varphi_{i'j'}\rangle, \qquad &[\mathbf{A}_{\mathcal{D}}]_{i'j';ij} = a_{\mathcal{D}}(\varphi_{ij}, \varphi_{i'j'}), \qquad &[\boldsymbol{B}_{\varepsilon}]_{i'j'} = b_{\varepsilon}(\boldsymbol{u}_h^*, \varphi_{i'j'}),\\
&[\boldsymbol{B}_{\mathcal{M}}]_{i'j'} = \langle \rho_h^{\text{H}} e_h^{\text{H}}, \varphi_{i'j'}\rangle, \qquad &[\boldsymbol{B}_{\mathcal{D}}]_{i'j'} = b_{\mathcal{D}}(\varphi_{i'j'}), \qquad &[\boldsymbol{B}_{\lambda}]_{i'j'} = b_{\lambda}(\boldsymbol{u}_h^*, \varphi_{i'j'}).
\end{aligned}
$$

Then, the matrix formulation of (9c) reads: find vector $\boldsymbol{X}_e^{\text{P}}$, where $[\boldsymbol{X}_e^{\text{P}}]_{ij} = e_{ij}^{\text{P}}$, such that:

$$
(\mathbf{A}_{\mathcal{M}} + \frac{\Delta t\lambda}{\text{Re}}\mathbf{A}_{\mathcal{D}})\boldsymbol{X}_e^{\text{P}} = \boldsymbol{B}_{\mathcal{M}} + \frac{\Delta t}{\text{Re}}\boldsymbol{B}_{\varepsilon} + \frac{2\Delta t}{3\text{Re}}\boldsymbol{B}_{\lambda} + \frac{\Delta t\lambda}{\text{Re}}\boldsymbol{B}_{\mathcal{D}}. \tag{13}
$$

Recall we use $(k + 1)^d$-point Gauss–Lobatto quadrature rule to compute all of the numerical integrals in parabolic subproblem and the bases are numerically orthogonal. The matrix $\mathbf{A}_{\mathcal{M}}$ is diagonal with strictly positive diagonal entries. The e_{ij}^{P} represents the value of solution polynomial e_h^{P} evaluated at Gauss–Lobatto point $\boldsymbol{q}_{ij}$. The following lemma shows that the right-hand side of system (13) is positive.

Lemma 3. *On each cell $K_i \in \mathcal{T}_h$, if $\rho_h^{\text{H}}(\boldsymbol{q}_{ij}) > 0$ and $e_h^{\text{H}}(\boldsymbol{q}_{ij}) > 0$, for all $\boldsymbol{q}_{ij} \in S_{K_i}^{\text{P}}$. Then, under $(k+1)^d$-point Gauss–Lobatto quadrature rule, for any penalty $\sigma \geq 0$ and $\tilde{\sigma} \geq 0$, each entry of the right-hand side of* (13) *is positive.*

Proof. By numerical orthogonality of the Lagrange bases with respect to the $(k + 1)^d$-point Gauss–Lobatto quadrature rule, the condition $\rho_h^{\text{H}}(\boldsymbol{q}_{ij}) > 0$ and $e_h^{\text{H}}(\boldsymbol{q}_{ij}) > 0$, for all $\boldsymbol{q}_{ij} \in S_{K_i}^{\text{P}}$, implies $[\boldsymbol{B}_{\mathcal{M}}]_{ij} = \Delta x^2\hat{\omega}_j\rho_{ij}^{\text{H}}e_{ij}^{\text{H}} > 0$. Here, $\hat{\omega}_j$ denotes the j^{th} Gauss–Lobatto quadrature weight on the reference element.

For the second and third terms on the right-hand side of (13), we recall the support of DG basis function φ_{ij} is cell K_i and

$$b_{\varepsilon}(\boldsymbol{u}_h^*, \varphi_{ij}) + \frac{2}{3} b_{\lambda}(\boldsymbol{u}_h^*, \varphi_{ij}) = 2 \int_{K_i} \Big(\boldsymbol{\varepsilon}(\boldsymbol{u}_h^*) : \boldsymbol{\varepsilon}(\boldsymbol{u}_h^*) - \frac{1}{3} (\boldsymbol{\nabla} \cdot \boldsymbol{u}_h^*)^2\Big) \varphi_{ij} \\ + \frac{\sigma}{h} \int_{\partial K_i \subset \Gamma_h} [\![\boldsymbol{u}_h^*]\!] \cdot [\![\boldsymbol{u}_h^*]\!] \{\!\!\{ \varphi_{ij} \}\!\!\} + \frac{\sigma}{h} \int_{\partial K_i \subset \partial\Omega_{\mathrm{D}}} (\boldsymbol{u}_h^* - \boldsymbol{u}_{\mathrm{D}}) \cdot (\boldsymbol{u}_h^* - \boldsymbol{u}_{\mathrm{D}}) \varphi_{ij}. \quad (14)$$

Then, the term $[\boldsymbol{B}_{\varepsilon}]_{ij} + \frac{2}{3}[\boldsymbol{B}_{\lambda}]_{ij}$ equals to $(k+1)^d$-point Gauss–Lobatto integral of (14). By tensor inequality $\boldsymbol{\varepsilon}(\boldsymbol{u}) : \boldsymbol{\varepsilon}(\boldsymbol{u}) \geq \frac{1}{d}(\boldsymbol{\nabla} \cdot \boldsymbol{u})^2$, we obtain $\left(\boldsymbol{\varepsilon}(\boldsymbol{u}_h^*) : \boldsymbol{\varepsilon}(\boldsymbol{u}_h^*) - \frac{1}{3}(\boldsymbol{\nabla} \cdot \boldsymbol{u}_h^*)^2\right)\Big|_{\boldsymbol{q}_{ij}} \geq 0$, for all $\boldsymbol{q}_{ij} \in S_{K_i}^{\mathrm{P}}$ when dimension $d \leq 3$. Notice, from the bases construction, we always have $\varphi_{i_1 j}(\boldsymbol{q}_{i_2 \nu}) = \delta_{i_1 i_2} \delta_{j\nu} \geq 0$. Thus, as long as the penalty $\sigma \geq 0$, we have $[\boldsymbol{B}_{\varepsilon}]_{ij} + \frac{2}{3}[\boldsymbol{B}_{\lambda}]_{ij} \geq 0$.

Finally, it is straightforward to see the last term on the right-hand side of (13) is always non-negative, since the Dirichlet boundary condition $e_{\mathrm{D}} > 0$ and penalty $\tilde{\sigma} \geq 0$. □

In Step 3 of the fully discrete scheme, we have $\rho_h^{\mathrm{P}} = \rho_h^{\mathrm{H}}$. Furthermore, the system matrix $\mathbf{A}_{\mathcal{M}} + \frac{\Delta t \lambda}{\mathrm{Re}} \mathbf{A}_{\mathcal{D}}$ associated with the $\mathbb{Q}^1$ IIPG discretization has an M-matrix structure unconditionally. We include the proof in Appendix A. Therefore, we obtain $e_h^{\mathrm{H}}(\boldsymbol{q}_{ij}) > 0 \Rightarrow e_h^{\mathrm{P}}(\boldsymbol{q}_{ij}) > 0$, for all of the Gauss–Lobatto points $\boldsymbol{q}_{ij} \in S_{K_i}^{\mathrm{P}}$.

3.3.2. Preserve positivity through the spectral element method

Except the fourth order compact finite difference scheme [17], no known high order schemes have an M-matrix structure. On the other hand, M-matrix structure is only a sufficient but rather than a necessary condition for monotonicity. In particular, a matrix is monotone if it is a product of some M-matrices. For example, $\mathbf{A} = \mathbf{M}_1 \mathbf{M}_2$ where $\mathbf{M}_1$ and $\mathbf{M}_2$ are both M-matrices, then $\mathbf{A}$ is still monotone since $\mathbf{A}^{-1} = \mathbf{M}_2^{-1} \mathbf{M}_1^{-1} \geq 0$.

The $\mathbb{Q}^k$ continuous finite element method implemented by $(k+1)^d$-point Gauss–Lobatto quadrature is also called the spectral element method [50]. In [20], it is proven that $\mathbb{Q}^2$ spectral element method is a product of two M-matrices thus is monotone for a variable coefficient elliptic operator $-\boldsymbol{\nabla} \cdot (a \boldsymbol{\nabla} u) + cu$ under suitable mesh constraints. In [21], $\mathbb{Q}^3$ spectral element method is proven to be a product of four M-matrices for the Laplacian operator thus monotone. The monotonicity of $\mathbb{Q}^2$ spectral element method has been used to construct high order accurate positivity-preserving schemes for Keller–Segel, Allen–Cahn, and Fokker–Planck equations [41, 42, 43].

In this paper, we simply apply the existing monotonicity results in $\mathbb{Q}^2$ spectral element method [20] and $\mathbb{Q}^3$ spectral element method [21] to the Laplacian operator in (6b) and couple it with the DG discretization (9a) in parabolic subproblem. For the sake of simplicity, consider the thermally insulating boundary condition $\boldsymbol{\nabla} e \cdot \boldsymbol{n} = 0$ on the entire boundary of domain Ω. Define continuous piecewise $\mathbb{Q}^k$ polynomial space

$$\tilde{M}_h^k = \big\{ \chi_h \in C(\Omega) : \ \forall K \in \mathcal{T}_h, \ \chi_h|_K \in \mathbb{Q}^k(K) \big\}.$$

Recall in Step 3 of the fully discrete scheme, when solving (9c), the ρ_h^{H}, ρ_h^{P}, $\boldsymbol{u}_h^*$, and e_h^{H} are given data. We replace (9c) by introducing the bilinear form $a_{\mathrm{CG}} : \tilde{M}_h^k \times \tilde{M}_h^k \longrightarrow \mathbb{R}$ and the linear form $b_{\mathrm{CG}} : \tilde{M}_h^k \longrightarrow \mathbb{R}$, as follows:

$$a_{\mathrm{CG}}(e_h, \chi_h) = \int_{\Omega} \rho_h^{\mathrm{P}} e_h \chi_h + \frac{\Delta t \lambda}{\mathrm{Re}} \int_{\Omega} \boldsymbol{\nabla} e_h \cdot \boldsymbol{\nabla} \chi_h,$$
$$b_{\mathrm{CG}}(\chi_h) = \int_{\Omega} \rho_h^{\mathrm{H}} e_h^{\mathrm{H}} \chi_h + \frac{\Delta t}{\mathrm{Re}} b_{\varepsilon}(\boldsymbol{u}_h^*, \chi_h) + \frac{2\Delta t}{3\mathrm{Re}} b_{\lambda}(\boldsymbol{u}_h^*, \chi_h).$$

Then, the variational formulation for solving (6b) becomes: find $e_h^{\mathrm{P}} \in \tilde{M}_h^k$, such that for all $\chi_h \in \tilde{M}_h^k$, the $a_{\mathrm{CG}}(e_h^{\mathrm{P}}, \chi_h) = b_{\mathrm{CG}}(\chi_h)$ holds. For $\mathbb{Q}^k$ scheme, applying $(k+1)^d$-point Gauss–Lobatto quadrature to compute integrals, the (9c) is replaced by: given $(\rho_h^{\mathrm{H}}, \rho_h^{\mathrm{P}}, \boldsymbol{u}_h^*, e_h^{\mathrm{H}}) \in M_h^k \times M_h^k \times \mathbf{X}_h^k \times M_h^k$, solve for $e_h^{\mathrm{P}} \in \tilde{M}_h^k$, such that for all $\chi_h \in \tilde{M}_h^k$,

$$\langle \rho_h^{\mathrm{P}} e_h^{\mathrm{P}}, \chi_h \rangle + \frac{\Delta t \lambda}{\mathrm{Re}} \langle \nabla e_h^{\mathrm{P}}, \nabla \chi_h \rangle = \langle \rho_h^{\mathrm{H}} e_h^{\mathrm{H}}, \chi_h \rangle + \frac{\Delta t}{\mathrm{Re}} b_\varepsilon(\boldsymbol{u}_h^*, \chi_h) + \frac{2\Delta t}{3\mathrm{Re}} b_\lambda(\boldsymbol{u}_h^*, \chi_h). \tag{15}$$

Remark 1. *For two dimensional problems, if we set penalty $\sigma = 0$, namely employ the NIPG0 method in $\mathbb{Q}^2$ and $\mathbb{Q}^3$ discretization for $\nabla \cdot \boldsymbol{\tau}(\boldsymbol{u})$ and $\boldsymbol{\tau}(\boldsymbol{u}) : \nabla \boldsymbol{u}$, then* (15) *is further simplified. We have*

$$\langle \rho_h^{\mathrm{P}} e_h^{\mathrm{P}}, \chi_h \rangle + \frac{\Delta t \lambda}{\mathrm{Re}} \langle \nabla e_h^{\mathrm{P}}, \nabla \chi_h \rangle = \langle \rho_h^{\mathrm{H}} e_h^{\mathrm{H}}, \chi_h \rangle + \frac{2\Delta t}{\mathrm{Re}} \langle \boldsymbol{\varepsilon}(\boldsymbol{u}_h^*) : \boldsymbol{\varepsilon}(\boldsymbol{u}_h^*), \chi_h \rangle - \frac{2\Delta t}{3\mathrm{Re}} \langle (\nabla \cdot \boldsymbol{u}_h^*)(\nabla \cdot \boldsymbol{u}_h^*), \chi_h \rangle.$$

The above formula only involves volume integrals, which is convenient for implementation. And more importantly, with the NIPG0 method, we get rid of the face penalties, which minimizes the numerical viscosity.

The identity $\langle \nabla e_h^{\mathrm{P}}, \nabla 1 \rangle = 0$ acts in the same role as $a_{\mathcal{D}}(e_h^{\mathrm{P}}, 1) = 0$ in proving the conservation of total energy. Replacing (9c) with (15) does not affect the proof of Theorem 1. Therefore, the conservations of density, momentum, and total energy still hold. Similar to Lemma 3, it is straightforward to verify the right-hand side vector stems from (15) is still positive. For $\mathbb{Q}^2$ spectral element scheme, by the results in Section 4 in [20], we obtain a sufficient condition of monotonicity of the system matrix of (15) as follows:

$$\Delta t > \frac{\mathrm{Re}}{3\lambda} \max_{i,j} \rho_{ij}^{\mathrm{H}} \Delta x^2. \tag{16a}$$

For $\mathbb{Q}^3$ spectral element scheme, in principle it is possible to extend the same proof for $-\Delta u$ in Section 6 in [21] to an operator like $-\Delta u + cu$ with a variable coefficent c. Thus in principle the monotonicity of the system matrix of (15) using $\mathbb{Q}^3$ spectral element holds under a time step contraint like

$$\Delta t > C(\mathrm{Re}, \lambda, \rho_h^{\mathrm{H}}) \Delta x^2, \tag{16b}$$

where C is a constant depending on $\mathrm{Re}, \lambda, \rho_h^{\mathrm{H}}$.

We emphasize that the time step constraints (16a) and (16b) are lower bounds, i.e., the time step cannot be as small as Δx^2, which is a practical constraint, rather than an impossible one to implement.

Finally, the unique existence of e_h^{P} is a conclusion from the monotonicity of the system matrix, since it is invertible. Therefore, we obtain $e_h^{\mathrm{H}}(\boldsymbol{q}_{ij}) > 0 \Rightarrow e_h^{\mathrm{P}}(\boldsymbol{q}_{ij}) > 0$, for all of the Gauss–Lobatto points $\boldsymbol{q}_{ij} \in S_{K_i}^{\mathrm{P}}$.

3.4. Adaptive time-stepping strategy and implementation

We use SSP Runge–Kutta method in the fully discrete scheme to solve the hyperbolic subproblem. By [1, 5], for the compressible Euler equations on a structure mesh, a sufficient condition on preserving positivity in a single forward Euler step with step size Δt^{H} is

$$\frac{\Delta t^{\mathrm{H}}}{\Delta x} \max_e \alpha_e \le \frac{1}{2} \hat{\omega} = \frac{1}{2} \frac{1}{N(N-1)}, \tag{17}$$

where N is smallest integer satisfying $2N - 3 \ge k$ for $\mathbb{Q}^k$ basis. For the parabolic subproblem, the $\mathbb{Q}^k$ ($k = 2, 3$) scheme is positivity-preserving under the condition (16), which is a lower bound on the time step size. These constraints together imply that for a simulation the mesh resolution Δx should be small enough such that a feasible time step size exist when solving subproblem (H) followed by subproblem (P) in Strang splitting sequentially. However, we should not simply use a time step suggested by these constraints for the compressible NS equations because of the following reasons.

1. Mathematically, the (16) and (17) can be achieved at the same time if Δx is small enough. However, (16) and (17) are only sufficient, but not necessary for preserving positivity in practice.

2. To enforce (17) in SSP Runge–Kutta method, we need to estimate $\max_e \alpha_e$ for each stage. However, it is difficult to accurately estimate this quantity for the two inner time stages in a third order SSP Runge–Kutta method.

3. The wave speed contains $\sqrt{\gamma p/\rho}$, which will be inaccurate for extremely low density problems due to the round-off errors.

Instead, we can apply the following simple adaptive time-stepping strategy. At each time step t^n, given $\boldsymbol{U}_h^n(\boldsymbol{x}) \in G^\epsilon$ for all $\boldsymbol{x} \in S_h$, we start with a trial step size Δt^{trial} by

$$\Delta t^{\text{trial}} = a\hat{\omega}\frac{1}{\max_e \alpha_e}\Delta x, \tag{18}$$

where a is a parameter. We will specify its value in our experiments, see Section 4. For solving hyperbolic subproblem, the time-stepping strategy is the same as in Section 3.2 in [48], which is listed below for completeness:

Algorithm H. At time t^n, select a trial hyperbolic step size Δt^{H}. The input DG polynomial $\boldsymbol{U}_h^n$ satisfies $\boldsymbol{U}_h^n(\boldsymbol{x}) \in G^\epsilon$, for all $\boldsymbol{x} \in S_h$. The parameter ϵ can be set as $\epsilon = \min\{10^{-13}, \overline{\rho}_K^n, \overline{\rho e}_K^n\}$.

Step H1. Given DG polynomial $\boldsymbol{U}_h^n$, compute the first stage to obtain $\boldsymbol{U}_h^{(1)}$.

- If the cell averages $\overline{\boldsymbol{U}}_K^{(1)} \in G^\epsilon$, for all $K \in \mathcal{T}_h$, then apply a positivity-preserving limiter to obtain $\widetilde{\boldsymbol{U}}_h^{(1)}$ and go to Step H2.
- Otherwise, recompute the first stage with halved step size $\Delta t^{\text{H}} \longleftarrow \frac{1}{2}\Delta t^{\text{H}}$. Notice, when Δt^{H} satisfies the hyperbolic CFL (17), the $\overline{\boldsymbol{U}}_K^{(1)} \in G^\epsilon$ is guaranteed.

Step H2. Given DG polynomial $\widetilde{\boldsymbol{U}}_h^{(1)}$, compute the second stage to obtain $\boldsymbol{U}_h^{(2)}$.

- If the cell averages $\overline{\boldsymbol{U}}_K^{(2)} \in G^\epsilon$, for all $K \in \mathcal{T}_h$, then apply a positivity-preserving limiter to obtain $\widetilde{\boldsymbol{U}}_h^{(2)}$ and go to Step H3.
- Otherwise, return to Step H1 and restart the computation with halved step size $\Delta t^{\text{H}} \longleftarrow \frac{1}{2}\Delta t^{\text{H}}$. Notice, even if Δt^{H} satisfies the constraint (17) in Step H1, the $\overline{\boldsymbol{U}}_K^{(2)}$ still may not belong to set G^ϵ, since (17) is based on $\boldsymbol{U}_h^n$ rather than $\widetilde{\boldsymbol{U}}_h^{(1)}$.

Step H3. Given DG polynomial $\widetilde{\boldsymbol{U}}_h^{(2)}$, compute the third stage to obtain $\boldsymbol{U}_h^{(3)}$.

- If the cell averages $\overline{\boldsymbol{U}}_K^{(3)} \in G^\epsilon$, for all $K \in \mathcal{T}_h$, then apply a positivity-preserving limiter to obtain $\boldsymbol{U}_h^{\text{H}}$. We finish the current SSP Runge–Kutta.
- Otherwise, return to Step H1 and restart the computation with halved step size $\Delta t^{\text{H}} \longleftarrow \frac{1}{2}\Delta t^{\text{H}}$. Notice, even if Δt^{H} satisfies the constraint (17) in Step H1, the $\overline{\boldsymbol{U}}_K^{(3)}$ still may not belong to set G^ϵ, since (17) is based on $\boldsymbol{U}_h^n$ rather than $\widetilde{\boldsymbol{U}}_h^{(2)}$.

The adaptive time-stepping strategy for solving the compressible NS equations can be now defined as follows. At initial, the $\boldsymbol{U}_h^0$ is constructed by L^2 projection of $\boldsymbol{U}_0$ with a positive-preserving limiter on S_h, e.g., we have $\boldsymbol{U}_h^0(\boldsymbol{x}) \in G^\epsilon$, for all $\boldsymbol{x} \in S_h$.

Algorithm CNS. At time t^n, select $\Delta t = \Delta t^{\text{trial}}$ as a desired time step size. The input DG polynomial $\boldsymbol{U}_h^n$ satisfies $\boldsymbol{U}_h^n(\boldsymbol{x}) \in G^\epsilon$, for all $\boldsymbol{x} \in S_h$. The parameter ϵ is taken as $\epsilon = \min\{10^{-13}, \overline{\rho}_K^n, \overline{\rho e}_K^n\}$.

Step CNS1. Given DG polynomial $\boldsymbol{U}_h^n$, solve subproblem (H) form time t^n to $t^n + \frac{\Delta t}{2}$.

- Set $m = 0$. Let $t^{n,0} = t^n$ and $\boldsymbol{U}_h^{n,0} = \boldsymbol{U}_h^n$.
- Given $\boldsymbol{U}_h^{n,m}$ at time $t^{n,m}$, solve (H) to compute $\boldsymbol{U}_h^{n,m+1}$ by the Algorithm H. Let $t^{n,m+1} = t^{n,m} + \Delta t^{\mathrm{H}}$. If $t^{n,m+1} = t^n + \frac{\Delta t}{2}$, then apply a positive-preserving limiter for $\boldsymbol{U}_h^{n,m+1}$ on all Gauss–Lobatto points in S_K^{P}, for all $K \in \mathcal{T}_h$, we obtain $\boldsymbol{U}_h^{\mathrm{H}}$. Go to Step CNS2. Otherwise, set $m \longleftarrow m + 1$ and repeat solving (H) by Algorithm H until reach $t^n + \frac{\Delta t}{2}$. Notice, when compute $\boldsymbol{U}_h^{n,m+1}$, we can take the minimum of $\Delta t^{\mathrm{trial}}$ and $t^n + \frac{\Delta t}{2} - t^{n,m}$ as a trail Δt^{H} to start Algorithm H.

Step CNS2. Given DG polynomial $\boldsymbol{U}_h^{\mathrm{H}}$, take L^2 projection to compute $(\boldsymbol{u}_h^{\mathrm{H}}, e_h^{\mathrm{H}})$.

Step CNS3. Given DG polynomials $(\rho_h^{\mathrm{H}}, \boldsymbol{u}_h^{\mathrm{H}}, e_h^{\mathrm{H}})$, solve subproblem (P) form time t^n to $t^n + \Delta t$.

- If a negative internal energy $e_h^{\mathrm{P}}(\boldsymbol{q}_{ij})$ emerge, then goto Step CNS1 and restart the computation with doubled time step size $\Delta t \longleftarrow 2\Delta t$.
- Otherwise, go to Step CNS4. Notice, for $\mathbb{Q}^k$ $(k = 2, 3)$ scheme, when Δt satisfies (16), the positivity of internal energy is guaranteed.

Step CNS4. Given DG polynomials $(\rho_h^{\mathrm{P}}, \boldsymbol{u}_h^{\mathrm{P}}, e_h^{\mathrm{P}})$, take L^2 projection follows by applying a positivity-preserving limiter on all points in S_h to compute $\boldsymbol{U}_h^{\mathrm{P}}$.

Step CNS5. Given DG polynomial $\boldsymbol{U}_h^{\mathrm{P}}$, use adaptive time-stepping strategy to solve subproblem (H) form time $t^n + \frac{\Delta t}{2}$ to $t^n + \Delta t$.

Notice that the time-stepping strategy above can easily result in an endless loop for a general spatial discretization. However, since (16) and (17) are sufficient conditions for positivity, (16) and (17) ensure that there will be no endless loops when using this time-stepping strategy with the fully discretized scheme in this paper.

Remark 2. *Our* $\mathbb{Q}^1$ *DG scheme for solving subproblem* (P) *is unconditional positivity-preserving, since the associated system matrix enjoys an M-matrix structure unconditionally, see Appendix A. Therefore, for the* $\mathbb{Q}^1$ *DG scheme, we do not need to adapt time step size with respect to the parabolic subproblem, i.e., Step CNS3 always passes without recomputation. In practice, we can relax the condition for doubling time step size in Step CNS3, since it is not necessary to request the internal energy to be positive at each Gauss–Lobatto point. We can double the time step size only when a negative cell average* $\overline{\boldsymbol{U}_h^{\mathrm{P}}}_K$ *in Step CNS4 emerges. We only observed Step CNS3 recomputation in the first several time steps of* $\mathbb{Q}^2$ *and* $\mathbb{Q}^3$ *Sedov blast wave simulations. For all of the rest numerical experiments in Section 4, Step CNS3 recomputation is not triggered.*

4. Numerical tests

We consider some representative tests for validating our numerical scheme in one and two-dimensional spaces, including the Lax shock tube, the double rarefraction, Sedov blast wave, shock diffraction, shock reflection, and shock reflection-diffraction problems.

The parameters for all the tests are as follows. We use the ideal gas constants $\gamma = 1.4$ and Prandtl number $\mathrm{Pr} = 0.72$. For the penalty parameters in IPDG method for solving (P), in the $\mathbb{Q}^1$ scheme, we set $\sigma = 2$ on Γ_h, $\sigma = 4$ on $\partial\Omega$, and $\tilde{\sigma} = 2$; in the $\mathbb{Q}^2$ and $\mathbb{Q}^3$ schemes, we take NIPG0 method, namely set penalty $\sigma = 0$ on all faces. Since we use the continuous finite element to discretize the term $-\Delta e$ in $\mathbb{Q}^2$ and $\mathbb{Q}^3$ spaces, thus there is no $\tilde{\sigma}$ involved.

We emphasize that only the positivity-preserving limiter is used in the Runge–Kutta DG scheme for the hyperbolic subproblem, and no limiters are used in the parabolic subproblem, even though other limiters, such as TVB limiter [7] and WENO type limiters [51, 52, 53], for reducing oscillations could be used to improve quality of numerical solutions.

4.1. Spatial order of accuracy for smooth solutions in two dimensions

We test the accuracy in space for smooth solutions. We utilize the manufactured solution method on domain $\Omega = [0,1]^2$ and set the end time $T = 0.1024$. The prescribed density, velocity, and internal energy are as follows:

$$\rho = \exp(-t)\sin 2\pi(x+y) + 2,$$
$$\boldsymbol{u} = \begin{bmatrix} \exp(-t)\cos(2\pi x)\sin(2\pi y) + 2 \\ \exp(-t)\sin(2\pi x)\cos(2\pi y) + 2 \end{bmatrix},$$
$$e = \tfrac{1}{2}\exp(-t)\cos(2\pi x)\cos(2\pi y) + 1.$$

The total energy and pressure are computed by $E = \rho e + \frac{1}{2}\rho\|\boldsymbol{u}\|^2$ and $p = (\gamma - 1)\rho e$. The system right-hand side functions are evaluated from above manufactured solutions, as well as the initial and boundary conditions are imposed by the same prescribed solutions.

We choose $\mathrm{Re} = 1$ and $\lambda = 1$ and use the same IPDG penalties as in the physical simulations for solving (P), e.g., for $\mathbb{Q}^1$ scheme, we set $\sigma = 2$ on Γ_h, $\sigma = 4$ on $\partial\Omega$, and $\tilde{\sigma} = 2$; for $\mathbb{Q}^2$ and $\mathbb{Q}^3$ schemes, we take NIPG0 method by setting penalty $\sigma = 0$ on all faces. Note, there is no parameter $\tilde{\sigma}$ involved in $\mathbb{Q}^2$ and $\mathbb{Q}^3$ schemes, since we use the continuous finite element to discrete the term $-\Delta e$.

We obtain spatial convergence rates by computing the solutions on a sequence of uniformly refined meshes with fixed time step size $\Delta t = 2^{-4} \cdot 10^{-4}$. This time step size is small enough, such that the spatial error dominates and the hyperbolic CFL is satisfied. Define the discrete L^2_h error of density by

$$\|\rho_h^n - \rho(t^n)\|^2_{L^2_h} = \Delta x^2 \sum_{i=0}^{N_{\mathrm{el}}-1} \sum_{\nu=0}^{N_{\mathrm{q}}^{\mathrm{H,vol}}-1} \omega_\nu \left| \sum_{j=0}^{N_{\mathrm{loc}}-1} \rho_{ij}^n \hat{\varphi}_j(\hat{\boldsymbol{q}}_\nu) - \rho(t^n) \circ \boldsymbol{F}_i(\hat{\boldsymbol{q}}_\nu) \right|^2,$$

where ω_ν and $\hat{\boldsymbol{q}}_\nu$ are the Gauss quadrature weights and points used in evaluating volume integrals in (H). The discrete L^2_h errors for momentum and total energy are measured similarly. If $\mathtt{err}_{\Delta x}$ denotes the error on a mesh with resolution Δx, then the rate is given by $\ln(\mathtt{err}_{\Delta x}/\mathtt{err}_{\Delta x/2})/\ln 2$. When the time step size is sufficiently small, such that the spatial error dominates, we observe second order convergence for $\mathbb{Q}^1$ and $\mathbb{Q}^2$ schemes and fourth order convergence for $\mathbb{Q}^3$ scheme, see Table 1. For odd-order spaces, we obtain the optimal order of convergence. Since the NIPG method is suboptimal in even-order spaces, a second order convergence for $\mathbb{Q}^2$ scheme is as expected. ~~Even though $\mathbb{Q}^2$ scheme is only second order accurate, its error is obviously smaller than the error produced by $\mathbb{Q}^1$ scheme.~~

k	$\Delta x = \Delta y$	$\|\rho_h^{N_T} - \rho(T)\|_{L^2_h}$	rate	$\|\boldsymbol{m}_h^{N_T} - \boldsymbol{m}(T)\|_{L^2_h}$	rate	$\|E_h^{N_T} - E(T)\|_{L^2_h}$	rate
1	$1/2^3$	$6.397 \cdot 10^{-2}$	—	$2.144 \cdot 10^{-1}$	—	$4.392 \cdot 10^{-1}$	—
	$1/2^4$	$1.978 \cdot 10^{-2}$	1.693	$5.297 \cdot 10^{-2}$	2.017	$1.069 \cdot 10^{-1}$	2.039
	$1/2^5$	$5.194 \cdot 10^{-3}$	1.929	$1.288 \cdot 10^{-2}$	2.040	$2.729 \cdot 10^{-2}$	1.970
2	$1/2^4$	$9.257 \cdot 10^{-3}$	—	$2.519 \cdot 10^{-2}$	—	$4.538 \cdot 10^{-2}$	—
	$1/2^5$	$2.603 \cdot 10^{-3}$	1.830	$7.005 \cdot 10^{-3}$	1.847	$1.248 \cdot 10^{-2}$	1.863
	$1/2^6$	$6.847 \cdot 10^{-4}$	1.927	$1.838 \cdot 10^{-3}$	1.930	$3.327 \cdot 10^{-3}$	1.907
3	$1/2^1$	$1.100 \cdot 10^{-1}$	—	$3.353 \cdot 10^{-1}$	—	$5.739 \cdot 10^{-1}$	—
	$1/2^2$	$1.408 \cdot 10^{-2}$	2.996	$3.645 \cdot 10^{-2}$	3.202	$6.853 \cdot 10^{-2}$	3.066
	$1/2^3$	$9.518 \cdot 10^{-4}$	3.887	$2.360 \cdot 10^{-3}$	3.949	$4.663 \cdot 10^{-3}$	3.878

Table 1: Accuracy test: the $\mathbb{Q}^k$ scheme using a very small time step for a smooth solution, where $k \in \{1,2,3\}$, errors and convergence rates for density, momentum, and total energy.

4.2. Lax shock tube problem

The Lax shock tube problem a classical benchmark problem for gas dynamics equations. We choose the computational domain $\Omega = [-5, 5]$ and set the simulation end time $T = 1.3$. The initial condition is prescribed as follows:

$$[\rho_0, u_0, p_0]^{\mathrm{T}} = \begin{cases} [0.445,\ 0.698,\ 3.528]^{\mathrm{T}} & \text{if } \ x \in [-5, 0), \\ [0.5,\ 0,\ 0.571]^{\mathrm{T}} & \text{if } \ x \in [0, 5]. \end{cases}$$

In addition, the Dirichlet boundary conditions $[\rho, u, p]^{\mathrm{T}} = [0.445,\ 0.698,\ 3.528]^{\mathrm{T}}$ on the left end of domain Ω and $[\rho, u, p]^{\mathrm{T}} = [0.5,\ 0,\ 0.571]^{\mathrm{T}}$ on the right end of domain Ω are supplemented.

We uniformly partition domain Ω into 512 cells. For this one-dimensional problem, the $\mathbb{Q}^1$ scheme is considered. We take the parameter $a = 0.125$ in (18) for adaptive time step size. The Figure 3 shows simulation results of Reynolds number Re = 100 and Re = 1000. The reference solution is generated by a second order finite difference scheme using a fifth order positivity-preserving WENO flux for $\boldsymbol{F}^{\mathrm{a}}$ with a second order approximation for diffusion on a mesh of 64000 points [5].

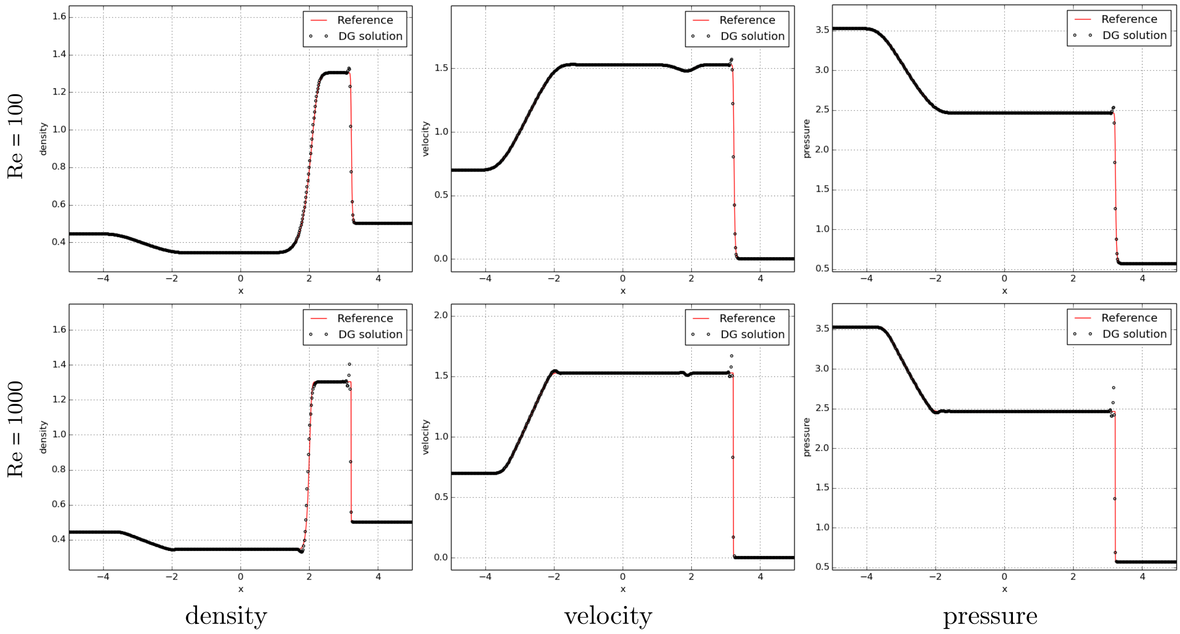

Figure 3: Lax shock tube: the $\mathbb{Q}^1$ scheme with only the positivity-preserving limiter on 512 uniform cells. The snapshots are taken at $T = 1.3$. Only cell averages are plotted.

4.3. Double rarefaction

This Riemann problem contains low density and low pressure. We choose the computational domain $\Omega = [-1, 1]$ and set the simulation end time $T = 0.6$. The initial condition is prescribed as follows:

$$[\rho_0, u_0, p_0]^{\mathrm{T}} = \begin{cases} [7,\ -1,\ 0.2]^{\mathrm{T}} & \text{if } \ x \in [-1, 0), \\ [7,\ 1,\ 0.2]^{\mathrm{T}} & \text{if } \ x \in [0, 1]. \end{cases}$$

In addition, the Dirichlet boundary conditions $[\rho, u, p]^{\mathrm{T}} = [7,\ -1,\ 0.2]^{\mathrm{T}}$ on the left end of domain Ω and $[\rho, u, p]^{\mathrm{T}} = [7,\ 1,\ 0.2]^{\mathrm{T}}$ on the right end of domain Ω are supplemented.

We uniformly partition domain Ω into 512 cells. For this one-dimensional problem, the $\mathbb{Q}^1$ scheme is considered. We take the parameter $a = 0.125$ in (18) for adaptive time step size. The Figure 4 shows simulation results of Reynolds number $\mathrm{Re} = 1000$. The reference solution is generated by a second order finite difference scheme on a mesh of 32000 points [5].

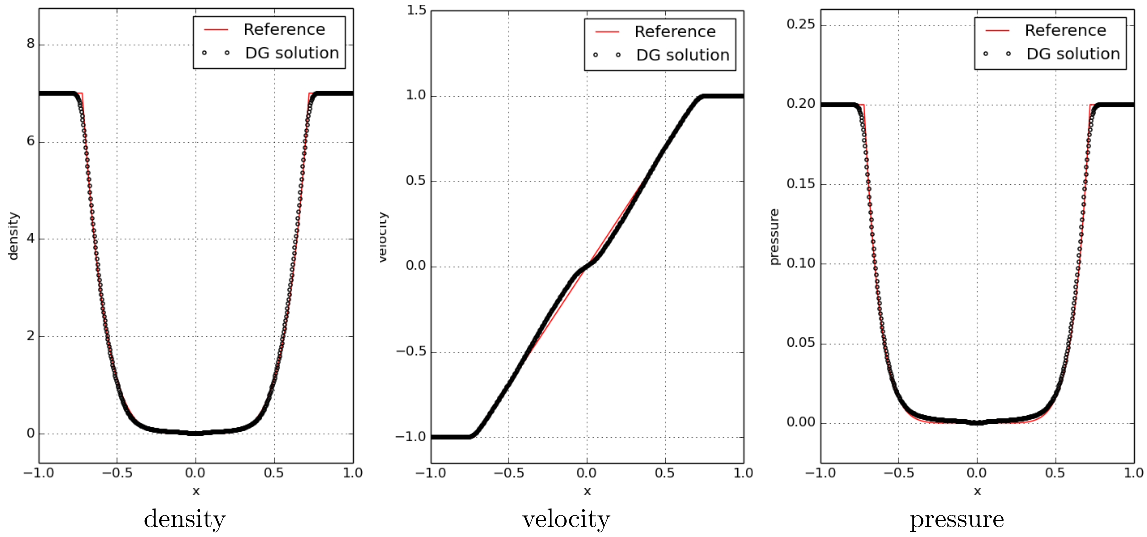

Figure 4: Double rarefaction: the $\mathbb{Q}^1$ scheme with only the positivity-preserving limiter on 512 uniform cells. The snapshots are taken at $T = 0.6$. Only cell averages are plotted.

4.4. Sedov blast wave

The Sedov blast wave involves low density, low pressure, and a strong shock, which is of great utility as a verification test for a positivity-preserving scheme.

We choose the computational domain $\Omega = [0, 1.1]^2$ and set the simulation end time $T = 1$. We uniformly partition domain Ω by square cells with mesh resolution $\Delta x = 1.1/320$. The initials are prescribed as piecewise constants: density $\rho_0 = 1$ and velocity $\boldsymbol{u}_0 = \mathbf{0}$, for all points in Ω; the total energy E_0 equals to 10^{-12} everywhere except the cell at the lower left corner, where $0.244816/\Delta x^2$ is used. The boundary conditions are as follows. In subproblem (H), we utilize reflective boundary condition on the left and bottom edges. The outflow boundary condition is employed on the right and top edges. In subproblem (P), we supplement Neumann-type boundary conditions for both velocity and internal energy.

We take parameter $a = 0.5$ in (18) for $\mathbb{Q}^1$ scheme and $a = 1$ in (18) for $\mathbb{Q}^2$ and $\mathbb{Q}^3$ schemes for adaptive time step size. The Figure 5 displays snapshots of the density field at time $T = 1$ with Reynolds number $\mathrm{Re} = 200$ and $\mathrm{Re} = 1000$. The results are comparable to those in literature, e.g., [5].

4.5. Shock diffraction

Let the computational domain Ω be the union of $[0, 1] \times [6, 12]$ and $[1, 13] \times [0, 12]$. We select the simulation end time $T = 2.3$. The initial condition is a pure right-moving shock of Mach number 5.09, initially located at $\{x = 0.5, 6 \le y \le 12\}$, moving into undisturbed air ahead of the shock with a density of 1.4 and a pressure of 1. For the hyperbolic subproblem, the left boundary of Ω is inflow, the right and bottom boundaries of Ω are outflow, the fluid–solid boundaries $\{y = 6, 0 \le x \le 1\}$ and $\{x = 1, 0 \le y \le 6\}$ are reflective, and the flow values on top boundary are set to describe the exact motion of the Mach 5.09 shock.

We uniformly partition Ω by square cells with mesh resolution $\Delta x = 1/96$ for $\mathbb{Q}^1$ scheme and $\Delta x = 1/64$ for $\mathbb{Q}^2$ and $\mathbb{Q}^3$ schemes, respectively. We take parameter $a = 0.5$ in (18) for $\mathbb{Q}^1$ scheme and $a = 1$ in (18)

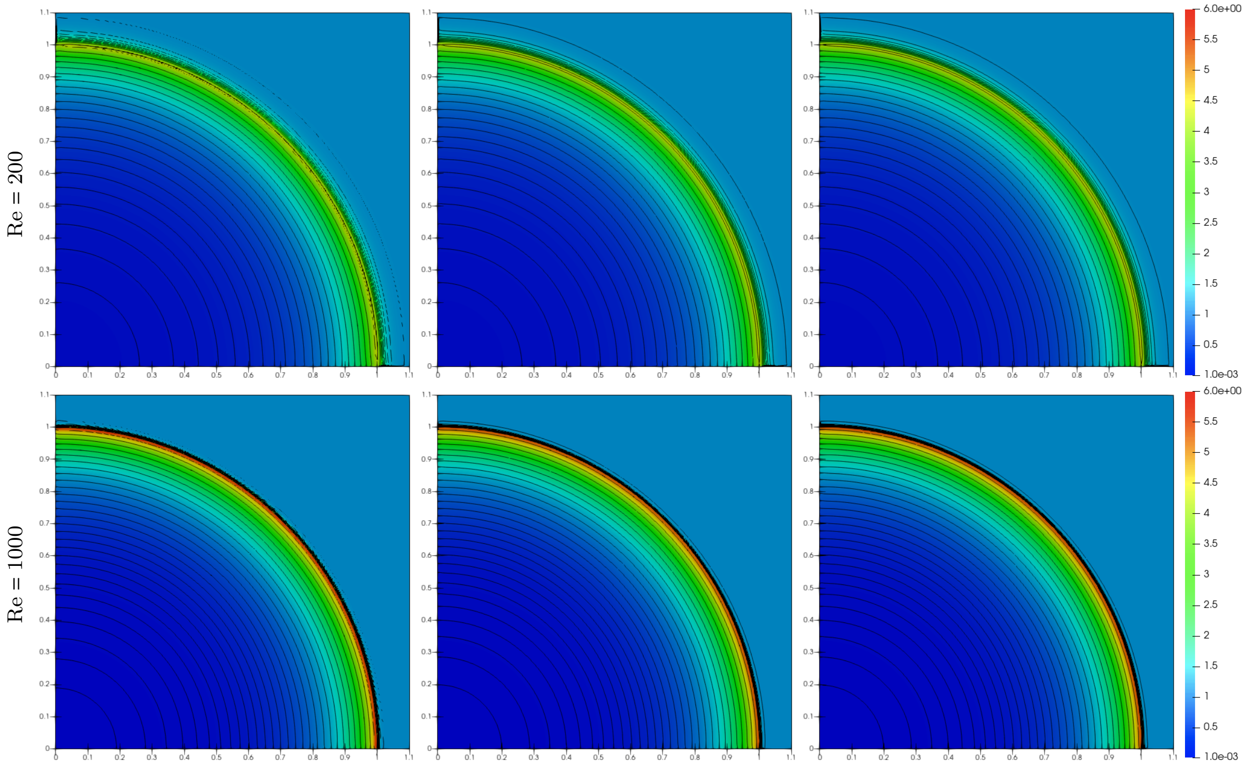

Figure 5: 2D Sedov blast wave. From left to right: the $\mathbb{Q}^1$, $\mathbb{Q}^2$, and $\mathbb{Q}^3$ schemes with only the positivity-preserving limiter on a 320×320 uniform mesh. The snapshots of density profile are taken at $T = 1$. Plot of density: 50 exponentially distributed contour lines of density from 0.001 to 6.

for $\mathbb{Q}^2$ and $\mathbb{Q}^3$ schemes for adaptive time step size. The diffraction of high-speed shocks at a sharp corner generates low density and low pressure. We compare two groups of simulations with Reynolds number Re = 200 and Re = 1000. See Figure 6 for a snapshots of the density field at time $T = 2.3$. We only employ the positivity-preserving limiter. No special treatment is taken at the corner.

4.6. Double Mach reflection of a Mach 10 *shock*

The double Mach reflection of a Mach 10 shock is a widely used benchmark test problem [54]. This experiment studies a planar shock flow in a tube, which contains an oblique wall of thirty degree. In the beginning, the planar shock is perpendicular to the tube surface and move to right. Later, when the shock meets the oblique wall a complicated shock reflection occurs. Following the numerical setup in [55], we tilt the incident shock rather than the solid surface and select the computational domain $\Omega = [0, 4] \times [0, 1]$. We set the simulation end time $T = 0.2$.

A Mach 10 shock initially is positioned at point $(\frac{1}{6}, 0)$ and makes a sixty degree angle with x-axis. The line $6x - 2\sqrt{3}y - 1 = 0$ denotes the shock location and separates domain Ω into left and right zones. For initials, the density equals to 8, the velocity equals to $[4.125\sqrt{3}, -4.125]^{\mathrm{T}}$, and the pressure equals to 116.5 in the post-shock region (left zone). And the undisturbed air ahead of the shock (right zone) has a density of 1.4 and a pressure of 1. For the hyperbolic subproblem, the left boundary of Ω is inflow, the right boundary of Ω is outflow, part of the bottom boundary of Ω on $\{y = 0, \frac{1}{6} \leq x \leq 4\}$ are reflective, and the post-shock condition is imposed at $\{y = 0, 0 \leq x < \frac{1}{6}\}$. On the boundary with post-shock condition, the density, velocity, and pressure are fixed in time with the initial values to make the reflected shock stick to the bottom wall. The flow values on top boundary are set to describe the exact motion of the Mach 10 shock.

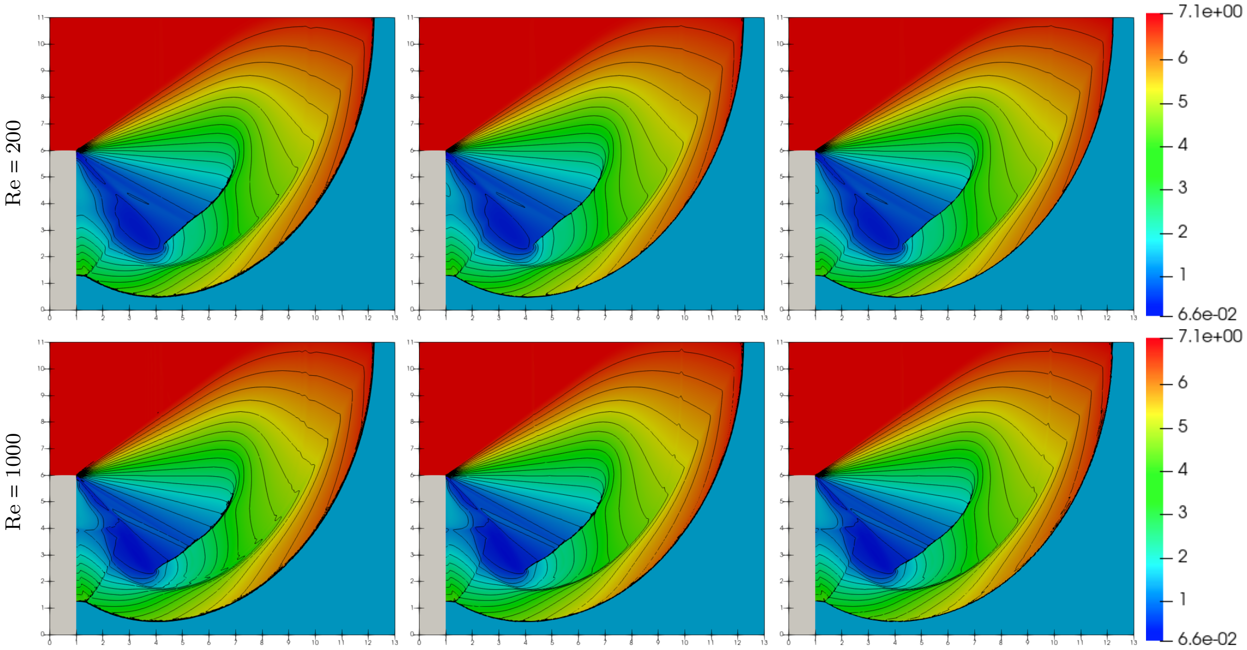

Figure 6: Shock diffraction: the $\mathbb{Q}^1$, $\mathbb{Q}^2$, and $\mathbb{Q}^3$ schemes with only the positivity-preserving limiter on a uniform mesh with resolution $\Delta x = 1/96$ for $\mathbb{Q}^1$ scheme and $\Delta x = 1/64$ for $\mathbb{Q}^2$ and $\mathbb{Q}^3$ schemes. The snapshots of density profile are taken at $T = 2.3$. Plot of density: 20 equally space contour lines from 0.066227 to 7.0668.

We uniformly partition Ω by square cells with the mesh resolution $\Delta x = 1/480$ for $\mathbb{Q}^1$ scheme and $\Delta x = 1/240$ for $\mathbb{Q}^2$ and $\mathbb{Q}^3$ schemes. We take parameter $a = 0.5$ in (18) for $\mathbb{Q}^1$ scheme and $a = 1$ in (18) for $\mathbb{Q}^2$ and $\mathbb{Q}^3$ schemes for adaptive time step size. We compare two groups of simulations with Reynolds number Re = 100 and Re = 1000. The Figure 7 and Figure 8 provide snapshots of the density fields at time $T = 0.2$. For high Reynolds number simulations, it is clear that the rollup is better-captured by the $\mathbb{Q}^3$ scheme than the $\mathbb{Q}^1$ scheme, see Figure 8.

4.7. Mach 10 *shock reflection and diffraction*

This is the same test as in [9]. Let the computational domain Ω be the union of $[1, 4] \times [-1, 0]$ and $[0, 4] \times [0, 1]$. We select the simulation end time $T = 0.2$. A Mach 10 shock initially is positioned at point $(\frac{1}{6}, 0)$ and makes a sixty degree angle with x-axis. The line $6x - 2\sqrt{3}y - 1 = 0$ denotes the initial shock location and separates domain Ω into left zone and right zone. For initials, the density equals to 8, the velocity equals to $[4.125\sqrt{3}, -4.125]^{\mathrm{T}}$, and the pressure equals to 116.5 in the post-shock region (left zone). And the undisturbed air ahead of the shock (right zone) has a density of 1.4 and a pressure of 1.

For the hyperbolic subproblem, the left boundary of Ω is inflow, the right and bottom boundaries of Ω are outflow, part of the fluid–solid boundaries of Ω on $\{y = 0, \frac{1}{6} \leq x \leq 1\}$ and $\{x = 1, -1 \leq y \leq 1\}$ are reflective, and the post-shock condition is imposed at $\{y = 0, 0 \leq x < \frac{1}{6}\}$. On the boundary with post-shock condition, the density, velocity, and pressure are fixed in time with the initial values to make the reflected shock stick to the solid wall. The flow values on top boundary are set to describe the exact motion of the Mach 10 shock.

We take the parameter $a = 0.5$ in (18) for $\mathbb{Q}^1$ scheme and $a = 1$ in (18) for $\mathbb{Q}^2$ and $\mathbb{Q}^3$ schemes for adaptive time step size. Consider three groups of numerical experiments. In the first group of tests, we choose $\mathbb{Q}^1$ scheme and uniformly partition Ω by square cells with the mesh resolution $\Delta x = 1/480$. We various the Reynolds number in three different levels: 100, 500, and 1000. From Figure 9, we see as the Reynolds number increases the rollup becomes stronger. In the second group of tests, we fix the Reynolds

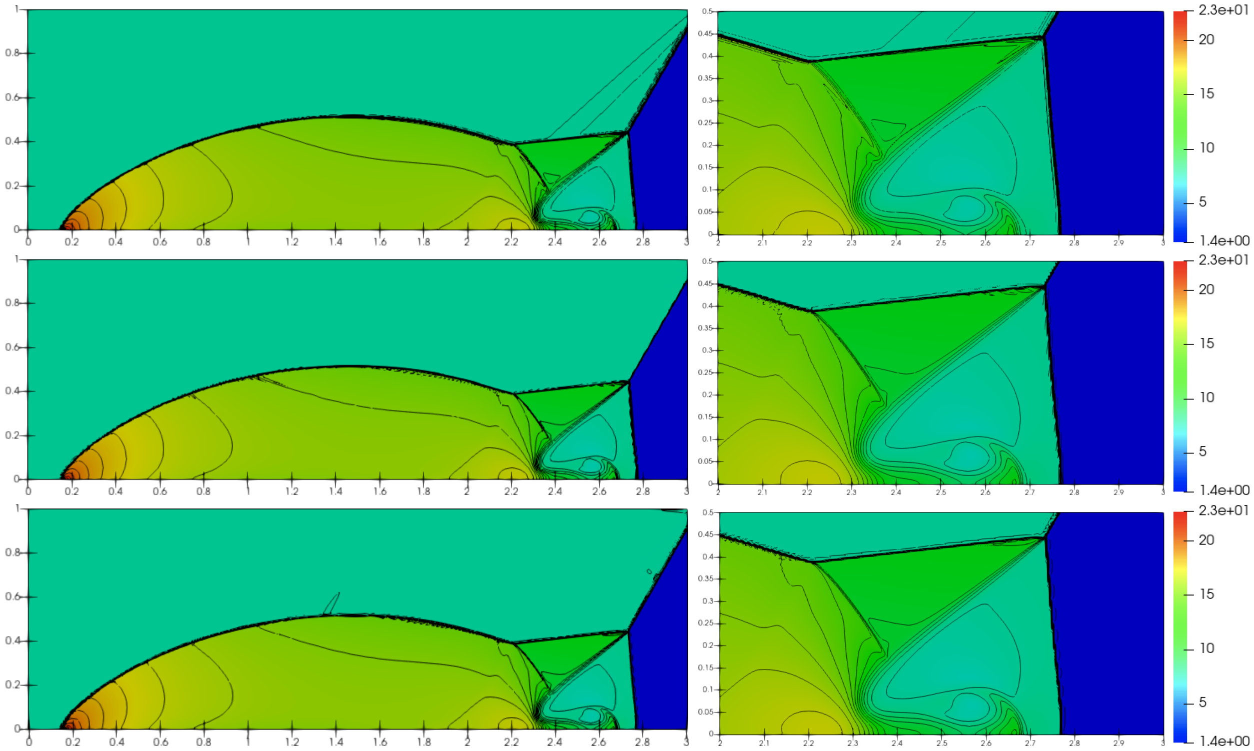

Figure 7: Shock reflection. From top to bottom: simulation results of $\mathbb{Q}^1$, $\mathbb{Q}^2$, and $\mathbb{Q}^3$ schemes for Re = 100 with only the positivity-preserving limiter. The snapshots of density profile are taken at $T = 0.2$. Plot of density: 30 equally space contour lines from 1.3965 to 22.682.

number Re = 1000 and compare the $\mathbb{Q}^1$, $\mathbb{Q}^2$, and $\mathbb{Q}^3$ schemes with mesh resolution $\Delta x = 1/480$, $1/240$, and $1/120$. From Figure 10, we see even though the degrees of freedom for $\mathbb{Q}^3$ simulation are significantly less than the $\mathbb{Q}^1$ simulation, the rollup is well-captured in the $\mathbb{Q}^3$ case. In the third group of tests, we take $\mathbb{Q}^3$ scheme and compare simulation results under different mesh resolutions $\Delta x = 1/120$, $1/180$, $1/240$. From Figure 11, we see as mesh refinement, our scheme produces satisfactory non-oscillatory solutions when the physical diffusion is accurately resolved, which is consistent with the observations for fully explicit high order accurate schemes in [5].

5. Concluding remarks

In this paper, we have constructed an implicit-explicit scheme with high order polynomial basis for solving the compressible NS equations. Our scheme preserves the local conservation of density, global conservation of momentum and total energy, and positivity of density and internal energy, under a CFL constraint like $\Delta t = \mathcal{O}(\Delta x)$. Even though the time accuracy is at most first order, numerical tests suggest that the $\mathbb{Q}^2$ scheme and $\mathbb{Q}^3$ scheme are not only robust but also producing better numerical solutions than the low order $\mathbb{Q}^1$ scheme. Numerical experiments also indicate that our $\mathbb{Q}^3$ scheme with only positivity-preserving limiter produces satisfactory non-oscillatory solutions when physical diffusion is accurately resolved.

Acknowledgments

Research is supported by NSF DMS-2208518.

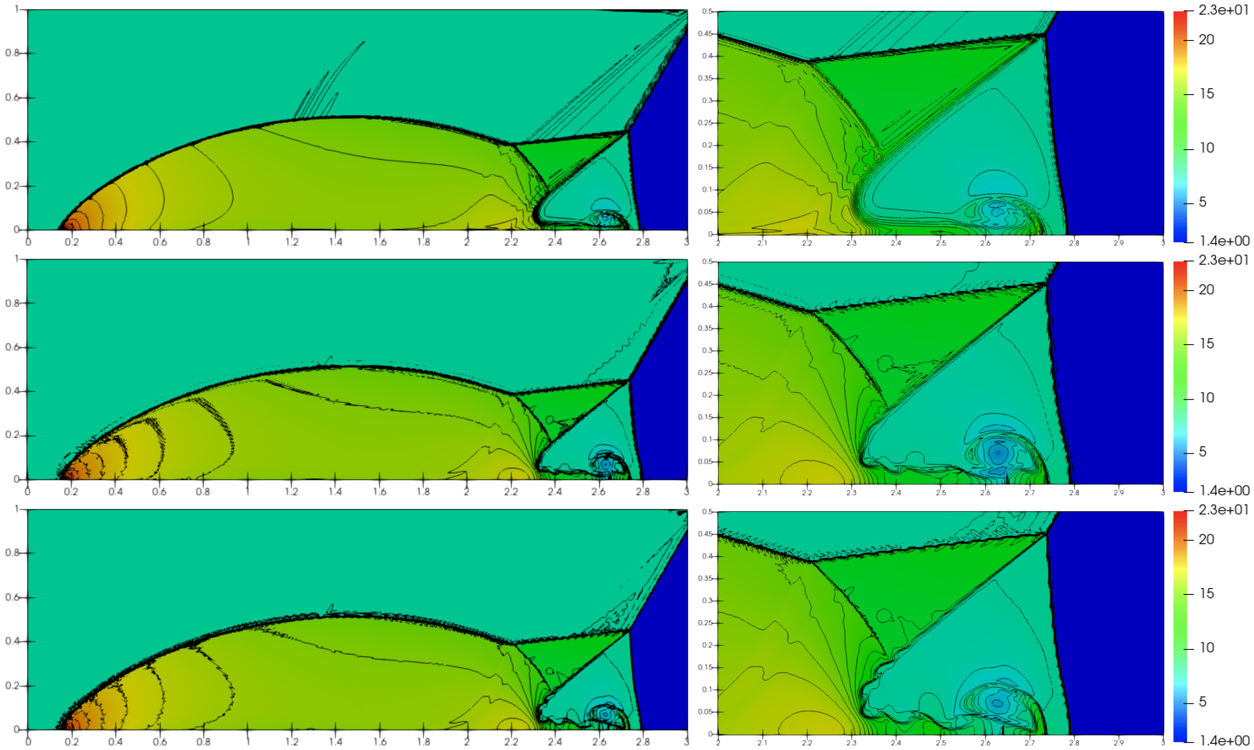

Figure 8: Shock reflection. From top to bottom: simulation results of $\mathbb{Q}^1$, $\mathbb{Q}^2$, and $\mathbb{Q}^3$ schemes for Re = 1000 with only the positivity-preserving limiter. The snapshots of density profile are taken at $T = 0.2$. Plot of density: 30 equally space contour lines from 1.3965 to 22.682.

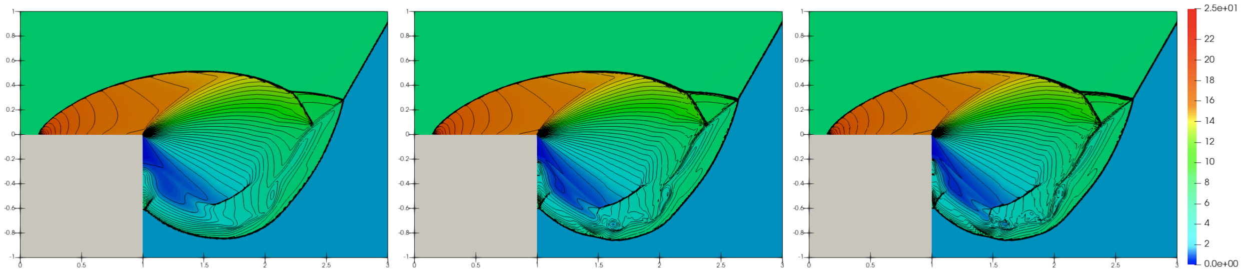

Figure 9: Mach 10 shock reflection and diffraction. The snapshots of density profile are taken at $T = 0.2$. Plot of density: 50 equally space contour lines from 0 to 25. From left to right: simulation results of Reynolds number Re = 100, 500, and 1000 with mesh resolution $\Delta x = 1/480$.

Appendix A. The M-matrix structure of the $\mathbb{Q}^1$ DG scheme for parabolic subproblem

The non-singular M-matrix is an inverse-positive matrix, which serves as a convenient tool for proving the positivity of internal energy. There are many equivalent definitions or characterizations of M-matrix. A comprehensive review of M-matrix can be found in [18]. Here, we state a sufficient but not necessary condition to verify the nonsingular M-matrix.

Lemma 4. *For a real square matrix* $\mathbf{A}$ *with positive diagonal entries and nonpositive off-diagonal entries, it is a nonsingular M-matrix if all the row sums of* $\mathbf{A}$ *are nonnegative and at least one row sum is positive.*

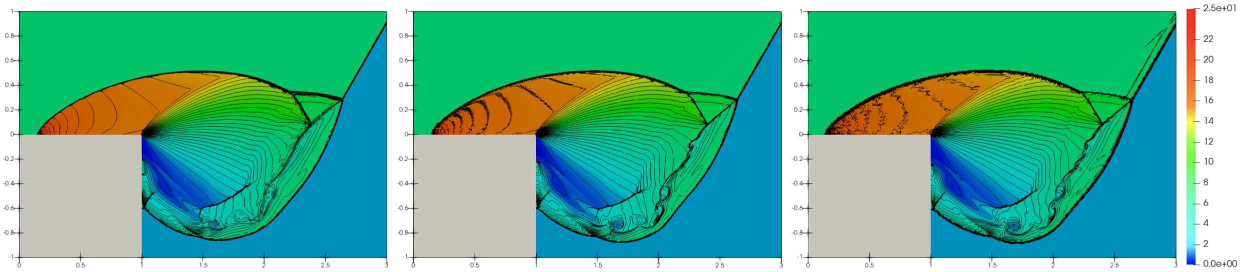

Figure 10: Mach 10 shock reflection and diffraction. The snapshots of density profile are taken at $T = 0.2$. Plot of density: 50 equally space contour lines from 0 to 25. From left to right: simulation results of $\mathbb{Q}^1$, $\mathbb{Q}^2$, and $\mathbb{Q}^3$ schemes with mesh resolution $\Delta x = 1/480$, $1/240$, and $1/120$.

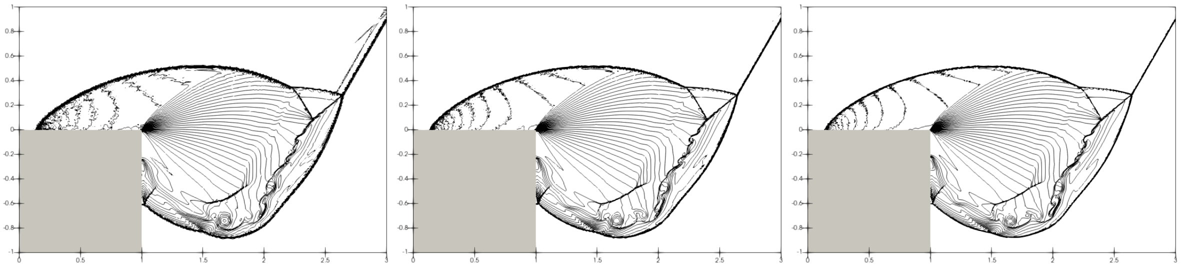

Figure 11: Mach 10 shock reflection and diffraction. The snapshots of density profile are taken at $T = 0.2$. Plot of density: 50 equally space contour lines from 0 to 25. Only contour lines are plotted. From left to right: simulation results of $\mathbb{Q}^3$ scheme with mesh resolution $\Delta x = 1/120$, $1/180$, and $1/240$.

One-dimensional case. Assume the computational domain $\Omega = [-L, L]$, where $L > 0$, is uniformly partitioned into N_{el} intervals (cells) with spacing Δx. Let $-L = x_0 < x_1 < \cdots < x_{N_{\text{el}}} = L$ denote the grid points. On cell $K_i = [x_i, x_{i+1}]$, where $i = 0, \cdots, N_{\text{el}} - 1$, the piecewise linear bases are defined as follows: $\varphi_{i0}(x) = \frac{1}{\Delta x}(x_{i+1} - x)$ and $\varphi_{i1}(x) = \frac{1}{\Delta x}(x - x_i)$. And if $x \notin K_i$, the φ_{i0} and φ_{i1} equal to 0.

In one dimension, the matrix from IIPG discretization of the Laplace operator evaluated by 2-point Gauss–Lobatto quadrature enjoys an M-matrix structure. This result is well-known in literature, for instance, see [56]. Let us present the matrix $\mathbf{A}_{\mathcal{D}}$ explicitly. For simplicity, we only show $\mathbf{A}_{\mathcal{D}}$ with respect to pure Neumann boundary condition. Enforcing part or entire Dirichlet boundary does not break the M-matrix structure.

$$\mathbf{A}_{\mathcal{D}} = \left(\begin{array}{cccccccccc}
{\color{red}\frac{1}{\Delta x}} & -\frac{1}{\Delta x} & 0 & 0 & & & & & & \\
-\frac{1}{2\Delta x} & {\color{red}\frac{1+2\tilde{\sigma}}{2\Delta x}} & \frac{1-2\tilde{\sigma}}{2\Delta x} & -\frac{1}{2\Delta x} & & & & & & \\
& \ddots & \ddots & \ddots & \ddots & & & & & \\
& & -\frac{1}{2\Delta x} & \frac{1-2\tilde{\sigma}}{2\Delta x} & {\color{red}\frac{1+2\tilde{\sigma}}{2\Delta x}} & -\frac{1}{2\Delta x} & & & & \\
& & & & -\frac{1}{2\Delta x} & {\color{red}\frac{1+2\tilde{\sigma}}{2\Delta x}} & \frac{1-2\tilde{\sigma}}{2\Delta x} & -\frac{1}{2\Delta x} & & \\
& & & & & \ddots & \ddots & \ddots & \ddots & \\
& & & & & & -\frac{1}{2\Delta x} & \frac{1-2\tilde{\sigma}}{2\Delta x} & {\color{red}\frac{1+2\tilde{\sigma}}{2\Delta x}} & -\frac{1}{2\Delta x} \\
& & & & & & 0 & 0 & -\frac{1}{\Delta x} & {\color{red}\frac{1}{\Delta x}}
\end{array}\right).$$

In above, we mark all diagonal entries in red color. Obviously, when the penalty parameter $\tilde{\sigma} > 1/2$, the diagonal entries of $\mathbf{A}_{\mathcal{D}}$ are positive. All the off-diagonal entries of $\mathbf{A}_{\mathcal{D}}$ are non-positive. The row sum of $\mathbf{A}_{\mathcal{D}}$ equals zero. In addition, since the Lagrange bases are numerically orthogonal with respect to the Gauss–Lobatto quadrature, the mass matrix is diagonal with positive diagonal entries $[\mathbf{A}_{\mathcal{M}}]_{ij;ij} = \Delta x \hat{\omega}_j \rho^{\text{P}}_{ij}$. Thus the row sum of matrix $\mathbf{A}_{\mathcal{M}} + \frac{\Delta t \lambda}{\text{Re}} \mathbf{A}_{\mathcal{D}}$ is positive. Above all, by Lemma 4, the system matrix $\mathbf{A}_{\mathcal{M}} + \frac{\Delta t \lambda}{\text{Re}} \mathbf{A}_{\mathcal{D}}$ is a non-singular M-matrix, therefore is monotone.

Two-dimensional case. In this part, we show the matrix corresponds to the IIPG discretization of $-\Delta e$ with 2^2-point Gauss–Lobatto quadrature enjoys the M-matrix structure. To the best knowledge of the authors, this is the first time that an M-matrix structure is reported with respect to IPDG method for the Laplace operator in two dimension.

Consider the computational domain Ω is uniformly partitioned into N_{el} square cells with side length Δx. The $\mathbb{Q}^1$ Lagrange bases on reference element $\hat{K} = [-\frac{1}{2}, \frac{1}{2}]^2$ are defined as follows: for $\hat{\boldsymbol{x}} = [\hat{x}, \hat{y}]^{\mathrm{T}} \in \hat{K}$,

$$\hat{\varphi}_0(\hat{\boldsymbol{x}}) = (\frac{1}{2} - \hat{x})(\frac{1}{2} - \hat{y}), \qquad \hat{\varphi}_1(\hat{\boldsymbol{x}}) = (\frac{1}{2} + \hat{x})(\frac{1}{2} - \hat{y}),$$
$$\hat{\varphi}_2(\hat{\boldsymbol{x}}) = (\frac{1}{2} - \hat{x})(\frac{1}{2} + \hat{y}), \qquad \hat{\varphi}_3(\hat{\boldsymbol{x}}) = (\frac{1}{2} + \hat{x})(\frac{1}{2} + \hat{y}).$$

Denote the lower left corner of a cell $K_i \in \mathcal{T}_h$ by $\boldsymbol{a}_{i0}$. The mapping $\boldsymbol{F}_i : \hat{K} \longrightarrow K_i$ and its inverse $\boldsymbol{F}_i^{-1} : K_i \longrightarrow \hat{K}$ are defined by

$$\boldsymbol{F}_i(\hat{\boldsymbol{x}}) = \Delta x\Big(\hat{\boldsymbol{x}} + \frac{1}{2}\begin{bmatrix}1\\1\end{bmatrix}\Big) + \boldsymbol{a}_{i0} \quad \text{and} \quad \boldsymbol{F}_i^{-1}(\boldsymbol{x}) = \frac{1}{\Delta x}(\boldsymbol{x} - \boldsymbol{a}_{i0}) - \frac{1}{2}\begin{bmatrix}1\\1\end{bmatrix}.$$

Then, the bases on cell K_i are $\varphi_{ij} = \hat{\varphi}_j \circ \boldsymbol{F}_i^{-1}$, where $j = 0, \cdots, 3$. Let $\hat{\boldsymbol{\nabla}} = [\hat{\partial}_{\hat{x}}, \hat{\partial}_{\hat{y}}]^{\mathrm{T}}$ denote the gradient on $\hat{K}$. We list the gradient of the basis functions on the reference element, as follows:

$$\hat{\boldsymbol{\nabla}}\hat{\varphi}_0 = \frac{1}{2}\begin{bmatrix}-1+2\hat{y}\\-1+2\hat{x}\end{bmatrix}, \qquad \hat{\boldsymbol{\nabla}}\hat{\varphi}_1 = \frac{1}{2}\begin{bmatrix}1-2\hat{y}\\-1-2\hat{x}\end{bmatrix}, \qquad \hat{\boldsymbol{\nabla}}\hat{\varphi}_2 = \frac{1}{2}\begin{bmatrix}-1-2\hat{y}\\1-2\hat{x}\end{bmatrix}, \qquad \hat{\boldsymbol{\nabla}}\hat{\varphi}_3 = \frac{1}{2}\begin{bmatrix}1+2\hat{y}\\1+2\hat{x}\end{bmatrix}.$$

We index the two faces of $\hat{K}$ which are perpendicular to x-axis by $\hat{e}_0$ and $\hat{e}_1$ and index the two faces which are perpendicular to y-axis by $\hat{e}_2$ and $\hat{e}_3$, namely

$$\hat{e}_0 = \{\hat{x} = -1/2,\ -1/2 \le \hat{y} \le 1/2\}, \qquad \hat{e}_1 = \{\hat{x} = 1/2,\ -1/2 \le \hat{y} \le 1/2\},$$
$$\hat{e}_2 = \{\hat{y} = -1/2,\ -1/2 \le \hat{x} \le 1/2\}, \qquad \hat{e}_3 = \{\hat{y} = 1/2,\ -1/2 \le \hat{x} \le 1/2\}.$$

Define shift mappings with respect to the faces of the reference element as follows:

$$\hat{\boldsymbol{\vartheta}}_0(\hat{\boldsymbol{x}}) = \hat{\boldsymbol{x}} + [1,0]^{\mathrm{T}} \quad \text{if } \hat{\boldsymbol{x}} \in \hat{e}_0, \qquad \hat{\boldsymbol{\vartheta}}_1(\hat{\boldsymbol{x}}) = \hat{\boldsymbol{x}} - [1,0]^{\mathrm{T}} \quad \text{if } \hat{\boldsymbol{x}} \in \hat{e}_1,$$
$$\hat{\boldsymbol{\vartheta}}_2(\hat{\boldsymbol{x}}) = \hat{\boldsymbol{x}} + [0,1]^{\mathrm{T}} \quad \text{if } \hat{\boldsymbol{x}} \in \hat{e}_2, \qquad \hat{\boldsymbol{\vartheta}}_3(\hat{\boldsymbol{x}}) = \hat{\boldsymbol{x}} - [0,1]^{\mathrm{T}} \quad \text{if } \hat{\boldsymbol{x}} \in \hat{e}_3.$$

Let us evaluate entries in matrix $\mathbf{A}_{\mathcal{D}}$. We consider the thermally insulating boundary condition $\boldsymbol{\nabla} e \cdot \boldsymbol{n} = 0$ on the entire boundary of domain Ω. Enforcing part or entire Dirichlet boundary does not break the M-matrix structure. Let matrix $\mathbf{D} = \mathrm{diag}(\mathbf{D}_0, \cdots, \mathbf{D}_{N_{\mathrm{el}}-1})$ be a block diagonal matrix, where each diagonal subblock $\mathbf{D}_{i'} \in \mathbb{R}^{4\times 4}$ is defined by: for any $j', j \in \{0, \cdots, 3\}$, the entry at j'^{th} row and j^{th} column of $\mathbf{D}_{i'}$ is the Gauss–Lobatto integral of the expression

$$\begin{aligned}&\int_{K_{i'}} \boldsymbol{\nabla}\varphi_{i'j} \cdot \boldsymbol{\nabla}\varphi_{i'j'} - \frac{1}{2}\sum_{m=0}^{3}\int_{e_m \in \Gamma_h} \boldsymbol{\nabla}\varphi_{i'j} \cdot \boldsymbol{n}_{K_{i'}}\, \varphi_{i'j'} + \frac{\tilde{\sigma}}{h}\sum_{m=0}^{3}\int_{e_m \in \Gamma_h} \varphi_{i'j}\varphi_{i'j'}\\ &= \int_{\hat{K}} \hat{\boldsymbol{\nabla}}\hat{\varphi}_j \cdot \hat{\boldsymbol{\nabla}}\hat{\varphi}_{j'} - \frac{1}{2}\sum_{m=0}^{3}\iota_m\int_{\hat{e}_m} \hat{\boldsymbol{\nabla}}\hat{\varphi}_j \cdot \hat{\boldsymbol{n}}_{\hat{K}}\, \hat{\varphi}_{j'} + \frac{\tilde{\sigma}}{\sqrt{2}}\sum_{m=0}^{3}\iota_m\int_{\hat{e}_m} \hat{\varphi}_j\hat{\varphi}_{j'}.\end{aligned}$$

In above, ι_m is an indicator, which equals to 1, if the face e_m of element $K_{i'}$ is an interior face, and otherwise equals to 0. We mark all the diagonal entries of $\mathbf{A}_{\mathcal{D}}$ in red color. The diagonal subblocks of $\mathbf{A}_{\mathcal{D}}$ are: for

$i' = 0, \cdots, N_{\mathrm{el}} - 1$,

$$\mathbf{D}_{i'} = \begin{pmatrix} 1 + (\iota_0 + \iota_2)(\frac{\tilde{\sigma}}{2\sqrt{2}} - \frac{1}{4}) & -\frac{1}{2} + \frac{\iota_0}{4} & -\frac{1}{2} + \frac{\iota_2}{4} & 0 \\ -\frac{1}{2} + \frac{\iota_1}{4} & 1 + (\iota_1 + \iota_2)(\frac{\tilde{\sigma}}{2\sqrt{2}} - \frac{1}{4}) & 0 & -\frac{1}{2} + \frac{\iota_2}{4} \\ -\frac{1}{2} + \frac{\iota_3}{4} & 0 & 1 + (\iota_0 + \iota_3)(\frac{\tilde{\sigma}}{2\sqrt{2}} - \frac{1}{4}) & -\frac{1}{2} + \frac{\iota_0}{4} \\ 0 & -\frac{1}{2} + \frac{\iota_3}{4} & -\frac{1}{2} + \frac{\iota_1}{4} & 1 + (\iota_1 + \iota_3)(\frac{\tilde{\sigma}}{2\sqrt{2}} - \frac{1}{4}) \end{pmatrix}. \quad \text{(A.1)}$$

Before computing the off-diagonal subblocks of $\mathbf{A}_{\mathcal{D}}$, let us take a look at an example of a square domain $\Omega = [0, L]^2$, where $L > 0$. For any patition of the domain Ω with more than 2×2 square cells, we divide all cells into three categories: all faces are interior faces; only one face is a boundary face; only two faces are boundary faces. See the blue, green, and red cells in the schematic Figure A.12. Using (A.1), we get if all faces of a cell $K_{i'}$ are interior faces, then the associated diagonal subblock

$$\mathbf{D}_{i'} = \begin{pmatrix} \frac{1}{2} + \frac{\tilde{\sigma}}{\sqrt{2}} & -\frac{1}{4} & -\frac{1}{4} & 0 \\ -\frac{1}{4} & \frac{1}{2} + \frac{\tilde{\sigma}}{\sqrt{2}} & 0 & -\frac{1}{4} \\ -\frac{1}{4} & 0 & \frac{1}{2} + \frac{\tilde{\sigma}}{\sqrt{2}} & -\frac{1}{4} \\ 0 & -\frac{1}{4} & -\frac{1}{4} & \frac{1}{2} + \frac{\tilde{\sigma}}{\sqrt{2}} \end{pmatrix}.$$

If only one face of a cell $K_{i'}$ is a boundary face, then dependents on the boundary face location, the associated diagonal subblock belongs to the following four cases.

$$e_0 \subset \partial\Omega\colon\ \mathbf{D}_{i'} = \begin{pmatrix} \frac{3}{4} + \frac{\tilde{\sigma}}{2\sqrt{2}} & -\frac{1}{2} & -\frac{1}{4} & 0 \\ -\frac{1}{4} & \frac{1}{2} + \frac{\tilde{\sigma}}{\sqrt{2}} & 0 & -\frac{1}{4} \\ -\frac{1}{4} & 0 & \frac{3}{4} + \frac{\tilde{\sigma}}{2\sqrt{2}} & -\frac{1}{2} \\ 0 & -\frac{1}{4} & -\frac{1}{4} & \frac{1}{2} + \frac{\tilde{\sigma}}{\sqrt{2}} \end{pmatrix}, \quad e_1 \subset \partial\Omega\colon\ \mathbf{D}_{i'} = \begin{pmatrix} \frac{1}{2} + \frac{\tilde{\sigma}}{\sqrt{2}} & -\frac{1}{4} & -\frac{1}{4} & 0 \\ -\frac{1}{2} & \frac{3}{4} + \frac{\tilde{\sigma}}{2\sqrt{2}} & 0 & -\frac{1}{4} \\ -\frac{1}{4} & 0 & \frac{1}{2} + \frac{\tilde{\sigma}}{\sqrt{2}} & -\frac{1}{4} \\ 0 & -\frac{1}{4} & -\frac{1}{2} & \frac{3}{4} + \frac{\tilde{\sigma}}{2\sqrt{2}} \end{pmatrix},$$

$$e_2 \subset \partial\Omega\colon\ \mathbf{D}_{i'} = \begin{pmatrix} \frac{3}{4} + \frac{\tilde{\sigma}}{2\sqrt{2}} & -\frac{1}{4} & -\frac{1}{2} & 0 \\ -\frac{1}{4} & \frac{3}{4} + \frac{\tilde{\sigma}}{2\sqrt{2}} & 0 & -\frac{1}{2} \\ -\frac{1}{4} & 0 & \frac{1}{2} + \frac{\tilde{\sigma}}{\sqrt{2}} & -\frac{1}{4} \\ 0 & -\frac{1}{4} & -\frac{1}{4} & \frac{1}{2} + \frac{\tilde{\sigma}}{\sqrt{2}} \end{pmatrix}, \quad e_3 \subset \partial\Omega\colon\ \mathbf{D}_{i'} = \begin{pmatrix} \frac{1}{2} + \frac{\tilde{\sigma}}{\sqrt{2}} & -\frac{1}{4} & -\frac{1}{4} & 0 \\ -\frac{1}{4} & \frac{1}{2} + \frac{\tilde{\sigma}}{\sqrt{2}} & 0 & -\frac{1}{4} \\ -\frac{1}{2} & 0 & \frac{3}{4} + \frac{\tilde{\sigma}}{2\sqrt{2}} & -\frac{1}{4} \\ 0 & -\frac{1}{2} & -\frac{1}{4} & \frac{3}{4} + \frac{\tilde{\sigma}}{2\sqrt{2}} \end{pmatrix}.$$

If only two faces of a cell $K_{i'}$ are boundary faces, then dependents on the boundary face location, the associated diagonal subblock belongs to the following four cases.

$$e_0, e_2 \subset \partial\Omega\colon\ \mathbf{D}_{i'} = \begin{pmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} & 0 \\ -\frac{1}{4} & \frac{3}{4} + \frac{\tilde{\sigma}}{2\sqrt{2}} & 0 & -\frac{1}{2} \\ -\frac{1}{4} & 0 & \frac{3}{4} + \frac{\tilde{\sigma}}{2\sqrt{2}} & -\frac{1}{2} \\ 0 & -\frac{1}{4} & -\frac{1}{4} & \frac{1}{2} + \frac{\tilde{\sigma}}{\sqrt{2}} \end{pmatrix}, \quad e_1, e_2 \subset \partial\Omega\colon\ \mathbf{D}_{i'} = \begin{pmatrix} \frac{3}{4} + \frac{\tilde{\sigma}}{2\sqrt{2}} & -\frac{1}{4} & -\frac{1}{2} & 0 \\ -\frac{1}{2} & 1 & 0 & -\frac{1}{2} \\ -\frac{1}{4} & 0 & \frac{1}{2} + \frac{\tilde{\sigma}}{\sqrt{2}} & -\frac{1}{4} \\ 0 & -\frac{1}{4} & -\frac{1}{2} & \frac{3}{4} + \frac{\tilde{\sigma}}{2\sqrt{2}} \end{pmatrix},$$

$$e_0, e_3 \subset \partial\Omega\colon\ \mathbf{D}_{i'} = \begin{pmatrix} \frac{3}{4} + \frac{\tilde{\sigma}}{2\sqrt{2}} & -\frac{1}{2} & -\frac{1}{4} & 0 \\ -\frac{1}{4} & \frac{1}{2} + \frac{\tilde{\sigma}}{\sqrt{2}} & 0 & -\frac{1}{4} \\ -\frac{1}{2} & 0 & 1 & -\frac{1}{2} \\ 0 & -\frac{1}{2} & -\frac{1}{4} & \frac{3}{4} + \frac{\tilde{\sigma}}{2\sqrt{2}} \end{pmatrix}, \quad e_1, e_3 \subset \partial\Omega\colon\ \mathbf{D}_{i'} = \begin{pmatrix} \frac{1}{2} + \frac{\tilde{\sigma}}{\sqrt{2}} & -\frac{1}{4} & -\frac{1}{4} & 0 \\ -\frac{1}{2} & \frac{3}{4} + \frac{\tilde{\sigma}}{2\sqrt{2}} & 0 & -\frac{1}{4} \\ -\frac{1}{2} & 0 & \frac{3}{4} + \frac{\tilde{\sigma}}{2\sqrt{2}} & -\frac{1}{4} \\ 0 & -\frac{1}{2} & -\frac{1}{2} & 1 \end{pmatrix}.$$

Let matrix $\mathbf{F} = \mathbf{A}_{\mathcal{D}} - \mathbf{D}$, namely $\mathbf{F}$ contains all the off-diagonal subblocks of $\mathbf{A}_{\mathcal{D}}$, where each off-diagonal subblock is associated with integrals on a cell face. To be more accurate, each off-diagonal subblock $\mathbf{F}^m_{i'i} \in \mathbb{R}^{4\times4}$, where $i' \neq i$ and $K_{i'} \cap K_i = e_m$ with $m \in \{0, \cdots, 3\}$, is defined by: for any $j', j \in \{0, \cdots, 3\}$, the entry on j'^{th} row and j^{th} column of $\mathbf{F}^m_{i'i}$ is the Gauss–Lobatto integral of the expression

$$-\frac{1}{2}\int_{e_m} \nabla\varphi_{ij} \cdot \boldsymbol{n}_{K_{i'}}\, \varphi_{i'j'} - \frac{\tilde{\sigma}}{h}\int_{e_m} \varphi_{ij}\varphi_{i'j'} = -\frac{1}{2}\int_{\hat{e}_m} \hat{\nabla}\hat{\varphi}_j \circ \boldsymbol{\vartheta}_m \cdot \hat{\boldsymbol{n}}_{\hat{K}}\, \hat{\varphi}_{j'} - \frac{\tilde{\sigma}}{\sqrt{2}}\int_{\hat{e}_m} \hat{\varphi}_j \circ \boldsymbol{\vartheta}_m \hat{\varphi}_{j'}.$$

Therefore, the matrix $\mathbf{F}$ only contains the following four types of non-zero off-diagonal subblocks, namely when $i' \neq i$ and $K_{i'} \cap K_i \neq \emptyset$,

$$\mathbf{F}^0_{i'i} = \begin{pmatrix} -\frac{1}{4} & \frac{1}{4} - \frac{\tilde{\sigma}}{2\sqrt{2}} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{4} & \frac{1}{4} - \frac{\tilde{\sigma}}{2\sqrt{2}} \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad \mathbf{F}^1_{i'i} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \frac{1}{4} - \frac{\tilde{\sigma}}{2\sqrt{2}} & -\frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{4} - \frac{\tilde{\sigma}}{2\sqrt{2}} & -\frac{1}{4} \end{pmatrix},$$

$$\mathbf{F}^2_{i'i} = \begin{pmatrix} -\frac{1}{4} & 0 & \frac{1}{4} - \frac{\tilde{\sigma}}{2\sqrt{2}} & 0 \\ 0 & -\frac{1}{4} & 0 & \frac{1}{4} - \frac{\tilde{\sigma}}{2\sqrt{2}} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad \mathbf{F}^3_{i'i} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \frac{1}{4} - \frac{\tilde{\sigma}}{2\sqrt{2}} & 0 & -\frac{1}{4} & 0 \\ 0 & \frac{1}{4} - \frac{\tilde{\sigma}}{2\sqrt{2}} & 0 & -\frac{1}{4} \end{pmatrix}.$$

Obviously, when the penalty parameter $\tilde{\sigma} > \frac{\sqrt{2}}{2}$, the diagonal entries of $\mathbf{A}_{\mathfrak{D}}$ are positive. All the off-diagonal entries of $\mathbf{A}_{\mathfrak{D}}$ are non-positive. The row sum of $\mathbf{A}_{\mathfrak{D}}$ equals zero. In addition, since the Lagrange bases are numerically orthogonal with respect to the Gauss–Lobatto quadrature, the mass matrix is diagonal with positive diagonal entries $[\mathbf{A}_{\mathcal{M}}]_{ij;ij} = \Delta x^2 \hat{\omega}_j \rho^{\mathrm{P}}_{ij}$. Thus the row sum of matrix $\mathbf{A}_{\mathcal{M}} + \frac{\Delta t \lambda}{\mathrm{Re}} \mathbf{A}_{\mathfrak{D}}$ is positive. Above all, by Lemma 4, the system matrix $\mathbf{A}_{\mathcal{M}} + \frac{\Delta t \lambda}{\mathrm{Re}} \mathbf{A}_{\mathfrak{D}}$ is a non-singular M-matrix, therefore is monotone. Here, we highlight our system matrix holds the M-matrix structure unconditionally.

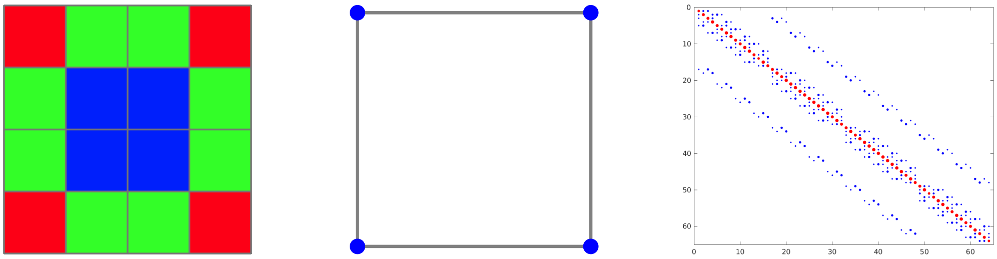

Figure A.12: A schematic graph of the domain partition, quadrature, and the M-matrix structure of $\mathbf{A}_{\mathfrak{D}}$. Left: a 4×4 mesh of domain $[0,1]^2$. The cells with zero, one, and two boundary faces are marked in blue, green, and red. Middle: 2^2-point Gauss–Lobatto quadrature used in $\mathbb{Q}^1$ scheme for computing integrals in parabolic subproblem. Right: sparsity pattern of $\mathbf{A}_{\mathfrak{D}}$ associated with a 4×4 mesh of the domain $[0,1]^2$. The positive and negative entries are plotted by red and blue dots.

References

[1] 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.

[2] N.-A. Lai, C. Liu, A. Tarfulea, Positivity of temperature for some non-isothermal fluid models, Journal of Differential Equations 339 (2022) 555–578. doi:https://doi.org/10.1016/j.jde.2022.08.025.
URL `https://www.sciencedirect.com/science/article/pii/S0022039622005034`

[3] C. Liu, J.-E. Sulzbach, The brinkman-fourier system with ideal gas equilibrium, Discrete and Continuous Dynamical Systems 42 (1) (2021) 425–462.

[4] D. Grapsas, R. Herbin, W. Kheriji, J.-C. Latché, An unconditionally stable staggered pressure correction scheme for the compressible Navier–Stokes equations, The SMAI journal of computational mathematics 2 (2016) 51–97.

[5] X. Zhang, On positivity-preserving high order discontinuous Galerkin schemes for compressible Navier–Stokes equations, Journal of Computational Physics 328 (2017) 301–343.

[6] 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.

[7] 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.

[8] C. Fan, X. Zhang, J. Qiu, Positivity-preserving high order finite volume hybrid Hermite WENO schemes for compressible Navier-Stokes equations, Journal of Computational Physics 445 (2021) 110596.

[9] C. Fan, X. Zhang, J. Qiu, Positivity-preserving high order finite difference WENO schemes for compressible Navier-Stokes equations, Journal of Computational Physics 467 (2022) 111446.

[10] J.-L. Guermond, M. Maier, B. Popov, I. Tomas, Second-order invariant domain preserving approximation of the compressible Navier–Stokes equations, Computer Methods in Applied Mechanics and Engineering 375 (2021) 113608.

[11] L. Demkowicz, J. Oden, W. Rachowicz, A new finite element method for solving compressible Navier–Stokes equations based on an operator splitting method and hp adaptivity, Computer methods in applied mechanics and engineering 84 (3) (1990) 275–326.

[12] X. Zhang, Y. Liu, C.-W. Shu, Maximum-principle-satisfying high order finite volume weighted essentially nonoscillatory schemes for convection-diffusion equations, SIAM Journal on Scientific Computing 34 (2) (2012) A627–A658.

[13] J. Yan, Maximum principle satisfying direct discontinuous Galerkin method and its variation for convection diffusion equations, Mathematics of Computation (2014).

[14] 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.

[15] S. Srinivasan, J. Poggie, X. Zhang, A positivity-preserving high order discontinuous Galerkin scheme for convection–diffusion equations, Journal of Computational Physics 366 (2018) 120–143.

[16] Z. Sun, J. A. Carrillo, C.-W. Shu, A discontinuous Galerkin method for nonlinear parabolic equations and gradient flow problems with interaction potentials, Journal of Computational Physics 352 (2018) 76–104.

[17] H. Li, S. Xie, X. Zhang, A high order accurate bound-preserving compact finite difference scheme for scalar convection diffusion equations, SIAM Journal on Numerical Analysis 56 (6) (2018) 3308–3345.

[18] R. J. Plemmons, M-matrix characterizations. I–nonsingular M-matrices, Linear Algebra and its Applications 18 (2) (1977) 175–188.

[19] W. Höhn, H. D. Mittelmann, Some remarks on the discrete maximum-principle for finite elements of higher order, Computing 27 (2) (1981) 145–154.

[20] H. Li, X. Zhang, On the monotonicity and discrete maximum principle of the finite difference implementation of C^0-Q^2 finite element method, Numerische Mathematik 145 (2) (2020) 437–472.

[21] L. Cross, X. Zhang, On the monotonicity of high order discrete Laplacian, arXiv preprint arXiv:2010.07282 (2020).

[22] 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.

[23] B. Cockburn, C.-W. Shu, The local discontinuous Galerkin method for time-dependent convection-diffusion systems, SIAM journal on numerical analysis 35 (6) (1998) 2440–2463.

[24] P. Castillo, B. Cockburn, I. Perugia, D. Schötzau, An a priori error analysis of the local discontinuous Galerkin method for elliptic problems, SIAM Journal on Numerical Analysis 38 (5) (2000) 1676–1706.

[25] J. Peraire, P.-O. Persson, The compact discontinuous Galerkin (CDG) method for elliptic problems, SIAM Journal on Scientific Computing 30 (4) (2008) 1806–1824.

[26] A. Uranga, P.-O. Persson, M. Drela, J. Peraire, Implicit large eddy simulation of transitional flows over airfoils and wings, in: 19th AIAA Computational Fluid Dynamics, American Institute of Aeronautics and Astronautics, Inc., 2009, p. 4131.

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

[28] 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.

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

[30] B. Cockburn, B. Dong, J. Guzmán, M. Restelli, R. Sacco, A hybridizable discontinuous Galerkin method for steady-state convection-diffusion-reaction problems, SIAM Journal on Scientific Computing 31 (5) (2009) 3827–3846.

[31] J. Peraire, N. Nguyen, B. Cockburn, A hybridizable discontinuous Galerkin method for the compressible Euler and Navier–Stokes equations, in: 48th AIAA aerospace sciences meeting including the new horizons forum and aerospace exposition, 2010, p. 363.

[32] N. C. Nguyen, J. Peraire, B. Cockburn, An implicit high-order hybridizable discontinuous Galerkin method for the incompressible Navier–Stokes equations, Journal of Computational Physics 230 (4) (2011) 1147–1170.

[33] B. Riviere, Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations: Theory and Implementation, Frontiers in Applied Mathematics, Society for Industrial and Applied Mathematics, 2008.

[34] B. Rivière, M. F. Wheeler, V. Girault, Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems. Part I, Computational Geosciences 3 (3) (1999) 337–360.

[35] B. Rivière, M. F. Wheeler, V. Girault, A priori error estimates for finite element methods based on discontinuous approximation spaces for elliptic problems, SIAM Journal on Numerical Analysis 39 (3) (2001) 902–931.

[36] R. Masri, C. Liu, B. Riviere, A discontinuous Galerkin pressure correction scheme for the incompressible Navier–Stokes equations: Stability and convergence, Mathematics of Computation 91 (336) (2022) 1625–1654.

[37] J. Wang, X. Ye, A weak Galerkin finite element method for second-order elliptic problems, Journal of Computational and Applied Mathematics 241 (2013) 103–115.

[38] J. Wang, X. Ye, A weak Galerkin finite element method for the Stokes equations, Advances in Computational Mathematics 42 (1) (2016) 155–174.

[39] D. N. Arnold, F. Brezzi, B. Cockburn, L. D. Marini, Unified analysis of discontinuous Galerkin methods for elliptic problems, SIAM journal on numerical analysis 39 (5) (2002) 1749–1779.

[40] C.-W. Shu, Discontinuous Galerkin method for time-dependent problems: survey and recent developments, Recent devel-

opments in discontinuous Galerkin finite element methods for partial differential equations (2014) 25–62.

[41] J. Hu, X. Zhang, Positivity-preserving and energy-dissipative finite difference schemes for the Fokker–Planck and Keller–Segel equations, IMA Journal of Numerical AnalysisDrac014 (05 2022). doi:10.1093/imanum/drac014.

[42] J. Shen, X. Zhang, Discrete maximum principle of a high order finite difference scheme for a generalized Allen–Cahn equation, Communications in Mathematical Sciences 20 (5) (2022) 1409–1436.

[43] C. Liu, Y. Gao, X. Zhang, Structure preserving schemes for Fokker-Planck equations of irreversible processes, arXiv preprint arXiv:2210.16628 (2022).

[44] C.-W. Shu, Total-variation-diminishing time discretizations, SIAM Journal on Scientific and Statistical Computing 9 (6) (1988) 1073–1084.

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

[46] 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.

[47] Y. Xing, X. Zhang, C.-W. Shu, Positivity-preserving high order well-balanced discontinuous Galerkin methods for the shallow water equations, Advances in Water Resources 33 (12) (2010) 1476–1493.

[48] 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.

[49] Z. Xu, X. Zhang, Bound-preserving high-order schemes, in: Handbook of Numerical Analysis, Vol. 18, Elsevier, 2017, pp. 81–102.

[50] Y. Maday, E. M. Rønquist, Optimal error analysis of spectral methods with emphasis on non-constant coefficients and deformed geometries, Computer Methods in Applied Mechanics and Engineering 80 (1-3) (1990) 91–115.

[51] J. Qiu, C.-W. Shu, Runge–Kutta discontinuous Galerkin method using WENO limiters, SIAM Journal on Scientific Computing 26 (3) (2005) 907–929.

[52] 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.

[53] 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.

[54] P. Woodward, P. Colella, The numerical simulation of two-dimensional fluid flow with strong shocks, Journal of computational physics 54 (1) (1984) 115–173.

[55] 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.

[56] T. L. Horváth, M. E. Mincsovics, Discrete maximum principle for interior penalty discontinuous Galerkin methods, Central European Journal of Mathematics 11 (4) (2013) 664–679.