

# HIGH ORDER NUMERICAL METHODS PRESERVING INVARIANT DOMAIN FOR HYPERBOLIC AND RELATED SYSTEMS \*

KAILIANG WU <sup>†</sup>, XIANGXIONG ZHANG<sup>‡</sup>, AND CHI-WANG SHU <sup>§</sup>

**Abstract.** Admissible states in hyperbolic systems and related equations often form a convex invariant domain. Numerical violations of this domain can lead to loss of hyperbolicity, resulting in ill-posedness and severe numerical instabilities. It is therefore crucial for numerical schemes to preserve the invariant domain to ensure both physically meaningful solutions and robust computations. For complex systems, constructing invariant-domain-preserving (IDP) schemes is highly nontrivial and particularly challenging for high-order accurate methods. This paper presents a comprehensive survey of IDP schemes for hyperbolic and related systems, with a focus on the most popular approaches for constructing provable IDP schemes. We first give a systematic review of the fundamental approaches for establishing the IDP property in first-order accurate schemes, covering finite difference, finite volume, finite element, and residual distribution methods. Then we focus on two widely used and actively developed classes of high order IDP schemes as well as their recent developments, most of which have emerged in the past decade. The first class of methods seeks an intrinsic weak IDP property in high-order schemes and then designs polynomial limiters to enforce a strong IDP property at the points of interest. This generic approach applies to high-order finite volume and discontinuous Galerkin schemes. The second class is based on the flux limiting approaches, which originated from the flux-corrected transport method and can be adapted to a broader range of spatial discretizations, including finite difference and continuous finite element methods. In this survey, we elucidate the main ideas and underlying principles that guide the construction of IDP schemes, unify several existing IDP analysis techniques and theories, and provide some new perspectives and insights on the existing approaches. We also illustrate these approaches through extensive examples, such as positivity-preserving schemes for the gas dynamics equations, and present numerical experiments drawn from interdisciplinary applications in gas dynamics and magnetohydrodynamics.

**Key words.** Convex invariant domains, positivity-preserving, bound-preserving, hyperbolic conservation laws, high order accurate schemes, invariant-domain-preserving limiters, gas dynamics, magnetohydrodynamics

**MSC codes.** 65M06, 65M08, 65M60, 65M12, 76N15, 35L65

## 1. Introduction.

**1.1. Motivation.** Consider the initial value problem of a time-dependent PDE system of  $N$  equations in  $d$  spatial dimensions,

$$(1.1) \quad \partial_t \mathbf{u} + \mathcal{L}(\mathbf{u}) = \mathbf{0}, \quad \mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d, \quad \mathbf{u} = \mathbf{u}(\mathbf{x}, t) \in \mathbb{R}^N,$$

defined in a bounded domain with appropriate boundary conditions, where  $\mathcal{L}$  denotes the spatial differential operator. A set  $G \subset \mathbb{R}^N$  is called an *invariant domain* of (1.1) if well posed solutions to (1.1) satisfy  $\mathbf{u}(\mathbf{x}, t) \in G$  for all  $\mathbf{x}$  and  $t > 0$  as long as  $\mathbf{u}(\mathbf{x}, 0) \in G$  for all  $\mathbf{x}$ . In many problems,  $G$  can be represented by the positivity or

---

\*Submitted on May 1, 2025; accepted for publication (in revised form) December 8, 2025.

**Funding:** Kailiang Wu is partially supported by Science Challenge Project (No. TZ2025007), Shenzhen Science and Technology Program (Grant No. RCJC20221008092757098), and National Natural Science Foundation of China (Grant No. 12171227). Xiangxiong Zhang is partially supported by NSF grant DMS-2208518 and Purdue University Seed Funding for review papers. Chi-Wang Shu is partially supported by NSF grant DMS-2309249.

<sup>†</sup>Department of Mathematics, Southern University of Science and Technology, Shenzhen, Guangdong 518055, China ([WUKL@sustech.edu.cn](mailto:WUKL@sustech.edu.cn)).

<sup>‡</sup>Department of Mathematics, Purdue University, West Lafayette, IN 47907, USA ([zhan1966@purdue.edu](mailto:zhan1966@purdue.edu)).

<sup>§</sup>Division of Applied Mathematics, Brown University Providence, RI 02906, USA ([Chi-Wang.Shu@brown.edu](mailto:Chi-Wang.Shu@brown.edu)).

non-negativity of several functions of  $\mathbf{u}$ :

$$(1.2) \quad G = \left\{ \mathbf{u} \in \mathbb{R}^N : g_i(\mathbf{u}) > 0 \quad \forall i \in \mathbb{I}, \quad g_i(\mathbf{u}) \geq 0 \quad \forall i \in \widehat{\mathbb{I}} \right\}.$$

For a given PDE, its invariant domain is not necessarily convex, but the invariant domain used in stabilizing many interesting and important hyperbolic systems is often a convex set. In this paper, we only consider convex invariant domains. Such convex invariant domains are typically found in hyperbolic conservation laws

$$(1.3) \quad \partial_t \mathbf{u} + \nabla \cdot \mathbf{f}(\mathbf{u}) = \mathbf{0},$$

as well as in other time-dependent PDEs, such as convection-diffusion equations [55], semilinear parabolic equations [73], and reaction-diffusion equations [80, 196], etc.

For example, for gas dynamics equations consisting of conservation of mass, momentum and total energy, the set of admissible states is defined by positive density and internal energy (or pressure for many common equations of state). Such a set of admissible states is needed for not only physically meaningful solutions but also maintaining the hyperbolicity of the governing equations, which will be reviewed in [Subsection 2.1](#). More importantly, numerically preserving positivity of density and pressure is critical for stabilizing computations of challenging problems such as high speed flows, especially for high order numerical schemes. Preserving an invariant domain defined by positivity of density and pressure in a conservative scheme can ensure  $L^1$  stability of mass and total energy [174].

For a generic hyperbolic problem, the set of admissible states such as positive water height in shallow water equations, usually defines a convex invariant domain, preserving which renders numerical schemes more robust. There are many such examples in applications including but not limited to weather modeling [168], radiative transfer [41, 170, 42], kinetic equations [50, 20], neutrino transport in core-collapse supernovae [165], relativistic hydrodynamics [182, 223, 179, 225, 214, 221], hydraulic engineering [227], astrophysics [125, 127], chemically reactive flows [161], etc.

**1.2. Scalar conservation laws and order barriers.** As a simplified model problem, consider a scalar conservation law  $\partial_t u + \nabla \cdot \mathbf{f}(u) = 0$ , whose entropy solution  $u(\mathbf{x}, t)$  is total-variation-diminishing (TVD) and satisfies the maximum principle  $\min_{\mathbf{x}} u(\mathbf{x}, t) \leq u(\mathbf{x}, s) \leq \max_{\mathbf{x}} u(\mathbf{x}, t)$  for any  $s > t \geq 0$ , implying a bound-preserving property  $U_{\min} := \min_{\mathbf{x} \in \Omega} u_0(\mathbf{x}) \leq u(\mathbf{x}, t) \leq \max_{\mathbf{x}} u_0(\mathbf{x}) =: U_{\max}$ . Hence, the scalar conservation law admits a convex invariant domain  $G = [U_{\min}, U_{\max}]$ .

For scalar conservation laws, there is extensive literature on enforcing stronger properties, such as the TVD property and monotonicity, which go beyond merely preserving the invariant domain  $G$ . Consider the one-dimensional scalar equation  $\partial_t u + \partial_x f(u) = 0$ . A numerical scheme of the form  $u_j^{n+1} = H(u_{j-k}^n, u_{j-k+1}^n, \dots, u_{j+k}^n)$  is called *monotone* if the function  $H$  is monotone increasing (non-decreasing) with respect to each of its arguments. This definition naturally extends to multi-dimensional problems and to convection-diffusion equations: the numerical solution at the next time level must be a monotone increasing function of each of its arguments within the stencil at the current time level.

It is well known that such a monotone scheme is at most first order accurate [88, 112]. There is a common misconception that a monotone scheme never produces new extrema. A monotone scheme is also TVD thus *monotonicity-preserving*:

$$\text{monotone} \Rightarrow \text{TVD} \Rightarrow \text{monotonicity-preserving},$$

none of which, however, implies that no new extrema can be generated. The definition of *monotonicity-preserving* is that if  $u_{j+1}^n \geq u_j^n$  for all  $j$ , then  $u_{j+1}^{n+1} \geq u_j^{n+1}$  for all  $j$ . Thus a monotone scheme cannot generate new extrema on an infinite domain if the initial condition is a monotone profile. For instance, the Lax-Friedrichs scheme (a.k.a. Rusanov scheme) is monotone, and one can easily construct a non-monotone initial condition for which the Lax-Friedrichs scheme produces new extrema [40].

The order barrier that a monotone scheme can only be first order accurate is also called *Godunov Theorem*. There are similar order barriers stated in the literature. For example, a TVD finite difference scheme satisfying  $\sum_j |u_{j+1}^{n+1} - u_j^{n+1}| \leq \sum_j |u_{j+1}^n - u_j^n|$  can be at most first order accurate in two dimensions [89]. Similarly, if seeking maximum principle in the form  $\min_j u_j^n \leq u_j^{n+1} \leq \max_j u_j^n$  in a finite difference or finite volume scheme, then such a scheme can be most second order accurate in spatial truncation error analysis, see [247] for a simple counterexample accredited to Harten. Central schemes [118, 129] satisfy such maximum principle and achieve second order accuracy.

To enforce the maximum principle or the TVD property in high order finite volume and finite difference schemes, various limiters can be designed, which however causes loss of high order accuracy near extrema due to the above-mentioned order barriers. Nonetheless, such schemes can still achieve high-order accuracy for smooth solutions that are monotonically increasing or decreasing, while delivering good resolution of discontinuities.

**1.3. First order schemes for systems.** Consider a one-dimensional hyperbolic system  $\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = \mathbf{0}$  as an example, then a 3-point-stencil first order *locally conservative* scheme can be written as

$$(1.4) \quad \mathbf{u}_j^{n+1} = \mathbf{u}_j^n - \lambda \left( \hat{\mathbf{f}}(\mathbf{u}_j^n, \mathbf{u}_{j+1}^n) - \hat{\mathbf{f}}(\mathbf{u}_{j-1}^n, \mathbf{u}_j^n) \right), \quad \lambda = \frac{\Delta t}{\Delta x},$$

where  $\mathbf{u}_j^n$  denotes the solution at a grid point  $x_j$  and  $n$ -th time step,  $\hat{\mathbf{f}}(\cdot, \cdot)$  denotes the numerical flux function, and  $\Delta t$  and  $\Delta x$  are the temporal and spatial step-sizes. Stability properties of scalar equations such as TVD and maximum principle do not directly extensible to systems of conservation laws. For instance, in a blast wave solution to gas dynamics equations, both the total variation and upper bound of density become much larger than their initial values (see Figure 7). Instead, first order monotone schemes like the Godunov scheme and the Lax-Friedrichs scheme (a.k.a. Rusanov scheme) for scalar conservation laws can be shown invariant-domain-preserving (IDP) for hyperbolic systems, which will be reviewed in Section 3.

For a given hyperbolic system, it is usually difficult to show that the exact solution preserves a given invariant domain from PDE analysis. If numerical solutions to a locally conservative scheme converge in the sense of bounded variation when refining the meshes, then the limit must be a weak solution by the Lax-Wendroff Theorem [142]. Schemes like the Godunov scheme also satisfy the entropy inequality, thus the converged limit is also an entropy solution. Therefore, numerical solutions of the first order Godunov scheme can be used to as a numerical evidence for whether the hyperbolic system should preserve a particular chosen convex domain [113, 114]. If  $\mathbf{u}_j^n$  produced by a first order IDP scheme (1.4) converges as mesh refines, then it is a strong numerical evidence that at least one exact solution to the PDE should preserve the same invariant domain.

**1.4. High order schemes for systems.** Although almost all desired stability properties can be proven for a first-order scheme (1.4) using either the Godunov flux

or the Lax–Friedrichs flux, their numerical resolution of fine structures is often unsatisfactory. High-resolution and high-order schemes are preferred for achieving better resolution. High-order schemes are typically defined as those that are at least third-order accurate for smooth solutions. Even though any Eulerian scheme on a uniform mesh can be at most half-order accurate in the  $L^2$ -norm for discontinuous solutions [143, §11.2], high-order schemes such as WENO (weighted essentially non-oscillatory) schemes [195] or discontinuous Galerkin (DG) methods [56] generally produce better resolution than low-order schemes for discontinuities such as shocks, due to the reduced artificial viscosity present in high-order schemes.

However, popular and practical high order accurate finite difference, finite volume, and finite element schemes are not robust for very challenging hyperbolic problems (e.g., high speed flows involving low density and pressure), often due to violation of the invariant domains (e.g., loss of positivity of density or pressure in compressible flows). With proper modifications or limiters, high order schemes can be rendered to preserve the invariant domain of admissible states, which may improve their robustness.

We emphasize that the order barriers such as the Godunov Theorem hold only in the sense as stated in Section 1.2. Under different definitions of discrete total variation or maximum, it is possible to avoid these order barriers. For example, if defining the total variation as the total variation of the piecewise polynomials reconstructed in a finite volume scheme, a very high order accurate finite volume TVD scheme can be constructed for scalar equations in one dimension [244]. If seeking a scheme preserving a simple invariant domain  $U_{\min} \leq u_j^n \leq U_{\max}$  instead of a strict maximum principle  $\min_j u_j^n \leq u_j^{n+1} \leq \max_j u_j^n$  for scalar equations, then in general it is still possible to achieve high order accuracy for a smooth solution. For a properly defined invariant domain (1.2), there should be no order barriers for systems.

For the sake of stabilizing high order Eulerian schemes without adding excessive artificial viscosity, we should consider a method that is high order accurate (at least for smooth solutions), conservative, and IDP. For a special scheme solving a special system, there might be many different ways to construct such a method. For general problems, there are two popular and flexible approaches which are described briefly as follows.

Take a high order accurate finite volume scheme on an interval  $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$  as an example. With forward Euler time stepping, it can be written as

$$(1.5) \quad \bar{\mathbf{u}}_j^{n+1} = \bar{\mathbf{u}}_j^n - \lambda \left( \hat{\mathbf{f}}(\mathbf{u}_{j+\frac{1}{2}}^-, \mathbf{u}_{j+\frac{1}{2}}^+) - \hat{\mathbf{f}}(\mathbf{u}_{j-\frac{1}{2}}^-, \mathbf{u}_{j-\frac{1}{2}}^+) \right), \quad \lambda = \frac{\Delta t}{\Delta x},$$

where  $\bar{\mathbf{u}}_j$  denotes the cell average,  $\hat{\mathbf{f}}$  is a numerical flux function and  $\mathbf{u}_{j+\frac{1}{2}}^-, \mathbf{u}_{j+\frac{1}{2}}^+$  are reconstructed values at  $x_{j+\frac{1}{2}}$  from the left and from the right, respectively. Let  $\hat{\mathbf{f}}_{j+\frac{1}{2}} = \hat{\mathbf{f}}(\mathbf{u}_{j+\frac{1}{2}}^-, \mathbf{u}_{j+\frac{1}{2}}^+)$  denote the numerical flux.

The first popular and easy-to-use approach is to consider a modification or limiting of the numerical fluxes  $\hat{\mathbf{f}}_{j\pm\frac{1}{2}}$  so that  $\bar{\mathbf{u}}_j^{n+1}$  is in the invariant domain, and the idea of flux limiting traces back to the seminal work of the flux corrected transport (FCT) by Boris and Book [35, 34, 36] and also Zalesak [239].

The second popular approach is to modify only the reconstructed values  $\mathbf{u}_{j\pm\frac{1}{2}}^\pm$  so that  $\bar{\mathbf{u}}_j^{n+1}$  is in the invariant domain under a suitable time step, and such an approach has been popularized by Zhang and Shu [246], which builds upon the idea by Perthame and Shu in [176] and the simple polynomial limiter analyzed in [155]. By the Godunov Theorem, a high order scheme like (1.5) cannot be monotone for solving

scalar equations, but it can still be *weakly monotone* (Theorem 4.1 in Section 4), which is the key of such an approach.

**1.5. Scope and organization of this paper.** The preservation of invariant domain is merely a partial characterization of nonlinear stability, which is not sufficient for convergence. For convergence to entropy solutions, discrete entropy inequality should also be considered. Moreover, more properties might also be desired such as energy stability, and well-balancedness for shallow water equations. It is possible to combine discussions of other properties with the IDP property. For simplicity, we focus on only how to preserve a convex invariant domain, and we do not discuss boundary conditions. We only discuss the numerical scheme in the interior of the domain, e.g., assuming periodic boundary conditions on a rectangular domain or zero inflow boundary conditions. For the organization of the rest of this paper, we first list some representative examples of invariant domains in Section 2. In Section 3, we discuss how to show IDP in classical first order schemes, on which IDP techniques in high order schemes depend heavily. In Section 4 and Section 5, we review two popular approaches for enforcing invariant domain in high order schemes, which can be used for many hyperbolic systems such as gas dynamics and shallow water equations. In Section 6, we briefly discuss other approaches and extensions, then survey recent breakthroughs and developments for the much more challenging MHD systems. The concluding remarks will be given in Section 7.

**2. Examples of invariant domains.** For a system  $\partial_t \mathbf{u} + \partial_x \mathbf{f}(\mathbf{u}) = \mathbf{0}$ , it is hyperbolic if the Jacobian matrix  $\mathbf{f}'(\mathbf{u})$  has real eigenvalues and a complete set of eigenvectors, from which usually an invariant domain like (1.2) can be defined. For nonlinear hyperbolic systems in multiple dimensions, it is in general difficult to prove that the exact solutions satisfy  $\mathbf{u}(\mathbf{x}, t) \in G$ , but violation of such an invariant domain usually causes blow-ups in numerical computation due to loss of hyperbolicity. Next, we list some examples of (1.2).

### 2.1. Gas dynamics equations.

EXAMPLE 2.1 (Positivity density and pressure for ideal gas). *Consider the compressible Euler equations, which can be written in the form of (1.3):*

$$(2.1a) \quad \partial_t \begin{pmatrix} \rho \\ \mathbf{m} \\ E \end{pmatrix} + \nabla \cdot \begin{pmatrix} \mathbf{m} \\ \rho^{-1} \mathbf{m} \otimes \mathbf{m} + p \mathbf{I} \\ \rho^{-1} (E + p) \mathbf{m} \end{pmatrix} = \mathbf{0}, \quad \mathbf{x} \in \mathbb{R}^d, \quad d = 1, 2, 3$$

where  $\rho$  is the density,  $\mathbf{m}$  denotes the momentum vector,  $p$  is the pressure, and  $E = \rho e + \frac{|\mathbf{m}|^2}{2\rho}$  is the total energy with  $e$  being the specific internal energy. The system (2.1) is closed with an equation of state (EOS), e.g., the EOS for ideal gas is

$$(2.1b) \quad p = (\gamma - 1)\rho e,$$

where  $\gamma > 1$  is the ratio of specific heats. The commonly considered invariant domain for (2.1) is defined by positive density and positive pressure

$$(2.2) \quad G = \left\{ \mathbf{u} = (\rho, \mathbf{m}, E)^\top \in \mathbb{R}^{d+2} : \rho > 0, p = (\gamma - 1)(E - \frac{|\mathbf{m}|^2}{2\rho}) > 0 \right\}.$$

The pressure function  $p(\mathbf{u})$  is concave in  $\mathbf{u}$ , which can be verified via its Hessian matrix. Thus  $p$  satisfies the Jensen's inequality

$$(2.3) \quad p(\lambda \mathbf{u}_1 + (1 - \lambda) \mathbf{u}_2) \geq \lambda p(\mathbf{u}_1) + (1 - \lambda) p(\mathbf{u}_2) \quad \forall \lambda \in (0, 1), \quad \forall \mathbf{u}_1, \mathbf{u}_2 \in G,$$

which implies the convexity of  $G$ .

Negative density or pressure leads not only to physically meaningless solutions but also to loss of hyperbolicity, since the local sound speed  $\sqrt{\frac{\gamma p}{\rho}}$  becomes imaginary and the eigenvalues of the Jacobian matrix of system (2.1) become complex. In practice, negative density or pressure almost always cause significant numerical instabilities when solving (2.1), for example in simulations of high-speed flows (see Figure 8). Therefore, the set (2.2) is also known as the *set of admissible states*.

EXAMPLE 2.2 (Invariant domain with minimum entropy principle). *For compressible Euler equations with ideal gas EOS (2.1), one may also consider adding the minimum entropy principle [201, 124],  $S(\mathbf{u}) \geq S_{\min} := \min_{\mathbf{x}} S(\mathbf{u}_0(\mathbf{x}))$ , for the specific entropy  $S = \ln \frac{p}{\rho^\gamma}$ , which is not a concave function but instead a quasi-concave function [248, Lemma 2.1] thus satisfies*

$$(2.4) \quad S(\lambda \mathbf{u}_1 + (1 - \lambda) \mathbf{u}_2) \geq \min\{S(\mathbf{u}_1), S(\mathbf{u}_2)\} \quad \forall \lambda \in (0, 1), \quad \forall \mathbf{u}_1, \mathbf{u}_2 \in G.$$

One can obtain another convex invariant domain:

$$(2.5) \quad G_S = \left\{ \mathbf{u} = (\rho, \mathbf{m}, E)^\top \in \mathbb{R}^{d+2} : \rho > 0, p > 0, S = \ln \frac{p}{\rho^\gamma} \geq S_{\min} \right\}.$$

EXAMPLE 2.3 (Invariant domain for any EOS). *The pressure function may no longer be a concave function in a general EOS, for which the internal energy  $pe = E - \frac{|\mathbf{m}|^2}{2\rho}$  is always concave for  $\rho > 0$ . The invariant domain defined by positivity of internal energy can be considered for both compressible Euler and Navier–Stokes equations [92, 242, 95]:*

$$(2.6) \quad G = \left\{ \mathbf{u} = (\rho, \mathbf{m}, E)^\top \in \mathbb{R}^{d+2} : \rho > 0, pe = E - \frac{|\mathbf{m}|^2}{2\rho} > 0 \right\}.$$

**2.2. Single layer shallow water equations.** The single layer shallow water equations with a bottom topography, which has been used to model flows in rivers and coastal areas for ocean and hydraulic engineering, can be written as

$$(2.7) \quad \partial_t \begin{pmatrix} h \\ \mathbf{m} \end{pmatrix} + \nabla \cdot \left( h^{-1} \mathbf{m} \otimes \mathbf{m} + \frac{1}{2} gh \mathbf{I} \right) = \begin{pmatrix} 0 \\ -gh \nabla b \end{pmatrix}, \quad \mathbf{x} \in \mathbb{R}^2,$$

where  $h$  is water height,  $\mathbf{m} = h(u, v)^\top$  is the momentum,  $b(\mathbf{x})$  is the bottom topography function, and  $g$  is the gravity constant. The eigenvalues of the Jacobian are related to  $\sqrt{gh}$ , thus the non-negativity of the water height function defines a convex invariant domain, which is crucial for numerical stability [28, 45, 128].

**2.3. Two-layer shallow water equations.** Consider the more complicated two-layer shallow water equations, which are widely used in the study of stratified flow motions such as salinity-driven exchange flow motions and layered flows. In one dimension, the equations for conservative variables can be written as

$$(2.8) \quad \partial_t \begin{pmatrix} h_1 \\ h_1 u_1 \\ h_2 \\ h_2 u_2 \end{pmatrix} + \partial_x \begin{pmatrix} h_1 u_1 \\ h_1 u_1^2 + \frac{1}{2} gh_1^2 \\ h_2 u_2 \\ h_2 u_2^2 + \frac{1}{2} gh_2^2 \end{pmatrix} = \begin{pmatrix} 0 \\ -gh_1(h_2 + b)_x \\ 0 \\ -rgh_2(h_1)_x - gh_2 b_x \end{pmatrix},$$

where the subscripts 1 and 2 denote the upper and the lower layers of the water respectively,  $h_i$  is the height of the  $i$ th layer,  $u_i$  denotes the velocity of the  $i$ th layer,  $b$  is the bottom topography, and  $r = \rho_1/\rho_2 \in (0, 1)$  is the ratio of density.

By using the expansion-based first order approximation [162], the eigenvalues of the Jacobian can be given as the external wavespeeds and internal wavespeeds

$$(2.9) \quad \lambda_{ext}^{\pm} \approx \frac{h_1 u_1 + h_2 u_2}{h_1 + h_2} \pm \sqrt{g(h_1 + h_2)},$$

$$(2.10) \quad \lambda_{int}^{\pm} \approx \frac{h_1 u_2 + h_2 u_1}{h_1 + h_2} \pm \sqrt{g' \frac{h_1 h_2}{h_1 + h_2} \left[ 1 - \frac{(u_1 - u_2)^2}{g'(h_1 + h_2)} \right]}, \quad g' = g(1 - r) > 0.$$

One can first consider an approximated invariant domain defined to ensure the approximated eigenvalues to be real numbers:

$$(2.11) \quad G = \left\{ (h_1, h_1 u_1, h_2, h_2 u_2)^{\top} : h_1 > 0, h_2 > 0, \frac{(u_1 - u_2)^2}{h_1 + h_2} \leq g' \right\},$$

which unfortunately is not a convex set because  $\frac{(u_1 - u_2)^2}{h_1 + h_2}$  is not a convex function of the conserved variables  $(h_1, h_1 u_1, h_2, h_2 u_2)^{\top}$ . However, the function  $\frac{(u_1 - u_2)^2}{h_1 + h_2}$  is a convex function of the primitive variables  $(h_1, u_1, h_2, u_2)^{\top}$ , which can be easily verified via its Hessian matrix. Thus we may consider rewriting the two-layer equations in these variables:

$$(2.12) \quad \begin{pmatrix} h_1 \\ u_1 \\ h_2 \\ u_2 \end{pmatrix}_t + \begin{pmatrix} h_1 u_1 \\ \frac{1}{2} u_1^2 + g(h_1 + h_2 + b) \\ h_2 u_2 \\ \frac{1}{2} u_2^2 + g(r h_1 + h_2 + b) \end{pmatrix}_x = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

and consider the following invariant domain for the same constraints:

$$(2.13) \quad G = \left\{ (h_1, u_1, h_2, u_2)^{\top} : h_1 > 0, h_2 > 0, \frac{(u_1 - u_2)^2}{h_1 + h_2} \leq g' \right\}.$$

Although (2.11) and (2.13) describe the same admissible states, the key difference is that only the set (2.13) is convex, which facilitates the construction of IDP schemes for the equivalent system (2.12); see [72]. However, the systems (2.8) and (2.12) are equivalent only for smooth solutions and they may produce different weak solutions. Moreover, there are some drawbacks of solving a system for non-conservative variables (2.12), e.g., it is nontrivial to obtain conservation of momentum.

**2.4. Ten-moment Gaussian closure model.** Gaussian closure models are alternatives of compressible Euler equations for modeling compressible flows with nonlocal thermodynamic equilibrium assumed, especially for extremely low pressure rarefied gas flows. The ten-moment Gaussian closure model [144, 27] in two dimensions are given by  $\partial_t \mathbf{u} + \partial_x \mathbf{f}_1(\mathbf{u}) + \partial_y \mathbf{f}_2(\mathbf{u}) = \mathbf{0}$  with

$$(2.14) \quad \mathbf{u} = \begin{pmatrix} \rho \\ m_1 \\ m_2 \\ E_{11} \\ E_{12} \\ E_{22} \end{pmatrix}, \quad \mathbf{f}_j(\mathbf{u}) = \begin{pmatrix} m_j \\ m_1 v_j + p_{1j} \\ m_2 v_j + p_{2j} \\ E_{11} v_j + p_{1j} v_1 \\ E_{12} v_j + \frac{1}{2}(p_{1j} v_2 + p_{2j} v_1) \\ E_{22} v_j + p_{2j} v_2 \end{pmatrix}, \quad j = 1, 2,$$

where  $\rho$  denotes the mass density,  $\mathbf{m} = (m_1, m_2)^\top$  is the momentum, and the velocity vector is given by  $\mathbf{v} = \mathbf{m}/\rho$ . The symmetric tensor  $\mathbf{E} = (E_{ij})$  represents the energy, and  $\mathbf{p} = (p_{ij})$  is the pressure tensor, which is also symmetric and anisotropic in nature.

The system (2.14) is closed by the relation  $\mathbf{p} = 2\mathbf{E} - \rho\mathbf{v} \otimes \mathbf{v}$ . This ensures that the model remains consistent with physical constraints. The set of admissible states is characterized by the positivity of both the density and the pressure tensor,

$$(2.15) \quad G = \left\{ \mathbf{u} \in \mathbb{R}^6 : \rho > 0, \mathbf{E} - \frac{\mathbf{m} \otimes \mathbf{m}}{2\rho} \text{ is positive-definite} \right\},$$

which is a convex set [163].

**2.5. Ideal MHD equations.** The system for ideal compressible magnetohydrodynamics (MHD) can be reformulated in conservative form as follows:

$$(2.16) \quad \partial_t \begin{pmatrix} \rho \\ \mathbf{m} \\ \mathbf{B} \\ E \end{pmatrix} + \nabla \cdot \begin{pmatrix} \mathbf{m} \\ \mathbf{m} \otimes \mathbf{v} - \mathbf{B} \otimes \mathbf{B} + p_{\text{tot}} \mathbf{I} \\ \mathbf{v} \otimes \mathbf{B} - \mathbf{B} \otimes \mathbf{v} \\ (E + p + \frac{1}{2}|\mathbf{B}|^2) \mathbf{v} - (\mathbf{v} \cdot \mathbf{B}) \mathbf{B} \end{pmatrix} = \mathbf{0},$$

where  $\rho$  denotes the mass density,  $\mathbf{m}$  is the momentum, and the velocity field is computed via  $\mathbf{v} = \mathbf{m}/\rho$ . The magnetic field is represented by  $\mathbf{B}$ , satisfying the solenoidal constraint  $\nabla \cdot \mathbf{B} = 0$  if it is satisfied at  $t = 0$ . The total energy  $E$  includes contributions from internal, kinetic, and magnetic energy,  $E = \rho e + \frac{1}{2}(\rho|\mathbf{v}|^2 + |\mathbf{B}|^2)$ , where  $e$  is the specific internal energy. Physically meaningful states must satisfy positivity of both the mass density and internal energy. This leads to a convex invariant domain:

$$(2.17) \quad G = \left\{ \mathbf{u} = (\rho, \mathbf{m}, \mathbf{B}, E)^\top \in \mathbb{R}^{2d+2} : \rho > 0, g(\mathbf{u}) := E - \frac{|\mathbf{m}|^2}{2\rho} - \frac{|\mathbf{B}|^2}{2} > 0 \right\}.$$

The positivity is crucial for stable computations [51, 67, 215].

**2.6. Relativistic MHD equations.** The relativistic magnetohydrodynamic (RMHD) system [225, 221] can be expressed in conservative form as follows:

$$(2.18) \quad \partial_t \begin{pmatrix} D \\ \mathbf{m} \\ \mathbf{B} \\ E \end{pmatrix} + \nabla \cdot \begin{pmatrix} D\mathbf{v} \\ \mathbf{m} \otimes \mathbf{v} - \mathbf{B} \otimes (W^{-2}\mathbf{B} + (\mathbf{v} \cdot \mathbf{B})\mathbf{v}) + p_{\text{tot}} \mathbf{I} \\ \mathbf{v} \otimes \mathbf{B} - \mathbf{B} \otimes \mathbf{v} \\ \mathbf{m} \end{pmatrix} = \mathbf{0}.$$

In this formulation, the relativistic mass density  $D = \rho W$ , where  $\rho$  is the rest-mass density and  $W = (1 - |\mathbf{v}|^2)^{-1/2}$  is the Lorentz factor. The momentum vector  $\mathbf{m}$  is defined as

$$\mathbf{m} = (\rho h W^2 + |\mathbf{B}|^2) \mathbf{v} - (\mathbf{v} \cdot \mathbf{B}) \mathbf{B},$$

where  $h$  is the specific enthalpy and  $\mathbf{v}$  is the velocity. The velocity is normalized such that the speed of light is unity. The total energy is given by

$$E = \rho h W^2 - p_{\text{tot}} + |\mathbf{B}|^2.$$

The magnetic field  $\mathbf{B}$  obeys the divergence-free condition  $\nabla \cdot \mathbf{B} = 0$  if it is satisfied at  $t = 0$ , as in ideal MHD. The total pressure  $p_{\text{tot}} = p + p_m$  is composed of the thermal component  $p$  and the magnetic contribution

$$p_m = \frac{1}{2} (W^{-2} |\mathbf{B}|^2 + (\mathbf{v} \cdot \mathbf{B})^2).$$

To ensure physical admissibility, the solution must lie within the invariant domain

$$(2.19) \quad G = \{ \mathbf{u} = (D, \mathbf{m}, \mathbf{B}, E)^\top \in \mathbb{R}^{2d+2} : D > 0, p(\mathbf{u}) > 0, |\mathbf{v}(\mathbf{u})| < 1 \},$$

where both  $\mathbf{v}(\mathbf{u})$  and  $p(\mathbf{u})$  are nonlinear functions of the conserved variables and cannot be explicitly written in closed form. These nonlinear implicit functions are frequently expressed in terms of another auxiliary function  $\hat{\phi}(\mathbf{u})$ , which is determined implicitly. Specifically, the expressions take the form:

$$(2.20) \quad p(\mathbf{u}) = \frac{\gamma - 1}{Z_{\mathbf{u}}^2(\hat{\phi}) \gamma} \left( \hat{\phi} - D Z_{\mathbf{u}}(\hat{\phi}) \right), \quad \mathbf{v}(\mathbf{u}) = \frac{\mathbf{m} + (\mathbf{m} \cdot \mathbf{B})\mathbf{B}/\hat{\phi}}{\hat{\phi} + |\mathbf{B}|^2}.$$

Here,  $\hat{\phi} = \hat{\phi}(\mathbf{u})$  is defined as the unique positive root of a nonlinear equation:

$$\phi - E + |\mathbf{B}|^2 - \frac{1}{2} \left( \frac{(\mathbf{m} \cdot \mathbf{B})^2}{\phi^2} + \frac{|\mathbf{B}|^2}{Z_{\mathbf{u}}^2(\phi)} \right) + \frac{\gamma - 1}{\gamma} \left( \frac{D}{Z_{\mathbf{u}}(\phi)} - \frac{\phi}{Z_{\mathbf{u}}^2(\phi)} \right) = 0,$$

where  $\gamma$  is the adiabatic index (ratio of specific heats), and  $Z_{\mathbf{u}}(\phi)$  is

$$Z_{\mathbf{u}}(\phi) := \left( \frac{\phi^2(\phi + |\mathbf{B}|^2)^2 - [\phi^2|\mathbf{m}|^2 + (2\phi + |\mathbf{B}|^2)(\mathbf{m} \cdot \mathbf{B})^2]}{\phi^2(\phi + |\mathbf{B}|^2)^2} \right)^{-1/2}.$$

**3. First order schemes preserving invariant domains.** We only consider schemes with forward Euler time stepping in this section.

**3.1. Monotone schemes for 1D scalar conservation laws.** We first consider the simplest example of enforcing bounds in solving scalar conservation laws. Consider the one-dimensional (1D) version of the scalar conservation law

$$(3.1) \quad \partial_t u + \partial_x f(u) = 0.$$

A 3-point-stencil first order monotone scheme for (3.1) can be written as

$$(3.2) \quad u_j^{n+1} = u_j^n - \lambda \left( \hat{f}(u_j^n, u_{j+1}^n) - \hat{f}(u_{j-1}^n, u_j^n) \right) =: H_\lambda(u_{j-1}^n, u_j^n, u_{j+1}^n),$$

where  $u_j^n$  denotes the numerical solution at the  $j$ -th grid point in finite difference or  $j$ -th cell in finite volume at the time level  $n$ ,  $\lambda = \Delta t / \Delta x$  with  $\Delta t$  and  $\Delta x$  denoting the temporal and spatial mesh sizes<sup>1</sup>, and  $\hat{f}(u^-, u^+)$  is a monotone flux, i.e., it is non-decreasing in the first argument  $u^-$  and non-increasing in its second argument  $u^+$ , and satisfies the consistency  $\hat{f}(u, u) = f(u)$ .

Under a suitable Courant–Friedrichs–Lewy (CFL) condition, typically of the form:

$$(3.3) \quad \lambda \alpha \leq 1 \quad \text{with} \quad \alpha := \max_u |f'(u)|,$$

the function  $H_\lambda(\cdot, \cdot, \cdot)$  is monotonically non-decreasing in all its three arguments. Note that the consistency of  $\hat{f}$  implies that  $H_\lambda(u, u, u) = u$ . Therefore, if

$$U_{\min} \leq u_{j-1}^n, \quad u_j^n, \quad u_{j+1}^n \leq U_{\max},$$

<sup>1</sup>Uniform mesh size is assumed here for simplicity in presentation, while all our discussions are extensible to non-uniform meshes.

then the monotonicity and consistency imply

$$\begin{aligned} u_j^{n+1} &= H_\lambda(u_{j-1}^n, u_j^n, u_{j+1}^n) \geq H_\lambda(U_{\min}, U_{\min}, U_{\min}) = U_{\min}, \\ u_j^{n+1} &= H_\lambda(u_{j-1}^n, u_j^n, u_{j+1}^n) \leq H_\lambda(U_{\max}, U_{\max}, U_{\max}) = U_{\max}. \end{aligned}$$

Hence, the monotone scheme (3.2) preserves the invariant domain  $G = [U_{\min}, U_{\max}]$  under the CFL condition (3.3). Examples of monotone numerical fluxes include:

- the Lax–Friedrichs, a.k.a. Rusanov flux

$$(3.4) \quad \hat{f}^{\text{LF}}(u^-, u^+) = \frac{1}{2} \left( f(u^-) + f(u^+) - \alpha(u^+ - u^-) \right);$$

- the Godunov flux

$$\hat{f}^G(u^-, u^+) = \begin{cases} \min_{u^- \leq u \leq u^+} f(u) & \text{if } u^- \leq u^+; \\ \max_{u^+ \leq u \leq u^-} f(u) & \text{if } u^- > u^+; \end{cases}$$

- the Engquist–Osher flux

$$\hat{f}^{\text{EO}}(u^-, u^+) = \frac{1}{2} \left( f(u^-) + f(u^+) - \int_{u^-}^{u^+} |f'(u)| du \right).$$

For instance, the scheme (3.2) with the Lax–Friedrichs flux (3.4), can be reformulated as

$$H_\lambda(u_{j-1}^n, u_j^n, u_{j+1}^n) = (1 - \alpha\lambda)u_j^n + \frac{1}{2}\lambda(\alpha u_{j+1}^n - f(u_{j+1}^n)) + \frac{1}{2}\lambda(\alpha u_{j-1}^n + f(u_{j-1}^n)),$$

which is non-decreasing with respect to  $u_{j-1}^n, u_j^n, u_{j+1}^n$ , thus preserves the bounds under (3.3). Although the monotone scheme is at most first order accurate (by the Godunov Theorem), it satisfies much stronger stability properties such as  $L^1$  contraction and it converges to the unique entropy solution for scalar conservation laws in multiple dimensions [58].

**REMARK 3.1.** *Although some implicit schemes can be shown to be monotone for scalar linear problems [16], the extensions to nonlinear equations can be difficult due to the algebraic nonlinear systems involved.*

**3.2. Basic assumptions for systems.** Many techniques of the first order IDP methods have been well established since 1980s, e.g., [113, 114, 201, 76, 176, 85]. For enforcing a convex invariant domain defined as (1.2), we need to make some basic assumptions about the flux function in (1.3), which are used in almost all classical IDP methods. We consider a nonlinear system (1.3) and first state basic assumptions for the 1D version:

$$(3.5) \quad \partial_t \mathbf{u} + \partial_x \mathbf{f}(\mathbf{u}) = \mathbf{0}.$$

Let  $\mathbf{U}^{\text{RP}}(x, t; \mathbf{u}_L, \mathbf{u}_R)$  be the exact solution of a Riemann problem to (3.5) with the initial data

$$(3.6) \quad \mathbf{u}_0(x) = \begin{cases} \mathbf{u}_L, & \text{if } x \leq 0, \\ \mathbf{u}_R, & \text{if } x > 0. \end{cases}$$

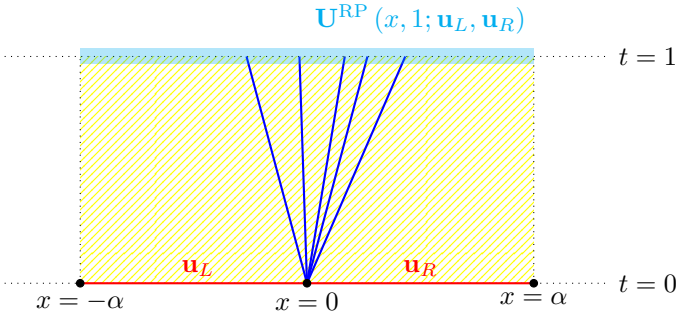
The exact solution of a Riemann problem is self-similar, i.e.,  $\mathbf{U}^{\text{RP}}(\alpha x, \alpha t; \mathbf{u}_L, \mathbf{u}_R) = \mathbf{U}^{\text{RP}}(x, t; \mathbf{u}_L, \mathbf{u}_R)$  for any constant  $\alpha > 0$ . Therefore, we only need to consider  $\mathbf{U}^{\text{RP}}(x, 1; \mathbf{u}_L, \mathbf{u}_R)$ .

ASSUMPTION 1. *The exact solution of the Riemann problem preserves the invariant domain: if  $\mathbf{u}_L, \mathbf{u}_R \in G$ , then  $\mathbf{U}^{\text{RP}}(x, t; \mathbf{u}_L, \mathbf{u}_R) \in G$  for any  $x \in \mathbb{R}$  and  $t > 0$ . And there exists a maximum wave speed  $a_{\max}(\mathbf{u}_L, \mathbf{u}_R) > 0$  such that*

$$(3.7) \quad \begin{aligned} \mathbf{U}^{\text{RP}}(x, t; \mathbf{u}_L, \mathbf{u}_R) &= \mathbf{u}_L & \forall x/t \leq -a_{\max}, \\ \mathbf{U}^{\text{RP}}(x, t; \mathbf{u}_L, \mathbf{u}_R) &= \mathbf{u}_R & \forall x/t \geq a_{\max}. \end{aligned}$$

Such an assumption can be verified for most systems of hyperbolic conservation laws and well studied equations such as scalar conservation laws, shallow water equations [123], and compressible Euler equations [205]. Assumption 1 does not hold in certain cases, such as the multidimensional ideal MHD and relativistic MHD systems with a jump in the normal component of magnetic field [117] or when the divergence-free constraint of magnetic field is violated [218, 221].

For any  $\alpha \geq a_{\max}$ , integrating (3.5) with the initial data (3.6) over  $[-\alpha, \alpha] \times [0, 1]$  in space-time domain with the Divergence Theorem yields



$$\begin{aligned} \int_{-\alpha}^{\alpha} \mathbf{U}^{\text{RP}}(x, 1; \mathbf{u}_L, \mathbf{u}_R) dx &= \int_{-\alpha}^{\alpha} \mathbf{u}_0(x) dx - [\mathbf{f}(\mathbf{U}^{\text{RP}}(\alpha, 1; \mathbf{u}_L, \mathbf{u}_R)) + \mathbf{f}(\mathbf{U}^{\text{RP}}(-\alpha, 1; \mathbf{u}_L, \mathbf{u}_R))] \\ &= \alpha(\mathbf{u}_L + \mathbf{u}_R) - \mathbf{f}(\mathbf{u}_R) + \mathbf{f}(\mathbf{u}_L), \end{aligned}$$

where we have used (3.7) in the second step. Dividing this identity by  $2\alpha$  implies

$$\frac{\mathbf{u}_L + \mathbf{u}_R}{2} + \frac{\mathbf{f}(\mathbf{u}_L) - \mathbf{f}(\mathbf{u}_R)}{2\alpha} = \frac{1}{2\alpha} \int_{-\alpha}^{\alpha} \mathbf{U}^{\text{RP}}(x, 1; \mathbf{u}_L, \mathbf{u}_R) dx.$$

Thus Assumption 1 implies:

$$(3.8) \quad \mathbf{u}_L, \mathbf{u}_R \in G \Rightarrow \frac{\mathbf{u}_L + \mathbf{u}_R}{2} + \frac{\mathbf{f}(\mathbf{u}_L) - \mathbf{f}(\mathbf{u}_R)}{2\alpha} \in G \quad \forall \alpha \geq a_{\max}(\mathbf{u}_L, \mathbf{u}_R).$$

We state the second assumption without using any exact solutions:

ASSUMPTION 2. *There exists a suitable function  $\hat{a}(\mathbf{u}) > 0$  such that*

$$(3.9) \quad \mathbf{u} \in G \implies \mathbf{u} \pm \frac{\mathbf{f}(\mathbf{u})}{\alpha} \in G \quad \forall \alpha \geq \hat{a}(\mathbf{u}).$$

In Assumption 2, a suitable  $\hat{a}$  should satisfy  $\hat{a}(\mathbf{u}) \leq \eta \hat{a}_{\max}(\mathbf{u})$  for some constant  $\eta > 0$ , where  $\hat{a}_{\max}(\mathbf{u})$  denotes the maximum wave speed at the state  $\mathbf{u}$ . If this requirement is removed, then for any open set  $G$ , one could always construct such a function by selecting sufficiently large values; however, this would lead to arbitrarily small time steps in an IDP scheme and thus be practically useless.

REMARK 3.2. For many problems, it often suffices to take  $\hat{a}(\mathbf{u})$  as the maximum wave speed to satisfy [Assumption 2](#). For example, for the 1D compressible Euler equations with the ideal gas EOS (2.1b), the maximum wave speed is given by the spectral radius of the Jacobian matrix  $\mathbf{f}'(\mathbf{u})$ , which is simply  $|u| + \sqrt{\gamma \frac{p}{\rho}}$ . For enforcing the invariant domain in (2.2), by [242, Lemma 6], to satisfy [Assumption 2](#), one can take  $\hat{a}(\mathbf{u}) = |u| + \sqrt{\frac{\gamma-1}{2} \frac{p}{\rho}} < |u| + \sqrt{\gamma \frac{p}{\rho}}$ . However, for other stability considerations such as entropy stability, the use of the maximum wave speed can be necessary.

[Assumption 2](#) is not a universal property and it may not hold for all convex invariant domains. For instance, it does not hold for many scalar conservation laws with the bound-preserving property or for the compressible MHD systems. When the entropy principle is considered for gas dynamics, e.g., the invariant domain (2.5), the property [Assumption 2](#) may not hold in general. In the relativistic hydrodynamic case, if the entropy principle is included in the invariant domain, then [Assumption 2](#) no longer holds, as observed in [216].

When the invariant domain involves highly nonlinear (or even implicit) constraints, e.g., (2.19), it is often very challenging to verify the above assumptions. In [222], Wu and Shu proposed a general approach, termed Geometric Quasi Linearization (GQL), which transforms the nonlinear constraints in (1.2) into *equivalent linear* constraints:

$$(3.10) \quad G^* := \left\{ \mathbf{u} \in \mathbb{R}^N : (\mathbf{u} - \mathbf{u}^*) \cdot \mathbf{n}_i^* \succ 0 \quad \forall \mathbf{u}^* \in \mathcal{S}_i, \forall i \in \mathbb{I} \cup \hat{\mathbb{I}} \right\},$$

where  $\mathcal{S}_i := \partial G \cap \partial G_i$  with  $\partial G_i := \{\mathbf{u} \in \mathbb{R}^N : g_i(\mathbf{u}) = 0\}$ ,  $\mathbf{n}_i^* := \nabla g_i(\mathbf{u}^*)$  is an inward normal vector of  $\partial G$  at  $\mathbf{u}^*$ , and the symbol  $\succ$  represents  $>$  for  $i \in \mathbb{I}$  and  $\geq$  for  $i \in \hat{\mathbb{I}}$ . Here,  $\mathbf{u}^*$  is independent of  $\mathbf{u}$  and is referred to as a *free auxiliary variable* in the GQL framework [222]. These variables are introduced to lift the dimension for linearity.

Under the GQL framework [222], we introduce a more general and weaker assumption than [Assumption 2](#), which is extensible to more equations such as MHD systems.

ASSUMPTION 3. *There exists a suitable function  $\hat{a}(\mathbf{u}) > 0$  and  $\zeta(\mathbf{u}^*)$  such that*

$$(3.11) \quad \mathbf{u} \in G \implies \alpha(\mathbf{u} - \mathbf{u}^*) \cdot \mathbf{n}_i^* \pm \mathbf{f}(\mathbf{u}) \cdot \mathbf{n}_i^* \succ \pm \zeta(\mathbf{u}^*) \quad \forall \mathbf{u}^* \in \mathcal{S}_i, \quad \forall \alpha \geq \hat{a}(\mathbf{u}).$$

Based on the equivalence  $G = G^*$ , one can show that [Assumption 3](#) also implies (3.8), with  $a_{\max}(\mathbf{u}_L, \mathbf{u}_R) = \max\{\hat{a}(\mathbf{u}_L), \hat{a}(\mathbf{u}_R)\}$ . Moreover, [Assumption 2](#) can be viewed as a special case of [Assumption 3](#) by taking  $\zeta(\mathbf{u}^*) = 0$ . With or without the entropy principle included in the invariant domain, [Assumption 3](#) holds for the relativistic hydrodynamic [216]. For MHD systems, where [Assumption 1](#) and [Assumption 2](#) are both inapplicable, a variant of [Assumption 3](#) can still be established by carefully constructing  $\zeta(\mathbf{u}^*)$  so that its related terms vanishes under the (discrete) divergence-free constraint [225, 215, 219, 221].

REMARK 3.3. *For special systems and specific invariant domains, it is possible to construct IDP schemes without employing these assumptions, e.g., kinetic schemes for compressible Euler equations [174, 175, 81, 203].*

In the rest of this section, for simplicity, we only focus on how to use [Assumption 1](#) or [Assumption 2](#) for proving that first order schemes are IDP for systems like gas dynamics equations. For harder problems like compressible MHD and relativistic hydrodynamics equations, [Assumption 3](#) applies and will be reviewed in [Subsections 6.7](#) and [6.8](#).

**3.3. First order schemes for 1D hyperbolic systems.** The monotonicity technique in Section 3.1 is useful for maximum principles, but it does not apply to system of hyperbolic conservation laws. Next we demonstrate three techniques based on the above assumptions, for provable IDP schemes in the form (1.4) for solving 1D hyperbolic system (3.5).

**DEFINITION 3.1** (IDP numerical flux). *Let  $\lambda = \frac{\Delta t}{\Delta x}$ . A numerical flux  $\hat{\mathbf{f}}(\cdot, \cdot)$  is said to be IDP, if the corresponding 1D three-point first order scheme (1.4) is IDP,*

$$\mathbf{u}_j^{n+1} := \mathbf{u}_j^n - \lambda \left( \hat{\mathbf{f}}(\mathbf{u}_j^n, \mathbf{u}_{j+1}^n) - \hat{\mathbf{f}}(\mathbf{u}_{j-1}^n, \mathbf{u}_j^n) \right) \in G \quad \forall \mathbf{u}_{j-1}^n, \mathbf{u}_j^n, \mathbf{u}_{j+1}^n \in G,$$

under a suitable CFL condition  $\lambda \alpha \leq c_0$ , where  $\alpha$  denote the estimated maximum wavespeed and  $c_0$  is the IDP CFL number.

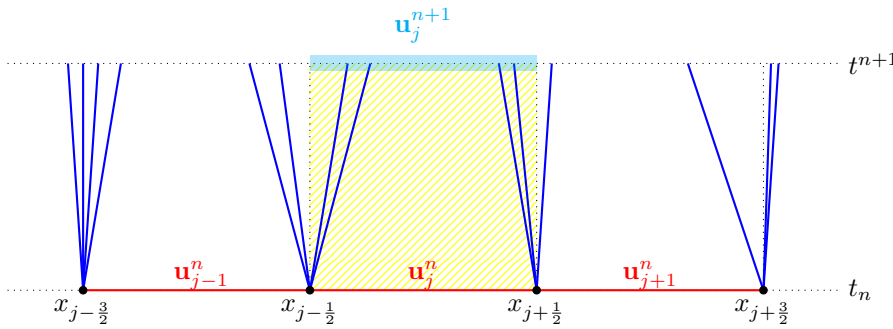
For compressible Euler equations with ideal gas EOS (2.1), the Lax–Friedrichs flux (a.k.a. Rusanov flux) can be shown IDP for  $G$  in (2.2) with  $c_0 = 1$  [202] and the Godunov and HLLE fluxes are IDP with  $c_0 = \frac{1}{2}$ ; see [76]. See [174, 81, 203] for the CFL of kinetic schemes to be IDP. Next, we demonstrate how the three assumptions can be used to show IDP in first order schemes by a few examples.

**3.3.1. Method 1 using Assumption 1.** This technique is a traditional approach and has been widely used; e.g., [76, 176]. Based on Assumption 1, we can investigate the IDP property of several numerical fluxes defined as exact or approximate Riemann solvers, such as the Godunov scheme, Lax–Friedrichs scheme, and HLL type schemes.

**EXAMPLE 3.1** (Godunov scheme). *Let the self similar solution to the Riemann problem be denoted by  $\mathbf{U}^{\text{RP}}(\xi; \mathbf{u}_L, \mathbf{u}_R)$  with  $\xi := x/t$ . Then the Godunov flux is*

$$\hat{\mathbf{f}}(\mathbf{u}^-, \mathbf{u}^+) = \mathbf{f}(\mathbf{U}^{\text{RP}}(0; \mathbf{u}^-, \mathbf{u}^+)).$$

Consider the scheme (1.4) using the Godunov flux with  $\mathbf{u}_j^n$  denoting the cell average on the interval  $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$ . Let  $a_{\max} = \max_j a_{\max}(\mathbf{u}_j^n, \mathbf{u}_{j+1}^n)$  be the maximum wavespeed.



By integrating (3.5) over a space-time rectangle  $I_j \times [t^n, t^{n+1}]$  with the Divergence Theorem as shown in the figure above, we obtain

$$\frac{1}{\Delta x} \int_{I_j} \mathbf{u}(x, t^{n+1}) dx = \frac{1}{\Delta x} \int_{I_j} \mathbf{u}(x, t^n) dx - \frac{\Delta t}{\Delta x} \int_{t^n}^{t^{n+1}} [\mathbf{f}(\mathbf{u}(x_{j+\frac{1}{2}}, t)) - \mathbf{f}(\mathbf{u}(x_{j-\frac{1}{2}}, t))] dt,$$

in which  $\mathbf{u}(x, t^n)$  is piecewise constant  $\mathbf{u}_j^n$ , the interface value  $\mathbf{u}(x_{j+\frac{1}{2}}, t)$  is equal to the self similar exact solution  $\mathbf{U}^{\text{RP}}(0; \mathbf{u}_j^n, \mathbf{u}_{j+1}^n)$  for any  $t \in [t^n, t^{n+1}]$ , if the local

Riemann problems do not intersect, ensured by the CFL  $\frac{\Delta t}{\Delta x} a_{\max} \leq \frac{1}{2}$  as shown in the figure. Thus in the Godunov scheme

$$\mathbf{u}_j^{n+1} = \mathbf{u}_j^n - \lambda [\hat{\mathbf{f}}(\mathbf{u}_j^n, \mathbf{u}_{j+1}^n) - \mathbf{f}(\mathbf{u}_{j-1}^n, \mathbf{u}_j^n)],$$

the numerical solution  $\mathbf{u}_j^{n+1}$  is exactly the average of the exact solution to two non-intersecting Riemann problems under the CFL  $\frac{\Delta t}{\Delta x} a_{\max} \leq \frac{1}{2}$ . The integral operator  $\frac{1}{\Delta} \int_{I_j} \cdot dx$  preserves a convex invariant domain, thus [Assumption 1](#) implies  $\mathbf{u}_j^{n+1} \in G$ .

EXAMPLE 3.2 (Global Lax–Friedrichs flux). For the Lax–Friedrichs flux (3.4) with a global estimate of the wavespeed  $\alpha$  at time  $t^n$ , the scheme (1.4) can be rewritten as a convex combination:

$$(3.12) \quad \mathbf{u}_j^{n+1} = (1 - \lambda\alpha)\mathbf{u}_j^n + \lambda\alpha \left( \frac{\mathbf{u}_{j-1}^n + \mathbf{u}_{j+1}^n}{2} + \frac{\mathbf{f}(\mathbf{u}_{j-1}^n) - \mathbf{f}(\mathbf{u}_{j+1}^n)}{2\alpha} \right)$$

under the CFL condition  $\lambda\alpha \leq 1$ . By (3.8) and the convexity of  $G$ , we have  $\mathbf{u}_j^{n+1} \in G$ , if  $\alpha$  in the Lax–Friedrichs flux (3.4) satisfies  $\alpha \geq \max_j a_{\max}(\mathbf{u}_{j-1}^n, \mathbf{u}_j^n)$ .

EXAMPLE 3.3 (Local Lax–Friedrichs flux). Consider the scheme (1.4) with a local Lax–Friedrichs flux defined by

$$(3.13) \quad \hat{\mathbf{f}}(\mathbf{u}_{j-1}^n, \mathbf{u}_j^n) = \frac{1}{2} [\mathbf{f}(\mathbf{u}_{j-1}^n) + \mathbf{f}(\mathbf{u}_j^n) - \alpha_{j-\frac{1}{2}}^n (\mathbf{u}_j^n - \mathbf{u}_{j-1}^n)], \quad \alpha_{j-\frac{1}{2}}^n = a_{\max}(\mathbf{u}_j, \mathbf{u}_{j-1}).$$

Then the first order local Lax–Friedrichs scheme is

$$\mathbf{u}_j^{n+1} = \mathbf{u}_j^n - \lambda \left[ \frac{\mathbf{f}(\mathbf{u}_j^n) + \mathbf{f}(\mathbf{u}_{j+1}^n) - \alpha_{j+\frac{1}{2}}^n (\mathbf{u}_{j+1}^n - \mathbf{u}_j^n)}{2} - \frac{\mathbf{f}(\mathbf{u}_{j-1}^n) + \mathbf{f}(\mathbf{u}_j^n) - \alpha_{j-\frac{1}{2}}^n (\mathbf{u}_j^n - \mathbf{u}_{j-1}^n)}{2} \right],$$

which is equivalent to

$$(3.14) \quad \begin{aligned} \mathbf{u}_j^{n+1} &= \left( 1 - \lambda\alpha_{j+\frac{1}{2}}^n - \lambda\alpha_{j-\frac{1}{2}}^n \right) \mathbf{u}_j^n + \lambda\alpha_{j+\frac{1}{2}}^n \left( \frac{\mathbf{u}_{j+1}^n + \mathbf{u}_j^n}{2} + \frac{\mathbf{f}(\mathbf{u}_j^n) - \mathbf{f}(\mathbf{u}_{j+1}^n)}{2\alpha_{j+\frac{1}{2}}^n} \right) \\ &+ \lambda\alpha_{j-\frac{1}{2}}^n \left( \frac{\mathbf{u}_j^n + \mathbf{u}_{j-1}^n}{2} + \frac{\mathbf{f}(\mathbf{u}_{j-1}^n) - \mathbf{f}(\mathbf{u}_j^n)}{2\alpha_{j-\frac{1}{2}}^n} \right). \end{aligned}$$

For compressible Euler equations with ideal gas EOS (2.1) and the invariant domain  $G_S$  including the minimum entropy principle (2.5), by (3.8), we have

$$\mathbf{u}_j^n, \mathbf{u}_{j-1}^n \in G_S \implies \frac{\mathbf{u}_j^n + \mathbf{u}_{j-1}^n}{2} + \frac{\mathbf{f}(\mathbf{u}_{j-1}^n) - \mathbf{f}(\mathbf{u}_j^n)}{2\alpha_{j-\frac{1}{2}}^n} \in G_S.$$

Therefore, by the convexity of  $G_S$ , we get  $\mathbf{u}_j^{n+1} \in G_S$  because the scheme (3.14) is a convex combination under the CFL condition

$$(3.15) \quad \lambda \max_j \alpha_{j+\frac{1}{2}}^n \leq \frac{1}{2},$$

which is the same CFL derived as in [176, Appendix] by a different yet essentially equivalent approach.

We remark that an estimate of a rigorous yet explicit upper bound  $a_{\max}$  is needed for the maximum wave speed in the Riemann problem in Method 1. A naive estimate based on the largest eigenvalues is not always adequate when fast shocks arise in the Riemann problem; see, e.g., [101] for a discussion in the context of the Euler equations. On the other hand, the eigenvalue-based estimate can be overly restrictive in some cases, for example, when simulating the Sod shock tube problem.

**3.3.2. Method 2 using Assumption 2.** This approach is more flexible, which will be frequently used in this paper. The basic idea is to decompose a scheme into a convex combination of several simpler functions of the form  $\mathbf{u}$  and  $\mathbf{u} \pm \mathbf{f}(\mathbf{u})/\alpha$ , which preserve the invariant domain  $G$ , then the convexity of  $G$  implies that the target scheme is IDP.

EXAMPLE 3.4 (Local Lax–Friedrichs flux). *For solving (2.1), the first order the local Lax–Friedrichs scheme can also be rewritten as*

$$(3.16) \quad \mathbf{u}_j^{n+1} = \frac{2 - \lambda \alpha_{j-\frac{1}{2}}^n - \lambda \alpha_{j+\frac{1}{2}}^n}{2} \mathbf{u}_j^n + \frac{\lambda \alpha_{j-\frac{1}{2}}^n}{2} \left( \mathbf{u}_{j-1}^n + \frac{\mathbf{f}(\mathbf{u}_{j-1}^n)}{\alpha_{j-\frac{1}{2}}^n} \right) + \frac{\lambda \alpha_{j+\frac{1}{2}}^n}{2} \left( \mathbf{u}_{j+1}^n - \frac{\mathbf{f}(\mathbf{u}_{j+1}^n)}{\alpha_{j+\frac{1}{2}}^n} \right).$$

For the invariant domain  $G$  in (2.2), by taking

$$\alpha_{j-\frac{1}{2}}^n = \max_{i=j, j-1} |\mathbf{f}'(\mathbf{u}_i^n)| = \max_{i=j, j-1} |v_i| + \sqrt{\frac{\gamma p_i}{\rho_i}} \quad \forall j,$$

if  $\mathbf{u}_j^n \in G$ , then  $\mathbf{u}_j^n \pm \frac{\mathbf{f}(\mathbf{u}_j^n)}{\alpha_{j \pm \frac{1}{2}}^n} \in G$  for all  $j$ , see [246, Remark 2.4]. Therefore, by the convexity of  $G$ , the scheme (3.16) preserves the invariant domain  $G$  under the CFL condition  $\lambda \max_j \alpha_{j+\frac{1}{2}}^n \leq 1$ , which allows a larger time step than (3.15) in Method 1.

Example 3.3 and Example 3.4 are two different approaches on the same scheme. By comparing these two examples, we can see that each method has its own advantages. Although Method 2 allows a larger time step for provable positivity-preserving property for density and pressure for the local Lax–Friedrichs scheme, it cannot be used for enforcing the minimum entropy principle since Assumption 2 may not hold for  $G_S$  in (2.5). On the other hand, for gas dynamics equations such as compressible Navier–Stokes equations with a generic EOS and  $G$  in (2.6), Method 2 is very flexible to use.

Note that Assumption 2 is commonly used to construct IDP schemes with Lax–Friedrichs type fluxes. However, for IDP schemes employing other numerical fluxes, it may not directly apply and must be appropriately adapted.

**3.4. Basic assumptions in multiple dimensions.** For a multi-dimensional system (1.3) and any given unit vector  $\mathbf{n} \in \mathbb{R}^d$ , let  $\mathbf{U}^{\text{RP}}(x, t; \mathbf{n}, \mathbf{u}_L, \mathbf{u}_R)$  be the exact solution of a Riemann problem to one dimensional equation  $\mathbf{u}_t + \partial_x [\mathbf{f}(\mathbf{u}) \cdot \mathbf{n}] = 0$  with the initial data

$$(3.17) \quad \mathbf{u}_0(x) = \begin{cases} \mathbf{u}_L, & \text{if } x \leq 0, \\ \mathbf{u}_R, & \text{if } x > 0. \end{cases}$$

ASSUMPTION 4 (Assumption 1 in multiple dimensions). *The exact solution of the Riemann problem preserves the invariant domain, namely,  $\mathbf{u}_L, \mathbf{u}_R \in G \Rightarrow \mathbf{U}^{\text{RP}}(x, t; \mathbf{n}, \mathbf{u}_L, \mathbf{u}_R) \in G$  for any  $x \in \mathbb{R}$  and  $t > 0$ . There exists a maximum wave*

speed  $a_{\max}(\mathbf{n}, \mathbf{u}_L, \mathbf{u}_R) > 0$  such that

$$\begin{aligned} \mathbf{U}^{\text{RP}}(x, t; \mathbf{n}, \mathbf{u}_L, \mathbf{u}_R) &= \mathbf{u}_L & \forall x/t \leq -a_{\max}, \\ \mathbf{U}^{\text{RP}}(x, t; \mathbf{n}, \mathbf{u}_L, \mathbf{u}_R) &= \mathbf{u}_R & \forall x/t \geq a_{\max}. \end{aligned}$$

Similar to the derivation of (3.8), this assumption yields

$$\mathbf{u}_L, \mathbf{u}_R \in G \implies \frac{\mathbf{u}_L + \mathbf{u}_R}{2} + \frac{\mathbf{f}(\mathbf{u}_L) \cdot \mathbf{n} - \mathbf{f}(\mathbf{u}_R) \cdot \mathbf{n}}{2\alpha} \in G \quad \forall \alpha \geq a_{\max}(\mathbf{n}, \mathbf{u}_L, \mathbf{u}_R).$$

ASSUMPTION 5 (Assumption 2 in multiple dimensions). *There exists a suitable function  $\hat{a}(\mathbf{u}, \mathbf{n}) > 0$  such that  $\mathbf{u} \in G \implies \mathbf{u} \pm \frac{\mathbf{f}(\mathbf{u}) \cdot \mathbf{n}}{\alpha} \in G \quad \forall \alpha \geq \hat{a}(\mathbf{u}, \mathbf{n})$ .*

Assumption 5 does hold for many interesting systems and commonly considered convex invariant domains such as the Euler and ten-moment Gaussian closure systems without considering minimum entropy principle. For instance, the following result implies that this assumption holds for compressible Euler and also Navier–Stokes equations, with the invariant domain defined in (2.6).

LEMMA 3.2 (Lemma 6 in [242]). *Consider any  $\mathbf{u} = (\rho, \mathbf{m}, E)^\top$ , and*

$$\mathbf{f}^a(\mathbf{u}) = \begin{pmatrix} \mathbf{m} \\ \rho^{-1} \mathbf{m} \otimes \mathbf{m} + p \mathbf{I} \\ \rho^{-1}(E + p) \mathbf{m} \end{pmatrix}, \quad \mathbf{f}^d(\mathbf{u}) = \begin{pmatrix} 0 \\ \boldsymbol{\tau} \\ \rho^{-1} \mathbf{m} \cdot \boldsymbol{\tau} - \mathbf{q} \end{pmatrix},$$

where  $p$ ,  $\boldsymbol{\tau}$ , and  $\mathbf{q}$  are not necessarily dependent on  $\mathbf{u}$ . Let  $e = \rho^{-1}E - \frac{1}{2}\rho^{-2}|\mathbf{m}|^2$ . For any unit vector  $\mathbf{n}$ , let  $v = \rho^{-1}\mathbf{m} \cdot \mathbf{n}$ ,  $q = \mathbf{q} \cdot \mathbf{n}$  and  $\tau = \mathbf{n} \cdot \boldsymbol{\tau}$ . Then we have the following for  $G = \{\mathbf{u} : \rho > 0, e = \rho^{-1}E - \frac{1}{2}\rho^{-2}|\mathbf{m}|^2 \geq 0\}$ ,

- (a)  $\mathbf{u} \pm \alpha^{-1} \mathbf{f}^a(\mathbf{u}) \cdot \mathbf{n} \in G$  if and only if  $\alpha > |v| + \sqrt{\frac{p^2}{2\rho^2 e}}$ ,
- (b)  $\mathbf{u} \pm \beta^{-1} (\mathbf{f}^a(\mathbf{u}) - \mathbf{f}^d(\mathbf{u})) \cdot \mathbf{n} \in G$  if and only if

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

REMARK 3.4. *For the compressible Navier–Stokes equations, we can write it as if it were a formal convection system  $\partial_t \mathbf{u} + \nabla \cdot \mathbf{f}(\mathbf{u}) = \mathbf{0}$  where  $\mathbf{f} = \mathbf{f}^a - \mathbf{f}^b$  and  $\mathbf{f}^a, \mathbf{f}^b$  are given in Lemma 3.2. Then  $\beta$  in Lemma 3.2 (b) gives one way to satisfy Assumption 2 for such a formal system. However, here  $\beta$  is not any approximation to wave speeds, but instead merely a quantity designed to satisfy Assumption 2. After all, the concept of wavespeed is not well defined for a convection diffusion system like the Navier–Stokes equations.*

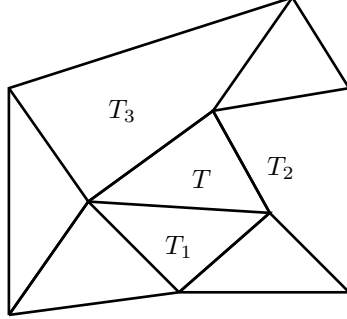
ASSUMPTION 6 (Assumption 3 in multiple dimensions). *There exists a suitable function  $\hat{a}(\mathbf{u}, \mathbf{n}) > 0$  and a vector function  $\boldsymbol{\zeta}(\mathbf{u}^*)$  such that, for any  $\alpha \geq \hat{a}(\mathbf{u}, \mathbf{n})$ ,*

$$(3.18) \quad \mathbf{u} \in G \implies \alpha(\mathbf{u} - \mathbf{u}^*) \cdot \mathbf{n}_i^* \pm (\mathbf{f}(\mathbf{u}) \cdot \mathbf{n}) \cdot \mathbf{n}_i^* \succ \pm \boldsymbol{\zeta}(\mathbf{u}^*) \cdot \mathbf{n} \quad \forall \mathbf{u}^* \in \mathcal{S}_i,$$

where  $\mathbf{n}_i^*$  is an inward normal vector of  $\partial G$  at  $\mathbf{u}^*$ ; see (3.10).

Assumption 5 can be regarded as a special case of Assumption 6 with  $\boldsymbol{\zeta}(\mathbf{u}^*) \equiv \mathbf{0}$ . Hence, Assumption 6 is generally weaker than Assumption 5. The application of using Assumption 6 will be reviewed in Subsections 4.6.3, 6.7 and 6.8.

**3.5. Finite volume scheme in multiple dimensions.** For simplicity, consider the two dimensional system (1.3) and a polygonal mesh as shown in the following figure.



Let  $T$  be a polygonal cell with edges  $E_i$  ( $i = 1, \dots, T_E$ ) and  $T_i$  be the adjacent cell which shares the edge  $E_i$  with  $T$ . Let  $|E_i|$  denote the length of the edge  $E_i$ . For solving (1.3), consider a first order finite volume scheme on the cell  $T$ ,

$$\mathbf{u}_T^{n+1} = \mathbf{u}_T^n - \frac{\Delta t}{|T|} \sum_{i=1}^{T_E} |E_i| \widehat{\mathbf{f} \cdot \mathbf{n}}(\mathbf{u}_T^n, \mathbf{u}_{T_i}^n),$$

with the Lax-Friedrichs or Rusanov flux defined by

$$\widehat{\mathbf{f} \cdot \mathbf{n}}(\mathbf{u}_T^n, \mathbf{u}_{T_i}^n) = \frac{1}{2} [\mathbf{f}(\mathbf{u}_T^n) \cdot \mathbf{n}_i + \mathbf{f}(\mathbf{u}_{T_i}^n) \cdot \mathbf{n}_i - \alpha_i (\mathbf{u}_{T_i}^n - \mathbf{u}_T^n)],$$

where  $\mathbf{u}_T^n$  is the approximation to the average of  $\mathbf{u}$  on  $T$  at time level  $n$ ,  $\mathbf{n}_i$  is the unit vector normal to the edge  $E_i$  pointing outward of  $T$ , and  $\alpha_i$  is a positive number dependent on  $\mathbf{u}_T^n$  and  $\mathbf{u}_{T_i}^n$ . With the assumption  $\mathbf{u}_T^n, \mathbf{u}_{T_i}^n \in G$ , we want to find a proper  $\alpha_i$  and a CFL condition so that  $\mathbf{u}_T^{n+1} \in G$ . A simple fact for any polygon  $T$  is

$$(3.19) \quad \sum_{i=1}^{T_E} \mathbf{n}_i |E_i| = \mathbf{0}.$$

EXAMPLE 3.5 (Method 1). By (3.19), we obtain  $\sum_{i=1}^{T_E} \mathbf{f}(\mathbf{u}_T^n) \cdot \mathbf{n}_i |E_i| = 0$ , thus the right hand side of the first order Lax-Friedrichs scheme can be rewritten as

$$\begin{aligned} & \mathbf{u}_T^n - \frac{\Delta t}{|T|} \sum_{i=1}^{T_E} |E_i| \widehat{\mathbf{f} \cdot \mathbf{n}}(\mathbf{u}_T^n, \mathbf{u}_{T_i}^n) + \frac{\Delta t}{|T|} \sum_{i=1}^{T_E} \mathbf{f}(\mathbf{u}_T^n) \cdot \mathbf{n}_i |E_i| \\ &= \mathbf{u}_T^n + \frac{\Delta t}{|T|} \sum_{i=1}^{T_E} |E_i| \frac{\mathbf{f}(\mathbf{u}_T^n) \cdot \mathbf{n}_i - \mathbf{f}(\mathbf{u}_{T_i}^n) \cdot \mathbf{n}_i + \alpha_i (\mathbf{u}_{T_i}^n - \mathbf{u}_T^n)}{2} \\ &= \left( 1 - \frac{\Delta t}{|T|} \sum_{i=1}^{T_E} |E_i| \alpha_i \right) \mathbf{u}_T^n + \frac{\Delta t}{|T|} \sum_{i=1}^{T_E} |E_i| \alpha_i \left( \frac{\mathbf{u}_{T_i}^n + \mathbf{u}_T^n}{2} + \frac{\mathbf{f}(\mathbf{u}_T^n) \cdot \mathbf{n}_i - \mathbf{f}(\mathbf{u}_{T_i}^n) \cdot \mathbf{n}_i}{2\alpha_i} \right), \end{aligned}$$

which is a convex combination under the CFL constraint

$$(3.20) \quad \Delta t \frac{|\partial T|}{|T|} \max_i \alpha_i \leq 1,$$

with  $|\partial T| = \sum_{i=1}^{T_E} |E_i|$ . Now Method 1 can be applied whenever Assumption 4 holds.

EXAMPLE 3.6 (Method 2). With (3.19), the first order Lax-Friedrichs or Rusanov finite volume scheme can be rewritten as

$$(3.21) \quad \mathbf{u}_T^{n+1} = \left(1 - \frac{1}{2} \frac{\Delta t}{|T|} \sum_{i=1}^{T_E} |E_i| \alpha_i\right) \mathbf{u}_T^n + \frac{1}{2} \frac{\Delta t}{|T|} \sum_{i=1}^{T_E} |E_i| \alpha_i [\mathbf{u}_{T_i}^n - \alpha_i^{-1} \mathbf{f}(\mathbf{u}_{T_i}^n) \cdot \mathbf{n}_i].$$

For gas dynamics equations with a generic EOS and  $G$  in (2.6), by Lemma 3.2, we have  $\mathbf{u}_{T_i}^n - \alpha_i^{-1} \mathbf{f}(\mathbf{u}_{T_i}^n) \cdot \mathbf{n}_i \in G$  if we use any viscosity parameter

$$\alpha_i > \max_{\mathbf{u}_T^n, \mathbf{u}_{T_i}^n} |\rho^{-1} \mathbf{m} \cdot \mathbf{n}_i| + \sqrt{\frac{p^2}{2\rho^2 e}}.$$

Notice that (3.21) is a convex combination of  $\mathbf{u}_T^n$  and  $\mathbf{u}_{T_i}^n - \alpha_i^{-1} \mathbf{f}(\mathbf{u}_{T_i}^n) \cdot \mathbf{n}_i$  thus  $\mathbf{u}_T^{n+1} \in G$ , under the CFL constraint  $\Delta t \frac{|\partial T|}{|T|} \max_i \alpha_i \leq 2$ , which twice of the CFL as in (3.20).

**3.6. Continuous finite element method.** We consider a first order accurate IDP continuous finite element method for hyperbolic problems [102]. Such a scheme is also known as the group finite element method [84, 192, 193, 24]. As an example, we consider solving (1.3) on a triangular mesh in two dimensions, and it can be extended to a mesh consisting of tetrahedra, parallelepipeds, and triangular prisms in three dimensions.

**3.6.1. Definition of the scheme.** Let  $\Omega$  be the two dimensional domain and  $\mathcal{T}_h$  be a triangular mesh. Let  $V^h$  be the continuous piecewise linear polynomial space on  $\mathcal{T}_h$ . Let  $\varphi_i(\mathbf{x}) \in V^h$  be the Lagrangian basis at  $i$ -th vertex  $\mathbf{x}_i$  ( $i = 1, \dots, N$ ) of the triangular mesh, then  $\sum_{i=1}^N \varphi_i \equiv 1$ . Define  $M$  as the mass matrix and  $M^L$  as the lumped mass matrix, i.e.,  $M = [M_{ij}]$  with  $M_{ij} = \iint_{\Omega} \varphi_i(\mathbf{x}) \varphi_j(\mathbf{x}) d\mathbf{x}$ , and  $M^L$  is a diagonal matrix with the diagonal entries  $m_i = \iint_{\Omega} \varphi_i(\mathbf{x}) d\mathbf{x} = \sum_j M_{ij}$ . Define  $\mathcal{N}_i = \{j : \varphi_i(\mathbf{x}) \varphi_j(\mathbf{x}) \text{ is not constant zero}\}$ , i.e.,  $\mathcal{N}_i = \{i\} \cup \{j : \mathbf{x}_j \text{ is connected } \mathbf{x}_i \text{ by an edge}\}$ , as shown in the Figure 1.

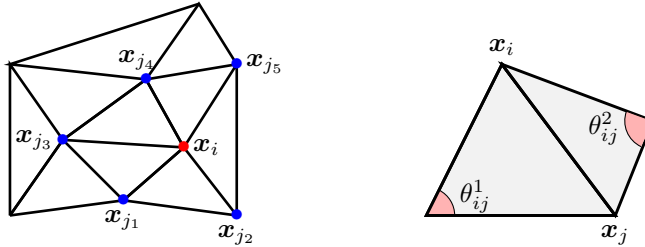


Fig. 1: Notation for continuous finite element method. Left:  $\mathcal{N}_i = \{i, j_1, j_2, j_3, j_4, j_5\}$ .

REMARK 3.5. Mass lumping is used not only in finite element methods for enforcing positivity [158] and but also used in the Petrov–Galerkin formulation of residual distribution approach to recover a monotone residual distribution scheme, see [16, Section 2]. For the Lagrange basis of piecewise linear polynomials, such a row-sum lumped mass matrix is also equal to approximating integrals  $\iint_{\Omega} \varphi_i(\mathbf{x}) \varphi_j(\mathbf{x}) d\mathbf{x}$  by the simple quadrature using only the Lagrange basis points, e.g., using the quadrature of only three vertices on each triangle in a triangular mesh for approximating the integral  $\iint_{\Omega} \varphi_i(\mathbf{x}) \varphi_j(\mathbf{x}) d\mathbf{x}$ .

Let  $\mathbf{u}_h^n$  denote the finite element solution at time step  $n$  on a mesh of size  $h$ , and  $\mathbf{u}_j^n = \mathbf{u}_h^n(\mathbf{x}_j)$ , then

$$\mathbf{u}_h^n(\mathbf{x}) = \sum_i \mathbf{u}_i^n \varphi_i(\mathbf{x}), \quad \iint_{\Omega} \mathbf{u}_h^n(\mathbf{x}) d\mathbf{x} = \sum_i \mathbf{u}_i^n m_i.$$

By the group finite element approximation  $\mathbf{f}(\mathbf{u}_h^n) \approx \sum_j \mathbf{f}(\mathbf{u}_j^n) \varphi_j(\mathbf{x})$ , we have

$$\iint_{\Omega} \nabla \cdot \mathbf{f}(\mathbf{u}_h^n) \varphi_i(\mathbf{x}) d\mathbf{x} \approx \sum_{j \in \mathcal{N}_i} \mathbf{f}(\mathbf{u}_j^n) \cdot \mathbf{c}_{ij}, \quad \mathbf{c}_{ij} = \iint_{\Omega} \varphi_i \nabla \varphi_j d\mathbf{x},$$

then one version of IDP continuous finite element method with forward Euler time stepping can be given as

$$(3.22a) \quad m_i \frac{\mathbf{u}_i^{n+1} - \mathbf{u}_i^n}{\Delta t} + \sum_{j \in \mathcal{N}_i} [\mathbf{f}(\mathbf{u}_j^n) \cdot \mathbf{c}_{ij} - d_{ij}^n \mathbf{u}_j^n] = 0,$$

where  $d_{ij}^n$  is the artificial viscosity coefficients designed to ensure stability including the IDP properties. Let  $D^n = [d_{ij}^n]$  be a sparse matrix with entries  $d_{ij}^n = 0$  for  $j \notin \mathcal{N}_i$ . Then  $D^n$  should be a symmetric matrix with zero row sum and non-negative off-diagonal entries,

$$(3.22b) \quad d_{ij}^n \geq 0, \quad d_{ij}^n = d_{ji}^n, \quad \forall i \neq j; \quad \sum_{j \in \mathcal{N}_i} d_{ij}^n = 0.$$

It is possible to discuss more properties for a first order scheme like (3.22) by choosing suitable  $d_{ij}$  or adding suitable viscosity terms, e.g., [100, 102]. Here we only review viscosity coefficients to achieve the IDP properties.

### 3.6.2. Examples of artificial viscosity.

We first give two examples of (3.22b).

**EXAMPLE 3.7** (Discrete Laplacian for Artificial Viscosity). *Note that the conditions in (3.22b) are met by the discrete Laplacian matrices of the linear finite element method on a simplicial mesh under some mild mesh constraints. On a 2D triangular mesh, for the edge connecting two interior vertices  $\mathbf{x}_i, \mathbf{x}_j$ , there are two angles  $\theta_{ij}^1$  and  $\theta_{ij}^2$  as shown in Figure 1. Let  $S$  with  $S_{ij} = \iint_{\Omega} \nabla \varphi_i \cdot \nabla \varphi_j d\mathbf{x}$  be the stiffness matrix in the continuous finite element method of Lagrange  $P^1$  basis on a triangular mesh for solving Laplace equation  $-\Delta u = 0$  in two dimensions, then  $S$  is a sparse symmetric matrix with zero row sums:*

$$S_{ij} = \begin{cases} 0, & j \notin \mathcal{N}_i, \\ -\frac{\cot \theta_{ij}^1 + \cot \theta_{ij}^2}{2}, & j \in \mathcal{N}_i, j \neq i, \\ -\sum_{j \neq i} S_{ij} & j = i. \end{cases}$$

Since the necessary and sufficient condition for  $\cot \theta_{ij}^1 + \cot \theta_{ij}^2 \geq 0$  is  $\theta_{ij}^1 + \theta_{ij}^2 \leq \pi$ , for satisfying (3.22b) up to a sign, it suffices to have  $\theta_{ij}^1 + \theta_{ij}^2 \leq \pi$ , which can be achieved in a Delaunay triangulation in two dimensions. See [232, Section 2] for similar formulae of simplicial meshes in higher dimensions. Then one choice of  $D^n$  is to set  $D^n = -\varepsilon S$ , which is an approximation to  $\varepsilon \Delta$  for some parameter  $\varepsilon > 0$ . With such a choice of  $d_{ij}^n$ , we can see that the scheme (3.22) is a first order approximation to  $\partial_t \mathbf{u} + \nabla \cdot \mathbf{f}(\mathbf{u}) = 0$  but a formally second order accurate approximation to the modified equation with extra artificial viscosity  $\partial_t \mathbf{u} + \nabla \cdot \mathbf{f}(\mathbf{u}) = \varepsilon \Delta \mathbf{u}$ . Such a modified equation approach is a standard tool for analyzing classical first order finite difference and finite volume schemes for conservation laws [143, Chapter 11].

The discrete Laplacian added here is essentially a multi-dimensional version of the Lax-Friedrichs type numerical dissipation, e.g., see the Lax-Friedrichs type scheme on unstructured grids in [2, Section 2.3.1.2].

EXAMPLE 3.8 (Graph Laplacian for Artificial Viscosity). *In the previous example, mesh constraints such as Delaunay triangulation are necessary in two dimensions, and such a mesh constraint will become more stringent in higher dimensions [232]. To remove mesh constraints for satisfying (3.22b), a graph Laplacian can be considered [96]; see also [2, Section 2.3.1.2] and [192, 141, 133]. For a given triangular mesh with nodes  $\mathbf{x}_i$  and edges  $E_{ij}$  connecting nodes  $\mathbf{x}_i$  and  $\mathbf{x}_j$ , we regard it as a weighted graph by defining the weight for  $E_{ij}$  as*

$$(3.23) \quad w_{ij} = \sum_{T \ni E_{ij}} \frac{a_T |T|}{2},$$

which is the weighted average of areas of two triangles sharing the edge  $E_{ij}$  with  $a_T$  denoting a viscosity constant for each cell  $T$ . Let  $j \sim i$  denote that  $\mathbf{x}_j$  is connected to  $\mathbf{x}_i$  by an edge. Then the graph Laplacian matrix  $L$  for such a weighted undirected graph can be given as

$$L_{ij} = \begin{cases} -w_{ij}, & j \sim i \\ \sum_{j \in \mathcal{N}_i} w_{ij}, & j = i \end{cases}.$$

That is,  $L$  is a symmetric sparse matrix with zero row sums, positive diagonal entries, and non-positive off-diagonal entries. The advantage of using graph Laplacian is the easiness to satisfy (3.22b) on any meshes, although graph Laplacian is a less accurate approximation to Laplacian compared to  $S_{ij}$ .

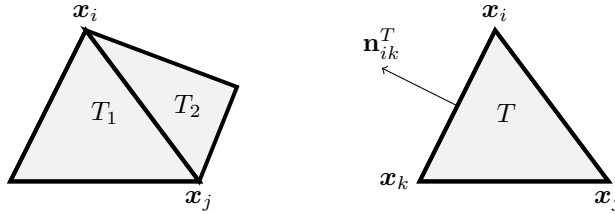


Fig. 2: An illustration of notations for computing  $\mathbf{c}_{ij}^T$  and  $\mathbf{c}_{ij}$ .

The viscosity here renders the scheme only first order accurate. See [103] for better viscosity constructed by the FCT method [141]. The residual distribution (RD) methods [2, 12, 16] can also provide improved viscosity constructions, see the next section for the connection between RD and the group finite element method.

**3.6.3. Explicit expressions of the scheme.** Next, we give more explicit expressions of the scheme. Let  $\mathbf{c}_{ij}^T = \iint_T \varphi_i \nabla \varphi_j d\mathbf{x}$ , where  $T$  is one triangle containing the edge  $E_{ij}$  connecting two nodes  $\mathbf{x}_i$  and  $\mathbf{x}_j$ . Then

$$\mathbf{c}_{ij} = \iint_{\Omega} \varphi_i \nabla \varphi_j d\mathbf{x} = \sum_{T \ni E_{ij}} \iint_T \varphi_i \nabla \varphi_j d\mathbf{x} = \sum_{T \ni E_{ij}} \mathbf{c}_{ij}^T = \mathbf{c}_{ij}^{T_1} + \mathbf{c}_{ij}^{T_2}, \quad j \neq i,$$

where  $T_1$  and  $T_2$  are two triangles sharing the edge  $E_{ij}$  as shown in Figure 2 (Left). Let  $T$  be a triangle with vertices  $\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k$ , and let  $\mathbf{n}_{ik}^T$  be unit normal vector to the edge

$E_{ik}$  outward to the triangle  $T$ , as shown in Figure 2 (Right). Then straightforward calculation (see [193, 135, 133]) gives

$$(3.24a) \quad \mathbf{c}_{ij}^T = -\frac{1}{6}|E_{ik}|\mathbf{n}_{ik}^T, \quad \mathbf{c}_{ij} = \begin{cases} \sum_{T \ni \mathbf{x}_i} \mathbf{c}_{ii}^T = \mathbf{0}, & i = j \\ \sum_{T \ni E_{ij}} \mathbf{c}_{ij}^T = -\mathbf{c}_{ji}, & i \neq j \end{cases}.$$

$$(3.24b) \quad \sum_{j \in \mathcal{N}_i, j \neq i} \mathbf{c}_{ij} = \sum_{j \in \mathcal{N}_i} \mathbf{c}_{ij} = \mathbf{0}, \quad \sum_{j \in \mathcal{N}_i} d_{ij}^n = \mathbf{0}.$$

Using (3.24b), the finite element scheme (3.22) can equivalently rewritten in a flux form as

$$(3.25) \quad m_i \frac{\mathbf{u}_i^{n+1} - \mathbf{u}_i^n}{\Delta t} + \sum_{j \in \mathcal{N}_i} [(\mathbf{f}(\mathbf{u}_j^n) + \mathbf{f}(\mathbf{u}_i^n)) \cdot \mathbf{c}_{ij} - d_{ij}^n (\mathbf{u}_j^n - \mathbf{u}_i^n)] = \mathbf{0},$$

and also rewritten as

$$(3.26) \quad m_i \frac{\mathbf{u}_i^{n+1} - \mathbf{u}_i^n}{\Delta t} + \sum_{j \in \mathcal{N}_i} [(\mathbf{f}(\mathbf{u}_j^n) - \mathbf{f}(\mathbf{u}_i^n)) \cdot \mathbf{c}_{ij} - d_{ij}^n (\mathbf{u}_j^n - \mathbf{u}_i^n)] = \mathbf{0}.$$

Below are a few special cases of the finite element scheme (3.22).

EXAMPLE 3.9 (Linear Advection). Consider a linear scalar PDE, i.e.,  $\mathbf{u} = u$  is a scalar and  $\mathbf{f}(u) = u\mathbf{v}$  with a constant vector  $\mathbf{v}$ , then (3.26) reduces to

$$m_i \frac{u_i^{n+1} - u_i^n}{\Delta t} = \sum_{j \in \mathcal{N}_i} e_{ij} (u_j^n - u_i^n), \quad e_{ij} =: d_{ij}^n - \mathbf{v} \cdot \mathbf{c}_{ij},$$

which is a monotone scheme [141, 2] if  $e_{ij}$  is non-negative for  $j \neq i$ , under the CFL condition  $\frac{\Delta t}{m_i} \sum_{j \in \mathcal{N}_i, j \neq i} e_{ij} \leq 1$ . The non-negativity of  $e_{ij}$  ( $j \neq i$ ) can be achieved by taking

$$d_{ij}^n = \max\{0, \mathbf{v} \cdot \mathbf{c}_{ij}, \mathbf{v} \cdot \mathbf{c}_{ji}\}, \quad j \neq i,$$

which gives an upwind feature [130, 60, 138]. Such a scheme is a popular choice in algebraic flux correction (AFC) method for linear advection, which is naturally connected with a convection diffusion problem [22]. The condition  $e_{ij} \geq 0$ ,  $j \neq i$  is also used in local extremum diminishing schemes [116] and residual distribution schemes [2]. A truly multidimensional upwind method is the N scheme [186, 66, 65].

EXAMPLE 3.10 (1D Problem). For the finite element scheme (3.22) with a graph Laplacian viscosity as in Example 3.8, we can consider a uniform mesh of intervals for a 1D problem with the viscosity coefficients  $\alpha_{i+\frac{1}{2}} = \frac{1}{\Delta x} \alpha_{i+\frac{1}{2}}$  for the interval  $I_{i+\frac{1}{2}} = [x_i, x_{i+1}]$ , then the graph Laplacian is a tridiagonal matrix

$$L_{ij} = \begin{cases} -\frac{1}{2}\alpha_{i+\frac{1}{2}}, & j = i+1 \\ -\frac{1}{2}\alpha_{i-\frac{1}{2}}, & j = i-1, \\ \frac{1}{2}[\alpha_{i-\frac{1}{2}} + \alpha_{i+\frac{1}{2}}], & j = i \end{cases}$$

and (3.25) becomes a first order finite difference or finite volume scheme with a local Lax-Friedrichs flux:

$$\mathbf{u}_i^{n+1} = \mathbf{u}_i^n - \frac{\Delta t}{\Delta x} \left[ \frac{\mathbf{f}(\mathbf{u}_i^n) + \mathbf{f}(\mathbf{u}_{i+1}^n) - \alpha_{i+\frac{1}{2}}^n (\mathbf{u}_{i+1}^n - \mathbf{u}_i^n)}{2} - \frac{\mathbf{f}(\mathbf{u}_{i-1}^n) + \mathbf{f}(\mathbf{u}_i^n) - \alpha_{i-\frac{1}{2}}^n (\mathbf{u}_i^n - \mathbf{u}_{i-1}^n)}{2} \right].$$

Similarly, the finite element scheme (3.22) on a uniform mesh of intervals for a 1D problem with viscosity defined in Example 3.7 reduces to the first order scheme (3.12).

**3.6.4. Basic properties of the scheme.** We discuss some basic properties of (3.22).

**Global conservation.** Let  $\tilde{\mathbf{f}} = \sum_j \mathbf{f}(\mathbf{u}_j^n) \varphi_j(\mathbf{x}) \in V^h$ , then by summing  $i$  in (3.22), we obtain conservation in the following sense:

$$\frac{\iint_{\Omega} \mathbf{u}_h^{n+1} d\mathbf{x} - \iint_{\Omega} \mathbf{u}_h^n d\mathbf{x}}{\Delta t} = - \sum_i \iint_{T_i} \nabla \cdot \tilde{\mathbf{f}} d\mathbf{x} = - \sum_i \oint_{\partial T_i} \tilde{\mathbf{f}} \cdot \mathbf{n} ds = - \oint_{\partial \Omega} \tilde{\mathbf{f}} \cdot \mathbf{n} ds,$$

where we have used the fact that  $\tilde{\mathbf{f}} \cdot \mathbf{n}$  is continuous across each edge of triangles.

**Local conservation.** It is not very obvious in what sense (3.22) is locally conservative. As a matter of fact, it is proven in [192, 193] that the group finite element method (3.22) can be written as a finite volume scheme on the median dual mesh shown in Figure 3, see [193, Section 6], thus the general version of Lax-Wendroff Theorem in [194] can be applied to show the convergence to weak solutions. In the next subsection, we will also show that (3.22) is exactly the first order Lax-Friedrichs scheme defined on unstructured grids via a definition of residual distribution scheme [2, Section 2.3.1.2], thus a Lax-Wendroff Theorem for residual distribution schemes can also be used [18, 5, 14]. Another proof of Lax-Wendroff Theorem for the continuous finite element method was given in [136].

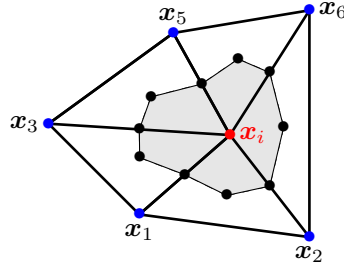


Fig. 3: The shaded polygon is the median dual cell  $C_i$  constructed by connecting triangle centroids to edges containing  $\mathbf{x}_i$ .

**Invariant domain.** Since we may regard the finite element scheme (3.22) as a Lax-Friedrichs type scheme on unstructured grids [2, Section 2.3.1.2], it is not a surprise that discussions in Example 3.3 and Example 3.4 can be applied here.

EXAMPLE 3.11 (Method 1). Since  $\mathbf{c}_{ii} = \mathbf{0}$ , the scheme (3.26) can be rewritten as

$$m_i \frac{\mathbf{u}_i^{n+1} - \mathbf{u}_i^n}{\Delta t} + \sum_{\substack{j \in \mathcal{N}_i \\ j \neq i}} [(\mathbf{f}(\mathbf{u}_j^n) - \mathbf{f}(\mathbf{u}_i^n)) \cdot \mathbf{c}_{ij} - d_{ij}^n (\mathbf{u}_j^n - \mathbf{u}_i^n)] = 0.$$

Since  $\sum_{j \in \mathcal{N}_i} d_{ij}^n = 0 \Rightarrow d_{ii}^n = - \sum_{\substack{j \in \mathcal{N}_i \\ j \neq i}} d_{ij}^n$ , we have

$$\sum_{j \in \mathcal{N}_i} d_{ij}^n (\mathbf{u}_j^n + \mathbf{u}_i^n) = \sum_{\substack{j \in \mathcal{N}_i \\ j \neq i}} d_{ij}^n (\mathbf{u}_j^n + \mathbf{u}_i^n) + \sum_{j=i} d_{ij}^n 2\mathbf{u}_i^n = \sum_{\substack{j \in \mathcal{N}_i \\ j \neq i}} d_{ij}^n (\mathbf{u}_j^n + \mathbf{u}_i^n) - 2\mathbf{u}_i^n \sum_{\substack{j \in \mathcal{N}_i \\ j \neq i}} d_{ij}^n,$$

thus the scheme can now be written as a convex combination,

$$\begin{aligned}\mathbf{u}_i^{n+1} &= \mathbf{u}_i^n \left( 1 - \sum_{\substack{j \in \mathcal{N}_i \\ j \neq i}} \frac{2\Delta t d_{ij}^n}{m_i} \right) + \sum_{\substack{j \in \mathcal{N}_i \\ j \neq i}} \frac{2\Delta t d_{ij}^n}{m_i} \bar{\mathbf{u}}_{ij}^{n+1}, \\ \bar{\mathbf{u}}_{ij}^{n+1} &= \frac{1}{2}(\mathbf{u}_j^n + \mathbf{u}_i^n) - [\mathbf{f}(\mathbf{u}_j^n) - \mathbf{f}(\mathbf{u}_i^n)] \cdot \frac{\mathbf{c}_{ij}}{2d_{ij}^n} = \frac{\mathbf{u}_j^n + \mathbf{u}_i^n}{2} - \frac{[\mathbf{f}(\mathbf{u}_j^n) - \mathbf{f}(\mathbf{u}_i^n)] \cdot \mathbf{n}_{ij}}{2\alpha_{ij}},\end{aligned}$$

where  $\mathbf{n}_{ij} = \frac{\mathbf{c}_{ij}}{|\mathbf{c}_{ij}|}$  and  $\alpha_{ij} = d_{ij}^n/|\mathbf{c}_{ij}|$ .

For [Assumption 4](#) to hold for compressible Euler equation (2.1) and invariant domain  $G_S$  in (2.5), we can take  $d_{ij}^n/|\mathbf{c}_{ij}| = \alpha_{ij}$  with a good estimate of the maximum wave speed  $\alpha_{ij}$ . One convenient estimate is

$$\alpha_{ij} = \max \{ |\mathbf{f}'(\mathbf{u}_i^n) \cdot \mathbf{n}_{ij}|, |\mathbf{f}'(\mathbf{u}_j^n) \cdot \mathbf{n}_{ij}| \}, \quad j \neq i,$$

where  $\mathbf{f}'(\mathbf{u}) \cdot \mathbf{n}$  denotes the Jacobian matrix of the flux along a given unit direction  $\mathbf{n}$  and  $|\mathbf{f}'(\mathbf{u}) \cdot \mathbf{n}|$  denotes its spectral radius. However, such an estimate may not ensure the IDP property in certain cases, see [\[101\]](#) for a way to rigorously estimate the maximum wave speed for compressible Euler equations. And the scheme is IDP since it is a convex combination of states in  $G_S$  under the CFL

$$\frac{\Delta t}{m_i}(-d_{ii}^n) = \frac{\Delta t}{m_i} \sum_{\substack{j \in \mathcal{N}_i \\ j \neq i}} d_{ij}^n \leq \frac{1}{2}.$$

EXAMPLE 3.12 (Method 2). By (3.24b) and the fact  $\mathbf{c}_{ii} = \mathbf{0}$ , the scheme (3.22) can also be rewritten as

$$m_i \frac{\mathbf{u}_i^{n+1} - \mathbf{u}_i^n}{\Delta t} + \sum_{\substack{j \in \mathcal{N}_i \\ j \neq i}} [\mathbf{f}(\mathbf{u}_j^n) \cdot \mathbf{c}_{ij} - d_{ij}^n(\mathbf{u}_j^n - \mathbf{u}_i^n)] = 0,$$

which is equivalent to

$$\mathbf{u}_i^{n+1} = \mathbf{u}_i^n \left( 1 - \sum_{\substack{j \in \mathcal{N}_i \\ j \neq i}} \frac{\Delta t d_{ij}^n}{m_i} \right) + \sum_{\substack{j \in \mathcal{N}_i \\ j \neq i}} \frac{\Delta t d_{ij}^n}{m_i} [\mathbf{u}_j^n - \alpha_{ij}^{-1} \mathbf{f}(\mathbf{u}_j^n) \cdot \mathbf{n}_{ij}],$$

where  $\mathbf{n}_{ij} = \frac{\mathbf{c}_{ij}}{|\mathbf{c}_{ij}|}$  and  $\alpha_{ij} = d_{ij}^n/|\mathbf{c}_{ij}|$ .

Therefore, to have  $\mathbf{u}_i^n \in G \Rightarrow \mathbf{u}_i^{n+1} \in G$  for compressible Euler equations with invariant domain (2.6), by Lemma 3.2, it suffices to take

$$\frac{\Delta t}{m_i}(-d_{ii}^n) = \frac{\Delta t}{m_i} \sum_{\substack{j \in \mathcal{N}_i \\ j \neq i}} d_{ij}^n \leq 1, \quad \frac{d_{ij}^n}{|\mathbf{c}_{ij}|} > \max_{\mathbf{u}_i^n, \mathbf{u}_j^n} \left( |\rho^{-1} \mathbf{m} \cdot \mathbf{n}_{ij}| + \sqrt{\frac{p^2}{2\rho^2 e}} \right).$$

**3.7. A special residual distribution scheme: the Lax–Friedrichs scheme on unstructured grids.** The concept of residual distribution schemes traces back to early work in 1980s [\[169\]](#), see also [\[199, 66, 65\]](#). Many schemes such as the streamline

diffusion method, the streamline upwind Petrov–Galerkin finite element methods and the cell vertex finite volume methods, as well as discontinuous Galerkin methods, can be rewritten as residual distribution schemes [5]. See [12, 183] for higher order accurate residual distribution schemes, [18, 2] for non-oscillatory residual distribution schemes and [17, 4] for entropy satisfying residual distribution schemes. Extensions to systems can be found in [8, 9, 3, 184]. Convergence including Lax-Wendroff Theorem for residual distribution schemes was discussed [18, 5, 14]. See [19] for higher order polynomial basis. In most references of residual distribution schemes, monotonicity for linear problems has been well studied, with which IDP can also be achieved using the methods reviewed in this section. As an example of illustrating the main ideas, we demonstrate how to construct  $P^1$  continuous finite element method (3.22) as a residual distribution scheme on a triangular mesh.

We use the same notation from the previous subsection. Let  $\mathbf{u}_i^n$  be the numerical solution value at each node  $\mathbf{x}_i$  at time step  $n$ . The dual cell in Figure 3 is constructed by connecting centroids of triangles to edges centers, which is also called median dual cell [192, 193]. Let  $|C_i|$  be the area of the dual cell volume of  $C_i$  around  $\mathbf{x}_i$ , then its area coincides with mass lumping in FEM:

$$|C_i| = \frac{1}{3} \sum_{T \ni \mathbf{x}_i} |T| = \iint_{\Omega} \varphi_i(\mathbf{x}) d\mathbf{x} = m_i.$$

A residual distribution scheme is of the form

$$(3.27) \quad |C_i| \frac{\mathbf{u}_i^{n+1} - \mathbf{u}_i^n}{\Delta t} = \sum_{T \ni \mathbf{x}_i} \phi_i^T,$$

where  $T \ni \mathbf{x}_i$  refers to summation over all triangles containing the given vertex  $\mathbf{x}_i$ , and  $\phi_i^T$  is the residual at  $\mathbf{x}_i$  for the triangle  $T$  satisfying

$$(3.28) \quad \sum_{\mathbf{x}_i \in T} \phi_i^T = - \oint_{\partial T} \widehat{\mathbf{f} \cdot \mathbf{n}} ds \approx - \oint_{\partial T} \mathbf{f}(\mathbf{u}) \cdot \mathbf{n} ds,$$

where  $\widehat{\mathbf{f} \cdot \mathbf{n}}$  is some numerical flux. In [5], it is shown that any scheme satisfying the local conservation relation (3.28) can be rewritten in flux form (algebraically equivalent), and the flux can be explicitly computed.

**3.7.1. The Lax-Friedrichs scheme on unstructured grids.** There are many methods to construct the residual for the same dual cell, e.g., see [1] for a finite volume approach. Here we give one special construction of  $\phi_i^T$  to recover exactly the same scheme as (3.22). By the Lax-Friedrichs scheme on unstructured grids in [2, Section 2.3.1.2], with the edge weight (3.23), We define the following residual

$$\begin{aligned} \phi_i^T = & -\frac{1}{3} \left[ \frac{\mathbf{f}(\mathbf{u}_i^n) + \mathbf{f}(\mathbf{u}_k^n)}{2} \cdot \mathbf{n}_{ik}^T |E_{ik}| - \frac{a_T |T|}{2} (\mathbf{u}_k^n - \mathbf{u}_i^n) \right] \\ & - \frac{1}{3} \left[ \frac{\mathbf{f}(\mathbf{u}_i^n) + \mathbf{f}(\mathbf{u}_j^n)}{2} \cdot \mathbf{n}_{ij}^T |E_{ij}| - \frac{a_T |T|}{2} (\mathbf{u}_j^n - \mathbf{u}_i^n) \right] - \frac{1}{3} \frac{\mathbf{f}(\mathbf{u}_j^n) + \mathbf{f}(\mathbf{u}_k^n)}{2} \cdot \mathbf{n}_{jk}^T |E_{jk}|. \end{aligned}$$

Here,  $(i, j, k)$  are the indices of the vertices of the triangle  $T$  as depicted in the right panel of Figure 2. Summing over all three vertices of a given triangle, we have

$$\begin{aligned}\phi^T &:= \sum_{\mathbf{x}_i \in T} \phi_i^T \\ &= -\frac{\mathbf{f}(\mathbf{u}_i^n) + \mathbf{f}(\mathbf{u}_k^n)}{2} \cdot \mathbf{n}_{ik}^T |E_{ik}| - \frac{\mathbf{f}(\mathbf{u}_i^n) + \mathbf{f}(\mathbf{u}_j^n)}{2} \cdot \mathbf{n}_{ij}^T |E_{ij}| - \frac{\mathbf{f}(\mathbf{u}_j^n) + \mathbf{f}(\mathbf{u}_k^n)}{2} \cdot \mathbf{n}_{jk}^T |E_{jk}| \\ &= -\oint_{\partial T} \widehat{\mathbf{f} \cdot \mathbf{n}} \, ds \approx -\oint_{\partial T} \mathbf{f}(\mathbf{u}) \cdot \mathbf{n} \, ds,\end{aligned}$$

where the flux is

$$\widehat{\mathbf{f} \cdot \mathbf{n}}|_{E_{ik}} = \frac{1}{2}[\mathbf{f}(\mathbf{u}_i^n) + \mathbf{f}(\mathbf{u}_k^n)] \cdot \mathbf{n}_{ik}^T.$$

By (3.24), in a triangle  $T$  with vertices  $\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k$ , we have  $\mathbf{c}_{ij}^T = -\frac{1}{6}\mathbf{n}_{ik}^T |E_{ik}|$ . With (3.19), we have  $\mathbf{c}_{ij}^T + \mathbf{c}_{ik}^T + \mathbf{c}_{ii}^T = \mathbf{0}$ , thus  $\mathbf{c}_{ij}^T = -\mathbf{c}_{ik}^T - \mathbf{c}_{ii}^T$ .

The residual can be rewritten as

$$\begin{aligned}\phi_i^T &= \frac{a_T |T|}{6}(\mathbf{u}_k^n - \mathbf{u}_i^n) + \frac{a_T |T|}{6}(\mathbf{u}_j^n - \mathbf{u}_i^n) \\ &\quad + [\mathbf{f}(\mathbf{u}_i^n) + \mathbf{f}(\mathbf{u}_k^n)] \cdot \mathbf{c}_{ij} + [\mathbf{f}(\mathbf{u}_i^n) + \mathbf{f}(\mathbf{u}_j^n)] \cdot \mathbf{c}_{ik} + [\mathbf{f}(\mathbf{u}_j^n) + \mathbf{f}(\mathbf{u}_k^n)] \cdot \mathbf{c}_{ii} \\ &= \frac{a_T |T|}{6}(\mathbf{u}_k^n - \mathbf{u}_i^n) + \frac{a_T |T|}{6}(\mathbf{u}_j^n - \mathbf{u}_i^n) \\ &\quad + [\mathbf{f}(\mathbf{u}_i^n) + \mathbf{f}(\mathbf{u}_k^n)] \cdot (-\mathbf{c}_{ik}^T - \mathbf{c}_{ii}^T) + [\mathbf{f}(\mathbf{u}_i^n) + \mathbf{f}(\mathbf{u}_j^n)] \cdot (-\mathbf{c}_{ij}^T - \mathbf{c}_{ii}^T) + [\mathbf{f}(\mathbf{u}_j^n) + \mathbf{f}(\mathbf{u}_k^n)] \cdot \mathbf{c}_{ii}^T \\ &= \frac{a_T |T|}{6}(\mathbf{u}_k^n - \mathbf{u}_i^n) + \frac{a_T |T|}{6}(\mathbf{u}_j^n - \mathbf{u}_i^n) \\ &\quad - [\mathbf{f}(\mathbf{u}_i^n) + \mathbf{f}(\mathbf{u}_k^n)] \cdot \mathbf{c}_{ik}^T - [\mathbf{f}(\mathbf{u}_i^n) + \mathbf{f}(\mathbf{u}_j^n)] \cdot \mathbf{c}_{ij}^T + 2\mathbf{f}(\mathbf{u}_i^n) \cdot \mathbf{c}_{ii}^T.\end{aligned}$$

Summing over all triangles  $T$  containing  $\mathbf{x}_i$ , with  $\mathbf{c}_{ii} = \mathbf{0}$ , by (3.24), we have

$$\begin{aligned}-\sum_{T \ni \mathbf{x}_i} \phi_i^T &= \sum_{T \ni \mathbf{x}_i} \left( \sum_{\substack{\mathbf{x}_j \in T \\ j \neq i}} [\mathbf{f}(\mathbf{u}_j^n) + \mathbf{f}(\mathbf{u}_i^n)] \cdot \mathbf{c}_{ij}^T - 2\mathbf{f}(\mathbf{u}_i^n) \cdot \mathbf{c}_{ii}^T - \frac{a_T |T|}{2} \sum_{\mathbf{x}_j \in T} (\mathbf{u}_j^n - \mathbf{u}_i^n) \right) \\ &= \sum_{\substack{j \in \mathcal{N}_i \\ j \neq i}} [\mathbf{f}(\mathbf{u}_j^n) + \mathbf{f}(\mathbf{u}_i^n)] \cdot \mathbf{c}_{ij} - 2\mathbf{f}(\mathbf{u}_i^n) \cdot \mathbf{c}_{ii} - \sum_{\substack{j \in \mathcal{N}_i \\ j \neq i}} w_{ij}(\mathbf{u}_j^n - \mathbf{u}_i^n) \\ &= \sum_{j \in \mathcal{N}_i} [\mathbf{f}(\mathbf{u}_j^n) + \mathbf{f}(\mathbf{u}_i^n)] \cdot \mathbf{c}_{ij} - \sum_{j \in \mathcal{N}_i} d_{ij}^n(\mathbf{u}_j^n - \mathbf{u}_i^n) = \sum_{j \in \mathcal{N}_i} [\mathbf{f}(\mathbf{u}_j^n) \cdot \mathbf{c}_{ij} - d_{ij}^n \mathbf{u}_j^n],\end{aligned}$$

where  $w_{ij}$  is defined in (3.23) and  $d_{ij}^n$  is the same graph Laplacian in Example 3.8. The residual distribution scheme (3.27) with such a residual is exactly the finite element scheme (3.22) with  $d_{ij}^n$  in Example 3.8. The Lax-Wendroff type Theorem can be proven to show convergence to weak solutions for residual distribution schemes [18, 5, 14, 15, 7]. Since an IDP finite element method (FEM) like (3.22) can be derived as a residual distribution scheme, the Lax-Wendroff Theorem for residual distribution schemes can be applied now to show its convergence to weak solutions. See also [136] for another proof of the Lax-Wendroff Theorem for FEM.

#### 4. Polynomial limiters for high order finite volume and DG schemes.

The Zhang-Shu approach introduced in [245, 246, 250] is a flexible IDP approach which can be easily applied to finite volume (FV) type and discontinuous Galerkin (DG) high order schemes. In this section, we review this approach and some recent related advances, e.g., [61, 62, 68].

**4.1. The main idea of the Zhang-Shu approach.** We demonstrate the basic idea for 1D problems. In a high order accurate finite volume (FV) scheme with forward Euler time stepping (1.5), the algorithmic structure follows these steps:

1. Given the cell averages  $\bar{\mathbf{u}}_j^n$  on intervals  $I_j$ .
2. Reconstruct a piecewise polynomial function  $\mathbf{u}_h^n(x)$  such that its cell average on  $I_j$  is  $\bar{\mathbf{u}}_j^n$ .
3. Evaluate the piecewise polynomial  $\mathbf{u}_h^n(x)$  at the cell ends  $x_{j+\frac{1}{2}}$  to obtain  $\mathbf{u}_{j+\frac{1}{2}}^\pm$ , using which (1.5) gives the cell averages at next time step  $\bar{\mathbf{u}}_j^{n+1}$ .

A high order DG scheme follows a similar algorithmic structure:

1. Given a piecewise polynomial solution  $\mathbf{u}_h^n(x)$  with cell averages equal to  $\bar{\mathbf{u}}_j^n$ .
2. Evolve the solution using a time discretization method to obtain  $\mathbf{u}_h^{n+1}(x)$  with the cell averages equal to  $\bar{\mathbf{u}}_j^{n+1}$ . In particular, with forward Euler time stepping, the cell average updates satisfy the same scheme (1.5).

For a given convex invariant domain  $G$ , instead of seeking  $\mathbf{u}_h^{n+1}(x) \in G$  for all  $x$ , the Zhang-Shu approach seeks to enforce the following IDP property in a finite volume or DG scheme IDP,

$$(4.1) \quad \mathbf{u}_h^n(x) \in G \quad \forall x \in \mathbb{S}_j, \forall j \implies \mathbf{u}_h^{n+1}(x) \in G \quad \forall x \in \mathbb{S}_j, \forall j,$$

where  $\mathbb{S}_j \subset I_j$  is a set of points for each cell  $I_j$  to be specified later. A flowchart for achieving (4.1) in a high order IDP finite volume or discontinuous Galerkin (DG) scheme follows three steps:

1. Start with  $\mathbf{u}_h^n(x)$ , which is high order accurate and satisfies

$$\bar{\mathbf{u}}_j^n \in G \quad \forall j, \quad \mathbf{u}_h^n(x) \in G \quad \forall x \in \mathbb{S}_j.$$

2. Evolve the solution forward in time to ensure

$$(4.2) \quad \bar{\mathbf{u}}_j^{n+1} \in G \quad \forall j.$$

This guarantees that the updated cell averages remain IDP, referred to as the **weak IDP** property. In general, such a step is nontrivial to achieve, which will be reviewed in this section.

3. Given the weak IDP condition (4.2), modify  $\mathbf{u}_h^{n+1}(x)$  without losing high order accuracy to enforce the **pointwise IDP at finitely many points**:

$$\mathbf{u}_h^{n+1}(x) \in G \quad \forall x \in \mathbb{S}_j,$$

which often can be enforced by a simple scaling limiter of polynomials.

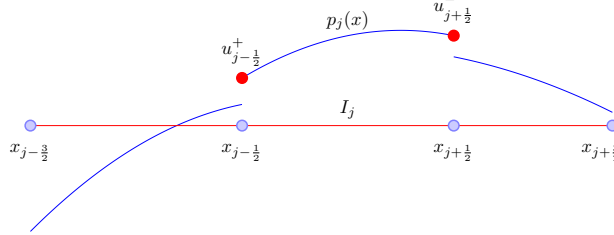
Most importantly, it can be proven that  $\mathbf{u}_h^n(x) \in G \quad \forall x \in \mathbb{S}_j$  is a sufficient condition to ensure (4.2), which holds for any high order finite volume scheme and DG scheme with an IDP numerical flux on any polygonal mesh in any dimension for a convex set  $G$ . This fact implies that any high order finite volume scheme and DG scheme with an IDP numerical flux using SSP time discretizations can be rendered IDP in the sense of (4.1), by adding a simple limiter to limit solution polynomials at some points within each cell, which allows not only easy implementation but also easy justification of the accuracy.

**4.2. One dimensional scalar conservation law.** We first demonstrate the method for solving 1D scalar conservation laws  $u_t + f(u)_x = 0$ . For convenience, we will first focus on the forward Euler time discretization, while high order time stepping methods will be discussed later in Subsection 4.2.4.

**4.2.1. Loss of monotonicity in high order schemes.** Let  $p_j(x)$  be a polynomial of degree  $k$  either evolved in a DG scheme or reconstructed in a FV method on a cell  $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$  with its cell average  $\bar{u}_j^n$ . The evolution equation of the cell averages in a high order FV or DG scheme can be written in a unified form as

$$(4.3) \quad \bar{u}_j^{n+1} = \bar{u}_j^n - \lambda \left( \hat{f}(u_{j+\frac{1}{2}}^-, u_{j+\frac{1}{2}}^+) - \hat{f}(u_{j-\frac{1}{2}}^-, u_{j-\frac{1}{2}}^+) \right) =: \hat{H}_\lambda \left( \bar{u}_j^n, u_{j+\frac{1}{2}}^-, u_{j+\frac{1}{2}}^+, u_{j-\frac{1}{2}}^-, u_{j-\frac{1}{2}}^+ \right),$$

where  $u_{j-\frac{1}{2}}^+ = p_j(x_{j-\frac{1}{2}})$ ,  $u_{j+\frac{1}{2}}^- = p_j(x_{j+\frac{1}{2}})$  are shown in the figure below.



Assume the numerical flux  $\hat{f}(\cdot, \cdot)$  is monotone as defined [Subsection 3.1](#), so that the corresponding first order scheme (3.2) is monotone under the CFL condition (3.3). By order barriers in [Subsection 1.2](#), the high order scheme (4.3) is in general not monotone. In particular, the function  $\hat{H}_\lambda$  is monotonically non-decreasing w.r.t.  $\bar{u}_j^n$  and  $u_{j+\frac{1}{2}}^+$ ,  $u_{j-\frac{1}{2}}^-$ , but non-increasing w.r.t.  $u_{j+\frac{1}{2}}^-$ ,  $u_{j-\frac{1}{2}}^+$ . This induces one challenge for achieving (4.2), which is explained by the following simple example.

**EXAMPLE 4.1.** For the linear advection equation  $u_t + u_x = 0$ , the high order scheme (4.3) with the upwind flux reduces to

$$(4.4) \quad \bar{u}_j^{n+1} = \bar{u}_j^n - \lambda \left( u_{j+\frac{1}{2}}^- - u_{j-\frac{1}{2}}^- \right),$$

which is decreasing w.r.t.  $u_{j+\frac{1}{2}}^-$ . Assume  $G := [m, M] = [0, 1]$ ,  $\bar{u}_j^n = u_{j-\frac{1}{2}}^- = 0$ , and  $u_{j+\frac{1}{2}}^- = 1$ , then  $\bar{u}_j^{n+1} = -\lambda < 0$  for any time step  $\Delta t > 0$ .

**4.2.2. Weak monotonicity of high order schemes.** From [Example 4.1](#), we can see that the loss of monotonicity in the high order scheme (4.3) implies that the scheme (4.3) may fail to preserve the bounds for any positive time step  $\Delta t > 0$  even if  $\bar{u}_j^n, u_{j+\frac{1}{2}}^-, u_{j+\frac{1}{2}}^+, u_{j-\frac{1}{2}}^-, u_{j-\frac{1}{2}}^+$  are within the desired bounds.

In other words, since the scheme (4.3) has the property  $\bar{u}_j^{n+1} = \hat{H}_\lambda(\uparrow, \downarrow, \uparrow, \uparrow, \downarrow)$ , requiring all of its input values to be in  $G = [m, M]$  is not enough to achieve  $\bar{u}_j^{n+1} \in G$ . Even though  $\bar{u}_j^{n+1} = \hat{H}_\lambda(\uparrow, \downarrow, \uparrow, \uparrow, \downarrow)$  is not a monotone function w.r.t. independent degree of freedoms  $\bar{u}_j^n, u_{j+\frac{1}{2}}^-, u_{j+\frac{1}{2}}^+, u_{j-\frac{1}{2}}^-, u_{j-\frac{1}{2}}^+$ , the key observation by Zhang and Shu in [245] is that (4.3) can still be rewritten as a monotone function w.r.t. some point values, which might be dependent degree of freedoms in general.

Notice that  $\hat{H}_\lambda$  is decreasing only w.r.t.  $u_{j+\frac{1}{2}}^-, u_{j-\frac{1}{2}}^+$ , which are degree of freedoms within the cell  $I_j$  thus can be controlled by  $\bar{u}_j^n$  if the cell average  $\bar{u}_j^n$  is decomposed into a convex combination of several point values including  $u_{j+\frac{1}{2}}^-, u_{j-\frac{1}{2}}^+$ . Such a decomposition can be achieved via  $L$ -point Gauss-Lobatto quadrature, which is exact for integrating polynomials of degree  $k$  with positive quadrature weights if  $L \geq \frac{k+3}{2}$ .

This implies

$$(4.5) \quad \bar{u}_j^n = \frac{1}{\Delta x} \int_{I_j} p_j(x) dx = \sum_{\mu=1}^L \omega_\mu p_j(x_j^{(\mu)}) = \sum_{\mu=2}^{L-1} \omega_\mu p_j(x_j^{(\mu)}) + \omega_1 u_{j-\frac{1}{2}}^+ + \omega_L u_{j+\frac{1}{2}}^-,$$

where  $\omega_\mu > 0$  are the Gauss–Lobatto quadrature weights for the interval  $[-\frac{1}{2}, \frac{1}{2}]$  satisfying  $\sum_{\mu} \omega_\mu = 1$  with  $\omega_1 = \omega_L = \frac{1}{L(L-1)}$ , and  $\{x_j^{(\mu)}\}$  are the quadrature nodes for  $I_j$  with  $x_j^{(1)} = x_{j-\frac{1}{2}}$  and  $x_j^{(L)} = x_{j+\frac{1}{2}}$ . Then the high order scheme (4.3) using the cell average decomposition (4.5) can be rewritten as follows,

$$(4.6) \quad \begin{aligned} \bar{u}_j^{n+1} &= \sum_{\mu=2}^{L-1} \omega_\mu p_j(x_j^{(\mu)}) + \omega_L \left( u_{j+\frac{1}{2}}^- - \frac{\lambda}{\omega_L} \left( \hat{f}(u_{j+\frac{1}{2}}^-, u_{j+\frac{1}{2}}^+) - \hat{f}(u_{j-\frac{1}{2}}^+, u_{j+\frac{1}{2}}^-) \right) \right) \\ &\quad + \omega_1 \left( u_{j-\frac{1}{2}}^+ - \frac{\lambda}{\omega_1} \left( \hat{f}(u_{j-\frac{1}{2}}^+, u_{j+\frac{1}{2}}^-) - \hat{f}(u_{j-\frac{1}{2}}^-, u_{j+\frac{1}{2}}^+) \right) \right) \\ &= \sum_{\mu=2}^{L-1} \omega_\mu p_j(x_j^{(\mu)}) + \omega_L H_{\frac{\lambda}{\omega_L}} \left( u_{j-\frac{1}{2}}^+, u_{j+\frac{1}{2}}^-, u_{j+\frac{1}{2}}^+ \right) + \omega_1 H_{\frac{\lambda}{\omega_1}} \left( u_{j-\frac{1}{2}}^-, u_{j-\frac{1}{2}}^+, u_{j+\frac{1}{2}}^- \right), \end{aligned}$$

which is a convex combination of

$$p_j(x_j^{(\mu)}), \quad H_{\frac{\lambda}{\omega_L}} \left( u_{j-\frac{1}{2}}^+, u_{j+\frac{1}{2}}^-, u_{j+\frac{1}{2}}^+ \right), \quad H_{\frac{\lambda}{\omega_1}} \left( u_{j-\frac{1}{2}}^-, u_{j-\frac{1}{2}}^+, u_{j+\frac{1}{2}}^- \right).$$

Recall that  $H_\lambda(\cdot, \cdot, \cdot)$ , defined in the first order scheme (3.2), is monotonically non-decreasing in all its three arguments under the CFL condition (3.3). Thus,  $H_{\frac{\lambda}{\omega_1}}$  and  $H_{\frac{\lambda}{\omega_L}}$  are monotonically non-decreasing under a reduced CFL condition

$$(4.7) \quad \lambda\alpha \leq \omega := \omega_1 = \omega_L = \frac{1}{L(L-1)}.$$

Therefore,  $\bar{u}_j^{n+1}$  is a monotonically non-decreasing function of

$$u_{j-\frac{1}{2}}^-, \quad u_{j+\frac{1}{2}}^+, \quad p_j(x_j^{(\mu)}), \quad 1 \leq \mu \leq N.$$

It can be stated as the following theorem, as a sufficient condition for achieving (4.2).

**THEOREM 4.1** (Weak monotonicity of high order schemes). *For a finite volume scheme or the scheme satisfied by the cell averages of the DG method with forward Euler time discretization (4.3) using a monotone flux  $\hat{f}$ , let  $p_j(x)$  be the reconstructed or DG solution polynomial of degree  $k$  satisfying*

$$\bar{u}_j^n = \frac{1}{\Delta x} \int_{I_j} p_j(x) dx, \quad u_{j-\frac{1}{2}}^+ = p_j \left( x_{j-\frac{1}{2}} \right) \quad \text{and} \quad u_{j+\frac{1}{2}}^- = p_j \left( x_{j+\frac{1}{2}} \right).$$

*Then under the CFL condition (4.7),  $\bar{u}_j^{n+1}$  is monotone w.r.t.*

$$u_{j-\frac{1}{2}}^- = p_{j-1}(x_{j-\frac{1}{2}}), \quad u_{j+\frac{1}{2}}^+ = p_{j+1}(x_{j+\frac{1}{2}}), \quad p_j(x_j^{(\mu)}), \quad \mu = 1, \dots, L$$

*Therefore, with (4.7), a sufficient condition for  $\bar{u}_j^{n+1} \in G := [m, M]$  for all  $j$  is*

$$(4.8) \quad p_j(x_j^{(\mu)}) \in G, \quad 1 \leq \mu \leq N \quad \forall j.$$

**Theorem 4.1** can be regarded as a complementary result to the classical Godunov Theorem. In a high order accurate linear finite volume scheme (4.3),  $\bar{u}_j^{n+1}$  is not monotone w.r.t.  $\bar{u}_j^n$ , but  $\bar{u}_j^{n+1}$  can still be a monotone function w.r.t. some quadrature point values.

**REMARK 4.1** (Optimality of Gauss–Lobatto quadrature). *The cell average decomposition (4.5) based on the Gauss–Lobatto quadrature plays a critical role in revealing the weak monotonicity. In fact, any quadrature rule with positive weights and nodes including the two endpoints:*

$$(4.9) \quad \frac{1}{\Delta x} \int_{I_j} p(x) dx = \sum_{\mu=2}^{N-1} \omega_\mu p(x_j^{(\mu)}) + \omega_1 p(x_{j-\frac{1}{2}}) + \omega_N p(x_{j+\frac{1}{2}})$$

is applicable, as long as it is exact for integrating polynomials of degree  $k$ . Different feasible quadrature rules would lead to different IDP CFL conditions, for example, (4.9) leads to  $\lambda\alpha \leq \min\{\omega_1, \omega_N\}$ . It is highly desirable to choose the optimal feasible quadrature to maximize the CFL number  $\min\{\omega_1, \omega_N\}$ . The Gauss–Lobatto quadrature was proven to be the optimal one in the sense of the largest CFL in 1D [62].

**REMARK 4.2.** *For high order DG methods, the CFL condition (4.7) is close to the CFL condition needed for linear stability. For high order FV schemes, the CFL condition (4.7) is usually smaller than the commonly used ones, e.g., for FV WENO schemes. On the other hand, (4.7) is only a convenient sufficient condition but not a necessary condition for achieving (4.2).*

**4.2.3. A simple scaling limiter for enforcing pointwise bounds.** The weak monotonicity in **Theorem 4.1** indicates that a high order conservative finite volume or DG scheme (4.3) preserves  $\bar{u}_j^{n+1} \in [m, M]$  if the approximation polynomials  $p_j(x)$  satisfy (4.8), which can be enforced by the following simple limiter for a polynomial with cell average  $\bar{u}_j^n \in [m, M]$ :

$$(4.10a) \quad \tilde{p}_j(x) = \theta(p_j(x) - \bar{u}_j^n) + \bar{u}_j^n,$$

$$(4.10b) \quad \theta = \min \left\{ \left| \frac{M - \bar{u}_j^n}{M_j - \bar{u}_j^n} \right|, \left| \frac{m - \bar{u}_j^n}{m_j - \bar{u}_j^n} \right|, 1 \right\}, M_j = \max_{x \in \mathbb{S}_j} p_j(x), m_j = \min_{x \in \mathbb{S}_j} p_j(x).$$

The limiter is a simplified version of those used in [25, 155], and it satisfies the following properties.

*Conservation.* Since  $\tilde{p}_j(x)$  is a convex combination of  $p_j(x)$  with its own cell average,  $\tilde{p}_j(x)$  has the same cell average  $\bar{u}_j^n$  on  $I_j$ .

*Bounds at  $\mathbb{S}_j$ .* By the definition of  $\theta$  in (4.10),  $\tilde{p}_j(x)$  satisfies (4.8), if the cell average is within the bounds  $\bar{u}_j^n \in [m, M]$ .

*Easiness of implementation.* This limiter is local to each cell and only requires point values at  $\mathbb{S}_j$ , thus its implementation is easy and friendly for parallel computing.

*High order accuracy.* This limiter does not destroy the high order approximation accuracy of  $p_j(x)$  under suitable assumptions.

**THEOREM 4.2** (Accuracy of Zhang–Shu limiter). *Assume the cell average of  $p_j(x)$  is  $\bar{u}_j^n \in [m, M]$ , then for any function  $u(x) \in [m, M]$ , the limiter (4.10) for polynomials of degree  $k$  satisfies*

$$|p_j(x) - \tilde{p}_j(x)| \leq C_k \max_{x \in I_j} |p_j(x) - u(x)|,$$

where  $C_k$  is a constant that depends only on the polynomial degree  $k$ .

REMARK 4.3. If replacing  $M_j$  and  $m_j$  in (4.10) by the maximum and minimum of  $p_j(x)$  in the cell  $I_j$ , then it is a more restrictive thus less accurate limiter. For such a more restrictive limiter on a reference cell of any shape in any dimension, the same result holds with  $C_k$  depending only on the polynomial degree  $k$  and the reference cell.

*Proof.* We review the key arguments of the proof here. Without loss of generality, we only need to discuss the case that  $p_j(x)$  is not a constant and  $\theta = \frac{M - \bar{u}_j^n}{M_j - \bar{u}_j^n}$  with  $M_j > M$ . For convenience, let  $\bar{p}_j = \bar{u}_j^n$ , then  $\tilde{p}_j(x) - p_j(x) = (M - M_j) \frac{p_j(x) - \bar{p}_j}{M_j - \bar{p}_j}$ .

First,  $|M - M_j| \leq \max_{x \in I_j} |p_j(x) - u(x)|$  because  $\max_{x \in \mathbb{S}_j} p_j(x) = M_j > M \geq u(x)$ , see [234]. Thus we only need to prove  $\left| \frac{p_j(x) - \bar{p}_j}{M_j - \bar{p}_j} \right| \leq C_k$ . Define  $q(\xi) = p_j\left(\xi \Delta x + x_{j-\frac{1}{2}}\right) - \bar{p}_j$  with  $\xi \in [0, 1]$ , then  $\bar{q} = \int_0^1 q(\xi) d\xi = 0$ ,  $\max_{\xi \in [0, 1]} q(\xi) = \max_{x \in I_j} p_j(x) - \bar{p}_j$  and  $\min_{\xi \in [0, 1]} q(\xi) = \min_{x \in I_j} p_j(x) - \bar{p}_j$ . We have

$$\left| \frac{p_j(x) - \bar{p}_j}{M_j - \bar{p}_j} \right| \leq \frac{|q(\xi)|}{\max_{\xi \in [0, 1]} q(\xi)} \leq \frac{\max_{\xi \in [0, 1]} |q(\xi)|}{\max_{\xi \in [0, 1]} q(\xi)}.$$

Thus we only need to prove  $\frac{\max_{\xi \in [0, 1]} |q(\xi)|}{\max_{\xi \in [0, 1]} q(\xi)} \leq C_k$  for any polynomial of degree  $k$  satisfying

$\int_0^1 q(\xi) d\xi = 0$ . For quadratic polynomials in one dimension,  $C_2 = 3$  was proven by explicit calculations in [155]. For higher order polynomials in one dimension,  $C_k \leq (k^2 + k - 1)\Lambda_{k+1}[0, 1]$ , where  $\Lambda_{k+1}[0, 1]$  is the Lebesgue constant on the interval  $[0, 1]$ , see [242, Lemma 7]. For general  $k$  and higher dimensions, the existence of the constant  $C_k$  can be established by an abstract proof similar to proving the equivalence of two norms in a finite-dimensional Banach space [242, Lemma 8], which can be used to prove the multi-dimensional case as mentioned Remark 4.3.  $\square$

**4.2.4. The bound-preserving algorithm flowchart.** Assuming  $\bar{u}_j^n \in [m, M]$ , then using the simple limiter (4.10) at time step  $n$  can achieve the sufficient condition in Theorem 4.1 to ensure  $\bar{u}_j^{n+1} \in [m, M]$ , with which the simple limiter (4.10) can again enforce bounds at time step  $n + 1$ .

Such a method can be easily extended from forward Euler to high order strong stability preserving (SSP) explicit Runge–Kutta or multistep methods [91, 90], which are convex combinations of forward Euler steps. For a semi-discrete scheme  $\frac{d}{dt}u_h = \mathcal{L}(u_h)$ , e.g.,  $\mathcal{L}$  denotes high order spatial discretization, the classic third order explicit SSP Runge–Kutta method is

$$\begin{aligned} u_h^{n,*} &= u_h^n + \Delta t \mathcal{L}(u_h^n), \\ (4.11) \quad u_h^{n,**} &= \frac{3}{4}u_h^n + \frac{1}{4}(u_h^{n,*} + \Delta t \mathcal{L}(u_h^{n,*})), \\ u_h^{n+1} &= \frac{1}{3}u_h^n + \frac{2}{3}(u_h^{n,**} + \Delta t \mathcal{L}(u_h^{n,**})), \end{aligned}$$

and the explicit SSP third order multistep method is

$$u_h^{n+1} = \frac{16}{27}(u_h^n + 3\Delta t \mathcal{L}(u_h^n)) + \frac{11}{27}\left(u_h^{n-3} + \frac{12}{11}\Delta t \mathcal{L}(u_h^{n-3})\right),$$

where  $u_h^n$  represents the numerical solution at the  $n$ -th time step.

To construct a high order accurate bound-preserving scheme, we can use SSP high order time discretizations with high order FV or DG methods in space with a monotone numerical flux, with the limiter (4.10) applied in each time step in a SSP multistep methods or each time stage in a SSP RK method. Then under suitable CFL conditions, Theorem 4.1 and Theorem 4.2 imply that the full scheme is conservative, bound-preserving and high order accurate.

The main advantages of such a method include easy extensions to systems and higher dimensions, easy implementation and easy justification of accuracy, which is due to not only Theorem 4.2 but also, more importantly, the fact that this approach is built upon an intrinsic weak monotonicity property of the high order spatial discretization (4.3).

On the other hand, although the weak monotonicity can be established for FV and DG schemes, in general it does not hold for high order finite difference (FD) schemes. For special compact FD schemes, weak monotonicity may hold and bound-preserving schemes can be constructed [145].

**4.2.5. A simplified weak monotonicity and limiter.** The limiter (4.10) involves the polynomial  $p_j(x)$  which is not available in ENO (essentially non-oscillatory) and WENO (weighted ENO) finite volume reconstructions. One way is to use interpolation to construct an approximation polynomial  $p_j(x)$  in ENO and WENO schemes to apply (4.10). An easier alternative provided in [247] is to avoid explicitly using point values  $p_j(x_j^{(\mu)})$  for  $\mu = 2, \dots, N-1$ . Since  $\sum_{\mu=2}^{N-1} \frac{\omega_\mu}{1-2\omega_1} p_j(x_j^{(\mu)})$  is a convex combination of point values  $p_j(x_j^{(\mu)})$  for  $\mu = 2, \dots, N-1$ , by the Mean Value Theorem, there exists some point  $x_j^* \in I_j$  such that  $\sum_{\mu=2}^{N-1} \frac{\omega_\mu}{1-2\omega_1} p_j(x_j^{(\mu)}) = p_j(x_j^*)$ . We can rewrite (4.6) as

$$\bar{u}_j^{n+1} = (1-2\omega_1)p_j(x_j^*) + \omega_1 H_{\frac{\Delta}{\omega_1}}(u_{j-\frac{1}{2}}^+, u_{j+\frac{1}{2}}^-, u_{j+\frac{1}{2}}^+) + \omega_1 H_{\frac{\Delta}{\omega_1}}(u_{j-\frac{1}{2}}^-, u_{j-\frac{1}{2}}^+, u_{j+\frac{1}{2}}^-),$$

thus in Theorem 4.1 we can use the following weaker sufficient condition to replace (4.8),

$$(4.12) \quad u_{j-\frac{1}{2}}^\pm, u_{j+\frac{1}{2}}^\pm, \frac{\bar{u}_j^n - \omega_1 u_{j-\frac{1}{2}}^+ - \omega_1 u_{j+\frac{1}{2}}^-}{1-2\omega_1} = \sum_{\mu=2}^{N-1} \frac{\omega_\mu}{1-2\omega_1} p_j(x_j^{(\mu)}) = p_j(x_j^*) \in [m, M].$$

A simplified limiter to enforce (4.12) in ENO and WENO finite volume schemes is

$$(4.13a) \quad \tilde{p}_j(x) = \theta (p_j(x) - \bar{p}_j) + \bar{p}_j, \quad \theta = \min \left\{ 1, \left| \frac{M - \bar{p}_j}{M_j - \bar{p}_j} \right|, \left| \frac{m - \bar{p}_j}{m_j - \bar{p}_j} \right| \right\},$$

(4.13b)

$$M_j = \max_{x_j^{(1)}, x_j^{(N)}, x_j^*} p_j(x), m_j = \min_{x_j^{(1)}, x_j^{(N)}, x_j^*} p_j(x), \quad p_j(x_j^*) = \frac{\bar{u}_j^n - \omega_1 u_{j-\frac{1}{2}}^+ - \omega_1 u_{j+\frac{1}{2}}^-}{1-2\omega_1}.$$

Theorem 4.2 still applies to (4.13) since it is a more relaxed limiter than (4.10).

REMARK 4.4. If (4.12) is replaced by requiring  $u_{j-\frac{1}{2}}^\pm, u_{j+\frac{1}{2}}^\pm, \frac{\bar{u}_j^n - a u_{j-\frac{1}{2}}^+ - a u_{j+\frac{1}{2}}^-}{1-a} \in [m, M]$  with  $a \in (0, \frac{1}{2})$ , then this still provides a sufficient condition for ensuring  $\bar{u}_j^{n+1} \in [m, M]$ , as first proven in [176]. However, it is difficult to justify the accuracy

of enforcing  $\frac{\bar{u}_j^n - au_{j-\frac{1}{2}}^+ - au_{j+\frac{1}{2}}^-}{1-a} \in [m, M]$ , unless  $a$  corresponds to a quadrature weight such that  $\frac{\bar{u}_j^n - au_{j-\frac{1}{2}}^+ - au_{j+\frac{1}{2}}^-}{1-a} = p_j(x_j^*)$ , which allows one to invoke [Theorem 4.2](#).

**4.3. One dimensional hyperbolic systems.** We now discuss the extension to 1D systems of the form (3.5) with a convex invariant domain  $G$  defined in (1.2). Consider the evolution equation of the cell averages for a high order finite volume or DG scheme for the system (3.5), which can be written as (1.5) with a numerical flux  $\hat{\mathbf{f}}(\cdot, \cdot)$ . Assume  $\hat{\mathbf{f}}$  is IDP as defined in [Definition 3.1](#) under a CFL condition  $\lambda\alpha \leq c_0$ .

**4.3.1. The weak IDP property of high order schemes.** The monotonicity is no longer directly applicable for hyperbolic systems. Instead, we can use either the convex decomposition technique [245, 246] or the GQL approach [222]. In particular, let  $\mathbf{p}_j(x)$  be the approximation polynomial in  $I_j$ , then (4.6) implies that a high order scheme (1.5) is a convex combination of two formal first order IDP schemes:

$$\begin{aligned} \bar{\mathbf{u}}_j^{n+1} &= \sum_{\mu=2}^{L-1} \omega_\mu \mathbf{p}_j(x_j^{(\mu)}) + \omega_L \left( \mathbf{u}_{j+\frac{1}{2}}^- - \frac{\lambda}{\omega_L} \left( \hat{\mathbf{f}}(\mathbf{u}_{j+\frac{1}{2}}^-, \mathbf{u}_{j+\frac{1}{2}}^+) - \hat{\mathbf{f}}(\mathbf{u}_{j-\frac{1}{2}}^+, \mathbf{u}_{j+\frac{1}{2}}^-) \right) \right) \\ &\quad + \omega_1 \left( \mathbf{u}_{j-\frac{1}{2}}^+ - \frac{\lambda}{\omega_1} \left( \hat{\mathbf{f}}(\mathbf{u}_{j-\frac{1}{2}}^+, \mathbf{u}_{j+\frac{1}{2}}^-) - \hat{\mathbf{f}}(\mathbf{u}_{j-\frac{1}{2}}^-, \mathbf{u}_{j-\frac{1}{2}}^+) \right) \right) \\ &= \sum_{\mu=2}^{L-1} \omega_\mu \mathbf{p}_j(x_j^{(\mu)}) + \omega_L \mathbf{H}_{\frac{\lambda}{\omega_L}} \left( \mathbf{u}_{j-\frac{1}{2}}^+, \mathbf{u}_{j+\frac{1}{2}}^-, \mathbf{u}_{j+\frac{1}{2}}^+ \right) + \omega_1 \mathbf{H}_{\frac{\lambda}{\omega_1}} \left( \mathbf{u}_{j-\frac{1}{2}}^-, \mathbf{u}_{j-\frac{1}{2}}^+, \mathbf{u}_{j+\frac{1}{2}}^- \right), \end{aligned}$$

Since  $\mathbf{H}$  is a first order scheme thus is IDP under suitable CFL, e.g.,

$$\mathbf{u}_{j-\frac{1}{2}}^+, \mathbf{u}_{j+\frac{1}{2}}^-, \mathbf{u}_{j+\frac{1}{2}}^+ \in G \Rightarrow \mathbf{H}_{\frac{\lambda}{\omega_L}} \left( \mathbf{u}_{j-\frac{1}{2}}^+, \mathbf{u}_{j+\frac{1}{2}}^-, \mathbf{u}_{j+\frac{1}{2}}^+ \right) \in G, \quad \text{if } \frac{\lambda}{\omega_L} \alpha \leq c_0.$$

thus [Theorem 4.1](#) can be extended as follows.

**THEOREM 4.3 (Weak IDP property).** *For a finite volume scheme or the scheme satisfied by the cell averages of the DG method in the form (1.5) using an IDP flux  $\hat{\mathbf{f}}$ , with an approximation polynomial vector  $\mathbf{p}_j(x)$  of degree  $k$  satisfying*

$$\bar{\mathbf{u}}_j^n = \frac{1}{\Delta x} \int_{I_j} \mathbf{p}_j(x) dx, \quad \mathbf{u}_{j-\frac{1}{2}}^+ = \mathbf{p}_j \left( x_{j-\frac{1}{2}} \right) \quad \text{and} \quad \mathbf{u}_{j+\frac{1}{2}}^- = \mathbf{p}_j \left( x_{j+\frac{1}{2}} \right),$$

then  $\bar{\mathbf{u}}_j^{n+1} \in G$  under the CFL condition  $\lambda\alpha \leq \omega c_0$  with  $\omega = \frac{1}{L(L-1)}$ , if

$$(4.14) \quad \mathbf{p}_{j-1}(x_{j-\frac{1}{2}}) = \mathbf{u}_{j-\frac{1}{2}}^-, \quad \mathbf{p}_{j+1}(x_{j+\frac{1}{2}}) = \mathbf{u}_{j+\frac{1}{2}}^+, \quad \mathbf{p}_j(x_j^{(\mu)}) \in G, \quad 1 \leq \mu \leq L.$$

**4.3.2. A simple scaling limiter.** Similar to the scalar case, given  $\mathbf{p}_j(x)$  with a cell average  $\bar{\mathbf{u}}_j^n \in G$ , for enforcing (4.14), on each cell  $I_j$  a simple scaling limiter can be designed as follows:

$$(4.15) \quad \tilde{\mathbf{p}}_j(x) = \theta(\mathbf{p}_j(x) - \bar{\mathbf{u}}_j^n) + \bar{\mathbf{u}}_j^n,$$

$$\theta = \min_{\mu} \{\theta_j^{(\mu)}\}, \quad \text{with} \quad \theta_j^{(\mu)} = \begin{cases} 1, & \text{if } \mathbf{p}_j(x_j^{(\mu)}) \in G, \\ \frac{|\mathbf{u}_* - \bar{\mathbf{u}}_j^n|}{|\mathbf{p}_j(x_j^{(\mu)}) - \bar{\mathbf{u}}_j^n|}, & \text{otherwise,} \end{cases}$$

where  $\mathbf{u}_*$  is the intersection point of  $\partial G$  and the line segment that connects  $\bar{\mathbf{u}}_j^n \in G$  and  $\mathbf{p}_j(x_j^{(\mu)}) \notin G$ .

In many cases, the computation of  $\theta_j^{(\mu)}$  is cumbersome and may require a root-finding procedure. If the invariant domain can be reformulated such that all the functions  $g_i(\mathbf{u})$  in (1.2) are linear or concave with respect to  $\mathbf{u}$ , then a different but easier alternative of  $\theta$  can be considered,

$$\theta = \min_i \{\theta_i\}, \quad \text{with} \quad \theta_i = \min \left\{ \left| \frac{g_i(\bar{\mathbf{u}}_j^n) - \varepsilon_i}{g_i(\bar{\mathbf{u}}_j^n) - \min_{\mu} g_i(\mathbf{p}_j(x_j^{(\mu)}))} \right|, 1 \right\},$$

where the small number  $\varepsilon_i \geq 0$  is introduced to mitigate the effect of round-off errors. Jensen inequality for concave functions  $g_i(\mathbf{u})$  implies  $\tilde{\mathbf{p}}_j(x) \in G \quad \forall x \in \mathbb{S}_j$ , e.g., see [246, 247, 248, 213, 242] for compressible Euler equations and [227] for shallow water equations. Such a limiter does not increase entropy for compressible Euler equations [46, 147].

A simplified limiter for finite volume schemes without directly using reconstruction polynomial  $\mathbf{p}_j(x)$  can also be constructed, similar to Subsection 4.2.5, see [247, Section 5] for details.

In practice, one may want to preserve the invariant domain of the numerical solution at more points, such as Gauss quadrature points used in DG methods. It can be easily enforced by the same limiter (4.15) at these points.

**4.3.3. Robust and efficient implementations.** The full algorithm flowchart follows similarly as in Subsection 4.2.4. Since the exact solutions to scalar conservation laws satisfy the maximum principle, it is not difficult to estimate the maximum wave speed  $\alpha$  in Definition 3.1 for scalar conservation laws. However, for a hyperbolic system, the maximum wave speed may grow in time. It is nontrivial to enforce the CFL needed for Theorem 4.3 in each time stage of a SSP Runge-Kutta method by computing a time step  $\Delta t$  based only on information of  $\mathbf{u}^n$ .

On the other hand, the CFL condition in Theorem 4.3 is only a sufficient but not a necessary condition for achieving  $\bar{\mathbf{u}}_j^{n+1} \in G$  thus one convenient and efficient implementation is described by the following two steps.

First, use a SSP Runge-Kutta method with a high order finite volume or DG method in space using an IDP numerical flux, e.g., the Lax-Friedrichs flux

$$\hat{\mathbf{f}}(\mathbf{u}_{j+\frac{1}{2}}^-, \mathbf{u}_{j+\frac{1}{2}}^+) = \frac{1}{2}[\mathbf{f}(\mathbf{u}_{j+\frac{1}{2}}^-) + \mathbf{f}(\mathbf{u}_{j+\frac{1}{2}}^+) - \alpha_{j+\frac{1}{2}}(\mathbf{u}_{j+\frac{1}{2}}^+ - \mathbf{u}_{j+\frac{1}{2}}^-)], \alpha_{j+\frac{1}{2}} = \max \left| \mathbf{f}'(\mathbf{u}_{j+\frac{1}{2}}^\pm) \right|.$$

Second, use any commonly used time step in a SSP Runge-Kutta method to evolve from time step  $n$  to step  $n+1$ , with the simple limiter (4.15) used on each time stage. If  $\bar{\mathbf{u}}_j^{n+1} \notin G$  happens at time step  $n+1$  or any inner time step of the Runge-Kutta method, then it means that the CFL condition in Theorem 4.3 is not met at that time stage or time step, thus go back to time step  $n$  to recompute with halved time step. See [213, 242] for more details.

Notice that the implementation of recomputing with halved time step whenever invariant domain is violated can be used for any numerical scheme, which however may result in an infinite loop of recomputing, e.g., a high order with any positive time step can violate bounds in Example 4.1. Only when the simple limiter (4.15) is added to control point values at the proper locations, this is not an infinite loop because Theorem 4.3 ensures  $\bar{\mathbf{u}}_j^{n+1} \in G$  when time step is small enough.

REMARK 4.5. *The Zhang-Shu approach can be used with any finite volume and DG methods to construct high order IDP schemes. For finite difference schemes, this approach can still be used if the FD scheme is defined via a pseudo finite volume scheme. For example, the classical Jiang-Shu FD WENO scheme can be rendered positivity-preserving by this approach [249, 82].*

**4.4. Multi-dimensional extensions.** We summarize the two key and essential ingredients for extending from 1D results above to multiple dimensions:

*A first order IDP flux or scheme.* It is available in multiple dimensions as reviewed in Section 3.

*A decomposition of a cell average into a convex combination of point values including point values on the cell boundary.* In one dimension, the cell boundary values are simply  $\mathbf{u}_{j+\frac{1}{2}}^\pm$ . In multiple dimensions, these cell boundary values should be the quadrature point values used for computing numerical flux along cell boundaries.

With these two ingredients, it is possible to extend Theorem 4.3 to multiple dimensions, which will be reviewed in the next two subsections, and the point values can be controlled and corrected by a similar simple limiter like (4.15). The desired decomposition of a cell average into point values is achieved by Gauss–Lobatto quadrature in 1D. In multiple dimensions, it can be done by a quadrature with positive quadrature weights. On a given cell, such a quadrature may or may not exist. On a polygonal cell, such a quadrature can be constructed and is not unique. In [250, 160, 46], different choices of such quadrature rules were constructed on triangles and simplices, by from which specialized quadrature rules on polygons can be obtained by partitioning the polygon into a union of triangles [212]. Quadrature rules on polygons with fewer points can be found in [197]. See [77] for the curvilinear elements. Figure 4 shows the special quadrature in two dimensions used in [246, 250], which is however not an optimal choice. In Figure 4, such a special quadrature consists of unnecessarily

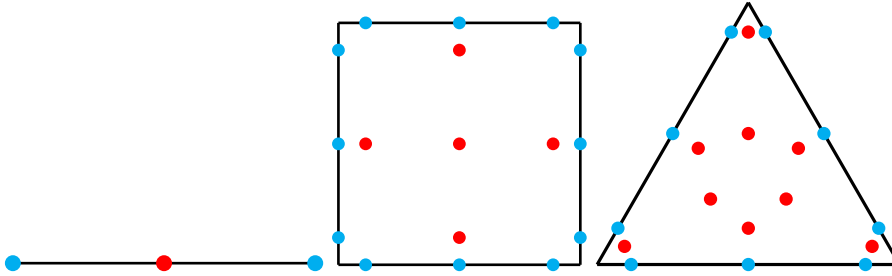


Fig. 4: One example of the special quadrature for quadratic polynomials. Left: 1D cell. Middle: 2D rectangle. Right: 2D triangle. For 1D, it is simply 3-point Gauss–Lobatto quadrature. The points in cyan color are the Gauss quadrature points for computing numerical flux integrals in high order schemes, and red points and cyan points together form a special quadrature that is exact for quadratic polynomials with positive weight.

too many points for integrating quadratic polynomials. We emphasize that such a quadrature is not used for computing any integrals, and we only need the following from this special quadrature:

- We need its existence to establish weak IDP properties like Theorem 4.3.
- We need the quadrature points and weights for defining and implementing

the limiter (4.15).

- The smallest weight on the cyan color points along the boundary also gives the CFL number in Theorem 4.3, which will be discussed later in this section.

REMARK 4.6. *If implementing the Zhang–Shu method using the simplified limiter as stated in Subsection 4.2.5, then evaluation of DG or FV polynomials at the highly redundant red points in Figure 4 can be avoided. Similar to Theorem 4.3, only quadrature weights of the cyan points are needed to state a weak IDP theorem and define a simplified limiter.*

**4.5. 2D hyperbolic systems on rectangular meshes.** We now review the Zhang–Shu approach for 2D hyperbolic system

$$(4.16) \quad \partial_t \mathbf{u} + \partial_x \mathbf{f}_1(u) + \partial_y \mathbf{f}_2(u) = \mathbf{0}$$

on rectangular meshes [246], and some recent advances [61, 62] on seeking optimal quadrature for IDP high order schemes.

We first review how to establish the weak IDP property for the updated cell averages. For (4.16) with the forward Euler time discretization on a rectangular cell  $I_{ij} := [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$ , a finite volume scheme or the cell average of DG scheme can be formulated as

$$(4.17) \quad \begin{aligned} \bar{\mathbf{u}}_{ij}^{n+1} = \bar{\mathbf{u}}_{ij}^n - \frac{\Delta t}{\Delta x} \sum_{q=1}^Q \tilde{\omega}_q \left[ \hat{\mathbf{f}}_1(\mathbf{u}_{i+\frac{1}{2},q}^-, \mathbf{u}_{i+\frac{1}{2},q}^+) - \hat{\mathbf{f}}_1(\mathbf{u}_{i-\frac{1}{2},q}^-, \mathbf{u}_{i-\frac{1}{2},q}^+) \right] \\ - \frac{\Delta t}{\Delta y} \sum_{q=1}^Q \tilde{\omega}_q \left[ \hat{\mathbf{f}}_2(\mathbf{u}_{q,j+\frac{1}{2}}^-, \mathbf{u}_{q,j+\frac{1}{2}}^+) - \hat{\mathbf{f}}_2(\mathbf{u}_{q,j-\frac{1}{2}}^-, \mathbf{u}_{q,j-\frac{1}{2}}^+) \right] \end{aligned}$$

with

$$\begin{aligned} \mathbf{u}_{i-\frac{1}{2},q}^+ &= \mathbf{p}_{ij}(x_{i-\frac{1}{2}}, \tilde{y}_j^{(q)}), & \mathbf{u}_{i+\frac{1}{2},q}^- &= \mathbf{p}_{ij}(x_{i+\frac{1}{2}}, \tilde{y}_j^{(q)}), \\ \mathbf{u}_{q,j-\frac{1}{2}}^+ &= \mathbf{p}_{ij}(\tilde{x}_i^{(q)}, y_{j-\frac{1}{2}}), & \mathbf{u}_{q,j+\frac{1}{2}}^- &= \mathbf{p}_{ij}(\tilde{x}_i^{(q)}, y_{j+\frac{1}{2}}), \end{aligned}$$

where  $\mathbf{p}_{ij}(x, y)$  is the approximate solution polynomial vector (reconstructed in finite volume methods or evolved in DG methods) on  $I_{ij}$  at time level  $n$  and its cell average over  $I_{ij}$  equals  $\bar{\mathbf{u}}_{ij}^n$ . Here,  $\{\tilde{x}_i^{(q)}\}_{q=1}^Q$  and  $\{\tilde{y}_j^{(q)}\}_{q=1}^Q$  denote the nodes of a  $Q$ -point Gauss quadrature of sufficiently high order accuracy in the intervals  $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$  and  $[y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$ , respectively, with the normalized weights  $\{\tilde{\omega}_q\}$  satisfying  $\sum_{q=1}^Q \tilde{\omega}_q = 1$ . We use tildes to distinguish this Gauss quadrature from the Gauss–Lobatto quadrature introduced earlier; both rules will be employed below.

**4.5.1. Weak IDP property.** In order to obtain an IDP scheme, the numerical fluxes  $\hat{\mathbf{f}}_1$  and  $\hat{\mathbf{f}}_2$  in (4.17) are taken as IDP fluxes, with which the corresponding 1D three-point first order schemes are IDP, i.e., for any  $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3 \in G$  it holds that

$$\mathbf{u}_2 - \frac{\Delta t}{\Delta x} \left( \hat{\mathbf{f}}_1(\mathbf{u}_2, \mathbf{u}_3) - \hat{\mathbf{f}}_1(\mathbf{u}_1, \mathbf{u}_2) \right) \in G, \quad \mathbf{u}_2 - \frac{\Delta t}{\Delta y} \left( \hat{\mathbf{f}}_2(\mathbf{u}_2, \mathbf{u}_3) - \hat{\mathbf{f}}_2(\mathbf{u}_1, \mathbf{u}_2) \right) \in G$$

under a suitable CFL condition  $\max\{\alpha_1 \Delta t / \Delta x, \alpha_2 \Delta t / \Delta y\} \leq c_0$ , where  $\alpha_1$  and  $\alpha_2$  denote the maximum characteristic speeds in the  $x$ - and  $y$ -directions, and  $c_0$  is the maximum allowable CFL number for the 1D first order schemes.

Similar to the role of the Gauss–Lobatto quadrature in the 1D case, in order to decompose the cell average  $\bar{\mathbf{u}}_{ij}^n$  into a convex combination of some point values of  $\mathbf{p}_{ij}$ , we need a special quadrature on the cell  $I_{ij}$  of the form

$$(4.18) \quad \frac{1}{\Delta x \Delta y} \int_{I_{ij}} p(x, y) dx dy = \sum_{q=1}^Q \tilde{\omega}_q \left[ \omega_1^- p(x_{i-\frac{1}{2}}, \tilde{y}_j^{(q)}) + \omega_1^+ p(x_{i+\frac{1}{2}}, \tilde{y}_j^{(q)}) \right. \\ \left. + \omega_2^- p(\tilde{x}_i^{(q)}, y_{j-\frac{1}{2}}) + \omega_2^+ p(\tilde{x}_i^{(q)}, y_{j+\frac{1}{2}}) \right] + \sum_{s=1}^S \omega_s^* p(x_s^*, y_s^*),$$

which should satisfy three requirements:

- (i) The quadrature (4.18) is exact for all polynomials  $p \in \mathbb{V}$ , where  $\mathbb{V}$  is the approximate solution space, e.g.,  $\mathbb{P}^k$  or  $\mathbb{Q}^k$ ;
- (ii) The weights  $\{\omega_1^\pm, \omega_2^\pm, \omega_s^*\}$  are positive and they sum to one;
- (iii) The internal node set  $\mathcal{I}_K = \{(x_s^*, y_s^*)\}_{s=1}^S \subset I_{ij}$ .

With such a quadrature, Theorem 4.3 can be easily extended to rectangular cells [245, 246, 250, 61, 62]. We state one version of such an extension in [61] as follows.

**THEOREM 4.4** (Weak IDP property on 2D rectangular mesh). *If the solution polynomials  $\{\mathbf{p}_{ij}\}$  satisfy*

$$(4.19) \quad \mathbf{p}_{ij}(x, y) \in G \quad \forall (x, y) \in \mathbb{S}_{ij}, \quad \forall i, j,$$

where  $\mathbb{S}_{ij}$  denotes the set of all the quadrature points in (4.18), then the high order scheme (4.17) preserves  $\bar{\mathbf{u}}_{ij}^{n+1} \in G$  under the CFL condition

$$(4.20) \quad \Delta t \leq c_0 \min \left\{ \frac{\omega_1^- \Delta x}{\alpha_1}, \frac{\omega_1^+ \Delta x}{\alpha_1}, \frac{\omega_2^- \Delta y}{\alpha_2}, \frac{\omega_2^+ \Delta y}{\alpha_2} \right\}.$$

**4.5.2. Quadrature and CFL.** The special 2D quadrature of the form (4.18) plays a critical role in constructing above 2D high order IDP schemes. Such quadrature rules are not unique. Below are several examples.

**EXAMPLE 4.2** (Zhang–Shu quadrature [245, 246]). *Based on the tensor product of the  $L$ -point Gauss–Lobatto quadrature (with  $L = \lceil \frac{k+3}{2} \rceil$ ) and the  $Q$ -point Gauss quadrature, Zhang and Shu [245, 246] proposed the following quadrature:*

$$(4.21) \quad \frac{1}{\Delta x \Delta y} \int_{I_{ij}} p(x, y) dx dy = \sum_{q=1}^Q \tilde{\omega}_q \left[ \kappa_1 \omega_1 p(x_{i-\frac{1}{2}}, \tilde{y}_j^{(q)}) + \kappa_1 \omega_1 p(x_{i+\frac{1}{2}}, \tilde{y}_j^{(q)}) \right. \\ \left. + \kappa_2 \omega_1 p(\tilde{x}_i^{(q)}, y_{j-\frac{1}{2}}) + \kappa_2 \omega_1 p(\tilde{x}_i^{(q)}, y_{j+\frac{1}{2}}) \right] \\ + \sum_{\mu=2}^{L-1} \sum_{q=1}^Q \omega_\mu \tilde{\omega}_q \left[ \kappa_1 p(x_i^{(\mu)}, \tilde{y}_j^{(q)}) + \kappa_2 p(\tilde{x}_i^{(q)}, y_j^{(\mu)}) \right],$$

where  $\{x_i^{(\mu)}\}_{\mu=1}^L$  and  $\{y_j^{(\mu)}\}_{\mu=1}^L$  denote the nodes of the  $L$ -point Gauss–Lobatto quadrature in the intervals  $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$  and  $[y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$ , respectively, with the weights  $\{\omega_\mu\}$  satisfying  $\sum_{\mu=1}^L \omega_\mu = 1$ , and

$$\kappa_1 := \frac{\alpha_1 / \Delta x}{\alpha_1 / \Delta x + \alpha_2 / \Delta y}, \quad \kappa_2 := \frac{\alpha_2 / \Delta y}{\alpha_1 / \Delta x + \alpha_2 / \Delta y}.$$

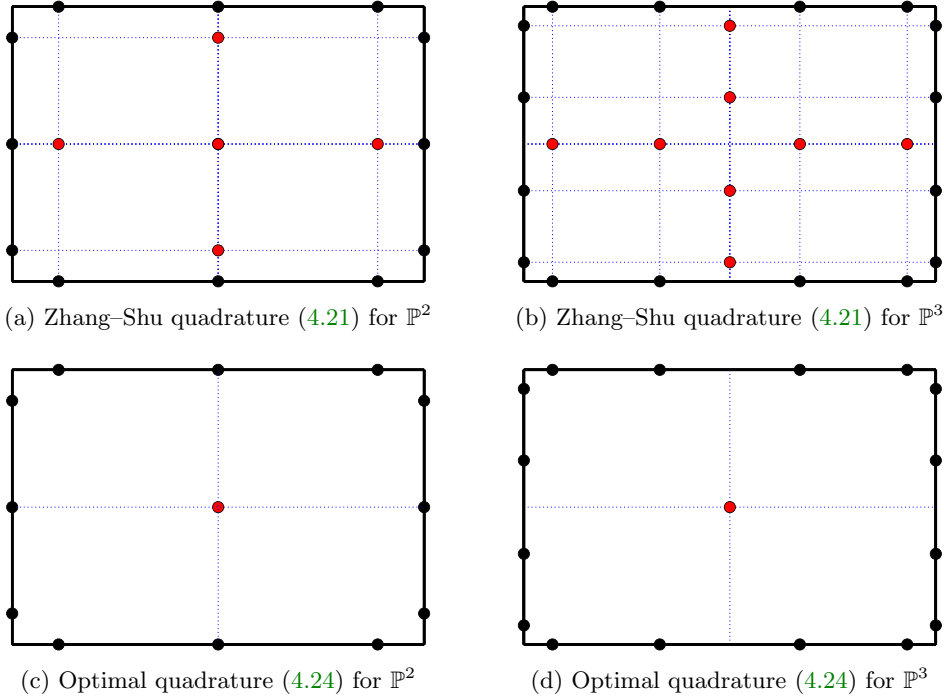


Fig. 5: Boundary nodes (black) and internal nodes (red) of the special 2D quadrature on a rectangular cell  $I_{ij}$ , for the  $\mathbb{P}^2$ - and  $\mathbb{P}^3$ -based methods, in the case of  $\frac{\Delta x}{\alpha_1} = \frac{\Delta y}{\alpha_2}$ .

The set of quadrature nodes in the quadrature rule (4.21) can be expressed as

$$\mathbb{S}_{ij}^{\text{ZS}} = \left( \mathbb{S}_i^x \otimes \tilde{\mathbb{S}}_j^y \right) \cup \left( \tilde{\mathbb{S}}_i^x \otimes \mathbb{S}_j^y \right),$$

where  $\tilde{\mathbb{S}}_i^x := \{\tilde{x}_i^{(q)}\}_{q=1}^Q$ ,  $\tilde{\mathbb{S}}_j^y := \{\tilde{y}_j^{(\mu)}\}_{\mu=1}^Q$ ,  $\mathbb{S}_i^x := \{x_i^{(\mu)}\}_{\mu=1}^L$ , and  $\mathbb{S}_j^y := \{y_j^{(\mu)}\}_{\mu=1}^L$  are the Gauss quadrature nodes for flux integration and the Gauss-Lobatto quadrature nodes for convex decomposition of cell averages, respectively. The quadrature nodes of the Zhang-Shu quadrature (4.21) are illustrated in Figure 5 for  $\mathbb{P}^2$  and  $\mathbb{P}^3$  based methods.

As a corollary of Theorem 4.4, we have the following result.

**THEOREM 4.5** (Weak IDP via Zhang-Shu quadrature). *If the solution polynomials  $\{\mathbf{p}_{ij}\}$  satisfy (4.19) with  $\mathbb{S}_{ij} = \mathbb{S}_{ij}^{\text{ZS}}$ , then the high order scheme (4.17) preserves  $\bar{\mathbf{u}}_{ij}^{n+1} \in G$  under the CFL condition*

$$(4.22) \quad \left( \frac{\alpha_1}{\Delta x} + \frac{\alpha_2}{\Delta y} \right) \Delta t \leq \omega_1 c_0 = \frac{c_0}{L(L-1)} \quad \text{with} \quad L = \left\lceil \frac{k+3}{2} \right\rceil.$$

**REMARK 4.7.** A similar quadrature was used by Jiang and Liu in [119]. The only difference from the Zhang-Shu quadrature is that  $\kappa_1 = \kappa_2 = \frac{1}{2}$ . In this case, the CFL condition for the weak IDP property becomes

$$(4.23) \quad 2 \max \left\{ \frac{\alpha_1}{\Delta x}, \frac{\alpha_2}{\Delta y} \right\} \Delta t \leq \omega_1 c_0,$$

which is more restrictive than (4.22).

Since such special quadrature rules are not unique, it is natural to seek the *optimal quadrature* such that the resulting IDP CFL condition (4.20) is the mildest. It was proven in [62] that the Zhang–Shu quadrature (4.21) is optimal for high order schemes based on the  $\mathbb{Q}^k$  space with any positive integer  $k$ . However, for the  $\mathbb{P}^k$ -based methods, the quadrature (4.21) is generally not optimal [61]. The optimal 2D quadrature rules for  $\mathbb{P}^k$ -based methods were systematically studied in [61, 62].

EXAMPLE 4.3 (Optimal quadrature [61, 62]). *For the  $\mathbb{P}^2$ - and  $\mathbb{P}^3$ -based methods, the optimal quadrature for weak IDP property is given by*

$$(4.24) \quad \begin{aligned} \frac{1}{\Delta x \Delta y} \int_{I_{ij}} p(x, y) \, dx dy &= \frac{\mu_1}{2} \sum_{q=1}^Q \tilde{\omega}_q \left[ p(x_{i-\frac{1}{2}}, \tilde{y}_j^{(q)}) + p(x_{i+\frac{1}{2}}, \tilde{y}_j^{(q)}) \right] \\ &+ \frac{\mu_2}{2} \sum_{q=1}^Q \tilde{\omega}_q \left[ p(\tilde{x}_i^{(q)}, y_{j-\frac{1}{2}}) + p(\tilde{x}_i^{(q)}, y_{j+\frac{1}{2}}) \right] + \omega^* \sum_{s=1}^2 p(x_s^*, y_s^*) \end{aligned}$$

with the internal nodes

$$(4.25) \quad \{x_s^*, y_s^*\}_{s=1}^2 = \begin{cases} \left( x_i, y_j \pm \frac{\Delta y}{2\sqrt{3}} \sqrt{\frac{\phi_* - \phi_2}{\phi_*}} \right), & \text{if } \phi_1 \geq \phi_2, \\ \left( x_i \pm \frac{\Delta x}{2\sqrt{3}} \sqrt{\frac{\phi_* - \phi_1}{\phi_*}}, y_j \right), & \text{if } \phi_1 < \phi_2, \end{cases}$$

where

$$\begin{aligned} \phi_1 &= \frac{\alpha_1}{\Delta x}, \quad \phi_2 = \frac{\alpha_2}{\Delta y}, \quad \phi^* = \max\{\phi_1, \phi_2\}, \\ \psi &= \phi_1 + \phi_2 + 2\phi^*, \quad \mu_1 = \frac{\phi_1}{\psi}, \quad \mu_2 = \frac{\phi_2}{\psi}, \quad \omega^* = \frac{\phi^*}{\psi}. \end{aligned}$$

There are only two internal nodes, which merge to a single node  $(x_i, y_j)$  in case of  $\phi_1 = \phi_2$ . The quadrature nodes of the optimal quadrature (4.24) are illustrated in Figure 5 for  $\mathbb{P}^2$ - and  $\mathbb{P}^3$ -based methods. For the optimal quadrature for the  $\mathbb{P}^k$ -based methods with higher  $k \geq 4$ , we refer the readers to [62].

THEOREM 4.6 (Weak IDP via optimal quadrature [61, 62]). *If the solution polynomials  $\{\mathbf{p}_{ij}\}$  are in the space  $\mathbb{P}^2$  or  $\mathbb{P}^3$  and satisfy (4.19) with  $\mathbb{S}_{ij}$  being the set of all the nodes in (4.24), then the high order scheme (4.17) preserves  $\bar{\mathbf{u}}_{ij}^{n+1} \in G$  under the CFL condition*

$$(4.26) \quad \left( 2 \frac{\alpha_1}{\Delta x} + 2 \frac{\alpha_2}{\Delta y} + 4 \max \left\{ \frac{\alpha_1}{\Delta x}, \frac{\alpha_2}{\Delta y} \right\} \right) \Delta t \leq c_0.$$

Table 1 lists a comparison of the IDP CFL conditions and the internal nodes for different 2D special quadrature rules.

The simple limiter (4.15) can be easily extended to multiple dimensions as follows

$$\tilde{\mathbf{p}}_{ij}(x, y) = \theta(\mathbf{p}_{ij}(x, y) - \bar{\mathbf{u}}_{ij}^n) + \bar{\mathbf{u}}_{ij}^n,$$

where  $\theta$  is computed via either point values in the special quadrature  $\mathbb{S}_{ij}$  or only the boundary quadrature point values in a simplified fashion as in Subsection 4.2.5. See [246, 250] for details.

Table 1: IDP CFL conditions and the numbers of internal nodes of different 2D quadrature for the  $\mathbb{P}^2$ - and  $\mathbb{P}^3$ -based methods.

	IDP CFL	IDP CFL	# Nodes	# Nodes
	general case	$\frac{\Delta x}{\alpha_1} = \frac{\Delta y}{\alpha_2} = h$	$\mathbb{P}^2$	$\mathbb{P}^3$
Optimal [61]	(4.26)	$\Delta t \leq \frac{c_0}{8} h$	$1 \sim 2$	$1 \sim 2$
Zhang-Shu [246]	(4.22)	$\Delta t \leq \frac{c_0}{12} h$	5	8
Jiang-Liu [119]	(4.23)	$\Delta t \leq \frac{c_0}{12} h$	5	8

**4.6. Extensions to unstructured triangular meshes.** For simplicity, we focus only on the special quadrature and extensions of [Theorem 4.3](#).

**4.6.1. The special quadrature for convex decompositions of the cell average.** Let  $K$  denote an arbitrary triangular cell with edge length denoted by  $l_K^{(i)}$  ( $i = 1, 2, 3$ ). Let  $(x_K^{i,q}, y_K^{i,q})$  be the  $q$ th node of the  $Q$ -point Gauss quadrature on the  $i$ th edge  $e_K^{(i)}$  and  $\tilde{\omega}_q$  be the weight. The first task is to find a special 2D quadrature on  $K$ :

$$\begin{aligned}
 (4.27) \quad \frac{1}{|K|} \iint_K p(x, y) \, dx dy &= \sum_{i=1}^3 \frac{w_i}{l_K^{(i)}} \int_{e_K^{(i)}} p(x, y) \, ds + \sum_{s=1}^S \omega_s^* p(x_s^*, y_s^*) \\
 &= \sum_{i=1}^3 \sum_{q=1}^Q w_i \tilde{\omega}_q p(x_K^{i,q}, y_K^{i,q}) + \sum_{s=1}^S \omega_s^* p(x_s^*, y_s^*)
 \end{aligned}$$

such that

- (i) The quadrature (4.27) holds exactly for all  $p(x, y) \in \mathbb{P}^k$ ;
- (ii) The edge weights  $\{w_i\}_{i=1}^3$  and the internal node weights  $\{\omega_s^*\}_{s=1}^S$  are all positive, and  $\sum_{i=1}^3 \sum_{q=1}^Q w_i \tilde{\omega}_q + \sum_{s=1}^S \omega_s^* = 1$ ;
- (iii) The internal node set  $\mathcal{I}_K = \{(x_s^*, y_s^*)\}_{s=1}^S \subset K$ .

EXAMPLE 4.4 (Zhang–Xia–Shu quadrature [250]). *Zhang, Xia, and Shu proposed in [250] the following quadrature for the  $\mathbb{P}^k$  space on a triangular cell  $K$ :*

$$\begin{aligned}
 (4.28) \quad \frac{1}{|K|} \iint_K p(x, y) \, dx dy &= \sum_{i=1}^3 \frac{2\omega_1}{3l_K^{(i)}} \int_{e_K^{(i)}} p(x, y) \, ds + \sum_{s=1}^{S^{\text{ZXS}}} \omega_s^{\text{ZXS}} p(x_s^{\text{ZXS}}, y_s^{\text{ZXS}}), \\
 &= \sum_{i=1}^3 \sum_{q=1}^Q \frac{2\omega_1 \tilde{\omega}_q}{3} p(x_K^{i,q}, y_K^{i,q}) + \sum_{s=1}^{S^{\text{ZXS}}} \omega_s^{\text{ZXS}} p(x_s^{\text{ZXS}}, y_s^{\text{ZXS}}),
 \end{aligned}$$

where  $\omega_1 = \frac{1}{L(L-1)}$  is the first Gauss–Lobatto quadrature weight with  $L = \lceil \frac{k+3}{2} \rceil$ ,  $\{(x_s^{\text{ZXS}}, y_s^{\text{ZXS}})\}$  denote the coordinates of  $S^{\text{ZXS}} = 3\lceil \frac{k-1}{2} \rceil(k+1)$  internal nodes (see [250] for more details). This quadrature was constructed by the average of three different mappings from the Zhang–Shu quadrature (4.21) on a rectangular cell to the triangular cell  $K$ . In practice, one may want to use different Gauss quadrature for each edge in

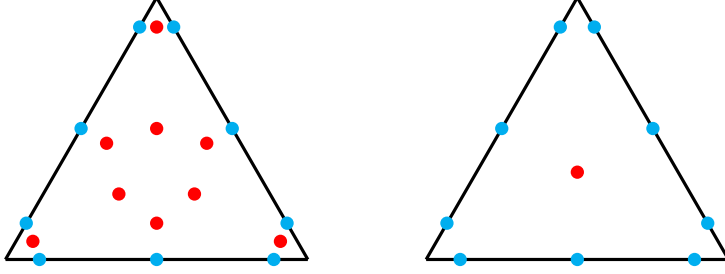


Fig. 6: Two special quadratures for  $\mathbb{P}^2$  on an equilateral triangle with area 1. Left is Zhang–Xia–Shu quadrature, with weights for three cyan points of each edge being  $(\frac{5}{162}, \frac{8}{162}, \frac{5}{162})$ . Right is Chen–Shu quadrature [46, Tale C.2 (b)] which is much less redundant, with weights for three cyan points of each edge being  $(\frac{1}{24}, \frac{1}{10}, \frac{1}{24})$ .

the high order schemes. For instance, see Figure 4 for the Zhang–Xia–Shu quadrature for  $\mathbb{P}^2$  with 3-point Gauss quadrature for each edge, which is in general enough for the  $\mathbb{P}^2$  DG method [250]. In [226], Zhang–Xia–Shu special quadrature (4.28) with 4-point Gauss quadrature for each edge of a triangle is used for the  $\mathbb{P}^2$  DG method in order to achieve well-balanced property for shallow water equations.

EXAMPLE 4.5 (Chen–Shu quadrature). In [46], Chen and Shu used another series of quadrature rules on triangular cells for entropy-stable DG methods, constructed by the quadrature method [241], which was also used in [160]. These quadrature rules are also qualified for designing  $\mathbb{P}^k$ -based IDP schemes on triangular cells; see [46, Appendix C] for further details on these quadrature rules, which is much less redundant than Zhang–Xia–Shu quadrature, as shown in Figure 6.

REMARK 4.8. Neither Zhang–Xia–Shu quadrature nor Chen–Shu quadrature is optimal for IDP studies in general, as demonstrated in [68], where the optimal quadrature rules for  $\mathbb{P}^1$ - and  $\mathbb{P}^2$ -based IDP schemes were found. We emphasize that the interior nodes (red nodes in Figure 6) in the special quadrature can be avoided in the final implementation, similar to Subsection 4.2.5; see [247].

**4.6.2. The weak IDP property on triangular meshes.** The cell average scheme of a high order finite volume or DG scheme with forward Euler time discretization on a triangular cell  $K$  is given by

$$(4.29) \quad \bar{\mathbf{u}}_K^{n+1} = \bar{\mathbf{u}}_K^n - \frac{\Delta t}{2|K|} \sum_{i=1}^3 l_K^{(i)} \sum_{q=1}^Q \tilde{\omega}_q \hat{\mathbf{f}}(\mathbf{u}_{i,q}^{\text{int}}, \mathbf{u}_{i,q}^{\text{ext}}, \mathbf{n}_K^i),$$

where  $|K|$  is the area of  $K$ ,  $l_K^{(i)}$  stands for the length of edge  $e_K^{(i)}$  (the  $i$ -th edge of element  $K$ ),  $\mathbf{n}_K^i = (n_K^{i,1}, n_K^{i,2})$  is the outward unit normal vector of the edge  $e_K^{(i)}$ , and

$$\mathbf{u}_{i,q}^{\text{int}} := \mathbf{u}_h^n(x_K^{i,q}, y_K^{i,q})|_K, \quad \mathbf{u}_{i,q}^{\text{ext}} := \mathbf{u}_h^n(x_K^{i,q}, y_K^{i,q})|_{K^i}$$

are approximations to edge values from interior and exterior of  $K$  respectively, with  $K^i$  denoting the neighboring cell that shares the edge  $e_K^{(i)}$  with  $K$ ,  $(x_K^{i,q}, y_K^{i,q})$  being the  $q$ th node of the  $Q$ -point Gauss quadrature on  $e_K^{(i)}$ , and  $\mathbf{u}_h^n$  denoting the piecewise polynomial solution either reconstructed in finite volume or evolved in DG methods.

Using the Zhang–Xia–Shu quadrature (4.28), one obtains the following convex decomposition for the cell averages:

$$(4.30) \quad \bar{\mathbf{u}}_K^n = \sum_{i=1}^3 \frac{2\omega_1}{3} \sum_{q=1}^Q \tilde{\omega}_q \mathbf{u}_{i,q}^{\text{int}} + \sum_{s=1}^{S^{\text{ZXS}}} \omega_s^{\text{ZXS}} \mathbf{u}_s^{\text{ZXS}}.$$

Substituting (4.30) into (4.29), one can reformulate the scheme (4.29) as follows (cf. [250]):

$$(4.31) \quad \bar{\mathbf{u}}_K^{n+1} = \sum_{s=1}^{S^{\text{ZXS}}} \omega_s^{\text{ZXS}} \mathbf{u}_s^{\text{ZXS}} + \frac{2\omega_1}{3} \sum_{q=1}^Q \tilde{\omega}_q (\mathbf{\Pi}_{1,q} + \mathbf{\Pi}_{2,q} + \mathbf{\Pi}_{3,q}),$$

where

$$\begin{aligned} \mathbf{\Pi}_{1,q} &= \mathbf{u}_{2,q}^{\text{int}} - \frac{3\Delta t}{2\omega_1|K|} \left[ \hat{\mathbf{f}}(\mathbf{u}_{2,q}^{\text{int}}, \mathbf{u}_{1,q}^{\text{int}}, \mathbf{n}_K^1) l_K^{(1)} + \hat{\mathbf{f}}(\mathbf{u}_{2,q}^{\text{int}}, \mathbf{u}_{2,q}^{\text{ext}}, \mathbf{n}_K^2) l_K^{(2)} + \hat{\mathbf{f}}(\mathbf{u}_{2,q}^{\text{int}}, \mathbf{u}_{3,q}^{\text{int}}, \mathbf{n}_K^3) l_K^{(3)} \right], \\ \mathbf{\Pi}_{2,q} &= \mathbf{u}_{1,q}^{\text{int}} - \frac{3\Delta t}{2\omega_1|K|} \left[ \hat{\mathbf{f}}(\mathbf{u}_{1,q}^{\text{int}}, \mathbf{u}_{1,q}^{\text{ext}}, \mathbf{n}_K^1) l_K^{(1)} + \hat{\mathbf{f}}(\mathbf{u}_{1,q}^{\text{int}}, \mathbf{u}_{2,q}^{\text{int}}, -\mathbf{n}_K^1) l_K^{(1)} \right], \\ \mathbf{\Pi}_{3,q} &= \mathbf{u}_{3,q}^{\text{int}} - \frac{3\Delta t}{2\omega_1|K|} \left[ \hat{\mathbf{f}}(\mathbf{u}_{3,q}^{\text{int}}, \mathbf{u}_{2,q}^{\text{int}}, -\mathbf{n}_K^3) l_K^{(3)} + \hat{\mathbf{f}}(\mathbf{u}_{3,q}^{\text{int}}, \mathbf{u}_{3,q}^{\text{ext}}, \mathbf{n}_K^3) l_K^{(3)} \right]. \end{aligned}$$

As an example, consider the Lax–Friedrichs flux:

$$\hat{\mathbf{f}}(\mathbf{u}_{i,q}^{\text{int}}, \mathbf{u}_{i,q}^{\text{ext}}, \mathbf{n}_K^i) = \frac{1}{2} \left( \mathbf{f}(\mathbf{u}_{i,q}^{\text{int}}) \cdot \mathbf{n}_K^i + \mathbf{f}(\mathbf{u}_{i,q}^{\text{ext}}) \cdot \mathbf{n}_K^i + \alpha \mathbf{u}_{i,q}^{\text{int}} - \alpha \mathbf{u}_{i,q}^{\text{ext}} \right),$$

which is IDP if the CFL number is less than or equal to one and

$$(4.32) \quad \alpha^{\text{LF}} := \max_{\mathbf{x} \in e_K^{(i)}, i \in \{1,2,3\}, K \in \mathcal{T}_h} \tilde{\alpha}(\mathbf{u}_h(\mathbf{x}), \mathbf{n}_K^i)$$

Note that  $\mathbf{\Pi}_{1,q}$  is formally a first order IDP scheme on the triangular cell  $K$  under the CFL condition

$$(4.33) \quad \alpha \frac{\Delta t}{|K|} \sum_{i=1}^3 l_K^{(i)} \leq \frac{2}{3} \omega_1,$$

and  $\mathbf{\Pi}_{2,q}$  and  $\mathbf{\Pi}_{3,q}$  are formally one-dimensional IDP schemes under the CFL conditions  $\frac{3\Delta t}{2\omega_1|K|} l_K^{(1)} \leq 1$  and  $\frac{3\Delta t}{2\omega_1|K|} l_K^{(3)} \leq 1$ . Therefore, if

$$\mathbf{u}_{i,q}^{\text{int}} \in G, \quad \mathbf{u}_{i,q}^{\text{ext}} \in G \quad \forall i, q,$$

then  $\mathbf{\Pi}_{i,q} \in G$  for all  $i$  and  $q$  under the CFL condition (4.33).

As observed from (4.31),  $\bar{\mathbf{u}}_K^{n+1}$  is a convex combination of  $\mathbf{u}_s^{\text{ZXS}}$  and  $\mathbf{\Pi}_{i,q}$ . Thanks to the convexity of  $G$ , we obtain the following theorem [250].

**THEOREM 4.7** (IDP via Zhang–Xia–Shu quadrature). *If  $\mathbf{u}_h^n(x, y) \in G \forall (x, y) \in \mathbb{S}_K^{\text{ZXS}}, \forall K$ , where  $\mathbb{S}_K^{\text{ZXS}}$  denotes the set of quadrature nodes on cell  $K$  in (4.28), then the scheme (4.29) preserves  $\bar{\mathbf{u}}_K^{n+1} \in G$  under the CFL condition (4.33).*

With the weak IDP property, a simple limiter similar to the one presented in Subsection 4.2.3 can be designed, and the extension of Theorem 4.2 to triangles or any cells is straightforward using [242, Lemma 8].

**4.6.3. IDP with larger CFL via optimal quadrature and GQL.** This subsection introduces the improvements of the theoretical IDP CFL condition by seeking an optimal quadrature [68] and the GQL approach [222] under [Assumption 6](#), which is weaker than [Assumption 5](#).

In order to precisely define the optimality of the special quadrature, we first show the IDP result [68] with an arbitrarily feasible quadrature (4.27).

**THEOREM 4.8** (IDP via general feasible quadrature). *If the solution  $\mathbf{u}_h^n$  satisfies*

$$(4.34) \quad \mathbf{u}_h^n(x, y) \in G \quad \forall (x, y) \in \mathbb{S}_K, \quad \forall K,$$

where  $\mathbb{S}_K$  denotes the set of all the nodes in the quadrature (4.27) on a triangle  $K$ , then the scheme (4.29) preserves  $\bar{\mathbf{u}}_K^{n+1} \in G$  under the CFL condition

$$(4.35) \quad \alpha \frac{\Delta t}{|K|} \leq \mathcal{C}_{\text{IDP}} := \min \left\{ \frac{w_1}{l_K^{(1)}}, \frac{w_2}{l_K^{(2)}}, \frac{w_3}{l_K^{(3)}} \right\}.$$

*Proof.* Using the equivalent linear GQL representation (3.10) of  $G$ , we derive for any  $\mathbf{u}^* \in \mathcal{S}_j$  and  $j \in \mathbb{I} \cup \hat{\mathbb{I}}$  that

$$\begin{aligned} & (\bar{\mathbf{u}}_K^{n+1} - \mathbf{u}^*) \cdot \mathbf{n}_j^* \\ & \stackrel{(4.27)}{=} \sum_{i=1}^3 l_K^{(i)} \sum_{q=1}^Q \tilde{\omega}_q \left[ \left( \frac{w_i}{l_K^{(i)}} - \frac{\alpha \Delta t}{2|K|} \right) (\mathbf{u}_{i,q}^{\text{int}} - \mathbf{u}^*) \cdot \mathbf{n}_j^* - \frac{\Delta t}{2|K|} (\mathbf{f}(\mathbf{u}_{i,q}^{\text{int}}) \cdot \mathbf{n}_K^i) \cdot \mathbf{n}_j^* \right] \\ & \quad + \sum_{i=1}^3 l_K^{(i)} \sum_{q=1}^Q \tilde{\omega}_q \left[ \frac{\alpha \Delta t}{2|K|} (\mathbf{u}_{i,q}^{\text{ext}} - \mathbf{u}^*) \cdot \mathbf{n}_j^* - \frac{\Delta t}{2|K|} (\mathbf{f}(\mathbf{u}_{i,q}^{\text{ext}}) \cdot \mathbf{n}_K^i) \cdot \mathbf{n}_j^* \right] \\ & \quad + \sum_{s=1}^S \omega_s^* (\mathbf{u}_h^n(x_s^*, y_s^*) - \mathbf{u}^*) \cdot \mathbf{n}_j^* \\ & \stackrel{(4.34), (4.35)}{\geq} \sum_{i=1}^3 l_K^{(i)} \sum_{q=1}^Q \tilde{\omega}_q \left[ \frac{\alpha \Delta t}{2|K|} (\mathbf{u}_{i,q}^{\text{int}} - \mathbf{u}^*) \cdot \mathbf{n}_j^* - \frac{\Delta t}{2|K|} (\mathbf{f}(\mathbf{u}_{i,q}^{\text{int}}) \cdot \mathbf{n}_K^i) \cdot \mathbf{n}_j^* \right] \\ & \quad + \sum_{i=1}^3 l_K^{(i)} \sum_{q=1}^Q \tilde{\omega}_q \left[ \frac{\alpha \Delta t}{2|K|} (\mathbf{u}_{i,q}^{\text{ext}} - \mathbf{u}^*) \cdot \mathbf{n}_j^* - \frac{\Delta t}{2|K|} (\mathbf{f}(\mathbf{u}_{i,q}^{\text{ext}}) \cdot \mathbf{n}_K^i) \cdot \mathbf{n}_j^* \right] \\ & \stackrel{(3.18)}{\succ} \sum_{i=1}^3 l_K^{(i)} \sum_{q=1}^Q \tilde{\omega}_q \frac{\Delta t}{2|K|} [\zeta(\mathbf{u}^*) \cdot \mathbf{n}_K^i] + \sum_{i=1}^3 l_K^{(i)} \sum_{q=1}^Q \tilde{\omega}_q \frac{\Delta t}{2|K|} [\zeta(\mathbf{u}^*) \cdot \mathbf{n}_K^i] \\ & = \frac{\Delta t}{|K|} \zeta(\mathbf{u}^*) \cdot \left( \sum_{i=1}^3 l_K^{(i)} \mathbf{n}_K^i \right) \stackrel{(3.19)}{=} \frac{\Delta t}{2|K|} \zeta(\mathbf{u}^*) \cdot \mathbf{0} = 0. \quad \square \end{aligned}$$

As shown in [Theorem 4.8](#), the CFL condition (4.35) depends on  $\mathcal{C}_{\text{IDP}}$ , which is determined by a chosen quadrature (4.27). It is therefore natural to seek the **optimal** quadrature of the form (4.27) that **maximizes**  $\mathcal{C}_{\text{IDP}}$ , thereby yielding the most lenient IDP CFL condition. This allows for larger stable time step sizes and improves the efficiency of high-order IDP schemes.

Such optimal quadrature rules were recently discovered in [68] for  $\mathbb{P}^1$ - and  $\mathbb{P}^2$ -based IDP schemes on arbitrary triangular meshes. For convenience, we consider an arbitrary triangular cell  $K$  and rearrange the indices of its edges  $\{e_K^{(i)}\}_{i=1}^3$  and vertices  $\{\mathbf{V}_K^{(i)}\}_{i=1}^3$  such that  $l_K^{(1)} \geq l_K^{(2)} \geq l_K^{(3)}$ .

EXAMPLE 4.6 (Optimal quadrature [68] for  $\mathbb{P}^1$ ). *The optimal quadrature of the form (4.27) for  $\mathbb{P}^1$ -based IDP schemes on any triangular cell  $K$  is given by*

$$(4.36) \quad w_i = \frac{2l_K^{(i)}}{3l_K^{(1)} + 3l_K^{(2)}}, \quad i = 1, 2, 3,$$

with at most one internal node ( $S \leq 1$ ), whose weight and location are given by:

$$(4.37) \quad \omega_1^* = \frac{l_K^{(1)} + l_K^{(2)} - 2l_K^{(3)}}{3l_K^{(1)} + 3l_K^{(2)}}, \quad (x_1^*, y_1^*) = \frac{(l_K^{(1)} - l_K^{(3)}) \mathbf{V}_K^{(1)} + (l_K^{(2)} - l_K^{(3)}) \mathbf{V}_K^{(2)}}{l_K^{(1)} + l_K^{(2)} - 2l_K^{(3)}}.$$

Note that if the cell  $K$  is equilateral, i.e.,  $l_K^{(1)} = l_K^{(2)} = l_K^{(3)}$ , then the weight  $\omega_1^*$  becomes zero, and the optimal quadrature contains no internal node ( $S = 0$ ).

EXAMPLE 4.7 (Optimal quadrature [68] for  $\mathbb{P}^2$ ). *The optimal quadrature of the form (4.27) for  $\mathbb{P}^2$ -based IDP schemes on any triangular cell  $K$  has the boundary weights*

$$w_i = \frac{2l_K^{(i)}}{9\bar{l}_K + 3\hat{l}_K}, \quad i = 1, 2, 3,$$

and two internal nodes with weights and coordinates

$$\omega_s^* = \frac{\bar{l}_K + \hat{l}_K}{6\bar{l}_K + 2\hat{l}_K}, \quad (x_s^*, y_s^*)^\top = \sum_{i=1}^3 \beta_{s,i} \mathbf{V}_K^{(i)}, \quad \beta_{s,i} = \frac{\mathbf{l}_K^\top \mathbf{M}_{s,i} \mathbf{l}_K + 2c_{s,i} \hat{l}_K}{18(\bar{l}_K + \hat{l}_K)(l_K^{(2)} + \hat{l}_K)}, \quad s = 1, 2,$$

where  $\mathbf{l}_K := (l_K^{(1)}, l_K^{(2)}, l_K^{(3)})^\top$ ,  $\bar{l}_K := (l_K^{(1)} + l_K^{(2)} + l_K^{(3)})/3$ , and

$$\hat{l}_K := \sqrt{(l_K^{(1)})^2 + (l_K^{(2)})^2 + (l_K^{(3)})^2 - \frac{2}{3}(l_K^{(1)}l_K^{(2)} + l_K^{(2)}l_K^{(3)} + l_K^{(3)}l_K^{(1)})}.$$

The positive coefficients  $c_{s,i}$  and the positive definite matrices  $\mathbf{M}_{s,i}$  are given by

$$\begin{aligned} c_{1,1} &= 3l_K^{(1)} + 3l_K^{(2)} + \sqrt{3}l_K^{(2)} - \sqrt{3}l_K^{(3)}, & c_{2,1} &= 3l_K^{(1)} + 3l_K^{(2)} + \sqrt{3}l_K^{(3)} - \sqrt{3}l_K^{(2)}, \\ c_{1,2} &= 6l_K^{(2)} + \sqrt{3}l_K^{(3)} - \sqrt{3}l_K^{(1)}, & c_{2,2} &= 6l_K^{(2)} + \sqrt{3}l_K^{(1)} - \sqrt{3}l_K^{(3)}, \\ c_{1,3} &= 3l_K^{(2)} + 3l_K^{(3)} + \sqrt{3}l_K^{(1)} - \sqrt{3}l_K^{(2)}, & c_{2,3} &= 3l_K^{(2)} + 3l_K^{(3)} + \sqrt{3}l_K^{(2)} - \sqrt{3}l_K^{(1)}, \end{aligned}$$

$$\mathbf{M}_{1,1} = \begin{bmatrix} 6 & 1 & -2 \\ 1 & 2\sqrt{3}+6 & -\sqrt{3}-2 \\ -2 & -\sqrt{3}-2 & 6 \end{bmatrix}, \quad \mathbf{M}_{2,1} = \begin{bmatrix} 6 & 1 & -2 \\ 1 & 6-2\sqrt{3} & \sqrt{3}-2 \\ -2 & \sqrt{3}-2 & 6 \end{bmatrix},$$

$$\mathbf{M}_{1,2} = \begin{bmatrix} 6 & -\sqrt{3}-2 & -2 \\ -\sqrt{3}-2 & 12 & \sqrt{3}-2 \\ -2 & \sqrt{3}-2 & 6 \end{bmatrix}, \quad \mathbf{M}_{2,2} = \begin{bmatrix} 6 & \sqrt{3}-2 & -2 \\ \sqrt{3}-2 & 12 & -\sqrt{3}-2 \\ -2 & -\sqrt{3}-2 & 6 \end{bmatrix},$$

$$\mathbf{M}_{1,3} = \begin{bmatrix} 6 & \sqrt{3}-2 & -2 \\ \sqrt{3}-2 & 6-2\sqrt{3} & 1 \\ -2 & 1 & 6 \end{bmatrix}, \quad \mathbf{M}_{2,3} = \begin{bmatrix} 6 & -\sqrt{3}-2 & -2 \\ -\sqrt{3}-2 & 2\sqrt{3}+6 & 1 \\ -2 & 1 & 6 \end{bmatrix}.$$

If the above optimal quadrature is used for cell average decomposition, we have the following results, as a corollary of Theorem 4.8.

THEOREM 4.9 (IDP via optimal quadrature [68] for  $\mathbb{P}^1$  and  $\mathbb{P}^2$ ). Assume  $l_K^{(1)} \geq l_K^{(2)} \geq l_K^{(3)}$ . If a  $\mathbb{P}^m$ -based ( $m = 1$  or  $2$ ) solution  $\mathbf{u}_h^n$  satisfy

$$(4.38) \quad \mathbf{u}_h^n(x, y) \in G \quad \forall (x, y) \in \mathbb{S}_{K,m}^{\text{DCW}}, \quad \forall K,$$

where  $\mathbb{S}_{K,1}^{\text{DCW}}$  and  $\mathbb{S}_{K,2}^{\text{DCW}}$  denote the set of all the nodes in the optimal quadrature in [Example 4.6](#) and [Example 4.7](#), respectively, proposed by Ding, Cui, and Wu [68], then the scheme (4.29) preserves  $\bar{\mathbf{u}}_K^{n+1} \in G$  under the CFL condition

$$(4.39) \quad \alpha \frac{\Delta t}{|K|} \leq C_{K,m}^{\text{DCW}} \quad \forall K \in \mathcal{T}_h,$$

$$C_{K,1}^{\text{DCW}} := \frac{2}{3(l_K^{(1)} + l_K^{(2)})}, \quad C_{K,2}^{\text{DCW}} := \frac{2}{9\bar{l}_K + 3\hat{l}_K},$$

which is optimal for using any quadrature of the form (4.27).

#### 4.6.4. Comparison of different quadrature rules for IDP.

REMARK 4.9 (IDP via Chen–Shu quadrature in [46]). As a direct consequence of [Theorem 4.8](#), if the Chen–Shu quadrature in [46, Table C.2] is used for cell average decomposition, then we obtain a  $\mathbb{P}^m$ -based ( $m = 1, 2, 3, 4$ ) high-order IDP scheme under the CFL condition:

$$(4.40) \quad \alpha \frac{\Delta t}{|K|} \leq w_m^{\text{CS}} \min \left\{ \frac{1}{l_K^{(1)}}, \frac{1}{l_K^{(2)}}, \frac{1}{l_K^{(3)}} \right\} =: C_{K,m}^{\text{CS}} \quad \forall K \in \mathcal{T}_h,$$

where  $w_1^{\text{CS}} = \frac{1}{3}$ ,  $w_2^{\text{CS}} = \frac{3}{20}$ ,  $w_3^{\text{CS}} \approx 0.086812$ , and  $w_4^{\text{CS}} \approx 0.05572449$ .

REMARK 4.10 (Improved IDP via Zhang–Xia–Shu quadrature (4.28)). Applying [Theorem 4.8](#) to the case of using Zhang–Xia–Shu quadrature (4.28), we can improve the IDP CFL condition (4.33) to the following more relaxed one:

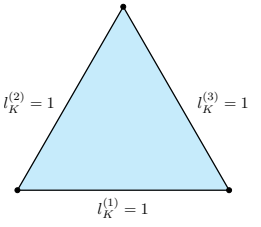
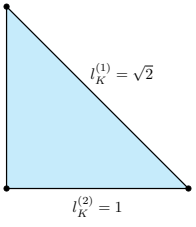
$$(4.41) \quad \alpha \frac{\Delta t}{|K|} \leq \frac{2}{3} \omega_1 \min \left\{ \frac{1}{l_K^{(1)}}, \frac{1}{l_K^{(2)}}, \frac{1}{l_K^{(3)}} \right\} =: C_{K,m}^{\text{ZXS}} \quad \forall K \in \mathcal{T}_h,$$

where  $\omega_1 = \frac{1}{L(L-1)}$  is the first weight of the  $L$ -point Gauss–Lobatto quadrature with  $L = \lceil \frac{m+3}{2} \rceil$ .

REMARK 4.11 (Lv–Ihme approach). A different decomposition for cell averages was proposed by Lv and Ihme in [160]. The idea is to begin with any quadrature rule that has positive weights and sufficiently high accuracy, and then solve an optimization problem to increase the decomposition weights at the boundary Gauss points. The primary advantage of this approach is its applicability to arbitrary polygonal or polyhedral cells. However, it requires solving computationally expensive optimization problems tailored to each specific cell geometry. Moreover, on triangular meshes, the resulting IDP CFL number is generally smaller than those obtained using the quadrature rules discussed above.

[Table 2](#) presents a comparison of the IDP CFL condition  $\alpha \frac{\Delta t}{|K|} \leq C_{\text{IDP}}$  obtained using different quadrature rules for  $\mathbb{P}^1$ - and  $\mathbb{P}^2$ -based schemes on two representative triangular cells. As expected,  $C_{\text{IDP}}$  derived using the optimal quadrature proposed in [68] is the largest among all the cases.

Table 2: Comparison of the IDP CFL condition  $\alpha \frac{\Delta t}{|K|} \leq \mathcal{C}_{\text{IDP}}$  and the number of internal nodes for various quadrature rules on two example triangular cells. (Note: The information in [68, Table 1] for the Chen–Shu quadrature was incorrect, and the corrected version is presented here.)

example cell $K$					
quadrature		$\mathcal{C}_{\text{IDP}}$	internal nodes	$\mathcal{C}_{\text{IDP}}$	internal nodes
$\mathbb{P}^1$	Optimal [68]	$\frac{1}{3} \approx 0.333$	0	$\frac{1}{3} \approx 0.333$	1
	ZXS [250]	$\frac{1}{9} \approx 0.111$	0	$\frac{1}{3(2+\sqrt{2})} \approx 0.0976$	0
	ZXS (4.41)	$\frac{1}{3} \approx 0.333$	0	$\frac{1}{3\sqrt{2}} \approx 0.236$	0
	Chen–Shu [46]	$\frac{1}{3} \approx 0.333$	0	$\frac{1}{3\sqrt{2}} \approx 0.236$	0
$\mathbb{P}^2$	Optimal [68]	$\frac{1}{6} \approx 0.167$	1	$\frac{2}{6+3\sqrt{2}+\sqrt{30-12\sqrt{2}}} \approx 0.144$	2
	ZXS [250]	$\frac{1}{27} \approx 0.037$	9	$\frac{1}{9(2+\sqrt{2})} \approx 0.0325$	9
	ZXS (4.41)	$\frac{1}{9} \approx 0.111$	9	$\frac{1}{9\sqrt{2}} \approx 0.0786$	9
	Chen–Shu [46]	$\frac{3}{20} = 0.15$	1	$\frac{3}{20\sqrt{2}} \approx 0.106$	1

**4.7. Numerical examples of high order DG schemes for gas dynamics equations.** We list a few benchmark tests in gas dynamics for verifying robustness of high order accurate schemes solving low density or low pressure problems, all of which are challenging tests for high order accurate DG schemes. Below are numerical results of high order DG schemes with the third order SSP Runge-Kutta method with only the simple limiter (4.15) for enforcing positivity of density and pressure.

EXAMPLE 4.8 (Sedov blast wave). *The blast wave generates low density and pressure. Figure 7 shows an IDP  $\mathbb{Q}^6$  DG method on a rectangular mesh for compressible Navier–Stokes equations. The parameters are chosen so that, at the final time, the shock front is a circle of radius 1. See [246, 242] for the problem setup, and the exact solution in Sedov’s monograph [191].*

EXAMPLE 4.9 (High speed astrophysical jets). *The extremely high speed renders small internal energy in the computation, which is a tough test even for many second order schemes, e.g., even a second order MUSCL scheme may blow up if positivity is not preserved. Figure 8 shows an IDP  $\mathbb{Q}^4$  DG method on a rectangular mesh for compressible Euler equations for a Mach 2000 jet with background pressure 0.4127, see [246] for the problem set up. Figure 9 shows an IDP  $\mathbb{Q}^6$  DG method on a rectangular*

mesh for compressible Navier–Stokes equations for a Mach 2000 jet with background pressure  $10^{-6}$ , see [204, 149] for the initial conditions.

EXAMPLE 4.10 (Mach 10 shock passing a sharp corner). In this test, a Mach 10 shock is first reflected, generating Kelvin–Helmholtz instability, exactly the same as those in the classical double Mach reflection test. Then the shock is diffracted at a sharp corner, which induces low density and low pressure, causing numerical instabilities in high order DG schemes. This test involves strong shocks, low density/pressure, as well as fine structures such as roll-ups from Kelvin–Helmholtz instability, which are often used as an indicator whether excessive artificial viscosity is added in numerical schemes for stabilization. Figure 9 shows results of IDP high order DG methods for solving compressible Navier–Stokes equations. Figure 9 (a) and (b) show results of  $\mathbb{P}^7$  DG on unstructured triangular meshes for a 60 degree corner, see [242] for the problem set up. Figure 9 (c) and (d) show results of  $\mathbb{Q}^6$  DG on rectangular meshes for a 90 degree corner, see [83, 149] for the problem setup. Limiters for enforcing non-oscillations can be added to reduce oscillations.

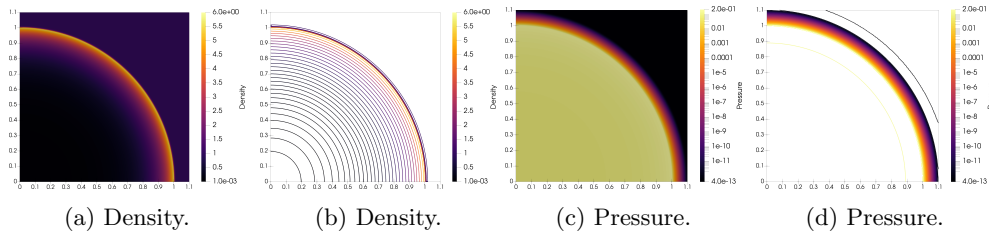


Fig. 7: 2D Sedov blast wave. Numerical results of a positivity-preserving high order DG method with  $\mathbb{Q}^6$  basis on a rectangular mesh of size  $\frac{1}{320}$  for compressible Navier–Stokes with Reynolds number 1000. Only positivity-preserving limiter is used and no other limiters are used.

## 5. Flux correction limiters and convex limiting for high order schemes.

Invoke flux correction limiters is also a very popular approach to enforce convex invariant domains in high order schemes of many spatial discretizations, including continuous finite element methods. The methodology of flux correction limiters for IDP is build upon the ideas of FCT by Boris and Book [35, 34, 36] and Zalesak [239]. We refer to [171, 141, 240, 135, 133] and references therein for histories and developments as well as comprehensive reviews of FCT methods. FCT type flux limiters can be considered for higher order PDEs and used to enforce more properties than a convex invariant domain. We mainly focus on the flux correction based methods for enforcing a convex invariant domain for systems of conservation laws, most of which emerged in the past 15 years.

Flux correction limiters can be designed for many different time discretizations. For the sake of simplicity, we focus on a high order accurate strong stability preserving (SSP) Runge–Kutta (RK) method (4.11), which is convex combination of forward Euler steps. Thus we only need to consider how to achieve IDP for the forward Euler step since a convex combination preserves a convex invariant domain.

**5.1. The main idea of flux corrections.** Consider a high order finite difference (or finite volume) spatial discretization with forward Euler time discretization

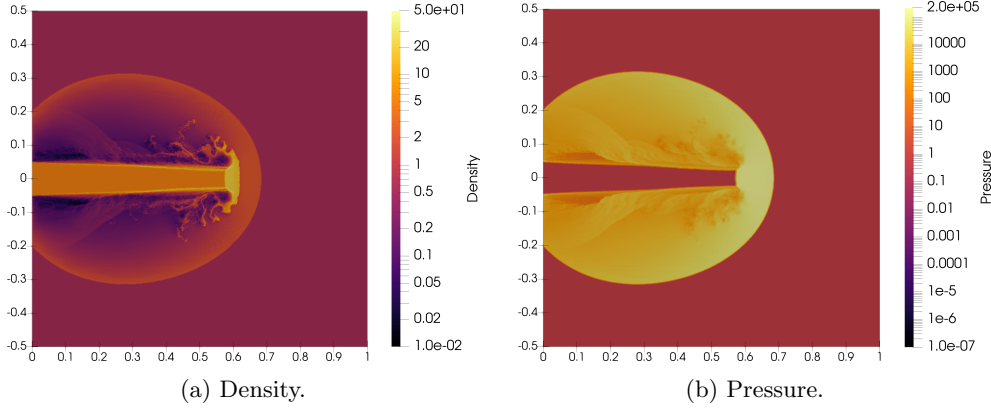


Fig. 8: *Mach 2000 jet with background pressure 0.4127*. Numerical results of a positivity-preserving high order DG method with  $\mathbb{Q}^4$  basis on a rectangular mesh of size  $\frac{1}{640}$  for compressible Euler equations. Only positivity-preserving limiter is used and no other limiters are used.

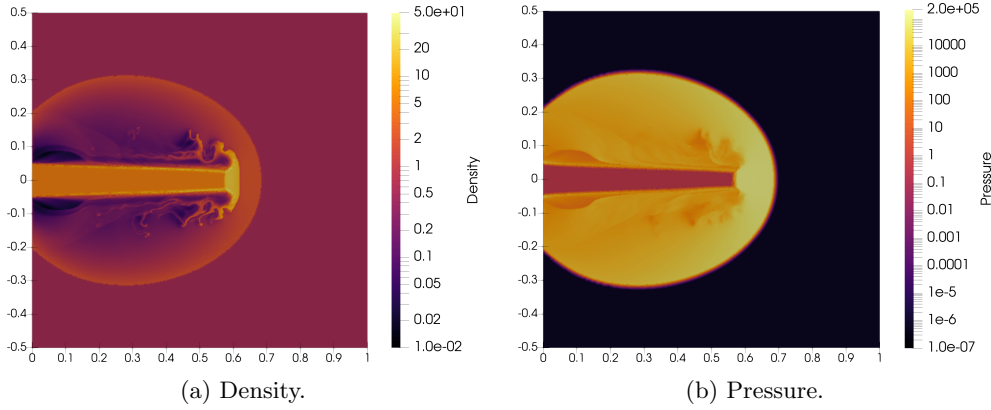


Fig. 9: *Mach 2000 jet with background pressure  $10^{-6}$* . Numerical results of a positivity-preserving high order DG method with  $\mathbb{Q}^6$  basis on a rectangular mesh of size  $\frac{1}{400}$  for compressible Navier-Stokes with Reynolds number 1000. Only positivity-preserving limiter is used and no other limiters are used.

as an example,

$$(5.1) \quad \mathbf{u}_j^{n+1,H} = \mathbf{u}_j^n - \lambda \left( \hat{\mathbf{f}}_{j+\frac{1}{2}}^H - \hat{\mathbf{f}}_{j-\frac{1}{2}}^H \right),$$

which is in general not IDP. Assume that there is a first order IDP scheme,

$$(5.2) \quad \mathbf{u}_j^{n+1,L} = \mathbf{u}_j^n - \lambda \left( \hat{\mathbf{f}}_{j+\frac{1}{2}}^L - \hat{\mathbf{f}}_{j-\frac{1}{2}}^L \right),$$

which is provably IDP under a CFL condition  $a_{\max} \Delta t / \Delta x \leq c_0$  as in Section 3. The idea of flux correction for constructing IDP high order schemes is to seek suitable

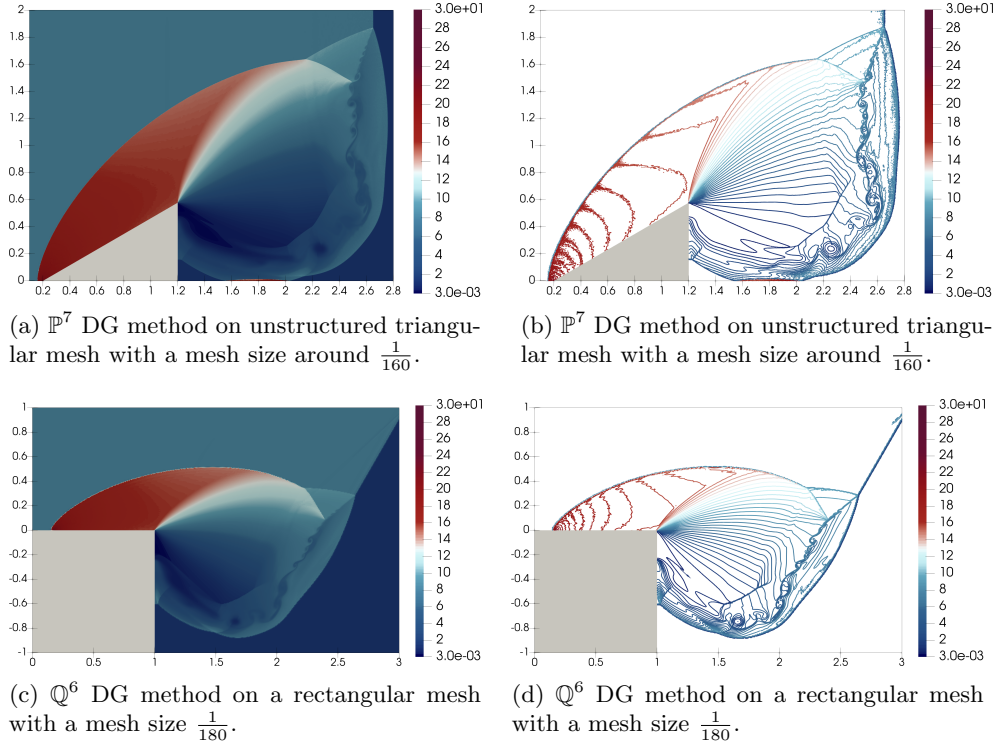


Fig. 10: *Mach 10 shock reflection and diffraction*. Numerical results of positivity-preserving high order DG methods for solving Compressible Navier–Stokes with Reynolds number 1000. Plot of Density. Only positivity-preserving limiter is used and no other limiters are used.

parameters  $\theta_{j+\frac{1}{2}} \in [0, 1]$  such that the scheme using a modified numerical flux is IDP,

$$(5.3a) \quad \mathbf{u}_j^{n+1} = \mathbf{u}_j^n - \lambda \left( \hat{\mathbf{f}}_{j+\frac{1}{2}} - \hat{\mathbf{f}}_{j-\frac{1}{2}} \right),$$

$$(5.3b) \quad \hat{\mathbf{f}}_{j+\frac{1}{2}} = \theta_{j+\frac{1}{2}} \left( \hat{\mathbf{f}}_{j+\frac{1}{2}}^H - \hat{\mathbf{f}}_{j+\frac{1}{2}}^L \right) + \hat{\mathbf{f}}_{j+\frac{1}{2}}^L.$$

Since  $\hat{\mathbf{f}}_{j+\frac{1}{2}}$  is a convex combination of  $\hat{\mathbf{f}}_{j+\frac{1}{2}}^H$  and  $\hat{\mathbf{f}}_{j+\frac{1}{2}}^L$ , the corrected numerical flux  $\hat{\mathbf{f}}_{j+\frac{1}{2}}$  is a consistent and locally conservative flux, thus the scheme (5.3) is still a consistent and locally conservative scheme. In order to maintain the high order accuracy,  $\theta_{j+\frac{1}{2}} \in [0, 1]$  should be as large as possible under the IDP constraint.

It is nontrivial to determine such parameters  $\{\theta_{j+\frac{1}{2}}\}$ , as they are coupled. For example, the IDP goal for preserving  $g(\mathbf{u}) > 0$  requires  $\{\theta_{j+\frac{1}{2}}\}$  to satisfy the globally coupled inequalities for all  $j$ :

$$(5.4) \quad g \left( \mathbf{u}_j^n - \lambda \left[ \left( \theta_{j+\frac{1}{2}} \left( \hat{\mathbf{f}}_{j+\frac{1}{2}}^H - \hat{\mathbf{f}}_{j+\frac{1}{2}}^L \right) + \hat{\mathbf{f}}_{j+\frac{1}{2}}^L \right) - \left( \theta_{j-\frac{1}{2}} \left( \hat{\mathbf{f}}_{j-\frac{1}{2}}^H - \hat{\mathbf{f}}_{j-\frac{1}{2}}^L \right) + \hat{\mathbf{f}}_{j-\frac{1}{2}}^L \right) \right] \right) > 0.$$

Defining  $\delta_{j+\frac{1}{2}} := \hat{\mathbf{f}}_{j+\frac{1}{2}}^H - \hat{\mathbf{f}}_{j+\frac{1}{2}}^L$  and using the notation in (5.2), the IDP inequalities (5.4) can be equivalently expressed as

$$(5.5) \quad g\left(\mathbf{u}_j^{n+1,L} - \lambda(\theta_{j+\frac{1}{2}}\delta_{j+\frac{1}{2}} - \theta_{j-\frac{1}{2}}\delta_{j-\frac{1}{2}})\right) > 0, \quad \forall j.$$

We list a simple fact implied by the Jensen's inequality:

LEMMA 5.1. *If  $g(\mathbf{u})$  is a concave or a quasi-concave function of  $\mathbf{u} \in \mathbb{R}^d$ , and  $\{\theta_{j+\frac{1}{2}}, j = 1, 2, \dots, n\}$  satisfies (5.5), then  $g\left(\mathbf{u}_j^{n+1,L}\right) \geq 0$  implies that  $\{a\theta_{j+\frac{1}{2}}, j = 1, 2, \dots, n\}$  satisfies (5.5) for any  $a \in [0, 1]$ .*

*Proof.* Since  $\mathbf{u}_j^{n+1} = \mathbf{u}_j^{n+1,L} - \lambda(\theta_{j+\frac{1}{2}}\delta_{j+\frac{1}{2}} - \theta_{j-\frac{1}{2}}\delta_{j-\frac{1}{2}})$ , we have a convex combination,

$$\mathbf{u}_j^{n+1,L} - \lambda(a\theta_{j+\frac{1}{2}}\delta_{j+\frac{1}{2}} - a\theta_{j-\frac{1}{2}}\delta_{j-\frac{1}{2}}) = (1-a)\mathbf{u}_j^{n+1,L} + a\mathbf{u}_j^{n+1},$$

If  $g(\mathbf{u})$  is concave, the Jensen's inequality gives

$$g\left(\mathbf{u}_j^{n+1,L} - \lambda(a\theta_{j+\frac{1}{2}}\delta_{j+\frac{1}{2}} - a\theta_{j-\frac{1}{2}}\delta_{j-\frac{1}{2}})\right) = (1-a)g\left(\mathbf{u}_j^{n+1,L}\right) + ag\left(\mathbf{u}_j^{n+1}\right) \geq 0.$$

If  $g(\mathbf{u})$  is quasi-concave, the same argument applies using the inequality (2.4).  $\square$

For the method (5.3), the largest  $\theta_{j+\frac{1}{2}} \in [0, 1]$  satisfying (5.5) would give the most accurate scheme, and they can be found by maximizing  $\theta_{j+\frac{1}{2}} \in [0, 1]$  under the constraints (5.5), which is a globally coupled constrained optimization thus expensive to solve. As pointed out in [148], the flux correction approach can be regarded as seeking easier alternatives for parameters  $\theta_{j+\frac{1}{2}}$ , to avoid solving the constrained global optimization problem. There are several different ways to efficiently compute limiting parameters  $\theta_{j+\frac{1}{2}}$  to satisfy (5.5), resulting in different flux-correction limiters and IDP schemes. In the following subsections, we will review several popular methods.

REMARK 5.1. *To improve accuracy in the flux corrected scheme (5.3), one method is to upgrade the low order IDP flux  $\mathbf{f}^L$  in (5.3) by any higher order accurate IDP flux. The iterative FCT method [190] is such a simple approach, by replacing the low order IDP flux  $\mathbf{f}^L$  in (5.3) by the IDP flux  $\hat{\mathbf{f}}$  in (5.3b) recursively as follows. With a given first order IDP flux  $\mathbf{f}^L$ , define  $\mathbf{f}^{L,0} = \mathbf{f}^L$ , and find  $\theta^m$  such that*

$$\hat{\mathbf{f}}_{j+\frac{1}{2}}^{L,m+1} = \theta_{j+\frac{1}{2}}^m \left( \hat{\mathbf{f}}_{j+\frac{1}{2}}^H - \hat{\mathbf{f}}_{j+\frac{1}{2}}^{L,m} \right) + \hat{\mathbf{f}}_{j+\frac{1}{2}}^{L,m},$$

*is an IDP flux, for  $m = 0, 1, 2, \dots, M$ , then  $\mathbf{u}_j^{n+1} = \mathbf{u}_j^n - \lambda(\hat{\mathbf{f}}_{j+\frac{1}{2}} - \hat{\mathbf{f}}_{j-\frac{1}{2}})$  with  $\hat{\mathbf{f}}_{j+\frac{1}{2}} := \hat{\mathbf{f}}_{j+\frac{1}{2}}^{L,M+1}$  gives a more accurate IDP scheme than (5.3), but with a higher computational cost.*

**5.2. Zalesak's FCT limiter for scalar conservation laws.** This method was originally proposed in [239] for preserving local maximum principle of scalar conservation laws. A detailed description of design principles behind FCT algorithms for structured grids can be found in Zalesak's book chapter [240]. See [135, 239, 139] for the general version for multidimensional problems and unstructured meshes. As mentioned in Section 1.2, if enforcing a maximum principle like  $\min_j u_j^n \leq u_j^{n+1} \leq \max_j u_j^n$ , then it would be at most second order accurate for smooth solutions. Thus for higher order accuracy, we consider enforcing the bound-preserving property, i.e., the invariant domain is a simple interval  $G = [m, M]$  with  $m$  and  $M$  being the lower and upper bounds of the initial condition.

**5.2.1. The original version of Zalesak's FCT limiter.** Zalesak's [239] classical FCT algorithm determines the parameter  $\theta_{j+\frac{1}{2}}$  as follows:

- Define

$$\begin{aligned} P_j^+ &= \max \{0, -\delta_{j+1/2}\} + \max \{0, \delta_{j-1/2}\}, \\ P_j^- &= \min \{0, -\delta_{j+1/2}\} + \min \{0, \delta_{j-1/2}\}. \end{aligned}$$

- Compute

$$(5.6) \quad Q_j^+ = \frac{1}{\lambda} (M - u_j^{n+1,L}), \quad Q_j^- = \frac{1}{\lambda} (m - u_j^{n+1,L}).$$

- Calculate

$$R_j^+ = \min \left\{ 1, \frac{Q_j^+}{P_j^+} \right\}, \quad R_j^- = \min \left\{ 1, \frac{Q_j^-}{P_j^-} \right\}.$$

- Obtain the IDP limiting parameter

$$(5.7) \quad \theta_{j+1/2} = \begin{cases} \min \{R_{j+1}^+, R_j^-\} & \text{if } \delta_{j+1/2} \geq 0, \\ \min \{R_{j+1}^-, R_j^+\} & \text{if } \delta_{j+1/2} < 0. \end{cases}$$

The formula (5.7) can be derived from enforcing constraints, e.g., see [148, Section 3]. Thus (5.3) with (5.7) gives the bound-preserving property  $u_j^{n+1} \in [m, M]$ .

**5.2.2. Parametrized flux limiters.** In order to have a better understanding of the Zalesak's algorithm for computing  $\theta_{j+\frac{1}{2}}$ , we first consider an alternative way to find  $\theta_{j+\frac{1}{2}}$  in (5.5). It is possible to decouple the constraints of  $\{\theta_{j+\frac{1}{2}}\}$  in (5.5) by a parametrized method, proposed in [120, 233]. Specifically, the parametrized method seeks a group of locally defined parameters  $\Lambda_{j+\frac{1}{2}}$  such that the IDP property is maintained for any  $\theta_{j+\frac{1}{2}} \in [0, \Lambda_{j+\frac{1}{2}}]$ . Then for the sake of minimal correction to maintain high order accuracy, one should simply take  $\theta_{j+\frac{1}{2}} = \Lambda_{j+\frac{1}{2}}$ .

For a finite difference scheme (5.1) at a grid point  $x_j$ , define an interval  $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$ . Parameters  $\Lambda_{-\frac{1}{2}, I_j}$  and  $\Lambda_{+\frac{1}{2}, I_j}$  will be constructed such that

$$(5.8) \quad \theta_{j+\frac{1}{2}} \in [0, \Lambda_{+\frac{1}{2}, I_j}] \cap [0, \Lambda_{-\frac{1}{2}, I_{j+1}}],$$

is sufficient for the scheme (5.3) to preserve the desired bounds. In (5.8),  $\theta_{j+\frac{1}{2}}$  at  $x_{j+\frac{1}{2}}$  satisfies some constraints in  $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$  and also  $I_{j+1} = [x_{j+\frac{1}{2}}, x_{j+\frac{3}{2}}]$ . The subscripts  $+\frac{1}{2}$  and  $-\frac{1}{2}$  in  $\Lambda_{\pm\frac{1}{2}, I_j}$  can be understood as a shift from the cell center of  $I_j$ , e.g., both  $\Lambda_{+\frac{1}{2}, I_j}$  and  $\Lambda_{-\frac{1}{2}, I_{j+1}}$  correspond to the grid point  $x_{j+\frac{1}{2}}$ .

Define

$$(5.9) \quad \Gamma_j^M = M - u_j^n + \lambda \left( \hat{f}_{j+\frac{1}{2}}^L - \hat{f}_{j-\frac{1}{2}}^L \right), \quad \Gamma_j^m = m - u_j^n + \lambda \left( \hat{f}_{j+\frac{1}{2}}^L - \hat{f}_{j-\frac{1}{2}}^L \right).$$

The IDP property of a first order monotone scheme yields

$$(5.10) \quad \Gamma_j^M \geq 0, \quad \Gamma_j^m \leq 0.$$

To maintain  $u_j^{n+1} \in [m, M]$ ,  $\theta_{j+\frac{1}{2}}$  must satisfy (5.5) with  $g(u) = M - u$  and  $g(u) = u - m$ , respectively, i.e.,

$$(5.11) \quad L_M(\theta_{j-\frac{1}{2}}, \theta_{j+\frac{1}{2}}) := \lambda \theta_{j-\frac{1}{2}} \delta_{j-\frac{1}{2}} - \lambda \theta_{j+\frac{1}{2}} \delta_{j+\frac{1}{2}} - \Gamma_j^M \leq 0,$$

$$(5.12) \quad L_m(\theta_{j-\frac{1}{2}}, \theta_{j+\frac{1}{2}}) := \lambda \theta_{j-\frac{1}{2}} \delta_{j-\frac{1}{2}} - \lambda \theta_{j+\frac{1}{2}} \delta_{j+\frac{1}{2}} - \Gamma_j^m \geq 0.$$

The parameters  $\Lambda_{\pm\frac{1}{2}, I_j}^M$  and  $\Lambda_{\pm\frac{1}{2}, I_j}^m$  are constructed as follows:

*Upper Bound Preservation.* The parameters  $\Lambda_{\pm\frac{1}{2},I_j}^M$  are defined to satisfy (5.11).

- (a) If  $\delta_{j-\frac{1}{2}} \leq 0$  and  $\delta_{j+\frac{1}{2}} \geq 0$ , define  $(\Lambda_{-\frac{1}{2},I_j}^M, \Lambda_{+\frac{1}{2},I_j}^M) = (1, 1)$ .  
 (b) If  $\delta_{j-\frac{1}{2}} \leq 0$  and  $\delta_{j+\frac{1}{2}} < 0$ , define

$$(5.13) \quad (\Lambda_{-\frac{1}{2},I_j}^M, \Lambda_{+\frac{1}{2},I_j}^M) = \left(1, \min \left\{1, \frac{\Gamma_j^M}{-\lambda\delta_{j+\frac{1}{2}} + \epsilon}\right\}\right).$$

- (c) If  $\delta_{j-\frac{1}{2}} > 0$  and  $\delta_{j+\frac{1}{2}} \geq 0$ , define

$$(5.14) \quad (\Lambda_{-\frac{1}{2},I_j}^M, \Lambda_{+\frac{1}{2},I_j}^M) = \left(\min \left\{1, \frac{\Gamma_j^M}{\lambda\delta_{j-\frac{1}{2}} + \epsilon}\right\}, 1\right).$$

- (d) If  $\delta_{j-\frac{1}{2}} > 0$  and  $\delta_{j+\frac{1}{2}} < 0$ :

$$(5.15) \quad (\Lambda_{-\frac{1}{2},I_j}^M, \Lambda_{+\frac{1}{2},I_j}^M) = (\Lambda_0, \Lambda_0), \quad \Lambda_0 = \min \left\{1, \frac{\Gamma_j^M}{\lambda\delta_{j-\frac{1}{2}} - \lambda\delta_{j+\frac{1}{2}} + \epsilon}\right\}.$$

Here,  $\epsilon$  is a small positive parameter, which is slightly above machine zero, to prevent division by zero. For convenience, they can also be equivalently written as

$$(5.16a) \quad \Lambda_{+\frac{1}{2},I_j}^M = \begin{cases} 1, & \text{if } \delta_{j+\frac{1}{2}} \geq 0 \\ \min \left\{1, \frac{\Gamma_j^M}{\lambda \max\{0, \delta_{j-\frac{1}{2}}\} - \lambda\delta_{j+\frac{1}{2}} + \epsilon}\right\}, & \text{if } \delta_{j+\frac{1}{2}} < 0 \end{cases},$$

$$(5.16b) \quad \Lambda_{-\frac{1}{2},I_j}^M = \begin{cases} 1, & \text{if } \delta_{j-\frac{1}{2}} \leq 0 \\ \min \left\{1, \frac{\Gamma_j^M}{\lambda\delta_{j-\frac{1}{2}} - \lambda \min\{0, \delta_{j+\frac{1}{2}}\} + \epsilon}\right\}, & \text{if } \delta_{j-\frac{1}{2}} > 0 \end{cases}.$$

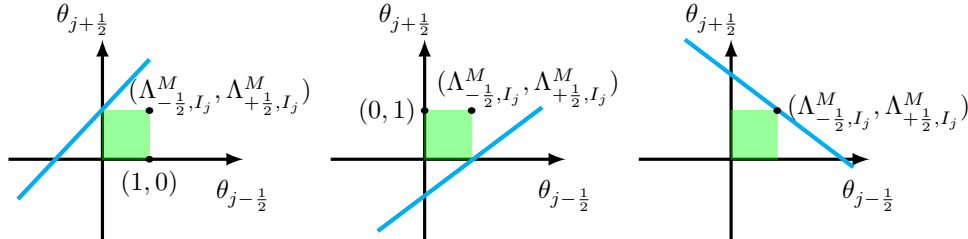
The first fact is that any smaller  $\theta_{j\pm\frac{1}{2}}$  than  $\Lambda_{\pm\frac{1}{2},I_j}^M$  also satisfy (5.11). Notice that Lemma 5.1 does not imply this result.

LEMMA 5.2. *The upper bound (5.11) is satisfied for  $\theta_{j-\frac{1}{2}} \in [0, \Lambda_{-\frac{1}{2},I_j}^M]$  and  $\theta_{j+\frac{1}{2}} \in [0, \Lambda_{+\frac{1}{2},I_j}^M]$ .*

*Proof.* Regard  $\theta_{j-\frac{1}{2}}$  and  $\theta_{j+\frac{1}{2}}$  as unknowns and consider a line equation

$$(5.17) \quad L_M(\theta_{j-\frac{1}{2}}, \theta_{j+\frac{1}{2}}) := \lambda\theta_{j-\frac{1}{2}}\delta_{j-\frac{1}{2}} - \lambda\theta_{j+\frac{1}{2}}\delta_{j+\frac{1}{2}} - \Gamma_j^M = 0.$$

Then we discuss it case by case. For case (a),  $L_M$  is a decreasing function, thus  $L_M(\theta_{j-\frac{1}{2}}, \theta_{j+\frac{1}{2}}) \leq L_M(0, 0) = -\Gamma_j^M \leq 0$ .



The cases (b), (c) and (d) are illustrated in the figures, in which the blue line (5.17) is the zero line, i.e.,  $L_M(\theta_{j-\frac{1}{2}}, \theta_{j+\frac{1}{2}}) = 0$ . Since  $L_M(0, 0) \leq 0$  and the green rectangle is on the same side as  $(0, 0)$  w.r.t. the zero line, any  $(\theta_{j-\frac{1}{2}}, \theta_{j+\frac{1}{2}})$  in the green rectangle achieves  $L_M(\theta_{j-\frac{1}{2}}, \theta_{j+\frac{1}{2}}) \leq 0$ . The proof is concluded.  $\square$

*Lower Bound Preservation.* Similarly, the parameters  $\Lambda_{\pm\frac{1}{2}, I_j}^m$  are defined to satisfy (5.12).

- (a) If  $\delta_{j-\frac{1}{2}} \geq 0, \delta_{j+\frac{1}{2}} \leq 0$ , define  $(\Lambda_{-\frac{1}{2}, I_j}^m, \Lambda_{+\frac{1}{2}, I_j}^m) = (1, 1)$ .
- (b) If  $\delta_{j-\frac{1}{2}} \geq 0, \delta_{j+\frac{1}{2}} > 0$ , define  $(\Lambda_{-\frac{1}{2}, I_j}^m, \Lambda_{+\frac{1}{2}, I_j}^m) = \left(1, \min \left\{1, \frac{\Gamma_j^m}{-\lambda\delta_{j+\frac{1}{2}} - \epsilon}\right\}\right)$ .
- (c) If  $\delta_{j-\frac{1}{2}} < 0, \delta_{j+\frac{1}{2}} \leq 0$ , define  $(\Lambda_{-\frac{1}{2}, I_j}^m, \Lambda_{+\frac{1}{2}, I_j}^m) = \left(\min \left\{1, \frac{\Gamma_j^m}{\lambda\delta_{j-\frac{1}{2}} - \epsilon}\right\}, 1\right)$ .
- (d) If  $\delta_{j-\frac{1}{2}} < 0, \delta_{j+\frac{1}{2}} > 0$ , define  $(\Lambda_{-\frac{1}{2}, I_j}^m, \Lambda_{+\frac{1}{2}, I_j}^m) = (\Lambda_0, \Lambda_0)$  with  $\Lambda_0 = \min \left\{1, \frac{\Gamma_j^m}{\lambda\delta_{j-\frac{1}{2}} - \lambda\delta_{j+\frac{1}{2}} - \epsilon}\right\}$ .

Similar to Lemma 5.2, any smaller  $\theta$  would also enforce the lower bound, i.e.,  $\theta_{j-\frac{1}{2}} \in [0, \Lambda_{-\frac{1}{2}, I_j}^m]$  and  $\theta_{j+\frac{1}{2}} \in [0, \Lambda_{+\frac{1}{2}, I_j}^m]$  satisfy (5.12).

The final limiting parameter combines upper and lower bounds

$$(5.18) \quad \Lambda_{+\frac{1}{2}, I_j} = \min \left\{ \Lambda_{+\frac{1}{2}, I_j}^M, \Lambda_{+\frac{1}{2}, I_j}^m \right\}, \quad \Lambda_{-\frac{1}{2}, I_{j+1}} = \min \left\{ \Lambda_{-\frac{1}{2}, I_{j+1}}^M, \Lambda_{-\frac{1}{2}, I_{j+1}}^m \right\}.$$

Implied by Lemma 5.2, we have

**THEOREM 5.2.** *Let  $\Lambda_{j+\frac{1}{2}} := \min\{\Lambda_{+\frac{1}{2}, I_j}, \Lambda_{-\frac{1}{2}, I_{j+1}}\}$ , then  $\theta_{j+\frac{1}{2}} \in [0, \Lambda_{j+\frac{1}{2}}]$  satisfy both (5.11) and (5.12). In particular, the modified flux (5.3b) with  $\theta_{j+\frac{1}{2}} = \Lambda_{j+\frac{1}{2}}$  is IDP for  $G = [m, M]$ .*

**5.2.3. Equivalence of Zalesak's FCT limiter and the parametrized flux limiter.** Although the parametrized flux limiter has a different formula, its final result for  $\theta_{j+\frac{1}{2}}$  is actually equivalent to the Zalesak's FCT limiter for enforcing bounds.

**THEOREM 5.3.** *The parameterized limiter is equivalent to Zalesak's FCT limiter, i.e., (5.7) is equal to  $\Lambda_{j+\frac{1}{2}} = \min\{\Lambda_{+\frac{1}{2}, I_j}, \Lambda_{-\frac{1}{2}, I_{j+1}}\}$ .*

*Proof.* Without loss of generality, we only consider the preservation of the upper bound  $u \leq M$ . For considering only the upper bound, the limiting parameter  $\theta_{j+\frac{1}{2}}$  given by Zalesak's FCT limiter becomes

$$(5.19) \quad \theta_{j+1/2} = \begin{cases} R_{j+1}^+, & \text{if } \delta_{j+1/2} \geq 0, \\ R_j^+, & \text{if } \delta_{j+1/2} < 0, \end{cases}$$

where  $R_j^+ = \min \left\{ 1, \frac{Q_j^+}{P_j^+} \right\}$ ,  $Q_j^+ = \frac{\Delta x}{\Delta t} (M - u_j^{n+1, L})$ , and

$$P_j^+ = \max \{0, -\delta_{j+1/2}\} + \max \{0, \delta_{j-1/2}\}.$$

First,  $\Gamma_j^M$  defined in (5.9) simply  $\Gamma_j^M = \lambda Q_j^+$ , with which (5.16) gives

$$\begin{aligned}\Lambda_{+\frac{1}{2}, I_j}^M &= 1, \quad \text{if } \delta_{j+\frac{1}{2}} \geq 0 \\ \Lambda_{-\frac{1}{2}, I_{j+1}}^M &= \min \left\{ 1, \frac{\Gamma_{j+1}^M}{\lambda \delta_{j+\frac{1}{2}} - \lambda \min\{0, \delta_{j+\frac{3}{2}}\} + \epsilon} \right\} \\ &= \min \left\{ 1, \frac{Q_{j+1}^+}{\delta_{j+\frac{1}{2}} + \max\{0, -\delta_{j+\frac{3}{2}}\}} \right\} = R_{j+1}^+, \quad \text{if } \delta_{j+\frac{1}{2}} \geq 0\end{aligned}$$

thus

$$\min \left\{ \Lambda_{+\frac{1}{2}, I_j}^M, \Lambda_{-\frac{1}{2}, I_{j+1}}^M \right\} = R_{j+1}^+,$$

which implies the equivalence between the parameterized limiter and Zalesak's FCT limiter for the case  $\delta_{j+\frac{1}{2}} \geq 0$ . Similarly, (5.16) gives

$$\begin{aligned}\Lambda_{+\frac{1}{2}, I_j}^M &= \min \left\{ 1, \frac{\Gamma_j^M}{\lambda \max\{0, \delta_{j-\frac{1}{2}}\} - \lambda \delta_{j+\frac{1}{2}} + \epsilon} \right\} \\ &= \min \left\{ 1, \frac{Q_j^+}{\max\{0, \delta_{j-\frac{1}{2}}\} - \delta_{j+\frac{1}{2}}} \right\} = R_j^+, \quad \text{if } \delta_{j+\frac{1}{2}} < 0 \\ \Lambda_{-\frac{1}{2}, I_{j+1}}^M &= 1, \quad \text{if } \delta_{j+\frac{1}{2}} < 0,\end{aligned}$$

thus  $\min \left\{ \Lambda_{+\frac{1}{2}, I_j}^M, \Lambda_{-\frac{1}{2}, I_{j+1}}^M \right\} = R_j^+$ , which implies the equivalence between the parameterized limiter and Zalesak's FCT limiter when  $\delta_{j+\frac{1}{2}} < 0$ . The proof is completed.  $\square$

**5.2.4. High order accuracy.** By Theorem 5.2, the scheme (5.3) with the Zalesak's FCT flux correction for scalar equations preserves  $G = [m, M]$ . In practice, such a flux correction can be applied to many high order schemes such as finite difference, finite volume and discontinuous Galerkin (DG) schemes, and high order accuracy can be observed numerically for sufficiently small time steps. For finite volume schemes solving 1D linear equation, the truncation error of the flux correction in this subsection can be shown high order accurate for smooth solutions under a reasonable time step constraint [233].

For high order finite difference schemes with SSP Runge-Kutta methods, if applying the flux correction to each time stage in Runge-Kutta methods, then high order accuracy can be maintained only under extremely small time steps such as  $\Delta t = \mathcal{O}(\Delta x^{1.5})$ , see [233]. To recover high order accuracy under a reasonable time step, one simple remedy is to apply the flux correction only at the final time stage of a Runge-Kutta method, with which third order accuracy of a finite difference scheme can be proven for solving 1D linear equation with a practical time step [228]. Such a flux correction can be applied to any high order time discretization such as Lax-Wendroff time stepping, and can also be extended to explicit time stepping for a convection dominated diffusion equation [229].

**5.3. Parametrized flux limiters for hyperbolic systems.** The parametrized flux limiting method can be extended to several hyperbolic systems. As an example, we review how it can be applied to a finite difference scheme for the compressible Euler equations [230], which is one kind of extension of Zalesak's FCT limiter to systems. See also [156] and references therein for another extension of Zalesak's FCT limiter to

systems. For simplicity, we only consider the one dimensional case since the multiple dimensional scheme can be defined similarly in a dimension by dimension fashion for a finite difference scheme.

For the positivity of density  $\rho$  and pressure  $p$  in the Euler equations, introduce two threshold parameters:  $\epsilon_\rho = \min_j(\rho_j^{n+1,L}, 10^{-13})$  and  $\epsilon_p = \min_j(p_j^{n+1,L}, 10^{-13})$ , where  $\rho_j^{n+1,L}$  and  $p_j^{n+1,L}$  denote density and pressure computed by a first order IDP scheme. Let  $(\hat{f}^{\rho,L}, \hat{f}^{m,L}, \hat{f}^{E,L})^\top$  represent the components of the first order IDP flux  $\hat{\mathbf{f}}^L$ . Similarly,  $\hat{\mathbf{f}}^H = (\hat{f}^{\rho,H}, \hat{f}^{m,H}, \hat{f}^{E,H})^\top$  denotes a high order numerical flux in a finite difference scheme, e.g., finite difference WENO scheme. Let  $\hat{\mathbf{f}} = (\hat{f}^\rho, \hat{f}^m, \hat{f}^E)^\top$  be the corrected flux. The method proceeds in two steps:

First, follow the scalar case to determine the limiting parameters  $\theta_{j\pm\frac{1}{2}}$  to enforce the positivity of density, i.e. to maintain

$$\rho_j^{n+1} = \rho_j^n - \lambda \left( \hat{f}_{j+\frac{1}{2}}^\rho - \hat{f}_{j-\frac{1}{2}}^\rho \right) \geq \epsilon_\rho.$$

Obtain  $(\Lambda_{-\frac{1}{2},I_j}^\rho, \Lambda_{+\frac{1}{2},I_j}^\rho)$  and define a rectangular box region:

$$(5.20) \quad S_\rho = \left\{ (\theta_{j-\frac{1}{2}}, \theta_{j+\frac{1}{2}}) : 0 \leq \theta_{j-\frac{1}{2}} \leq \Lambda_{-\frac{1}{2},I_j}^\rho, 0 \leq \theta_{j+\frac{1}{2}} \leq \Lambda_{+\frac{1}{2},I_j}^\rho \right\}.$$

Let  $A^1 = (0, \Lambda_{+\frac{1}{2},I_j}^\rho)$ ,  $A^2 = (\Lambda_{-\frac{1}{2},I_j}^\rho, 0)$ , and  $A^3 = (\Lambda_{-\frac{1}{2},I_j}^\rho, \Lambda_{+\frac{1}{2},I_j}^\rho)$  denote vertices of  $S_\rho$ , as shown in Figure 11.

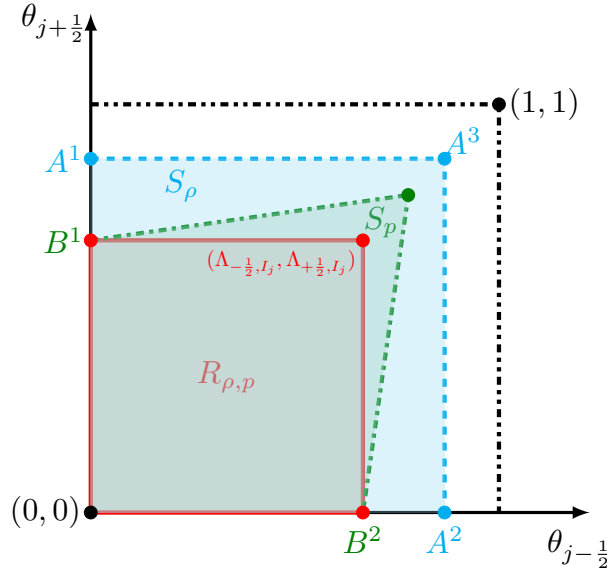


Fig. 11: An illustration of the parameters for enforcing positivity of both density and pressure: the first step is to find a box region  $S_\rho$  (cyan color rectangle bounded by dashed lines) with four vertices  $(0,0), A^1, A^2, A^3$ , the second step is to find  $B^i = rA^i$  with  $r \in [0,1]$  such that  $p(B^i) \geq \epsilon_p$  for each  $i$ . The convex hull of vertices  $(0,0), B^1, B^2, B^3$  is the green polygon  $S_p \subset S_{\rho,p}$ , and the largest rectangle inside  $S_p$  with two sides along axes is the red rectangle  $R_{\rho,p} \subset S_{\rho,p}$ .

The second step is to enforce pressure positivity without losing positivity of density. For the scheme with corrected flux (5.3), its numerical solution  $\mathbf{u}_j^{n+1}$  can be regarded as a function of  $\theta_{j\pm\frac{1}{2}}$ . For the ideal gas EOS (2.1b), the pressure function is concave w.r.t.  $\mathbf{u} = (\rho \quad m \quad E)^T$ , thus the proof of Lemma 5.1 implies the following set is a convex set,

$$(5.21) \quad S_{\rho,p} = \left\{ (\theta_{j-\frac{1}{2}}, \theta_{j+\frac{1}{2}}) \in S_\rho : p_j^{n+1}(\theta_{j-\frac{1}{2}}, \theta_{j+\frac{1}{2}}) = (\gamma - 1) \left( E_j^{n+1} - \frac{(m_j^{n+1})^2}{2\rho_j^{n+1}} \right) \geq \epsilon_p \right\}.$$

Next, a box region  $R_{\rho,p}$  in the plane of two variables  $(\theta_{j-\frac{1}{2}}, \theta_{j+\frac{1}{2}})$  for enforcing positivity of both density and pressure is constructed as follows and also illustrated in Figure 11:

1. For each vertex  $A^\ell$ , find  $B^\ell = rA^\ell$  with  $r \in [0, 1]$  such that  $p(B^\ell) \geq \epsilon_p$ . The convex hull of  $(0, 0)$  and  $B^\ell$  forms  $S_p \subset S_{\rho,p}$ .
2. Define the rectangle  $R_{\rho,p} = \left[0, \Lambda_{-\frac{1}{2}, I_j}\right] \times \left[0, \Lambda_{+\frac{1}{2}, I_j}\right]$ , with limiting parameters:

$$(5.22) \quad (\Lambda_{-\frac{1}{2}, I_j}, \Lambda_{+\frac{1}{2}, I_j}) = (\min(B_1^2, B_1^3), \min(B_2^1, B_2^3)).$$

The construction above and convexity of  $S_{\rho,p}$  ensure that  $R_{\rho,p} \subset S_{\rho,p}$ . Finally, the limiting parameter is given by  $\theta_{j+\frac{1}{2}} = \min(\Lambda_{-\frac{1}{2}, I_{j+1}}, \Lambda_{+\frac{1}{2}, I_j})$ , and such  $\theta_{j+\frac{1}{2}}$  is inside all  $R_{\rho,p}$  constructed thus IDP is achieved in (5.3).

Such a flux limiting method can be extended to other schemes such as DG methods [229] and finite volume schemes on unstructured meshes [53].

**5.4. A simple flux limiting based on Assumption 2.** In the literature, there are other simpler decoupling methods to find sufficient conditions for satisfying (5.4). We first review a simple flux limiting method introduced by Hu, Adams and Shu in [115] for compressible Euler equations, which can be extended to hyperbolic systems with  $G$  satisfying Assumption 2 (and Assumption 5 in multiple dimensions). It should be noted that scalar conservation laws with  $G = [U_{\min}, U_{\max}]$  do not satisfy Assumption 2 in general.

For the compressible Euler equations, Assumption 2 holds, implied by Lemma 3.2. In the Hu–Adams–Shu method [115], the low order IDP flux is chosen as the first order Lax–Friedrichs flux, e.g.,

$$(5.23) \quad \mathbf{f}_{j+\frac{1}{2}}^L = \frac{1}{2}[\mathbf{f}(\mathbf{u}_j^n) + \mathbf{f}(\mathbf{u}_{j+1}^n) - \alpha(\mathbf{u}_{j+1}^n - \mathbf{u}_j^n)].$$

The first order Lax–Friedrichs scheme can be written as

$$\mathbf{u}_j^{n+1,L} = \frac{1}{2}\mathbf{u}_j^{L,+} + \frac{1}{2}\mathbf{u}_j^{L,-}, \quad \mathbf{u}_j^{L,\pm} := \mathbf{u}_j^n \mp 2\lambda\hat{\mathbf{f}}_{j\pm\frac{1}{2}}^L.$$

Under the CFL condition  $\alpha\lambda \leq \frac{1}{2}$  and the assumption  $\mathbf{u}_j^n, \mathbf{u}_{j\pm 1}^n \in G$ , Assumption 2 implies

$$(5.24) \quad \mathbf{u}_j^{L,\pm} = (1 - 2\alpha\lambda)\mathbf{u}_j^n + \alpha\lambda \left( \mathbf{u}_j^n \mp \frac{\mathbf{f}(\mathbf{u}_j^n)}{\alpha} \right) + \alpha\lambda \left( \mathbf{u}_{j\pm 1}^n \mp \frac{\mathbf{f}(\mathbf{u}_{j\pm 1}^n)}{\alpha} \right) \in G.$$

In order to decouple the constraints of  $\{\theta_{j+\frac{1}{2}}\}$  in (5.5), the Hu–Adams–Shu method decomposes the original high order scheme (5.1) into two parts:

$$(5.25) \quad \mathbf{u}_j^{n+1,H} = \frac{1}{2} \left( \mathbf{u}_j^n + 2\lambda \hat{\mathbf{f}}_{j-\frac{1}{2}}^H \right) + \frac{1}{2} \left( \mathbf{u}_j^n - 2\lambda \hat{\mathbf{f}}_{j+\frac{1}{2}}^H \right) := \frac{1}{2} \mathbf{u}_j^{H,-} + \frac{1}{2} \mathbf{u}_j^{H,+}.$$

Then the scheme (5.3) can be written as

$$(5.26) \quad \mathbf{u}_j^{n+1} = \frac{1}{2} \left[ (1 - \theta_{j-\frac{1}{2}}) \mathbf{u}_j^{L,-} + \theta_{j-\frac{1}{2}} \mathbf{u}_j^{H,-} \right] + \frac{1}{2} \left[ (1 - \theta_{j+\frac{1}{2}}) \mathbf{u}_j^{L,+} + \theta_{j+\frac{1}{2}} \mathbf{u}_j^{H,+} \right],$$

and the basic idea of the Hu–Adams–Shu method is to use  $\mathbf{u}_j^{L,\pm} \in G$  to limit  $\mathbf{u}_j^{H,\pm}$ .

For simplicity, assume  $G$  in (1.2) is defined by concave functions  $g_i(\mathbf{u}) > 0, i = 1, \dots, N$ , thus  $g_i$  satisfies the Jensen inequality (2.3), such as the positivity of density or pressure in the compressible Euler equations. Define

$$(5.27a) \quad \theta_{j+\frac{1}{2}} = \min_i \theta_{j+\frac{1}{2}}^i, \quad \theta_{j+\frac{1}{2}}^i = \min \left\{ \theta_{j+\frac{1}{2}}^{i,+}, \theta_{j+\frac{1}{2}}^{i,-} \right\},$$

$$(5.27b) \quad \theta_{j+\frac{1}{2}}^{i,+} = \begin{cases} 1, & \text{if } g_i(\mathbf{u}_j^{H,+}) \geq \epsilon \\ \text{Solution to } (1 - \theta)g(\mathbf{u}_j^{L,+}) + \theta g(\mathbf{u}_j^{H,+}) = \epsilon, & \text{if } g_i(\mathbf{u}_j^{H,+}) < \epsilon \end{cases},$$

$$(5.27c) \quad \theta_{j+\frac{1}{2}}^{i,-} = \begin{cases} 1, & \text{if } g_i(\mathbf{u}_{j+1}^{H,-}) \geq \epsilon \\ \text{Solution to } (1 - \theta)g(\mathbf{u}_{j+1}^{L,-}) + \theta g(\mathbf{u}_{j+1}^{H,-}) = \epsilon, & \text{if } g_i(\mathbf{u}_{j+1}^{H,-}) < \epsilon \end{cases}.$$

Then (5.3) with  $\theta_{j+\frac{1}{2}}$  in (5.27) is IDP because of the following facts.

LEMMA 5.4. *Let  $\mathbf{u}^L, \mathbf{u}^H$  be two given vectors satisfying  $g(\mathbf{u}^L) \geq \epsilon > 0$  for a concave function  $g$ . Let  $\hat{\theta} \in [0, 1]$  be a solution to  $(1 - \hat{\theta})g(\mathbf{u}^L) + \hat{\theta}g(\mathbf{u}^H) = \epsilon$ , then*

$$g[(1 - \theta)\mathbf{u}^L + \theta\mathbf{u}^H] \geq \epsilon, \quad \forall \theta \in [0, \hat{\theta}].$$

*Proof.* Let  $a = \theta/\hat{\theta}$ , then  $a \in [0, 1]$ , thus Jensen's inequality from concavity gives

$$\begin{aligned} g[(1 - \theta)\mathbf{u}^L + \theta\mathbf{u}^H] &= g[(1 - a)\mathbf{u}^L + a(1 - \hat{\theta})\mathbf{u}^L + a\hat{\theta}\mathbf{u}^H] \\ &\geq (1 - a)g(\mathbf{u}^L) + a(1 - \hat{\theta})g(\mathbf{u}^L) + a\hat{\theta}g(\mathbf{u}^H) \geq \epsilon. \end{aligned} \quad \square$$

LEMMA 5.5. *If  $\theta_{j+\frac{1}{2}}^i$  is obtained from Equation (5.27), then (5.3) with any non-negative  $\theta_{j+\frac{1}{2}} \leq \theta_{j+\frac{1}{2}}^i$  satisfies  $g_i(\mathbf{u}^{n+1}) > 0$  for a concave function  $g_i$ .*

*Proof.* Lemma 5.4 implies

$$g_i[(1 - \theta_{j+\frac{1}{2}})\mathbf{u}_j^{L,+} + \theta_{j+\frac{1}{2}}\mathbf{u}_j^{H,+}] > 0, \quad g_i[(1 - \theta_{j-\frac{1}{2}})\mathbf{u}_j^{L,-} + \theta_{j-\frac{1}{2}}\mathbf{u}_j^{H,-}] > 0.$$

With (5.26), Jensen's inequality gives  $g(\mathbf{u}_j^{n+1}) > 0$ .  $\square$

**5.5. The convex limiting based on Assumption 1.** The convex limiting by Guermond, Popov and Tomas in [108] is another flux limiting approach adapted from the convex limiting technique for the continuous finite element method [103, 98], and applies to the DG, finite volume and finite difference schemes [108] solving general hyperbolic systems satisfying Assumption 1 (and Assumption 4 in multiple dimensions). See also [57, 137] for using flux limiters for conservation laws. In [131], Kuzmin introduced the monolithic convex limiting which can be applied to a semi-discrete scheme. As an example, we briefly review how these convex limiting methods apply to finite difference and continuous finite element methods.

**5.5.1. Finite difference schemes with monolithic convex limiting.** We describe the monolithic convex limiting approach in [131] which can be used with any IDP flux [133, Section 2.5.6.2]. By adding and subtracting  $\lambda \mathbf{f}(\mathbf{u}_j^n)$ , the original high order scheme (5.1) can be rewritten as

$$\mathbf{u}_j^{n+1,H} = \frac{1}{2} \left( \mathbf{u}_j^n + 2\lambda (\hat{\mathbf{f}}_{j-\frac{1}{2}}^H - \mathbf{f}(\mathbf{u}_j^n)) \right) + \frac{1}{2} \left( \mathbf{u}_j^n - 2\lambda (\hat{\mathbf{f}}_{j+\frac{1}{2}}^H - \mathbf{f}(\mathbf{u}_j^n)) \right) =: \frac{1}{2} \mathbf{u}_j^{H,-} + \frac{1}{2} \mathbf{u}_j^{H,+}.$$

This can be regarded as a decomposition of the high order scheme (5.1) into *residuals* as in the residual distribution schemes [2, 5], which is however different from the decomposition in the Hu-Adam-Shu approach.

Now consider any first order IDP scheme (5.2) with  $\hat{\mathbf{f}}_{j+\frac{1}{2}}^L = \hat{\mathbf{f}}(\mathbf{u}_j^n, \mathbf{u}_{j+1}^n)$  and  $\hat{\mathbf{f}}$  is any consistent IDP flux including Lax-Friedrichs, Godunov, and HLLE fluxes. Then it can be also rewritten as

$$\mathbf{u}_j^{n+1,L} = \frac{1}{2} \left( \mathbf{u}_j^n + 2\lambda (\hat{\mathbf{f}}_{j-\frac{1}{2}}^L - \mathbf{f}(\mathbf{u}_j^n)) \right) + \frac{1}{2} \left( \mathbf{u}_j^n - 2\lambda (\hat{\mathbf{f}}_{j+\frac{1}{2}}^L - \mathbf{f}(\mathbf{u}_j^n)) \right) =: \frac{1}{2} \mathbf{u}_j^{L,-} + \frac{1}{2} \mathbf{u}_j^{L,+}.$$

If the first order scheme (5.2) is IDP under CFL condition  $\lambda \max_{\mathbf{u}} |\mathbf{f}'(\mathbf{u})| \leq a_0$ , then the following holds under halved CFL condition,

$$(5.28) \quad \mathbf{u}_j^{L,\pm} := \mathbf{u}_j^n \mp 2\lambda (\hat{\mathbf{f}}_{j\pm\frac{1}{2}}^L - \mathbf{f}(\mathbf{u}_j^n)) \in G, \quad \lambda \max_{\mathbf{u}} |\mathbf{f}'(\mathbf{u})| \leq \frac{1}{2} a_0.$$

To see why (5.28) is true, we focus on  $\mathbf{u}_j^{L,-}$ . Notice that it can be written as another first order scheme

$$(5.29) \quad \mathbf{u}_j^{L,-} = \mathbf{u}_j^n + 2\lambda (\hat{\mathbf{f}}_{j-\frac{1}{2}}^L - \mathbf{f}(\mathbf{u}_j^n)) = \mathbf{u}_j^n - 2\lambda (\hat{\mathbf{f}}(\mathbf{u}_j^n, \mathbf{u}_j^n) - \hat{\mathbf{f}}(\mathbf{u}_{j-1}^n, \mathbf{u}_j^n)),$$

which is implied by  $\hat{\mathbf{f}}_{j-\frac{1}{2}}^L = \hat{\mathbf{f}}(\mathbf{u}_{j-1}^n, \mathbf{u}_j^n)$  and the consistency of the flux function  $\hat{\mathbf{f}}(\mathbf{u}_j^n, \mathbf{u}_j^n) = \mathbf{f}(\mathbf{u}_j^n)$ . Since (5.29) is in the same form as (1.4) but with doubled time step, it is also IDP under halved CFL.

Then the scheme (5.3) can be written as

$$(5.30) \quad \mathbf{u}_j^{n+1} = \frac{1}{2} \left[ (1 - \theta_{j-\frac{1}{2}}) \mathbf{u}_j^{L,-} + \theta_{j-\frac{1}{2}} \mathbf{u}_j^{H,-} \right] + \frac{1}{2} \left[ (1 - \theta_{j+\frac{1}{2}}) \mathbf{u}_j^{L,+} + \theta_{j+\frac{1}{2}} \mathbf{u}_j^{H,+} \right],$$

and we can use  $\mathbf{u}_j^{L,\pm} \in G$  to correct  $\mathbf{u}_j^{H,\pm}$  in the same way as in the Hu-Adam-Shu approach. For simplicity, assume  $G$  in (1.2) is defined by concave functions  $g_i(\mathbf{u}) > 0, i = 1, \dots, N$ . Then (5.3) with  $\theta_{j+\frac{1}{2}}$  in (5.27) is IDP due to Lemma 5.5.

**REMARK 5.3.** *The limiting approaches in Subsection 5.4 and Subsection 5.5.1 are defined via quantities of  $\mathbf{u}_j^n$  without involving  $\mathbf{u}_j^{n+1,L}$ , and such limiting methods are called monolithic [111, 131]. See [110, 188, 167] for more monolithic convex limiting techniques. Similar flux limiters were proposed in [93, 86] for the five-equation model.*

**5.5.2. Finite difference schemes with convex limiting of FCT type.** For a high order finite difference scheme (5.3), the convex limiting method in [98, 103, 108] provides a simpler sufficient solution for enforcing (5.4). The main idea is to rewrite the scheme (5.3) in the following form

$$\mathbf{u}_j^{n+1} = \frac{1}{2} \left[ \mathbf{u}_j^{n+1,L} - 2\lambda \theta_{j+\frac{1}{2}} \left( \hat{\mathbf{f}}_{j+\frac{1}{2}}^H - \hat{\mathbf{f}}_{j+\frac{1}{2}}^L \right) \right] + \frac{1}{2} \left[ \mathbf{u}_j^{n+1,L} + 2\lambda \theta_{j-\frac{1}{2}} \left( \hat{\mathbf{f}}_{j-\frac{1}{2}}^H - \hat{\mathbf{f}}_{j-\frac{1}{2}}^L \right) \right].$$

For a given concave function  $g(\mathbf{u}) > 0$ , the idea is to find  $\theta_{j+\frac{1}{2}}$  such that

$$(5.31) \quad g \left[ \mathbf{u}_j^{n+1,L} - 2\lambda\theta_{j+\frac{1}{2}} \left( \hat{\mathbf{f}}_{j+\frac{1}{2}}^H - \hat{\mathbf{f}}_{j+\frac{1}{2}}^L \right) \right] \geq \epsilon, \quad g \left[ \mathbf{u}_j^{n+1,L} + 2\lambda\theta_{j-\frac{1}{2}} \left( \hat{\mathbf{f}}_{j-\frac{1}{2}}^H - \hat{\mathbf{f}}_{j-\frac{1}{2}}^L \right) \right] \geq \epsilon.$$

Such  $\theta_{j+\frac{1}{2}}$  can be found in a way similar to (5.27), see [103, 108]. In such a convex limiting approach, one seeks to use  $\mathbf{u}_j^{n+1,L} \in G$  to correct  $\pm 2\lambda \left( \hat{\mathbf{f}}_{j\mp\frac{1}{2}}^H - \hat{\mathbf{f}}_{j\mp\frac{1}{2}}^L \right)$ , which is different from the previous approaches. As a comparison, the approaches in Subsection 5.4 and Subsection 5.5.1 use  $\mathbf{u}_j^{L,\pm} \in G$  to correct  $\mathbf{u}_j^{H,\pm}$ , and they satisfy

$$\mathbf{u}_j^{H,+} - \mathbf{u}_j^{L,+} = -\lambda(\hat{\mathbf{f}}_{j+\frac{1}{2}}^H - 2\hat{\mathbf{f}}_{j+\frac{1}{2}}^L), \quad \mathbf{u}_j^{H,-} - \mathbf{u}_j^{L,-} = 2\lambda(\hat{\mathbf{f}}_{j-\frac{1}{2}}^H - \hat{\mathbf{f}}_{j-\frac{1}{2}}^L).$$

**5.5.3. Convex limiting for continuous finite element methods.** Since  $\mathbf{u}_j^{n+1,L}$  is needed in (5.31), the convex limiting method (5.31) is different from a monolithic approach. On the other hand, such a method offers easiness for schemes on unstructured meshes in multiple dimensions. We briefly review the main idea in [98] for applying convex limiting to continuous finite element methods.

With the same notation as in Subsection 3.6, we consider the group finite element method with  $\mathbb{P}^1$  basis defined on unstructured triangular meshes. Recall the first order continuous finite element method with mass lumping (3.22) can be written in the flux form (3.25). For convenience, we denote (3.25) as

$$m_i \frac{\mathbf{u}_i^{n+1,L} - \mathbf{u}_i^n}{\Delta t} + \sum_{j \in \mathcal{N}_i} [(\mathbf{f}(\mathbf{u}_j^n) + \mathbf{f}(\mathbf{u}_i^n)) \cdot \mathbf{c}_{ij} - d_{ij}^L(\mathbf{u}_j^n - \mathbf{u}_i^n)] = 0,$$

where  $d_{ij}^L$  is the artificial viscosity coefficient in the first order IDP scheme.

A second order in space finite element scheme without mass lumping with forward Euler time discretization can be written as

$$(5.32) \quad \sum_{j \in \mathcal{N}_i} M_{ij} \frac{\mathbf{u}_j^{n+1,H} - \mathbf{u}_j^n}{\Delta t} + \sum_{j \in \mathcal{N}_i} [(\mathbf{f}(\mathbf{u}_j^n) + \mathbf{f}(\mathbf{u}_i^n)) \cdot \mathbf{c}_{ij} - d_{ij}^H(\mathbf{u}_j^n - \mathbf{u}_i^n)] = 0,$$

with smaller artificial viscosity coefficients  $d_{ij}^H$  satisfying the same symmetry and zero sum constraints (3.22b). We refer to [99, 103, 131] and references therein for how to compute  $d_{ij}^H$ . Let  $\delta_{ij}$  be the Kronecker delta, then  $\sum_{j \in \mathcal{N}_i} (M_{ij} - \delta_{ij}m_i) = 0$ , thus

$$\begin{aligned} \sum_{j \in \mathcal{N}_i} M_{ij} \frac{\mathbf{u}_j^{n+1,H} - \mathbf{u}_j^n}{\Delta t} &= \frac{m_i}{\Delta t} (\mathbf{u}_i^{n+1,H} - \mathbf{u}_i^n) + \sum_{j \in \mathcal{N}_i} \frac{M_{ij} - \delta_{ij}m_i}{\Delta t} (\mathbf{u}_j^{n+1,H} - \mathbf{u}_j^n) \\ &= \frac{m_i}{\Delta t} (\mathbf{u}_i^{n+1,H} - \mathbf{u}_i^n) + \sum_{j \in \mathcal{N}_i} \frac{M_{ij} - \delta_{ij}m_i}{\Delta t} (\mathbf{u}_j^{n+1,H} - \mathbf{u}_j^n - \mathbf{u}_i^{n+1,H} + \mathbf{u}_i^n). \end{aligned}$$

Together with properties of  $\mathbf{c}_{ij}$ , (5.32) can be rewritten as

$$\frac{m_i}{\Delta t} (\mathbf{u}_i^{n+1,H} - \mathbf{u}_i^n) + \sum_{j \in \mathcal{N}_i^*} \hat{\mathbf{f}}_{ij}^H = 0,$$

where  $j \in \mathcal{N}_i^*$  denotes  $j \in \mathcal{N}_i, j \neq i$ , and the numerical flux is

$$\hat{\mathbf{f}}_{ij}^H = \frac{M_{ij}}{\Delta t} (\mathbf{u}_j^{n+1,H} - \mathbf{u}_j^n - \mathbf{u}_i^{n+1,H} + \mathbf{u}_i^n) + (\mathbf{f}(\mathbf{u}_j^n) + \mathbf{f}(\mathbf{u}_i^n)) \cdot \mathbf{c}_{ij} - d_{ij}^H(\mathbf{u}_j^n - \mathbf{u}_i^n).$$

The first order IDP scheme can be written in a similar form:

$$\frac{m_i}{\Delta t}(\mathbf{u}_i^{n+1,L} - \mathbf{u}_i^n) + \sum_{j \in \mathcal{N}_i^*} \hat{\mathbf{f}}_{ij}^L = 0,$$

$$\hat{\mathbf{f}}_{ij}^L = (\mathbf{f}(\mathbf{u}_j^n) + \mathbf{f}(\mathbf{u}_i^n)) \cdot \mathbf{c}_{ij} - d_{ij}^L(\mathbf{u}_j^n - \mathbf{u}_i^n).$$

The flux corrected scheme can be written as

$$\mathbf{u}_i^{n+1} = \mathbf{u}_i^{n+1,L} - \frac{\Delta t}{m_i} \sum_{j \in \mathcal{N}_i^*} \theta_{ij} (\hat{\mathbf{f}}_{ij}^H - \hat{\mathbf{f}}_{ij}^L), \quad \theta_{ij} \in [0, 1].$$

As an easy approach to find  $\theta_{ij}$  to enforce convex invariant domains, the convex limiting method in [103, 98, 108] proposes to consider rewrite it as

$$\mathbf{u}_i^{n+1} = \sum_{j \in \mathcal{N}_i^*} a_j \left[ \mathbf{u}_i^{n+1,L} - \theta_{ij} \frac{\Delta t}{m_i a_j} (\hat{\mathbf{f}}_{ij}^H - \hat{\mathbf{f}}_{ij}^L) \right], \quad \theta_{ij} \in [0, 1],$$

where  $a_j > 0$  are any convex combination coefficients such that  $\sum_{j \in \mathcal{N}_i^*} a_j = 1$ . We refer to [108] for how to find  $\theta_{ij}$  ensuring

$$g \left[ \mathbf{u}_i^{n+1,L} - \theta_{ij} \frac{\Delta t}{m_i a_j} (\hat{\mathbf{f}}_{ij}^H - \hat{\mathbf{f}}_{ij}^L) \right] > 0,$$

which is sufficient for ensuring  $g(\mathbf{u}_i^{n+1}) > 0$  for a concave constraint function  $g$ .

**REMARK 5.4.** *The convex limiting becomes more diffusive with higher order polynomial basis, see [97, Figure 4]. Instead, a better way is to mix a high order finite element method with an invariant-domain preserving low order method on the closest neighbor stencil. Such continuous finite element methods with  $\mathbb{P}^2$  and  $\mathbb{P}^3$  bases on simplicial meshes were constructed in [97]. See also [157, 132, 111] for limiters on higher order finite element method.*

**5.5.4. Monolithic limiter via GQL representation.** A monolithic limiting method via GQL representation was presented in [11] for the Point-Average-Moment-Polynomial-interpreted (PAMPA) scheme [10]. For simplicity, we review the main idea of this method on a finite difference scheme, for which we write an alternative wave-speed-based convex decomposition of (5.3):

$$\begin{aligned} \mathbf{u}_j^{n+1} &= \mathbf{u}_j^n - \lambda \left( \hat{\mathbf{f}}_{j+\frac{1}{2}} - \hat{\mathbf{f}}_{j-\frac{1}{2}} \right) + \lambda \left( \alpha_{j-\frac{1}{2}} + \alpha_{j+\frac{1}{2}} \right) \mathbf{u}_j^n - \lambda \left( \alpha_{j-\frac{1}{2}} + \alpha_{j+\frac{1}{2}} \right) \mathbf{u}_j^n \\ &= (1 - \lambda \alpha_{j-\frac{1}{2}} - \lambda \alpha_{j+\frac{1}{2}}) \mathbf{u}_j^n + \lambda \alpha_{j-\frac{1}{2}} \mathbf{u}_j^- + \lambda \alpha_{j+\frac{1}{2}} \mathbf{u}_j^+, \end{aligned}$$

where  $\hat{\mathbf{f}}$  is any consistent IDP flux and

$$\begin{aligned} \mathbf{u}_j^- &= \mathbf{u}_j^n + \frac{\hat{\mathbf{f}}_{j-\frac{1}{2}}^L - \mathbf{f}(\mathbf{u}_j) + \theta_{j-\frac{1}{2}} \delta_{j-\frac{1}{2}}}{\alpha_{j-\frac{1}{2}}} =: \hat{\mathbf{u}}_j^{L,-} + \theta_{j-\frac{1}{2}} \frac{\delta_{j-\frac{1}{2}}}{\alpha_{j-\frac{1}{2}}}, \\ \mathbf{u}_j^+ &= \mathbf{u}_j^n - \frac{\hat{\mathbf{f}}_{j+\frac{1}{2}}^L - \mathbf{f}(\mathbf{u}_j) + \theta_{j+\frac{1}{2}} \delta_{j+\frac{1}{2}}}{\alpha_{j+\frac{1}{2}}} =: \hat{\mathbf{u}}_j^{L,+} - \theta_{j+\frac{1}{2}} \frac{\delta_{j+\frac{1}{2}}}{\alpha_{j+\frac{1}{2}}} \end{aligned}$$

with  $\delta_{j+\frac{1}{2}} := \hat{\mathbf{f}}_{j+\frac{1}{2}}^H - \hat{\mathbf{f}}_{j+\frac{1}{2}}^L$ . Similar to the previous discussion for (5.28), with sufficiently large  $\alpha_{j\pm\frac{1}{2}}$ , we have  $\hat{\mathbf{u}}_j^{L,\pm} \in G$ . Then there exist blending parameters  $\theta_{j\pm\frac{1}{2}} \in [0, 1]$  such that  $\mathbf{u}_j^\pm \in G$ . Consequently,  $\mathbf{u}_j^{n+1} \in G$  under the CFL condition  $\lambda(\alpha_{j-\frac{1}{2}} + \alpha_{j+\frac{1}{2}}) \leq 1$ .

The key feature of the method in [11] and its extension to polygonal meshes in [13] is a strategy to design effective blending parameters  $\theta_{j\pm\frac{1}{2}}$  for systems. We briefly review the idea for the 1D Euler equations. Based on the GQL framework [222], the invariant domain (2.2) for the 1D Euler system can be written equivalently as

$$G^* = \left\{ \mathbf{u} \in \mathbb{R}^3 : \mathbf{u} \cdot \mathbf{n}^* > 0 \quad \forall \mathbf{n}^* \in \mathcal{N} \right\}, \quad \mathcal{N} = \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \right\} \cup \left\{ \begin{pmatrix} v_*^2/2 \\ -v_* \\ 1 \end{pmatrix} : v_* \in \mathbb{R} \right\}.$$

To enforce  $\mathbf{u}_j^\pm \in G$  (equivalently  $G^*$ ), it suffices to require  $\mathbf{u}_j^\pm \cdot \mathbf{n}^* > 0$  for all  $\mathbf{n}^* \in \mathcal{N}$ , for which a convenient choice of blending parameter is

$$\theta_{j+\frac{1}{2}} = \min\{\theta_{j+\frac{1}{2}}^-, \theta_{j+\frac{1}{2}}^+\} \quad \text{with} \quad \theta_{j\pm\frac{1}{2}}^\mp = \min \left\{ 1, \alpha_{j\pm\frac{1}{2}} \min_{\mathbf{n}^* \in \mathcal{N}} \frac{\hat{\mathbf{u}}_j^{L,\pm} \cdot \mathbf{n}^*}{|\delta_{j\pm\frac{1}{2}} \cdot \mathbf{n}^*|} \right\}.$$

An explicit closed-form expression for the minimizer is given in [11].

**5.6. Numerical results.** We implement and test the fifth order finite difference WENO scheme solving compressible Euler equations with ideal gas EOS (2.1) with three flux correction methods for enforcing invariant domains:

1. The parametrized limiter in Subsection 5.3
2. The Hu–Adams–Shu simple flux limiting in Subsection 5.4
3. The monolithic convex limiting in Subsection 5.5.1.

EXAMPLE 5.1 (Leblanc shock tube). *This test is a 1D shock tube with  $\gamma = 1.4$ , the initial condition*

$$(\rho, v_1, p) = \begin{cases} (2, 0, 10^9), & \text{if } x < 0, \\ (0.001, 0, 10^{-12}), & \text{otherwise,} \end{cases}$$

*and outflow boundary conditions on the domain  $[-10, 10]$ . See Figure 12 for the plots of density and pressure.*

EXAMPLE 5.2 (Double rarefaction with low density and pressure). *This test is a 1D double rarefaction problem with  $\gamma = 1.4$  and the initial condition*

$$(\rho, v_1, p) = \begin{cases} (7, -100, 0.01), & \text{if } x < 0.5, \\ (7, 100, 0.01), & \text{otherwise.} \end{cases}$$

*The exact solution contains perfect vacuum for which high order schemes can easily produce negative density and pressure. The computational domain is  $[0, 1]$  divided into a quite coarse mesh of only 100 uniform cells. The outflow condition is applied on left and right boundaries. Figure 13 displays the density and velocity at  $t = 0.003$  obtained by the fifth order FD WENO scheme with IDP flux limiters on 100 grid points.*

EXAMPLE 5.3 (Shock vortex interaction). *This example simulates the interaction of shock and vortex, which involves very low density and low pressure, which was*

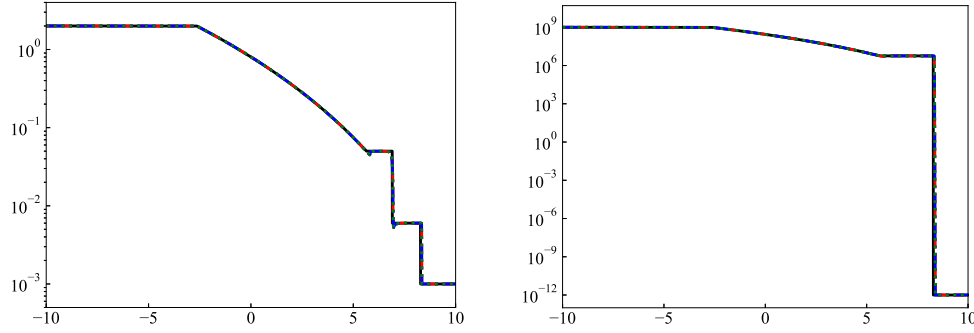


Fig. 12: **Example 5.1.** Leblanc shock tube: density (left) and pressure (right) at  $t = 0.001$  of the fifth order FD WENO scheme with IDP flux limiters on 4000 grid points. Red dashed line: the parametrized limiter; Blue dash-dot line: the Hu–Adams–Shu limiter; Green dotted line: the monolithic convex limiting.

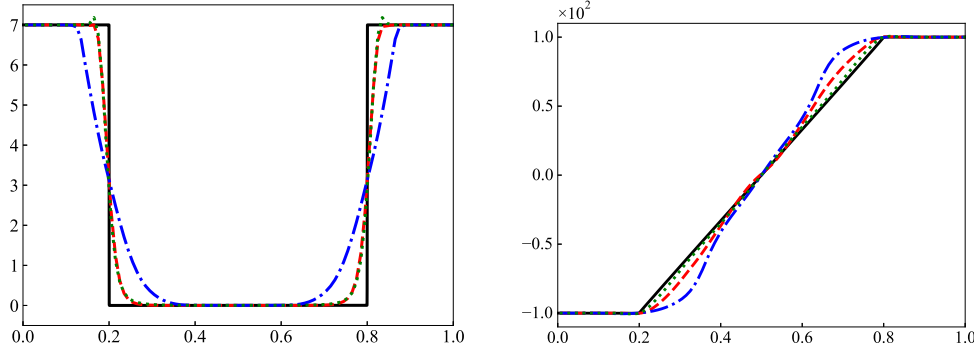


Fig. 13: **Example 5.2.** Double rarefaction: density (left) and velocity (right) at  $t = 0.003$  of the fifth order FD WENO scheme with IDP flux limiters on 100 grid points. Red dashed line: the parametrized limiter; Blue dash-dot line: the Hu–Adams–Shu limiter; Green dotted line: the monolithic convex limiting.

proposed in [62]. The computational domain is taken as  $[0, 2] \times [0, 1]$ , which is divided into  $450 \times 225$  uniform cells. Figure 14 shows the contours of the density and pressure obtained by the fifth order IDP finite difference WENO scheme using Hu–Adams–Shu limiter, convex limiting, and the parametrized limiter, respectively.

**EXAMPLE 5.4** (Relativistic axisymmetric jet). In this example, we simulate a very challenging astrophysical jet problem by solving the axisymmetric version of the relativistic hydrodynamic equations (6.1). The adiabatic index is taken as  $\gamma = 5/3$ . The computational domain is set as  $[0, 15] \times [0, 75]$ , which is divided into  $300 \times 1500$  uniform cells. Initially, the domain is full of the static uniform medium with

$$(\rho, u, v, p) = (1.0, 0.0, 0.0, 5.988006089640541 \times 10^{-11}).$$

A high-speed relativistic jet with state

$$(\rho_b, v_{1,b}, v_{2,b}, p_b) = (0.01, 0.0, 0.999, 5.988006089640541 \times 10^{-11})$$

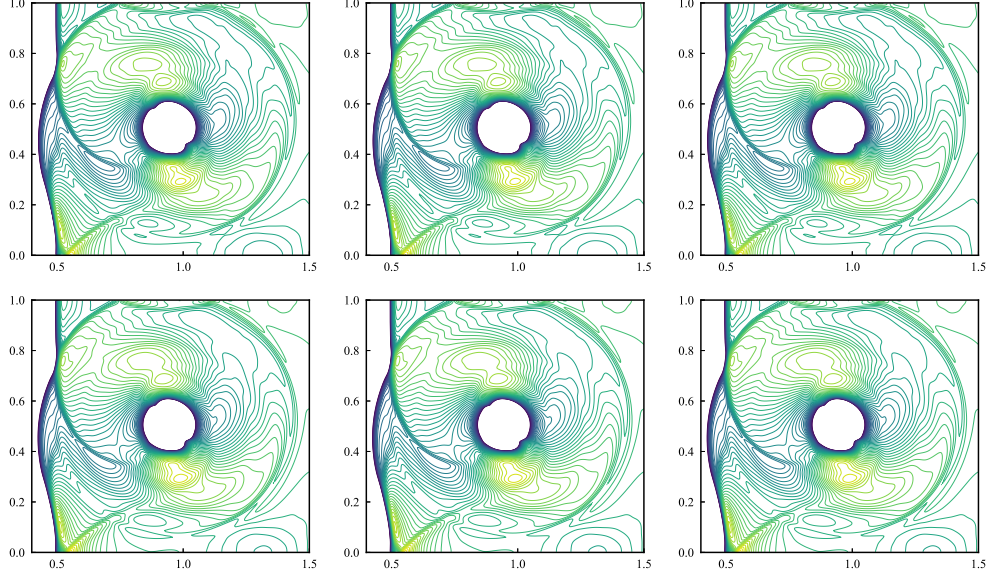


Fig. 14: [Example 5.3](#) The contour plots of density (top) and pressure (bottom) at  $t = 0.6$  of the fifth order FD WENO scheme with IDP flux limiters. 50 contour lines: from 1.03 to 1.39. Left: the parametrized limiter, middle: the Hu-Adams-Shu limiter, right: the monolithic convex limiting.

is injected in  $z$ -direction through nozzle ( $r \leq 1$ ) of the bottom boundary ( $z = 0$ ). In other words, the fixed inflow condition  $(\rho_b, v_{1,b}, v_{2,b}, p_b)$  is applied on  $\{r \leq 1, z = 0\}$  of the bottom boundary. The symmetrical condition is specified on the left boundary  $r = 0$ , outflow conditions are applied on other boundaries. For this jet, the classical Mach number is 10,000, and the relativistic Mach number is about 223,662.719, which is extremely high. [Figure 15](#) displays the contours of rest-mass density logarithm at  $t = 100$  obtained by the fifth order IDP finite difference WENO scheme using Hu-Adams-Shu limiter, convex limiting, and the parametrized limiter, respectively.

**6. Extensions and other approaches.** In this section, we review some extensions and generalizations of the approaches mentioned above, as well as other approaches to enforce invariant domains. In particular, for compressible MHD equations, the IDP method is more complicated due to the effect of the extra divergence free constraint of the magnetic field on the IDP property, which will be reviewed in [Subsection 6.8](#) with numerical examples shown in [Subsection 6.9](#).

**6.1. Other time discretizations.** The Zhang-Shu approach in [Section 4](#) can be used for any explicit SSP time discretizations. In order to use it in other time discretizations, one needs the weak monotonicity, which can be difficult to establish. The weak monotonicity of backward Euler time stepping for DG methods solving a linear advection was proven in [\[178\]](#). See [\[236\]](#) for scalar convection-diffusion equations. For DG methods, Lax-Wendroff time stepping was also considered in [\[166\]](#). For the flux limiters in [Section 5](#), it can be applied to the last time stage of any explicit Runge-Kutta method.

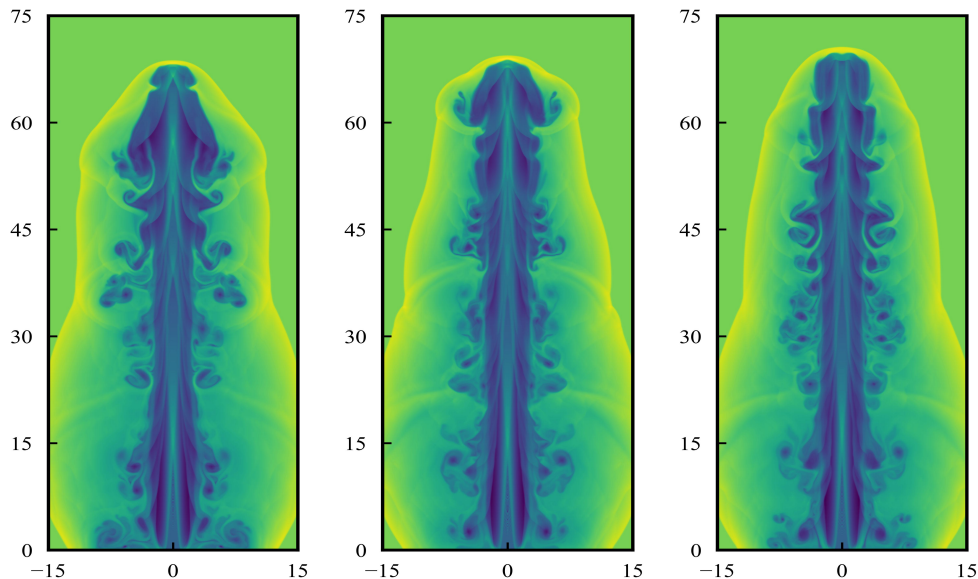


Fig. 15: (Example 5.4) The contour plots of rest mass density logarithm at  $t = 100$  of the fifth order FD WENO scheme with three IDP flux limiters. Left: the parametrized limiter, middle: the Hu–Adams–Shu limiter, right: the monolithic convex limiting.

On the other hand, flux limiters are more flexible to use and in general they can be applied to any explicit Runge–Kutta method, e.g., [228, 140]. In [78], every explicit Runge–Kutta method can be made IDP by limiting each stage by a forward Euler time step with IDP low order spatial discretization, e.g., the first order schemes in Section 3. Such a method can also be extended to IMEX (implicit-explicit) schemes [79], which can be used for convection diffusion equations if the first IMEX is IDP. See also [181] for the Diagonally Implicit Runge–Kutta method. In general, it can be quite difficult to establish an IDP result in an implicit scheme. For compressible Navier–Stokes equations, a few implicit and semi-implicit IDP schemes have been constructed in the literature, see e.g., the second order schemes [92, 95], and [152] with fourth order spatial accuracy.

**6.2. Lagrangian schemes.** All the schemes mentioned above are for Eulerian schemes defined on a given fixed mesh. The techniques and methods in Section 4 and Section 5 can be extended to Lagrangian type schemes including the semi-Lagrangian and arbitrary Lagrangian–Eulerian schemes, e.g., [180, 187, 231, 49, 211, 37, 106, 38, 107, 104, 122, 105].

**6.3. Subcell limiting methods.** For a finite volume scheme and DG scheme, a subcell limiting is to correct a bad cell solution violating given criteria such as invariant domain, by a first order IDP scheme on subcells of the bad cell. With enough number of subcells in one bad cell, such an IDP method gives a high order accurate correction. The subcell limiter in general can be used to preserve more properties such as entropy consideration. See [75, 74, 164, 208, 173, 110, 109, 210, 209] for subcell limiters for DG and FV methods.

**6.4. IDP reformation for point values in PAMPA scheme.** The PAMPA method [6] combines conservative and non-conservative formulations of hyperbolic conservation laws to evolve both cell averages and point values. By taking advantage of this flexibility, in [10] an automatic IDP reformulation of the non-conservative part was proposed, inspired by Softplus and clipped ReLU mappings from machine learning. This yields an unconditionally limiter-free IDP scheme for the point values. In this approach, the Zhang–Shu IDP limiter is still required but solely for the cell midpoint value so that the resulting PAMPA scheme is provably IDP for the updated cell averages.

**6.5. Weak monotonicity for convection-diffusion equations.** For preserving bounds or positivity in a scalar equation, the flux limiters in Section 5 can be easily extended from convection to convection-diffusion equations. For the Zhang–Shu method Section 4, extensions to convection-diffusion equations would require a weak monotonicity result for diffusion equations, which is in general not true for arbitrarily high order DG and FV schemes. For special high order schemes, weak monotonicity can still be proven for diffusion equations, including third order direct DG methods [48, 153], some high order compact finite difference schemes [145], and a nonstandard finite volume scheme using double cell averages [243], all of which are linear schemes, i.e., the scheme is linear when the equation is linear. Arbitrarily high order nonlinear DG schemes [200, 198] can be constructed to be weakly monotone so that the bound-preserving method in Section 4 still applies. All these schemes are explicit in time and can be applied to preserve bounds in nonlinear parabolic equations.

Such weak monotonicity for nonlinear parabolic equations is quite different from the discrete maximum principle finite element methods for linear elliptic and parabolic equations. For solving a Poisson equation  $-\Delta u = f$ , let  $-\Delta_h$  denote the discrete Laplacian. If  $(-\Delta_h)^{-1}$  is a matrix with positive entries, then we say a scheme is *monotone* which implies discrete maximum principle. Although high order finite element method is known to violate maximum principle on unstructured meshes, continuous finite element method with quadratic and cubic polynomial bases on uniform meshes can still be monotone for Laplacian, see [159, 146, 59] and references therein. We refer to [23, 21] for a recent comprehensive review for finite element methods satisfying discrete maximum principle for linear convection diffusion equations.

**6.6. Optimization based approaches for enforcing bounds.** In the literature, bound-preserving limiters and methods can be implemented or formulated as a constrained minimization problem, e.g., optimization based limiters for each cell [185, 26]. See also [32] and references therein. There is a natural connection between FCT methods and optimization based method for enforcing bounds and constraints [148]. In [32, Section 4.3], it was proven that Zalesak’s original formula (5.7) is the minimizer to a global optimization with a modified cost function under box constraints.

There are advantages of optimization based approaches such as easy treatment for implicit time stepping [207], flexibility for spatial discretizations [237], and easy generalizations to higher order PDEs [151]. For preserving bounds of a scalar variable, this has been well studied, e.g., the remap problem in arbitrary Lagrangian–Eulerian schemes [31, 30]. In particular, an efficient algorithm was used in [33] to find the minimizer in  $\ell^2$ -norm and a direct and cheap construction of one particular minimizer to the  $\ell^1$ -norm was given in [39]. Optimization based postprocessing was also considered by quadratic programming [94, 29, 235, 177] as well as first order splitting methods [151]. Gradient descent was used in [238]. In [189], Newton’s method was used to

solve a global optimization problem to find the optimal flux correction in the FCT method for enforcing bounds. A bound-preserving limiter for total energy was used to enforce positivity of pressure in [149]. See also [126, 121] for enforcing bounds in finite element methods via optimization or variational inequalities for scalar convection diffusion problems. On the other hand, for a general system, it is usually difficult to have an efficient optimization based approach with all desired properties enforced. See [150] for an optimization based limiter for enforcing the invariant domain set in gas dynamics and global conservation.

**6.7. Extensions to relativistic hydrodynamics.** Due to the strong nonlinearity and the effects of curved spacetime in general relativity, the design of IDP schemes encounters several unique challenges. The governing equations of special relativistic hydrodynamics, also known as the relativistic Euler equations, can be written in the form of (1.3):

$$(6.1) \quad \partial_t \begin{pmatrix} D \\ \mathbf{m} \\ E \end{pmatrix} + \nabla \cdot \begin{pmatrix} D\mathbf{v} \\ \mathbf{m} \otimes \mathbf{v} + p\mathbf{I} \\ \mathbf{m} \end{pmatrix} = \mathbf{0}, \quad \mathbf{x} \in \mathbb{R}^d,$$

where  $d = 1, 2, 3$  denotes the spatial dimension. The relativistic mass density is  $D = \rho W$ , with  $\rho$  being the rest-mass density and  $W = (1 - |\mathbf{v}|^2)^{-1/2}$  the Lorentz factor. The momentum vector is  $\mathbf{m} = \rho h W^2 \mathbf{v}$ , where  $h$  is the specific enthalpy, and the velocity is normalized such that the speed of light is unity. The total energy is given by  $E = \rho h W^2 - p$ .

An invariant domain of (6.1) is

$$(6.2) \quad G = \{ \mathbf{u} = (D, \mathbf{m}, E)^\top : D > 0, p(\mathbf{u}) > 0, |\mathbf{v}(\mathbf{u})| < 1 \}.$$

Here, both  $\mathbf{v}(\mathbf{u})$  and  $p(\mathbf{u})$  are nonlinear functions of the conserved variables and lack closed-form expressions, making the design and analysis of IDP schemes nontrivial. To address this difficulty, an equivalent characterization of the set (6.2) was proposed in [223]:

$$(6.3) \quad G = \{ \mathbf{u} = (D, \mathbf{m}, E)^\top : D > 0, q(\mathbf{u}) > 0 \},$$

where  $q(\mathbf{u}) := E - \sqrt{D^2 + |\mathbf{m}|^2}$  is concave in  $\mathbf{u}$ , implying the convexity of  $G$ . Based on this finding, Assumption 2 was rigorously proven in [223] for the relativistic Euler equations (6.1), and high-order finite difference IDP schemes were developed. Other extensions include high-order IDP (central) DG and finite volume schemes [179, 224, 47]. See Example 5.4 for IDP simulations of a challenging relativistic jet.

Similar to the non-relativistic case discussed in Example 2.2, it was shown in [216, 63] that the relativistic system (6.1) also satisfies the minimum entropy principle. Incorporating this leads to another invariant domain (see [63] for the general EOS case):

$$(6.4) \quad G_S = \left\{ \mathbf{u} = (D, \mathbf{m}, E)^\top : D > 0, q(\mathbf{u}) > 0, S = \ln \left( \frac{p}{\rho^\gamma} \right) \geq S_{\min} \right\},$$

where  $S(\mathbf{u})$  is not concave, but it was shown in [216] that  $D(S - S_{\min})$  is concave in  $\mathbf{u}$ , hence  $G_S$  remains convex. However, Assumption 2 (or Assumption 5 in 2D) does not hold for  $G_S$ , and verifying Assumption 1 (or Assumption 4) is highly nontrivial

due to the implicit form of  $S(\mathbf{u})$ . To overcome this difficulty, the GQL approach can be used to derive an equivalent linear representation of  $G_S$ :

$$(6.5) \quad G_S^* = \{\mathbf{u} = (D, \mathbf{m}, E)^\top : D > 0, \mathbf{u} \cdot \mathbf{n}^* + S_{\min} \rho_*^\gamma > 0 \quad \forall \rho_* > 0, \forall \mathbf{v}^* \in \mathbb{B}_1(\mathbf{0})\},$$

where  $\mathbf{n}^* := \left(-\sqrt{1 - |\mathbf{v}^*|^2} \left(1 + \frac{S_{\min} \gamma \rho_*^{\gamma-1}}{\gamma-1}\right), -\mathbf{v}^*, 1\right)^\top$  is the inward normal vector to  $\partial G_S$ , and  $\mathbb{B}_1(\mathbf{0})$  denotes the unit ball in  $\mathbb{R}^d$ . Thanks to the GQL technique, [Assumption 3](#) ([Assumption 6](#) in 2D) was rigorously proven in [216, Theorem 3.9], implying (3.8) and its multidimensional counterparts without using any assumptions on the exact Riemann solutions. Consequently, high-order numerical schemes preserving the invariant domain  $G_S$  in (6.4) were successfully developed in [216].

In general relativity, new challenges arise because the invariant domain depends on the spacetime metric and therefore varies across space, to which standard convex combination techniques do not directly apply. To address this issue, a new formulation (W-form) of the general relativistic hydrodynamic equations was introduced in [214], enabling the generalization of high-order IDP schemes to curved spacetimes [214, 44].

**6.8. Extensions to compressible MHD.** The MHD equations presents a non-linear coupling of fluid dynamics with Maxwell's equations; see (2.16) for the classical MHD system and (2.18) for the relativistic version. These systems involve two key structures: a divergence-free constraint on the magnetic field, and IDP constraints on fluid variables (positivity of density/pressure, subluminal velocity). While divergence-free schemes are well-established, constructing provably IDP methods—especially in multiple dimensions—has been an open problem. In the non-relativistic setting, IDP schemes are also called positivity-preserving schemes. Although many techniques and limiters (e.g., [51, 54, 52]) were adapted to enforce positivity in ideal MHD, few of the resulting schemes were rigorously and completely proven to be IDP in theory, even for the first-order schemes, especially in multiple dimensions [54].

The development of IDP schemes for MHD systems encounters unique challenges that do not typically arise in other hyperbolic systems. A fundamental difficulty lies in the lack of understanding of the intrinsic connection—if any—between the IDP property and the divergence-free constraint on the magnetic field. Do such connections exist? If so, how are they expressed mathematically, and how can they be established? These central questions remained open, until the work in [225, 215]. Identifying this connection is crucial, as it may provide the foundation for designing provably IDP methods for MHD. Notably, the IDP constraints are pointwise *algebraic* structure, whereas the divergence-free constraint is a *differential* structure in nature, making it inherently difficult to bridge the two.

This connection was first theoretically revealed in [225, 215] for cartesian meshes, in [219] for general unstructured meshes, and in [217] for central DG schemes on overlapping meshes. It was shown that a discrete divergence-free condition is essential for ensuring the IDP property. Even minor violations of this condition may lead to the loss of the IDP property. Moreover, because of the discrete divergence-free constraint, states at different quadrature points become strongly coupled, rendering classical convex decomposition techniques (decomposing multidimensional schemes into a convex combination of formally first-order schemes), such as (4.31), ineffective. Additionally, none of Assumptions 1–6 hold for multidimensional MHD when there is a discontinuity in the normal component of the magnetic field.

This difficulty was ultimately overcome using the GQL approach, yielding the

following representation of  $G$ :

$$(6.6) \quad G^* = \{ \mathbf{u} \in \mathbb{R}^{2d+2} : \rho > 0, (\mathbf{u} - \mathbf{u}^*) \cdot \mathbf{n}^* > 0 \quad \forall \mathbf{u}^* \},$$

where  $\mathbf{u}^* \in \partial G$  corresponds to an arbitrary state with thermal pressure  $p^* = 0$ , and  $\mathbf{n}^*$  is an inward normal vector to  $\partial G$  at  $\mathbf{u}^*$ ; see [215] for ideal MHD and [225] for relativistic MHD. GQL enables precise mathematical formulation of the relationship between the IDP property and the divergence-free condition [225, 215, 219, 217]. In particular, it was shown that a modified version of [Assumption 3](#) holds:

**THEOREM 6.1** (Validity of a modified [Assumption 3](#) in relativistic MHD). *If setting  $\zeta(\mathbf{u}^*) = -v_\ell^* p_m^*$ , where  $v_\ell^*$  and  $p_m^*$  are respectively the  $\ell$ th component of velocity vector  $\mathbf{v}^*$  and the magnetic pressure at the state  $\mathbf{u}^*$ , then we have*

$$(6.7) \quad \mathbf{u} \in G \implies \alpha(\mathbf{u} - \mathbf{u}^*) \cdot \mathbf{n}^* \pm \mathbf{f}_\ell(\mathbf{u}) \cdot \mathbf{n}^* > \pm \zeta(\mathbf{u}^*) \pm B_\ell(\mathbf{v}^* \cdot \mathbf{B}^*) \quad \forall \mathbf{u}^*, \forall \alpha \geq 1,$$

where  $\mathbf{f}_\ell$  is the  $\ell$ th component of the flux  $\mathbf{f} = (\mathbf{f}_1, \dots, \mathbf{f}_d)$ ,  $\ell = 1, \dots, d$ .

The additional term  $\pm B_\ell(\mathbf{v}^* \cdot \mathbf{B}^*)$  in (6.7) is essential; without it, the inequality reduces to the original [Assumption 3](#), which does not hold in general in the MHD case. This term captures the fundamental influence of the divergence-free condition on the IDP property.

**EXAMPLE 6.1.** *To illustrate the basic idea, consider the 1D case, where the divergence free condition simplifies to  $B_1 \equiv \text{const}$ . Assume  $\mathbf{u}_L, \mathbf{u}_R \in G$ , then it follows from (6.7) that*

$$\begin{aligned} \alpha(\mathbf{u}_L - \mathbf{u}^*) \cdot \mathbf{n}^* + \mathbf{f}_1(\mathbf{u}_L) \cdot \mathbf{n}^* &> \zeta(\mathbf{u}^*) + B_{1,L}(\mathbf{v}^* \cdot \mathbf{B}^*), \\ \alpha(\mathbf{u}_R - \mathbf{u}^*) \cdot \mathbf{n}^* - \mathbf{f}_1(\mathbf{u}_L) \cdot \mathbf{n}^* &> -\zeta(\mathbf{u}^*) - B_{1,R}(\mathbf{v}^* \cdot \mathbf{B}^*) \end{aligned}$$

If we further assume  $B_{1,L} = B_{1,R}$ , averaging the two inequalities yields

$$\alpha(\bar{\mathbf{u}} - \mathbf{u}^*) \cdot \mathbf{n}^* > \frac{1}{2}(B_{1,L} - B_{1,R})(\mathbf{v}^* \cdot \mathbf{B}^*) = 0,$$

where  $\bar{\mathbf{u}} = \frac{\mathbf{u}_L + \mathbf{u}_R}{2} + \frac{\mathbf{f}_1(\mathbf{u}_L) - \mathbf{f}_1(\mathbf{u}_R)}{2\alpha}$ . By the GQL representation (6.6), we conclude

$$(6.8) \quad \mathbf{u}_L, \mathbf{u}_R \in G, B_{1,L} = B_{1,R} \implies \frac{\mathbf{u}_L + \mathbf{u}_R}{2} + \frac{\mathbf{f}_1(\mathbf{u}_L) - \mathbf{f}_1(\mathbf{u}_R)}{2\alpha} \in G \quad \forall \alpha \geq 1.$$

This shows that (3.8) holds for relativistic MHD, but only under the additional discrete divergence-free constraint  $B_{1,L} = B_{1,R}$ . See [215] for the non-relativistic counterpart. While this discrete divergence-free condition is trivial in 1D, it becomes significantly more complex in multiple dimensions [219, 221]. For instance, a 2D version of (6.8) for first-order IDP schemes is given as follows:

**THEOREM 6.2.** *If  $\mathbf{u}_L, \mathbf{u}_R, \mathbf{u}_D, \mathbf{u}_U \in G$  and the 2D discrete divergence-free constraint  $\frac{B_{1,R} - B_{1,L}}{\Delta x} + \frac{B_{2,U} - B_{2,D}}{\Delta y} = 0$  holds, then*

$$\begin{aligned} \frac{1}{\frac{\alpha_x}{\Delta x} + \frac{\alpha_y}{\Delta y}} \left[ \frac{\alpha_x}{\Delta x} \left( \frac{\mathbf{u}_L + \mathbf{u}_R}{2} + \frac{\mathbf{f}_1(\mathbf{u}_L) - \mathbf{f}_1(\mathbf{u}_R)}{2\alpha_x} \right) \right. \\ \left. + \frac{\alpha_y}{\Delta y} \left( \frac{\mathbf{u}_D + \mathbf{u}_U}{2} + \frac{\mathbf{f}_2(\mathbf{u}_D) - \mathbf{f}_2(\mathbf{u}_U)}{2\alpha_y} \right) \right] \in G \quad \forall \alpha_x, \alpha_y \geq 1. \end{aligned}$$

It is worth noting that the GQL auxiliary variables in the additional term  $\pm B_\ell(\mathbf{v}^* \cdot \mathbf{B}^*)$  play a critical role in linking the IDP and discrete divergence-free properties.

We refer interested readers to [219, 221] for the multidimensional and higher-order versions of Theorems 6.1 and 6.2 and the corresponding discrete divergence-free constraints on general meshes. Notably, in multiple dimensions, these discrete divergence-free conditions are strongly coupled, involving information from neighboring cells. While global divergence-free discretizations can ensure these constraints, the use of local scaling IDP limiters, such as the Zhang–Shu limiter, typically destroys the global divergence-free property of the magnetic field. Consequently, it is nontrivial to construct high-order *strictly conservative* schemes that are both IDP and globally divergence-free.

Interestingly, at the continuous level, Wu and Shu discovered that the IDP property of the exact solution is also tightly linked to the divergence-free condition. In [218, 221], it was shown that if the magnetic field is not divergence-free, even the exact solution of the MHD system may violate pressure positivity. This implies that the set (2.17), and its relativistic counterpart (2.19), is no longer an invariant domain of the PDE. Therefore, when the numerical magnetic field fails to satisfy the divergence-free condition, even an exact PDE solver (assuming it were available) cannot guarantee IDP, highlighting the inherent complexity of constructing genuinely IDP schemes for multidimensional MHD.

Moreover, in [219], Wu and Shu observed that the symmetrizable modified formulation of the ideal MHD system—originally introduced by Godunov [87]—always admits the set (2.17) as an invariant domain, regardless of whether the magnetic field is divergence-free. For the relativistic system, the symmetrizable modified formulation enjoyed the same feature was recently shown in [220]. Motivated by this observation, Wu and Shu [218, 219, 221] proved that the IDP property of numerical schemes based on these symmetrizable modified formulations only depends on a (discrete) locally divergence-free constraint. Crucially, this local constraint is compatible with local scaling IDP limiters, such as the Zhang–Shu limiter. Based on these findings, a series of structure-preserving frameworks, provably IDP and locally divergence-free, have been systematically developed in [218, 219, 221, 217, 69, 70, 154, 71, 43] for compressible MHD systems using the symmetrizable modified formulation.

**REMARK 6.1.** *A localized element-based positivity-preserving FCT approach with divergence cleaning was proposed for continuous finite element discretization of the MHD system in [134]. A second-order structure-preserving finite element method for ideal MHD was recently proposed in [64]. This method combines convex limiting with a novel operator splitting technique. A constrained transport method provably preserving positivity and divergence-free constraint was further introduced for MHD in [172].*

## 6.9. Numerical results for ideal MHD and relativistic MHD equations.

**EXAMPLE 6.2** (Shock cloud interaction). *In this example, we solve the ideal MHD equations to simulate the interaction of a strong shock wave and a high density cloud with the adiabatic index  $\gamma = 5/3$ . The computational domain is chosen to be  $[0, 1]^2$ , as in [218, 219, 217]. The problem is simulated up to  $t = 0.06$  with  $400 \times 400$  uniform cells. Figure 16 presents the contours of the density, the thermal pressure, and the magnitude of magnetic field obtained by the third order IDP DG method and fifth order IDP finite volume method. It is worth noting that the schemes would produce nonphysical solutions without using the IDP limiter.*

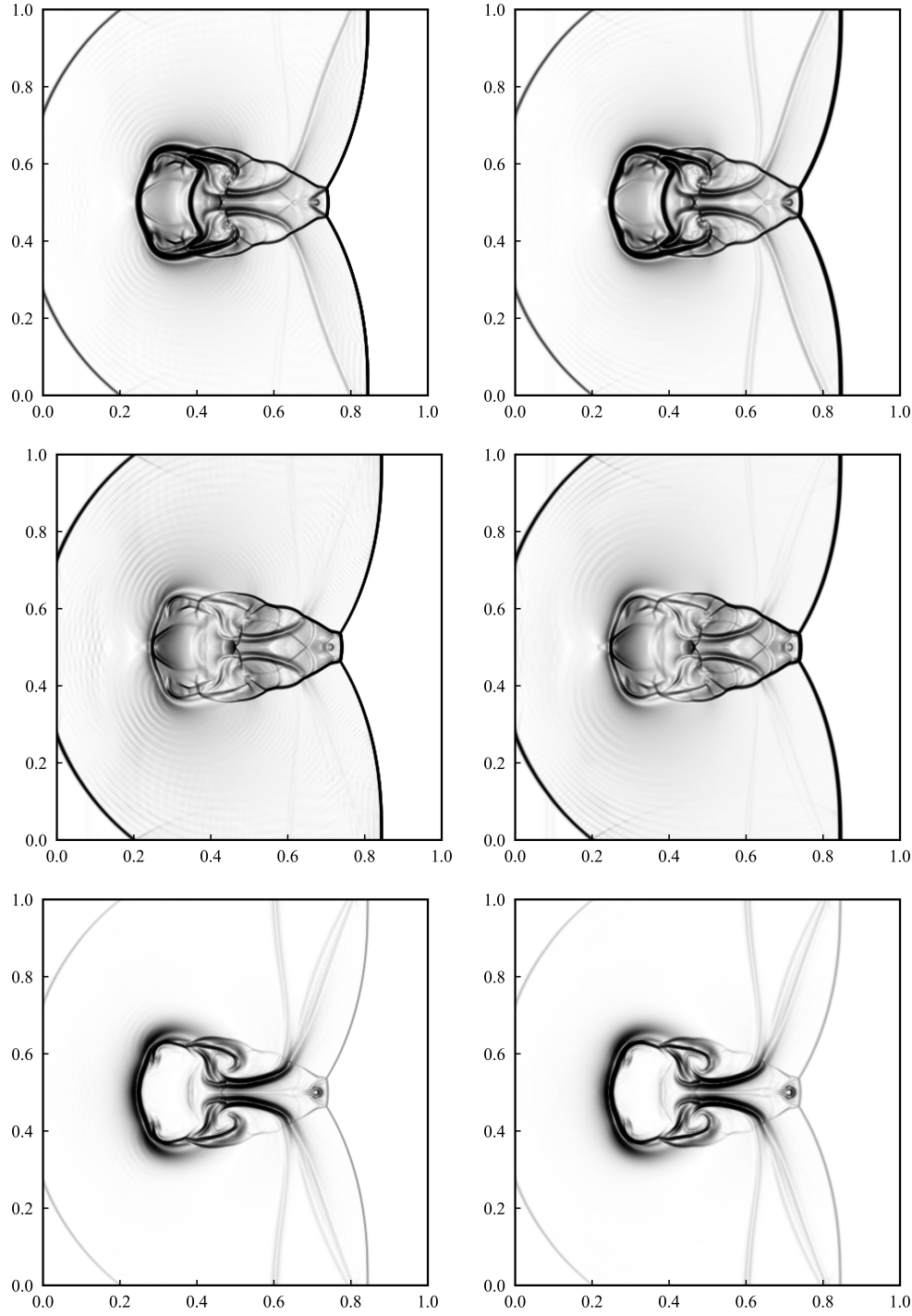


Fig. 16: (Example 6.2) The density logarithm (top), thermal pressure (middle), and magnitude of magnetic field (bottom). Left: third order IDP DG scheme. Right: fifth order IDP finite volume scheme.

EXAMPLE 6.3 (Astrophysical MHD jet with huge Mach number). *The high-speed MHD jets are proposed and first simulated in [218, 219, 221]. This test simulates an extremely high Mach number jet problem in a strong magnetic field. The adiabatic index is set to be  $\gamma = 1.4$ . Initially, the domain  $[-0.5, 0.5] \times [0, 1.5]$  is full of the ambient plasma with  $(\rho, \mathbf{v}, \mathbf{B}, p) = (0.1\gamma, 0, 0, 0, 2 \times 10^5, 0, 1)$ . A high-speed jet initially locates at  $x \in [-0.05, 0.05]$  and  $y = 0$ , it is injected in  $y$ -direction of the bottom boundary with the inflow jet condition  $(\rho, \mathbf{v}, \mathbf{B}, p) = (\gamma, 0, 10^6, 0, 0, 2 \times 10^5, 0, 1)$ . In our test, the computational domain is taken as  $[0, 0.5] \times [0, 1.5]$  and divided into  $200 \times 600$  cells. For the left boundary  $x = 0$ , the reflecting boundary condition is imposed. The outflow conditions are applied on other boundaries. The final time is  $t = 1.8 \times 10^{-6}$ . Figure 17 shows the schlieren images of density logarithm  $\log \rho$  obtained by the third order IDP DG method and fifth order IDP finite volume method.*

EXAMPLE 6.4 (Orszag–Tang problem). *We simulate the 2D Orszag–Tang problem of the relativistic MHD equations, following the setup in [206, 221]. In this problem, the initial maximum velocity reaches 0.99, close to the speed of light. Figure 18 presents the numerical results, obtained by the third order IDP DG method and fifth order IDP finite volume method with  $600 \times 600$  uniform cells in the domain  $[0, 2\pi]^2$ , for the logarithm of the rest-mass density  $\log \rho$  at  $t = 2.818127$  and  $6.8558$ . As time progresses, complex wave structures are generated and correctly captured by our method. The results agree with those reported in [221, 206].*

EXAMPLE 6.5 (Relativistic MHD blast problem). *Relativistic MHD blast wave problems [225, 221] are widely used to test the robustness of numerical schemes, as nonphysical solutions can easily be produced in numerical simulations. In this paper, we follow the setup in [221] and consider the blast problem with a huge strength of magnetic field 2000 in the  $x$ -direction. Figure 19 displays the rest-mass density logarithm, and magnitude of the magnetic field at  $t = 4$  obtained by the third order IDP DG method and fifth order IDP finite volume method with  $400 \times 400$  uniform cells in the domain  $\Omega = [-6, 6]^2$ .*

**7. Concluding remarks.** We have presented a comprehensive survey of numerical schemes which are invariant-domain-preserving (IDP) for hyperbolic systems and related equations. We have unified existing techniques and theories for establishing IDP properties in first-order accurate schemes, and given a systematical review of two popular approaches for constructing high-order IDP schemes, along with recent developments in the field. The Zhang–Shu approach leverages the intrinsic weak IDP property of high-order schemes, enabling the design of polynomial limiters that enforce a strong IDP property at point values for high-order finite volume and discontinuous Galerkin methods. The flux limiting approaches apply to a broader range of spatial discretizations, including finite difference and continuous finite element methods. In addition, we also discussed recent breakthroughs in constructing IDP schemes for more challenging systems, such as the magnetohydrodynamics equations, where maintaining the invariant domain is more delicate due to the complication from discrete divergence-free constraints. Throughout the paper, we have provided new perspectives and insights about existing IDP approaches. Extensive examples, including positivity-preserving schemes for gas dynamics and numerical experiments from gas dynamics and magnetohydrodynamics, were presented to illustrate the practical performance and importance of IDP schemes. In general, preserving the invariant domain remains a cornerstone for developing reliable and physically meaningful numerical methods for hyperbolic systems and related equations. The approaches surveyed in this paper

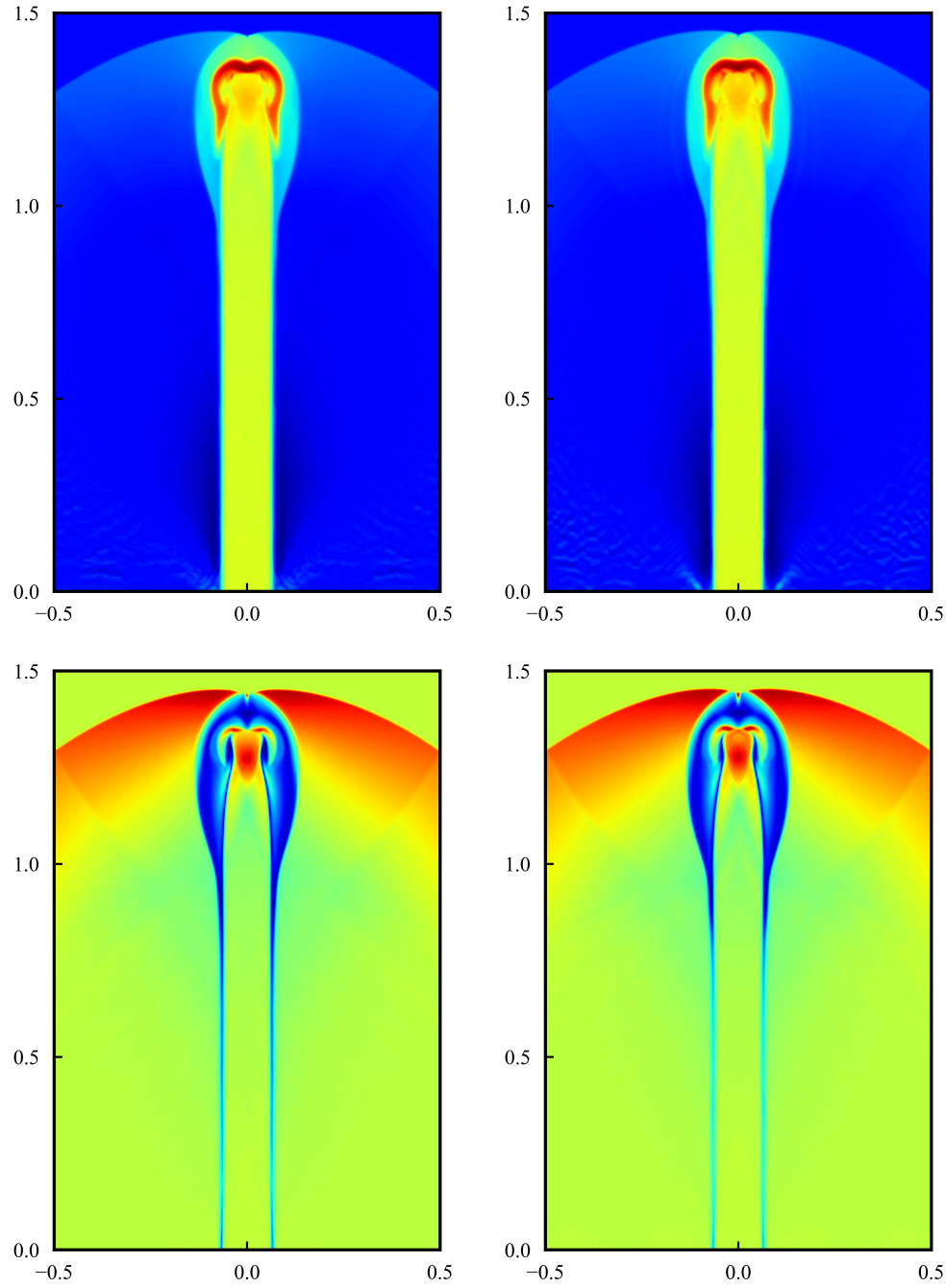


Fig. 17: (Example 6.3) Density logarithm (top) and the magnitude of magnetic field (bottom) at  $1.8 \times 10^{-6}$ . Left: third order IDP DG method, right: fifth order IDP finite volume method.

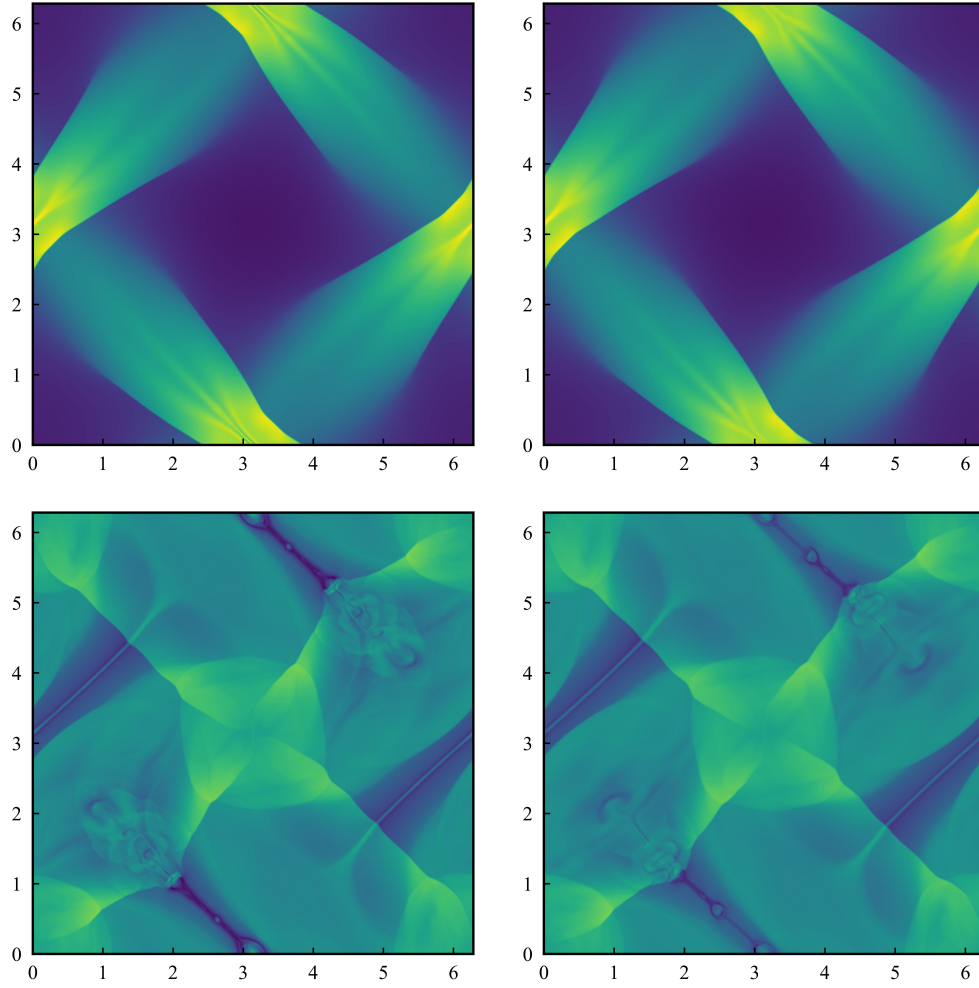


Fig. 18: (Example 6.4) Density logarithm  $\log \rho$  at  $t = 2.818127$  (top) and  $6.8558$  (bottom). Left: third order IDP DG method. Right: fifth order IDP finite volume method.

are useful for future research, particularly in the design of high-order, robust, and efficient solvers for more complex applications.

**Acknowledgement.** Xiangxiong Zhang is grateful to Professor Remi Abgrall for discussions on residual distribution schemes. The authors would like to thank Dr. Shengrong Ding and Dr. Chen Liu for the help on the visualization of some numerical results.

#### REFERENCES

- [1] R. ABGRALL, *Toward the ultimate conservative scheme: following the quest*, J. Comput. Phys., 167 (2001), pp. 277–315.

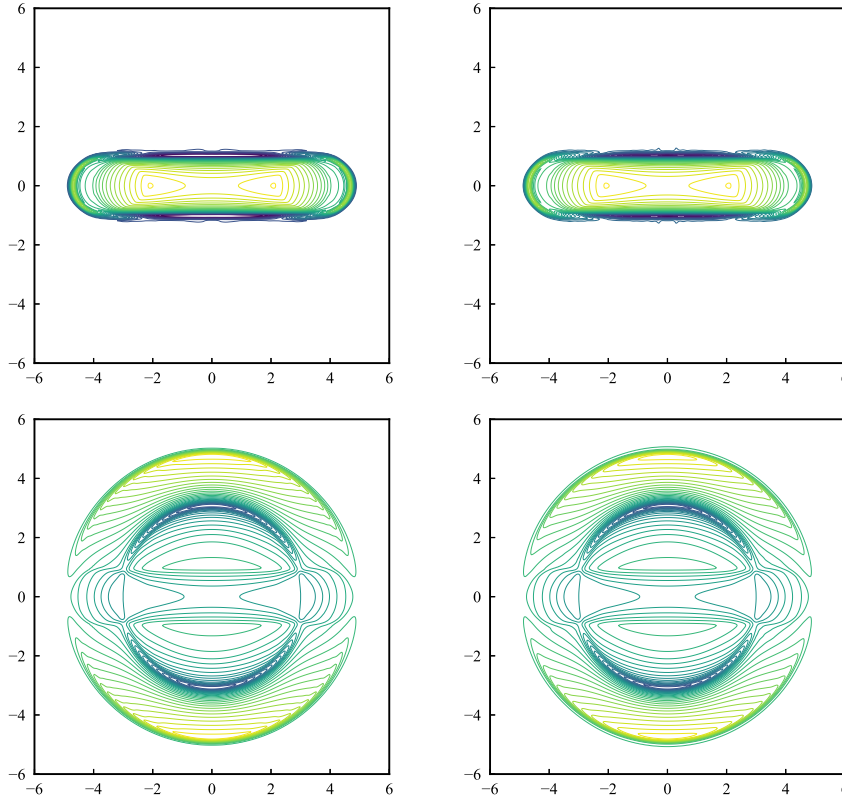


Fig. 19: (Example 6.5) Contour plots of the rest-mass density logarithm (top), thermal pressure (middle), and magnitude of magnetic field (bottom). Left: third order IDP DG method. Right: fifth order IDP finite volume method.

- [2] R. ABGRALL, *Essentially non-oscillatory residual distribution schemes for hyperbolic problems*, J. Comput. Phys., 214 (2006), pp. 773–808.
- [3] R. ABGRALL, *High order schemes for hyperbolic problems using globally continuous approximation and avoiding mass matrices*, J. Sci. Comput., 73 (2017), pp. 461–494.
- [4] R. ABGRALL, *A general framework to construct schemes satisfying additional conservation relations. Application to entropy conservative and entropy dissipative schemes*, J. Comput. Phys., 372 (2018), pp. 640–666.
- [5] R. ABGRALL, *Some remarks about conservation for residual distribution schemes*, Comput. Methods Appl. Math., 18 (2018), pp. 327–351.
- [6] R. ABGRALL, *A combination of residual distribution and the active flux formulations or a new class of schemes that can combine several writings of the same hyperbolic problem: application to the 1D Euler equations*, Commun. Appl. Math. Comput., 5 (2023), pp. 370–402.
- [7] R. ABGRALL AND T. BARTH, *Residual distribution schemes for conservation laws via adaptive quadrature*, SIAM J. Sci. Comput., 24 (2003), pp. 732–769.
- [8] R. ABGRALL AND D. DE SANTIS, *Linear and non-linear high order accurate residual distribution schemes for the discretization of the steady compressible Navier-Stokes equations*, J. Comput. Phys., 283 (2015), pp. 329–359.
- [9] R. ABGRALL, D. DE SANTIS, AND M. RICCHIUTO, *High-order preserving residual distribution schemes for advection-diffusion scalar problems on arbitrary grids*, SIAM J. Sci. Comput., 36 (2014), pp. a955–a983.
- [10] R. ABGRALL, M. JIAO, Y. LIU, AND K. WU, *A novel and simple invariant-domain-preserving framework for PAMPA scheme: 1D case*, to appear in SIAM J. Sci. Comput., <https://arxiv.org/abs/2412.03423>.

- [11] R. ABGRALL, M. JIAO, Y. LIU, AND K. WU, *Bound-preserving Point-Average-MomentPolynomiAl-Interpreted (PAMPA) scheme: one-dimensional case*, Commun. Comput. Phys., 39 (2026), pp. 29–58.
- [12] R. ABGRALL, A. LARAT, AND M. RICCHIUTO, *Construction of very high order residual distribution schemes for steady inviscid flow problems on hybrid unstructured meshes*, J. Comput. Phys., 230 (2011), pp. 4103–4136.
- [13] R. ABGRALL, Y. LIU, AND W. BOSCHERI, *Bound preserving Point-Average-Moment PolynomiAl-interpreted (PAMPA) on polygonal meshes*, Feb. 2025, <https://arxiv.org/abs/2502.10069>.
- [14] R. ABGRALL, M. LUKÁČOVA-MEDVID'OVÁ, AND P. ÖFFNER, *On the convergence of residual distribution schemes for the compressible Euler equations via dissipative weak solutions*, Math. Models Methods Appl. Sci., 33 (2023), pp. 139–173.
- [15] R. ABGRALL, K. MER, AND B. NKONGA, *A Lax–Wendroff type theorem for residual schemes*, in Innovative Methods for Numerical Solution of Partial Differential Equations, World Scientific, 2002, pp. 243–266.
- [16] R. ABGRALL AND M. MEZINE, *Construction of second order accurate monotone and stable residual distribution schemes for unsteady flow problems*, J. Comput. Phys., 188 (2003), pp. 16–55.
- [17] R. ABGRALL, P. ÖFFNER, AND H. RANOCHA, *Reinterpretation and extension of entropy correction terms for residual distribution and discontinuous Galerkin schemes: application to structure preserving discretization*, J. Comput. Phys., 453 (2022), p. 110955.
- [18] R. ABGRALL AND P. L. ROE, *High order fluctuation schemes on triangular meshes*, J. Sci. Comput., 19 (2003), pp. 3–36.
- [19] R. ABGRALL, Q. VIVILLE, H. BEAUGENDRE, AND C. DOBRZYNSKI, *Construction of ap-adaptive continuous residual distribution scheme*, J. Sci. Comput., 72 (2017), pp. 1232–1268.
- [20] G. ALLDREDGE AND F. SCHNEIDER, *A realizability-preserving discontinuous Galerkin scheme for entropy-based moment closures for linear kinetic equations in one space dimension*, J. Comput. Phys., 295 (2015), pp. 665–684.
- [21] G. R. BARRENECHEA, *Monotone Discretizations for Elliptic Second Order Partial Differential Equations*, Springer Nature, 2025.
- [22] G. R. BARRENECHEA, E. BURMAN, AND F. KARAKATSANI, *Edge-based nonlinear diffusion for finite element approximations of convection–diffusion equations and its relation to algebraic flux-correction schemes*, Numer. Math., 135 (2017), pp. 521–545.
- [23] G. R. BARRENECHEA, V. JOHN, AND P. KNOBLOCH, *Finite element methods respecting the discrete maximum principle for convection–diffusion equations*, SIAM Rev., 66 (2024), pp. 3–88.
- [24] G. R. BARRENECHEA AND P. KNOBLOCH, *Analysis of a group finite element formulation*, Appl. Numer. Math., 118 (2017), pp. 238–248.
- [25] T. J. BARTH AND D. C. JESPERSEN, *The design and application of upwind schemes on unstructured meshes*, in 27th AIAA Aerospace Sciences Meeting, Reno, NV, Jan. 1989, American Institute of Aeronautics and Astronautics. AIAA Paper 89-0366.
- [26] M. BERGER, M. AFTOSMIS, AND S. MUMAN, *Analysis of slope limiters on irregular grids*, in 43rd AIAA Aerospace Sciences Meeting and Exhibit, 2005, p. 490.
- [27] C. BERTHON, *Numerical approximations of the 10-moment Gaussian closure*, Math. Comp., 75 (2006), pp. 1809–1831.
- [28] C. BERTHON AND F. MARCHE, *A positive preserving high order VFRoe scheme for shallow water equations: a class of relaxation schemes*, SIAM J. Sci. Comput., 30 (2008), pp. 2587–2612.
- [29] P. BOCHEV, D. RIDZAL, M. D'ELIA, M. PEREGO, AND K. PETERSON, *Optimization-based, property-preserving finite element methods for scalar advection equations and their connection to algebraic flux correction*, Comput. Methods Appl. Mech. Engrg., 367 (2020), p. 112982.
- [30] P. BOCHEV, D. RIDZAL, AND K. PETERSON, *Optimization-based remap and transport: A divide and conquer strategy for feature-preserving discretizations*, J. Comput. Phys., 257 (2014), pp. 1113–1139.
- [31] P. BOCHEV, D. RIDZAL, G. SCOVAZZI, AND M. SHASHKOV, *Formulation, analysis and numerical study of an optimization-based conservative interpolation (remap) of scalar fields for arbitrary Lagrangian–Eulerian methods*, J. Comput. Phys., 230 (2011), pp. 5199–5225.
- [32] P. BOCHEV, D. RIDZAL, G. SCOVAZZI, AND M. SHASHKOV, *Constrained-optimization based data transfer*, in Flux-Corrected Transport: Principles, Algorithms, and Applications, D. Kuzmin, R. Löhner, and S. Turek, eds., Springer Netherlands, Dordrecht, 2012, pp. 345–398.
- [33] P. BOCHEV, D. RIDZAL, AND M. SHASHKOV, *Fast optimization-based conservative remap of scalar fields through aggregate mass transfer*, J. Comput. Phys., 246 (2013), pp. 37–57.
- [34] D. BOOK, J. BORIS, AND K. HAIN, *Flux-corrected transport II: Generalizations of the method*, J. Comput. Phys., 18 (1975), pp. 248–283.
- [35] J. P. BORIS AND D. L. BOOK, *Flux-corrected transport. I. SHASTA, a fluid transport algo-*

- rithm that works, *J. Comput. Phys.*, 11 (1973), pp. 38–69.
- [36] J. P. BORIS AND D. L. BOOK, *Flux-corrected transport. III. Minimal-error FCT algorithms*, *J. Comput. Phys.*, 20 (1976), pp. 397–431.
  - [37] W. BOSCHERI AND M. DUMBSER, *Arbitrary-Lagrangian–Eulerian discontinuous Galerkin schemes with a posteriori subcell finite volume limiting on moving unstructured meshes*, *J. Comput. Phys.*, 346 (2017), pp. 449–479.
  - [38] W. BOSCHERI, M. DUMBSER, R. LOUBÈRE, AND P.-H. MAIRE, *A second-order cell-centered Lagrangian ADER-MOOD finite volume scheme on multidimensional unstructured meshes for hydrodynamics*, *J. Comput. Phys.*, 358 (2018), pp. 103–129.
  - [39] A. M. BRADLEY, P. A. BOSLER, O. GUBA, M. A. TAYLOR, AND G. A. BARNETT, *Communication-efficient property preservation in tracer transport*, *SIAM J. Sci. Comput.*, 41 (2019), pp. C161–C193.
  - [40] M. BREUSS, *The correct use of the Lax–Friedrichs method*, *ESAIM Math. Model. Numer. Anal.*, 38 (2004), pp. 519–540.
  - [41] C. BUET AND B. DESPRES, *Asymptotic preserving and positive schemes for radiation hydrodynamics*, *J. Comput. Phys.*, 215 (2006), pp. 717–740.
  - [42] C. BUET, B. DESPRÉS, AND E. FRANCK, *An asymptotic preserving scheme with the maximum principle for the  $M_1$  model on distorted meshes*, *C. R. Math. Acad. Sci. Paris*, 350 (2012), pp. 633–638.
  - [43] C. CAI, J. QIU, AND K. WU, *Provably convergent Newton–Raphson method: Theoretically robust recovery of primitive variables in relativistic MHD*, *SIAM J. Numer. Anal.*, 63 (2025), pp. 1128–1159.
  - [44] H. CAO, M. PENG, AND K. WU, *Robust discontinuous Galerkin methods maintaining physical constraints for general relativistic hydrodynamics*, *J. Comput. Phys.*, (2025), p. 113770.
  - [45] M. J. CASTRO, J. M. GONZÁLEZ-VIDA, AND C. PARÉS, *Numerical treatment of wet/dry fronts in shallow flows with a modified Roe scheme*, *Math. Models Methods Appl. Sci.*, 16 (2006), pp. 897–931.
  - [46] T. CHEN AND C.-W. SHU, *Entropy stable high order discontinuous Galerkin methods with suitable quadrature rules for hyperbolic conservation laws*, *J. Comput. Phys.*, 345 (2017), pp. 427–461.
  - [47] Y. CHEN AND K. WU, *A physical-constraint-preserving finite volume WENO method for special relativistic hydrodynamics on unstructured meshes*, *J. Comput. Phys.*, 466 (2022), p. 111398.
  - [48] Z. CHEN, H. HUANG, AND J. YAN, *Third order maximum-principle-satisfying direct discontinuous Galerkin methods for time dependent convection diffusion equations on unstructured triangular meshes*, *J. Comput. Phys.*, 308 (2016), pp. 198–217.
  - [49] J. CHENG AND C.-W. SHU, *Positivity-preserving Lagrangian scheme for multi-material compressible flow*, *J. Comput. Phys.*, 257 (2014), pp. 143–168.
  - [50] Y. CHENG, I. M. GAMBA, AND P. J. MORRISON, *Study of conservation and recurrence of Runge–Kutta discontinuous Galerkin schemes for Vlasov–Poisson systems*, *J. Sci. Comput.*, 56 (2013), pp. 319–349.
  - [51] Y. CHENG, F. LI, J. QIU, AND L. XU, *Positivity-preserving DG and central DG methods for ideal MHD equations*, *J. Comput. Phys.*, 238 (2013), pp. 255–280.
  - [52] A. J. CHRISTLIEB, X. FENG, D. C. SEAL, AND Q. TANG, *A high-order positivity-preserving single-stage single-step method for the ideal magnetohydrodynamic equations*, *J. Comput. Phys.*, 316 (2016), pp. 218–242.
  - [53] A. J. CHRISTLIEB, Y. LIU, Q. TANG, AND Z. XU, *High order parametrized maximum-principle-preserving and positivity-preserving WENO schemes on unstructured meshes*, *J. Comput. Phys.*, 281 (2015), pp. 334–351.
  - [54] A. J. CHRISTLIEB, Y. LIU, Q. TANG, AND Z. XU, *Positivity-preserving finite difference weighted ENO schemes with constrained transport for ideal magnetohydrodynamic equations*, *SIAM J. Sci. Comput.*, 37 (2015), pp. A1825–A1845.
  - [55] K. N. CHUEH, C. C. CONLEY, AND J. A. SMOLLER, *Positively invariant regions for systems of nonlinear diffusion equations*, *Indiana Univ. Math. J.*, 26 (1977), pp. 373–392.
  - [56] B. COCKBURN, S. HOU, AND C.-W. SHU, *The Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws. IV. The multidimensional case*, *Math. Comp.*, 54 (1990), pp. 545–581.
  - [57] C. J. COTTER AND D. KUZMIN, *Embedded discontinuous Galerkin transport schemes with localised limiters*, *Journal of Computational Physics*, 311 (2016), pp. 363–373.
  - [58] M. G. CRANDALL AND A. MAJDA, *Monotone difference approximations for scalar conservation laws*, *Math. Comp.*, 34 (1980), pp. 1–21.
  - [59] L. J. CROSS AND X. ZHANG, *On the monotonicity of Q3 spectral element method for Laplacian*, *Ann. Appl. Math.*, 40 (2024), pp. 161–190.
  - [60] A. CSIK, M. RICCHIUTO, AND H. DECONINCK, *A conservative formulation of the multidimensional upwind residual distribution schemes for general nonlinear conservation laws*, *J. Comput. Phys.*, 179 (2002), pp. 286–312.
  - [61] S. CUI, S. DING, AND K. WU, *Is the classic convex decomposition optimal for bound-preserving*

- schemes in multiple dimensions?, *J. Comput. Phys.*, 476 (2023), p. 111882.
- [62] S. CUI, S. DING, AND K. WU, *On optimal cell average decomposition for high-order bound-preserving schemes of hyperbolic conservation laws*, *SIAM J. Numer. Anal.*, 62 (2024), pp. 775–810.
  - [63] S. CUI, K. WU, AND L. XU, *On local minimum entropy principle of high-order schemes for relativistic Euler equations*, *Math. Comp.*, (2025), <https://doi.org/10.1090/mcom/4139>.
  - [64] T. A. DAO, M. NAZAROV, AND I. TOMAS, *A structure preserving numerical method for the ideal compressible MHD system*, *J. Comput. Phys.*, 508 (2024), p. 113009.
  - [65] H. DECONINCK, P. L. ROE, AND R. STRUIJS, *A multidimensional generalization of Roe's flux difference splitter for the Euler equations*, *Comput. & Fluids*, 22 (1993), pp. 215–222.
  - [66] H. DECONINCK, R. STRUIJS, G. BOURGEOIS, AND P. L. ROE, *Compact advection schemes on unstructured meshes*, in *Computational Fluid Dynamics, VKI Lecture Series 1993-04*, 1993.
  - [67] D. DERIGS, A. R. WINTERS, G. J. GASSNER, AND S. WALCH, *A novel high-order, entropy stable, 3D AMR MHD solver with guaranteed positive pressure*, *J. Comput. Phys.*, 317 (2016), pp. 223–256.
  - [68] S. DING, S. CUI, AND K. WU, *Robust DG schemes on unstructured triangular meshes: Oscillation elimination and bound preservation via optimal convex decomposition*, *J. Comput. Phys.*, (2025), p. 113769.
  - [69] S. DING AND K. WU, *A new discretely divergence-free positivity-preserving high-order finite volume method for ideal MHD equations*, *SIAM J. Sci. Comput.*, 46 (2024), pp. A50–A79.
  - [70] S. DING AND K. WU, *GQL-based bound-preserving and locally divergence-free central discontinuous Galerkin schemes for relativistic magnetohydrodynamics*, *J. Comput. Phys.*, 514 (2024), p. 113208.
  - [71] S. DING, K. WU, AND C. YUAN, *Divergence-free finite volume WENO scheme for relativistic magnetohydrodynamics preserving positivity and subluminal velocity*, *Monthly Not. Roy. Astr. Soc.*, (2025), p. 1167–1190.
  - [72] C. DU AND M. LI, *A high-order domain preserving dg method for the two-layer shallow water equations*, *Comput. & Fluids*, 269 (2024), p. 106140.
  - [73] Q. DU, L. JU, X. LI, AND Z. QIAO, *Maximum bound principles for a class of semilinear parabolic equations and exponential time-differencing schemes*, *SIAM Rev.*, 63 (2021), pp. 317–359.
  - [74] M. DUMBSER AND R. LOUBÈRE, *A simple robust and accurate a posteriori sub-cell finite volume limiter for the discontinuous Galerkin method on unstructured meshes*, *J. Comput. Phys.*, 319 (2016), pp. 163–199.
  - [75] M. DUMBSER, O. ZANOTTI, R. LOUBÈRE, AND S. DIOT, *A posteriori subcell limiting of the discontinuous Galerkin finite element method for hyperbolic conservation laws*, *J. Comput. Phys.*, 278 (2014), pp. 47–75.
  - [76] B. EINFELDT, C.-D. MUNZ, P. L. ROE, AND B. SJÖGREEN, *On Godunov-type methods near low densities*, *J. Comput. Phys.*, 92 (1991), pp. 273–295.
  - [77] E. ENDEVE, C. D. HAUCK, Y. XING, AND A. MEZZACAPPA, *Bound-preserving discontinuous Galerkin methods for conservative phase space advection in curvilinear coordinates*, *J. Comput. Phys.*, 287 (2015), pp. 151–183.
  - [78] A. ERN AND J.-L. GUERMOND, *Invariant-domain-preserving high-order time stepping: I. Explicit Runge-Kutta schemes*, *SIAM J. Sci. Comput.*, 44 (2022), pp. A3366–A3392.
  - [79] A. ERN AND J.-L. GUERMOND, *Invariant-domain preserving high-order time stepping: II. IMEX schemes*, *SIAM J. Sci. Comput.*, 45 (2023), pp. A2511–A2538.
  - [80] D. J. ESTEP, M. G. LARSON, AND R. D. WILLIAMS, *Estimating the error of numerical solutions of systems of reaction-diffusion equations*, *Mem. Amer. Math. Soc.*, 146 (2000).
  - [81] J. ESTIVALEZES AND P. VILLEDIEU, *High-order positivity-preserving kinetic schemes for the compressible Euler equations*, *SIAM J. Numer. Anal.*, 33 (1996), pp. 2050–2067.
  - [82] C. FAN, X. ZHANG, AND J. QIU, *Positivity-preserving high order finite volume hybrid Hermite WENO schemes for compressible Navier-Stokes equations*, *J. Comput. Phys.*, 445 (2021), p. 110596.
  - [83] C. FAN, X. ZHANG, AND J. QIU, *Positivity-preserving high order finite difference WENO schemes for compressible Navier-Stokes equations*, *J. Comput. Phys.*, 467 (2022), p. 111446.
  - [84] C. A. FLETCHER, *The group finite element formulation*, *Comput. Methods Appl. Mech. Engrg.*, 37 (1983), pp. 225–244.
  - [85] H. FRID, *Maps of convex sets and invariant regions for finite-difference systems of conservation laws*, *Arch. Ration. Mech. Anal.*, 160 (2001), pp. 245–269.
  - [86] Q. FU, Y. GU, A. KURGANOV, AND B.-S. WANG, *Bound- and positivity-preserving path-conservative central-upwind AWENO scheme for the five-equation model of compressible two-component flows*, *J. Sci. Comput.*, 104 (2025). Article Number 94.
  - [87] S. K. GODUNOV, *Symmetric form of the equations of magnetohydrodynamics*, *Numerical Methods for Mechanics of Continuum Medium*, 1 (1972), pp. 26–34.
  - [88] S. K. GODUNOV AND I. BOHACHEVSKY, *Finite difference method for numerical computation*

- of discontinuous solutions of the equations of fluid dynamics, *Matematičeskij Sbornik*, 47 (1959), pp. 271–306.
- [89] J. B. GOODMAN AND R. J. LEVEQUE, *On the accuracy of stable schemes for 2D scalar conservation laws*, *Math. Comp.*, (1985), pp. 15–21.
  - [90] S. GOTTLIEB, D. I. KETCHESON, AND C.-W. SHU, *High order strong stability preserving time discretizations*, *J. Sci. Comput.*, 38 (2009), pp. 251–289.
  - [91] S. GOTTLIEB, C.-W. SHU, AND E. TADMOR, *Strong stability-preserving high-order time discretization methods*, *SIAM Rev.*, 43 (2001), pp. 89–112.
  - [92] D. GRAPSAS, R. HERBIN, W. KHERIJI, AND J.-C. LATCHÉ, *An unconditionally stable staggered pressure correction scheme for the compressible Navier-Stokes equations*, *SMAI J. Comput. Math.*, 2 (2016), pp. 51–97.
  - [93] Y. GU, Z. GAO, G. HU, P. LI, AND L. WANG, *A robust high order alternative WENO scheme for the five-equation model*, *J. Sci. Comput.*, 88 (2021). Article Number 12.
  - [94] O. GUBA, M. TAYLOR, AND A. ST-CYR, *Optimization-based limiters for the spectral element method*, *J. Comput. Phys.*, 267 (2014), pp. 176–195.
  - [95] J.-L. GUERMOND, M. MAIER, B. POPOV, AND I. TOMAS, *Second-order invariant domain preserving approximation of the compressible Navier-Stokes equations*, *Comput. Methods Appl. Mech. Engrg.*, 375 (2021), p. 113608.
  - [96] J.-L. GUERMOND AND M. NAZAROV, *A maximum-principle preserving C0 finite element method for scalar conservation equations*, *Comput. Methods Appl. Mech. Engrg.*, 272 (2014), pp. 198–213.
  - [97] J.-L. GUERMOND, M. NAZAROV, AND B. POPOV, *Finite element-based invariant-domain preserving approximation of hyperbolic systems: Beyond second-order accuracy in space*, *Comput. Methods Appl. Mech. Engrg.*, 418 (2024), p. 116470.
  - [98] J.-L. GUERMOND, M. NAZAROV, B. POPOV, AND I. TOMAS, *Second-order invariant domain preserving approximation of the Euler equations using convex limiting*, *SIAM J. Sci. Comput.*, 40 (2018), pp. A3211–A3239.
  - [99] J.-L. GUERMOND, M. NAZAROV, B. POPOV, AND Y. YANG, *A second-order maximum principle preserving Lagrange finite element technique for nonlinear scalar conservation equations*, *SIAM J. Numer. Anal.*, 52 (2014), pp. 2163–2182.
  - [100] J.-L. GUERMOND, R. PASQUETTI, AND B. POPOV, *Entropy viscosity method for nonlinear conservation laws*, *J. Comput. Phys.*, 230 (2011), pp. 4248–4267.
  - [101] J.-L. GUERMOND AND B. POPOV, *Fast estimation from above of the maximum wave speed in the Riemann problem for the Euler equations*, *J. Comput. Phys.*, 321 (2016), pp. 908–926.
  - [102] J.-L. GUERMOND AND B. POPOV, *Invariant domains and first-order continuous finite element approximation for hyperbolic systems*, *SIAM J. Numer. Anal.*, 54 (2016), pp. 2466–2489.
  - [103] J.-L. GUERMOND AND B. POPOV, *Invariant domains and second-order continuous finite element approximation for scalar conservation equations*, *SIAM J. Numer. Anal.*, 55 (2017), pp. 3120–3146.
  - [104] J.-L. GUERMOND, B. POPOV, AND L. SAAVEDRA, *Second-order invariant domain preserving ALE approximation of hyperbolic systems*, *J. Comput. Phys.*, 401 (2020), p. 108927.
  - [105] J.-L. GUERMOND, B. POPOV, AND L. SAAVEDRA, *Second-order invariant domain preserving ALE approximation of Euler equations*, *Commun. Appl. Math. Comput.*, 5 (2023), pp. 923–945.
  - [106] J.-L. GUERMOND, B. POPOV, L. SAAVEDRA, AND Y. YANG, *Invariant domains preserving arbitrary Lagrangian Eulerian approximation of hyperbolic systems with continuous finite elements*, *SIAM J. Sci. Comput.*, 39 (2017), pp. A385–A414.
  - [107] J.-L. GUERMOND, B. POPOV, L. SAAVEDRA, AND Y. YANG, *Arbitrary Lagrangian-Eulerian finite element method preserving convex invariants of hyperbolic systems*, in *Contributions to Partial Differential Equations and Applications*, B. N. Chetverushkin, W. Fitzgibbon, Y. Kuznetsov, P. Neittaanmäki, J. Periaux, and O. Pironneau, eds., Springer International Publishing, Cham, 2019, pp. 251–272.
  - [108] J.-L. GUERMOND, B. POPOV, AND I. TOMAS, *Invariant domain preserving discretization-independent schemes and convex limiting for hyperbolic systems*, *Comput. Methods Appl. Mech. Engrg.*, 347 (2019), pp. 143–175.
  - [109] A. HAIDAR, F. MARCHE, AND F. VILAR, *A posteriori finite-volume local subcell correction of high-order discontinuous Galerkin schemes for the nonlinear shallow-water equations*, *J. Comput. Phys.*, 452 (2022), p. 110902.
  - [110] H. HAJDUK, *Monolithic convex limiting in discontinuous Galerkin discretizations of hyperbolic conservation laws*, *Comput. Math. Appl.*, 87 (2021), pp. 120–138.
  - [111] H. HAJDUK, D. KUZMIN, T. KOLEV, V. TOMOV, I. TOMAS, AND J. N. SHADID, *Matrix-free subcell residual distribution for Bernstein finite elements: Monolithic limiting*, *Comput. & Fluids*, 200 (2020), p. 104451.
  - [112] A. HARTEN, J. M. HYMAN, P. D. LAX, AND B. KEYFITZ, *On finite-difference approximations and entropy conditions for shocks*, *Comm. Pure Appl. Math.*, 29 (1976), pp. 297–322.
  - [113] D. HOFF, *A finite difference scheme for a system of two conservation laws with artificial viscosity*, *Math. Comp.*, 33 (1979), pp. 1171–1193.

- [114] D. HOFF, *Invariant regions for systems of conservation laws*, Trans. Amer. Math. Soc., 289 (1985), pp. 591–610.
- [115] X. Y. HU, N. A. ADAMS, AND C.-W. SHU, *Positivity-preserving method for high-order conservative schemes solving compressible Euler equations*, J. Comput. Phys., 242 (2013), pp. 169–180.
- [116] A. JAMESON, *Positive schemes and shock modelling for compressible flows*, Internat. J. Numer. Methods Fluids, 20 (1995), pp. 743–776.
- [117] P. JANHUNEN, *A positive conservative method for magnetohydrodynamics based on HLL and Roe methods*, J. Comput. Phys., 160 (2000), pp. 649–661.
- [118] G.-S. JIANG AND E. TADMOR, *Nonoscillatory central schemes for multidimensional hyperbolic conservation laws*, SIAM J. Sci. Comput., 19 (1998), pp. 1892–1917.
- [119] Y. JIANG AND H. LIU, *Invariant-region-preserving DG methods for multi-dimensional hyperbolic conservation law systems, with an application to compressible Euler equations*, J. Comput. Phys., 373 (2018), pp. 385–409.
- [120] Y. JIANG AND Z. XU, *Parametrized maximum principle preserving limiter for finite difference WENO schemes solving convection-dominated diffusion equations*, SIAM J. Sci. Comput., 35 (2013), pp. A2524–A2553.
- [121] B. KEITH AND T. M. SUROWIEC, *Proximal Galerkin: A structure-preserving finite element method for pointwise bound constraints*, Found. Comput. Math., (2024), pp. 1–97.
- [122] M. KENAMOND, D. KUZMIN, AND M. SHASHKOV, *A positivity-preserving and conservative intersection-distribution-based remapping algorithm for staggered ALE hydrodynamics on arbitrary meshes*, J. Comput. Phys., 435 (2021), p. 110254.
- [123] D. I. KETCHESON, R. J. LEVEQUE, AND M. J. DEL RAZO, *Riemann problems and Jupiter solutions*, vol. 16, SIAM, 2020.
- [124] B. KHOBALATTE AND B. PERTHAME, *Maximum principle on the entropy and second-order kinetic schemes*, Math. Comp., 62 (1994), pp. 119–131.
- [125] L. E. KIDDER, S. E. FIELD, F. FOUCART, E. SCHNETTER, S. A. TEUKOLSKY, A. BOHN, N. DEPPE, P. DIENER, F. HÉBERT, J. LIPPUNER, J. MILLER, C. D. OTT, M. A. SCHEEL, AND T. VINCENT, *SpECTRE: A task-based discontinuous Galerkin code for relativistic astrophysics*, J. Comput. Phys., 335 (2017), pp. 84–114.
- [126] R. C. KIRBY AND D. SHAPERO, *High-order bounds-satisfying approximation of partial differential equations via finite element variational inequalities*, Numer. Math., 156 (2024), pp. 927–947.
- [127] C. KLINGENBERG, *Numerical methods for astrophysics*, in Handbook of Numerical Analysis, vol. 18, Elsevier, 2017, pp. 465–477.
- [128] A. KURGANOV AND G. PETROVA, *A second-order well-balanced positivity preserving central-upwind scheme for the Saint-Venant system*, Commun. Math. Sci., 5 (2007).
- [129] A. KURGANOV AND E. TADMOR, *New high-resolution central schemes for nonlinear conservation laws and convection–diffusion equations*, J. Comput. Phys., 160 (2000), pp. 241–282.
- [130] D. KUZMIN, *Positive finite element schemes based on the flux-corrected transport procedure*, Computational Fluid and Solid Mechanics, Elsevier, (2001), pp. 887–888.
- [131] D. KUZMIN, *Monolithic convex limiting for continuous finite element discretizations of hyperbolic conservation laws*, Comput. Methods Appl. Mech. Engrg., 361 (2020), p. 112804.
- [132] D. KUZMIN AND M. Q. DE LUNA, *Subcell flux limiting for high-order Bernstein finite element discretizations of scalar hyperbolic conservation laws*, Journal of Computational Physics, 411 (2020), p. 109411.
- [133] D. KUZMIN AND H. HAJDUK, *Property-preserving numerical schemes for conservation laws*, World Scientific, 2024.
- [134] D. KUZMIN AND N. KLYUSHNEV, *Limiting and divergence cleaning for continuous finite element discretizations of the MHD equations*, J. Comput. Phys., 407 (2020), p. 109230.
- [135] D. KUZMIN, R. LÖHNER, AND S. TUREK, *Flux-corrected transport: principles, algorithms, and applications*, Springer Science & Business Media, 2012.
- [136] D. KUZMIN, LUKÁČOVÁ-MEDVID'OVÁ, AND P. ÖFFNER, *Consistency and convergence of flux-corrected finite element methods for nonlinear hyperbolic problems*, J. Numer. Math., (2025), <https://doi.org/10.1515/jnma-2024-0123>.
- [137] D. KUZMIN, M. MÖLLER, J. N. SHADID, AND M. SHASHKOV, *Failsafe flux limiting and constrained data projections for equations of gas dynamics*, Journal of Computational physics, 229 (2010), pp. 8766–8779.
- [138] D. KUZMIN, M. MÖLLER, AND S. TUREK, *High-resolution FEM–FCT schemes for multidimensional conservation laws*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 4915–4946.
- [139] D. KUZMIN, M. QUEZADA DE LUNA, D. I. KETCHESON, AND J. GRÜLL, *Bound-preserving flux limiting for high-order explicit runge–kutta time discretizations of hyperbolic conservation laws*, J. Sci. Comput., 91 (2022), p. 21.
- [140] D. KUZMIN, M. QUEZADA DE LUNA, D. I. KETCHESON, AND J. GRÜLL, *Bound-preserving flux limiting for high-order explicit Runge–Kutta time discretizations of hyperbolic conservation laws*, Journal of Scientific Computing, 91 (2022). Article Number 21.

- [141] D. KUZMIN AND S. TUREK, *Flux correction tools for finite elements*, J. Comput. Phys., 175 (2002), pp. 525–558.
- [142] P. LAX AND B. WENDROFF, *Systems of conservation laws*, Comm. Pure Appl. Math., 13 (1960), pp. 217–237.
- [143] R. J. LEVEQUE, *Numerical methods for conservation laws*, vol. 132, Springer, 1992.
- [144] C. D. LEVERMORE AND W. J. MOROKOFF, *The Gaussian moment closure for gas dynamics*, SIAM J. Appl. Math., 59 (1998), pp. 72–96.
- [145] H. LI, S. XIE, AND X. ZHANG, *A high order accurate bound-preserving compact finite difference scheme for scalar convection diffusion equations*, SIAM J. Numer. Anal., 56 (2018), pp. 3308–3345.
- [146] H. LI AND X. ZHANG, *On the monotonicity and discrete maximum principle of the finite difference implementation of C0-Q2 finite element method*, Numer. Math., 145 (2020), pp. 437–472.
- [147] Y. LIN, J. CHAN, AND I. TOMAS, *A positivity preserving strategy for entropy stable discontinuous Galerkin discretizations of the compressible Euler and Navier-Stokes equations*, J. Comput. Phys., 475 (2023), p. 111850.
- [148] R. LISKA, M. SHASHKOV, P. VÁCHAL, AND B. WENDROFF, *Optimization-based synchronized flux-corrected conservative interpolation (remapping) of mass and momentum for arbitrary Lagrangian–Eulerian methods*, J. Comput. Phys., 229 (2010), pp. 1467–1497.
- [149] C. LIU, G. T. BUZZARD, AND X. ZHANG, *An optimization based limiter for enforcing positivity in a semi-implicit discontinuous Galerkin scheme for compressible Navier–Stokes equations*, J. Comput. Phys., 519 (2024), p. 113440.
- [150] C. LIU, D. MILESIS, C.-W. SHU, AND X. ZHANG, *Efficient optimization-based invariant-domain-preserving limiters in solving gas dynamics equations*, 2025, <https://arxiv.org/abs/2510.21080>.
- [151] C. LIU, B. RIVIERE, J. SHEN, AND X. ZHANG, *A simple and efficient convex optimization based bound-preserving high order accurate limiter for Cahn–Hilliard–Navier–Stokes system*, SIAM J. Sci. Comput., 46 (2024), pp. A1923–A1948.
- [152] C. LIU AND X. ZHANG, *A positivity-preserving implicit-explicit scheme with high order polynomial basis for compressible Navier–Stokes equations*, J. Comput. Phys., 493 (2023), p. 112496.
- [153] H. LIU AND H. YU, *Maximum-principle-satisfying third order discontinuous Galerkin schemes for Fokker–Planck equations*, SIAM J. Sci. Comput., 36 (2014), pp. A2296–A2325.
- [154] M. LIU AND K. WU, *Structure-preserving oscillation-eliminating discontinuous Galerkin schemes for ideal MHD equations: Locally divergence-free and positivity-preserving*, J. Comput. Phys., (2025), p. 113795.
- [155] X.-D. LIU AND S. OSHER, *Nonoscillatory high order accurate self-similar maximum principle satisfying shock capturing schemes I*, SIAM J. Numer. Anal., 33 (1996), pp. 760–779.
- [156] C. LOHMANN AND D. KUZMIN, *Synchronized flux limiting for gas dynamics variables*, J. Comput. Phys., 326 (2016), pp. 973–990.
- [157] C. LOHMANN, D. KUZMIN, J. N. SHADID, AND S. MABUZA, *Flux-corrected transport algorithms for continuous Galerkin methods based on high order Bernstein finite elements*, Journal of Computational Physics, 344 (2017), pp. 151–186.
- [158] R. LÖHNER, K. MORGAN, J. PERAIRE, AND M. VAHDATI, *Finite element flux-corrected transport (FEM–FCT) for the Euler and Navier–Stokes equations*, Internat. J. Numer. Methods Fluids, 7 (1987), pp. 1093–1109.
- [159] J. LORENZ, *Zur inversmonotonie diskreter probleme*, Numer. Math., 27 (1977), pp. 227–238.
- [160] Y. LV AND M. IHME, *Entropy-bounded discontinuous Galerkin scheme for Euler equations*, J. Comput. Phys., 295 (2015), pp. 715–739.
- [161] Y. LV AND M. IHME, *High-order discontinuous Galerkin method for applications to multicomponent and chemically reacting flows*, Acta Mech. Sin., 33 (2017), pp. 486–499.
- [162] K. T. MANDLI, *A numerical method for the two layer shallow water equations with dry states*, Ocean Modelling, 72 (2013), pp. 80–91.
- [163] A. K. MEENA, H. KUMAR, AND P. CHANDRASHEKAR, *Positivity-preserving high-order discontinuous Galerkin schemes for ten-moment Gaussian closure equations*, J. Comput. Phys., 339 (2017), pp. 370–395.
- [164] A. MEISTER AND S. ORTLEB, *A positivity preserving and well-balanced DG scheme using finite volume subcells in almost dry regions*, Appl. Math. Comput., 272 (2016), pp. 259–273.
- [165] A. MEZZACAPPA, E. ENDEVE, O. B. MESSER, AND S. W. BRUENN, *Physical, numerical, and computational challenges of modeling neutrino transport in core-collapse supernovae*, Living Reviews in Computational Astrophysics, 6 (2020), pp. 1–174.
- [166] S. A. MOE, J. A. ROSSMANITH, AND D. C. SEAL, *Positivity-preserving discontinuous Galerkin methods with Lax–Wendroff time discretizations*, J. Sci. Comput., 71 (2017), pp. 44–70.
- [167] P. MOUJAES AND D. KUZMIN, *Monolithic convex limiting and implicit pseudo-time stepping for calculating steady-state solutions of the Euler equations*, J. Comput. Phys., 523 (2025), p. 113687.
- [168] R. D. NAIR, M. N. LEVY, AND P. H. LAURITZEN, *Emerging numerical methods for atmo-*

- spheric modeling*, in Numerical Techniques for Global Atmospheric Models, P. Lauritzen, C. Jablonowski, M. Taylor, and R. Nair, eds., Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 251–311.
- [169] R.-H. NI, *A multiple-grid scheme for solving the Euler equations*, AIAA journal, 20 (1982), pp. 1565–1571.
  - [170] E. OLBRANT, C. D. HAUCK, AND M. FRANK, *A realizability-preserving discontinuous Galerkin method for the M1 model of radiative transfer*, J. Comput. Phys., 231 (2012), pp. 5612–5639.
  - [171] E. S. ORAN, J. P. BORIS, AND J. P. BORIS, *Numerical simulation of reactive flow*, vol. 2, Cambridge University Press, 2001.
  - [172] D. PANG AND K. WU, *Provably positivity-preserving constrained transport scheme for 2D and 3D ideal magnetohydrodynamics*, J. Comput. Phys., 541 (2025), p. 114312.
  - [173] W. PAZNER, *Sparse invariant domain preserving discontinuous Galerkin methods with subcell convex limiting*, Comput. Methods Appl. Mech. Engrg., 382 (2021), p. 113876.
  - [174] B. PERTHAME, *Second-order Boltzmann schemes for compressible Euler equations in one and two space dimensions*, SIAM J. Numer. Anal., 29 (1992), pp. 1–19.
  - [175] B. PERTHAME AND Y. QIU, *A variant of Van Leer’s method for multidimensional systems of conservation laws*, J. Comput. Phys., 112 (1994), pp. 370–381.
  - [176] B. PERTHAME AND C.-W. SHU, *On positivity preserving finite volume schemes for Euler equations*, Numer. Math., 73 (1996), pp. 119–130.
  - [177] K. PETERSON, P. BOCHEV, AND D. RIDZAL, *Optimization-based, property-preserving algorithm for passive tracer transport*, Comput. Math. Appl., 159 (2024), pp. 267–286.
  - [178] T. QIN AND C.-W. SHU, *Implicit positivity-preserving high-order discontinuous Galerkin methods for conservation laws*, SIAM J. Sci. Comput., 40 (2018), pp. A81–A107.
  - [179] T. QIN, C.-W. SHU, AND Y. YANG, *Bound-preserving discontinuous Galerkin methods for relativistic hydrodynamics*, J. Comput. Phys., 315 (2016), pp. 323–347.
  - [180] J.-M. QIU AND C.-W. SHU, *Positivity preserving semi-Lagrangian discontinuous Galerkin formulation: Theoretical analysis and application to the Vlasov–Poisson system*, J. Comput. Phys., 230 (2011), pp. 8386–8409.
  - [181] M. QUEZADA DE LUNA AND D. I. KETCHESON, *Maximum principle preserving space and time flux limiting for Diagonally Implicit Runge–Kutta discretizations of scalar convection-diffusion equations*, Journal of Scientific Computing, 92 (2022). Article Number 102.
  - [182] D. RADICE, L. REZZOLLA, AND F. GALEAZZI, *High-order fully general-relativistic hydrodynamics: New approaches and tests*, Classical and Quantum Gravity, 31 (2014), p. 075012.
  - [183] M. RICCHIUTO AND R. ABGRALL, *Explicit Runge-Kutta residual distribution schemes for time dependent problems: second order case*, J. Comput. Phys., 229 (2010), pp. 5653–5691.
  - [184] M. RICCHIUTO, Á. CSÍK, AND H. DECONINCK, *Residual distribution for general time-dependent conservation laws*, J. Comput. Phys., 209 (2005), pp. 249–289.
  - [185] W. RIDER, D. KOTHE, W. RIDER, AND D. KOTHE, *Constrained minimization for monotonic reconstruction*, in 13th Computational Fluid Dynamics Conference, 1997, p. 2036.
  - [186] P. L. ROE AND D. SIDILKOVER, *Optimum positive linear schemes for advection in two and three dimensions*, SIAM J. Numer. Anal., 29 (1992), pp. 1542–1568.
  - [187] J. A. ROSSMANITH AND D. C. SEAL, *A positivity-preserving high-order semi-Lagrangian discontinuous Galerkin scheme for the Vlasov–Poisson equations*, J. Comput. Phys., 230 (2011), pp. 6203–6232.
  - [188] A. M. RUEDA-RAMÍREZ, B. BOLM, D. KUZMIN, AND G. J. GASSNER, *Monolithic convex limiting for Legendre-Gauss-Lobatto discontinuous Galerkin spectral-element methods*, Commun. Appl. Math. Comput., 6 (2024), pp. 1860–1898.
  - [189] F. RUPPENTHAL AND D. KUZMIN, *Optimal control using flux potentials: A way to construct bound-preserving finite element schemes for conservation laws*, J. Comput. Appl. Math., 434 (2023), p. 115351.
  - [190] C. SCHÄR AND P. K. SMOLARKIEWICZ, *A synchronous and iterative flux-correction formalism for coupled transport equations*, J. Comput. Phys., 128 (1996), pp. 101–120.
  - [191] L. I. SEDOV, *Similarity and Dimensional Methods in Mechanics*, CRC Press, 10 ed., 1993, <https://doi.org/10.1201/9780203739730>.
  - [192] V. SELMIN, *The node-centred finite volume approach: bridge between finite differences and finite elements*, Comput. Methods Appl. Mech. Engrg., 102 (1993), pp. 107–138.
  - [193] V. SELMIN AND L. FORMAGGIA, *Unified construction of finite element and finite volume discretizations for compressible flows*, Internat. J. Numer. Methods Engrg., 39 (1996), pp. 1–32.
  - [194] C. SHI AND C.-W. SHU, *On local conservation of numerical methods for conservation laws*, Comput. & Fluids, 169 (2018), pp. 3–9.
  - [195] C.-W. SHU, *High order weighted essentially nonoscillatory schemes for convection dominated problems*, SIAM Rev., 51 (2009), pp. 82–126.
  - [196] J. SMOLLER, *Shock waves and reaction-diffusion equations*, vol. 258, Springer Science & Business Media, 2012.
  - [197] A. SOMMARIVA AND M. VIANELLO, *Gauss-Green cubature and moment computation over*

- arbitrary geometries, *J. Comput. Appl. Math.*, 231 (2009), pp. 886–896.
- [198] S. SRINIVASAN, J. POGGIE, AND X. ZHANG, *A positivity-preserving high order discontinuous Galerkin scheme for convection–diffusion equations*, *J. Comput. Phys.*, 366 (2018), pp. 120–143.
  - [199] R. STRUIJS, H. DECONINCK, AND P. ROE, *Fluctuation splitting schemes for the 2D Euler equations*, in *Computational Fluid Dynamics, VKI Lecture Series 1991-01*, 1991.
  - [200] Z. SUN, J. A. CARRILLO, AND C.-W. SHU, *A discontinuous Galerkin method for nonlinear parabolic equations and gradient flow problems with interaction potentials*, *J. Comput. Phys.*, 352 (2018), pp. 76–104.
  - [201] E. TADMOR, *A minimum entropy principle in the gas dynamics equations*, *Appl. Numer. Math.*, 2 (1986), pp. 211–219.
  - [202] H.-Z. TANG AND K. XU, *Positivity-preserving analysis of explicit and implicit Lax–Friedrichs schemes for compressible Euler equations*, *J. Sci. Comput.*, 15 (2000), pp. 19–28.
  - [203] T. TAO AND K. XU, *Gas-kinetic schemes for the compressible Euler equations: positivity-preserving analysis*, *Zeitschrift für angewandte Mathematik und Physik ZAMP*, 50 (1999), pp. 258–281.
  - [204] W. TONG, R. YAN, AND G. CHEN, *On a class of robust bound-preserving MUSCL–Hancock schemes*, *J. Comput. Phys.*, 474 (2023), p. 111805.
  - [205] E. F. TORO, *Riemann solvers and numerical methods for fluid dynamics: a practical introduction*, Springer Science & Business Media, 2013.
  - [206] B. VAN DER HOLST, R. KEPPENS, AND Z. MELIANI, *A multidimensional grid-adaptive relativistic magnetofluid code*, *Comput. Phys. Commun.*, 179 (2008), pp. 617–627.
  - [207] J. J. VAN DER VEGT, Y. XIA, AND Y. XU, *Positivity preserving limiters for time-implicit higher order accurate discontinuous Galerkin discretizations*, *SIAM J. Sci. Comput.*, 41 (2019), pp. A2037–A2063.
  - [208] F. VILAR, *A posteriori correction of high-order discontinuous Galerkin scheme through subcell finite volume formulation and flux reconstruction*, *J. Comput. Phys.*, 387 (2019), pp. 245–279.
  - [209] F. VILAR, *Local subcell monolithic DG/FV convex property preserving scheme on unstructured grids and entropy consideration*, *J. Comput. Phys.*, 521 (2025), p. 113535.
  - [210] F. VILAR AND R. ABGRALL, *A posteriori local subcell correction of high-order discontinuous Galerkin scheme for conservation laws on two-dimensional unstructured grids*, *SIAM J. Sci. Comput.*, 46 (2024), pp. A851–A883.
  - [211] F. VILAR, C.-W. SHU, AND P.-H. MAIRE, *Positivity-preserving cell-centered Lagrangian schemes for multi-material compressible flows: From first-order to high-orders. Part I: The one-dimensional case*, *J. Comput. Phys.*, 312 (2016), pp. 385–415.
  - [212] F. VILAR, C.-W. SHU, AND P.-H. MAIRE, *Positivity-preserving cell-centered Lagrangian schemes for multi-material compressible flows: From first-order to high-orders. Part II: The two-dimensional case*, *J. Comput. Phys.*, 312 (2016), pp. 416–442.
  - [213] C. WANG, X. ZHANG, C.-W. SHU, AND J. NING, *Robust high order discontinuous Galerkin schemes for two-dimensional gaseous detonations*, *J. Comput. Phys.*, 231 (2012), pp. 653–665.
  - [214] K. WU, *Design of provably physical-constraint-preserving methods for general relativistic hydrodynamics*, *Phys. Rev. D*, 95 (2017), p. 103001.
  - [215] K. WU, *Positivity-preserving analysis of numerical schemes for ideal magnetohydrodynamics*, *SIAM J. Numer. Anal.*, 56 (2018), pp. 2124–2147.
  - [216] K. WU, *Minimum principle on specific entropy and high-order accurate invariant-region-preserving numerical methods for relativistic hydrodynamics*, *SIAM J. Sci. Comput.*, 43 (2021), pp. B1164–B1197.
  - [217] K. WU, H. JIANG, AND C.-W. SHU, *Provably positive central discontinuous Galerkin schemes via geometric quasilinearization for ideal MHD equations*, *SIAM J. Numer. Anal.*, 61 (2023), pp. 250–285.
  - [218] K. WU AND C.-W. SHU, *A provably positive discontinuous Galerkin method for multidimensional ideal magnetohydrodynamics*, *SIAM J. Sci. Comput.*, 40 (2018), pp. B1302–B1329.
  - [219] K. WU AND C.-W. SHU, *Provably positive high-order schemes for ideal magnetohydrodynamics: analysis on general meshes*, *Numer. Math.*, 142 (2019), pp. 995–1047.
  - [220] K. WU AND C.-W. SHU, *Entropy symmetrization and high-order accurate entropy stable numerical schemes for relativistic MHD equations*, *SIAM J. Sci. Comput.*, 42 (2020), pp. A2230–A2261.
  - [221] K. WU AND C.-W. SHU, *Provably physical-constraint-preserving discontinuous Galerkin methods for multidimensional relativistic MHD equations*, *Numer. Math.*, 148 (2021), pp. 699–741.
  - [222] K. WU AND C.-W. SHU, *Geometric quasilinearization framework for analysis and design of bound-preserving schemes*, *SIAM Rev.*, 65 (2023), pp. 1031–1073.
  - [223] K. WU AND H. TANG, *High-order accurate physical-constraints-preserving finite difference WENO schemes for special relativistic hydrodynamics*, *J. Comput. Phys.*, 298 (2015), pp. 539–564.

- [224] K. WU AND H. TANG, *Physical-constraint-preserving central discontinuous Galerkin methods for special relativistic hydrodynamics with a general equation of state*, The Astrophysical Journal Supplement Series, 228 (2016), p. 3.
- [225] K. WU AND H. TANG, *Admissible states and physical-constraints-preserving schemes for relativistic magnetohydrodynamic equations*, Math. Models Methods Appl. Sci., 27 (2017), pp. 1871–1928.
- [226] Y. XING AND X. ZHANG, *Positivity-preserving well-balanced discontinuous Galerkin methods for the shallow water equations on unstructured triangular meshes*, J. Sci. Comput., 57 (2013), pp. 19–41.
- [227] Y. XING, X. ZHANG, AND C.-W. SHU, *Positivity-preserving high order well-balanced discontinuous Galerkin methods for the shallow water equations*, Advances in Water Resources, 33 (2010), pp. 1476–1493.
- [228] T. XIONG, J.-M. QIU, AND Z. XU, *A parametrized maximum principle preserving flux limiter for finite difference RK-WENO schemes with applications in incompressible flows*, J. Comput. Phys., 252 (2013), pp. 310–331.
- [229] T. XIONG, J.-M. QIU, AND Z. XU, *High order maximum-principle-preserving discontinuous Galerkin method for convection-diffusion equations*, SIAM J. Sci. Comput., 37 (2015), pp. A583–A608.
- [230] T. XIONG, J.-M. QIU, AND Z. XU, *Parametrized positivity preserving flux limiters for the high order finite difference WENO scheme solving compressible Euler equations*, J. Sci. Comput., 67 (2016), pp. 1066–1088.
- [231] T. XIONG, J.-M. QIU, Z. XU, AND A. CHRISTLIEB, *High order maximum principle preserving semi-Lagrangian finite difference WENO schemes for the Vlasov equation*, J. Comput. Phys., 273 (2014), pp. 618–639.
- [232] J. XU AND L. ZIKATANOV, *A monotone finite element scheme for convection-diffusion equations*, Math. Comp., 68 (1999), pp. 1429–1446.
- [233] Z. XU, *Parametrized maximum principle preserving flux limiters for high order schemes solving hyperbolic conservation laws: One-dimensional scalar problem*, Math. Comp., 83 (2014), pp. 2213–2238.
- [234] Z. XU AND X. ZHANG, *Bound-preserving high-order schemes*, in Handbook of Numerical Analysis, vol. 18, Elsevier, 2017, pp. 81–102.
- [235] B. C. YEE, S. S. OLIVIER, T. S. HAUT, M. HOLEC, V. Z. TOMOV, AND P. G. MAGINOT, *A quadratic programming flux correction method for high-order DG discretizations of SN transport*, J. Comput. Phys., 419 (2020), p. 109696.
- [236] K. YU, J. CHENG, Y. LIU, AND C.-W. SHU, *High-order implicit maximum-principle-preserving local discontinuous galerkin methods for convection–diffusion equations*, J. Comput. Appl. Math., (2025), p. 116660.
- [237] V. ZALA, R. M. KIRBY, AND A. NARAYAN, *Structure-preserving nonlinear filtering for continuous and discontinuous Galerkin spectral/hp element methods*, SIAM J. Sci. Comput., 43 (2021), pp. A3713–A3732.
- [238] V. ZALA, A. NARAYAN, AND R. M. KIRBY, *Convex optimization-based structure-preserving filter for multidimensional finite element simulations*, J. Comput. Phys., 492 (2023), p. 112364.
- [239] S. T. ZALESK, *Fully multidimensional flux-corrected transport algorithms for fluids*, J. Comput. Phys., 31 (1979), pp. 335–362.
- [240] S. T. ZALESK, *The design of Flux-Corrected Transport (FCT) algorithms for structured grids*, Springer, 2012.
- [241] L. ZHANG, T. CUI, AND H. LIU, *A set of symmetric quadrature rules on triangles and tetrahedra*, J. Comput. Math., 27 (2009), pp. 89–96.
- [242] X. ZHANG, *On positivity-preserving high order discontinuous Galerkin schemes for compressible Navier–Stokes equations*, J. Comput. Phys., 328 (2017), pp. 301–343.
- [243] X. ZHANG, Y. LIU, AND C.-W. SHU, *Maximum-principle-satisfying high order finite volume weighted essentially nonoscillatory schemes for convection-diffusion equations*, SIAM J. Sci. Comput., 34 (2012), pp. A627–A658.
- [244] X. ZHANG AND C.-W. SHU, *A genuinely high order total variation diminishing scheme for one-dimensional scalar conservation laws*, SIAM J. Numer. Anal., 48 (2010), pp. 772–795.
- [245] X. ZHANG AND C.-W. SHU, *On maximum-principle-satisfying high order schemes for scalar conservation laws*, J. Comput. Phys., 229 (2010), pp. 3091–3120.
- [246] X. ZHANG AND C.-W. SHU, *On positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations on rectangular meshes*, J. Comput. Phys., 229 (2010), pp. 8918–8934.
- [247] X. ZHANG AND C.-W. SHU, *Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: survey and new developments*, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 467 (2011), pp. 2752–2776.
- [248] X. ZHANG AND C.-W. SHU, *A minimum entropy principle of high order schemes for gas dynamics equations*, Numer. Math., 121 (2012), pp. 545–563.
- [249] X. ZHANG AND C.-W. SHU, *Positivity-preserving high order finite difference WENO schemes*

- for compressible Euler equations*, J. Comput. Phys., 231 (2012), pp. 2245–2258.
- [250] X. ZHANG, Y. XIA, AND C.-W. SHU, *Maximum-principle-satisfying and positivity-preserving high order discontinuous Galerkin schemes for conservation laws on triangular meshes*, J. Sci. Comput., 50 (2012), pp. 29–62.