ACTIVE FLUX METHODS FOR HYPERBOLIC CONSERVATION LAWS – FLUX VECTOR SPLITTING AND BOUND-PRESERVATION: ONE-DIMENSIONAL CASE *

JUNMING DUAN†, WASILIJ BARSUKOW‡, AND CHRISTIAN KLINGENBERG†

Abstract. The active flux (AF) method is a compact high-order finite volume method that evolves cell averages and point values at cell interfaces independently. Within the method of lines framework, the point value can be updated based on Jacobian splitting (JS), incorporating the upwind idea. However, such JS-based AF methods encounter transonic issues for nonlinear problems due to inaccurate upwind direction estimation. This paper proposes to use flux vector splitting for the point value update, offering a natural and uniform remedy to the transonic issue. To improve robustness, this paper also develops bound-preserving (BP) AF methods for one-dimensional hyperbolic conservation laws. Two cases are considered: preservation of the maximum principle for the scalar case, and preservation of positive density and pressure for the compressible Euler equations. The update of the cell average in high-order AF methods is rewritten as a convex combination of using the original high-order fluxes and robust low-order (local Lax-Friedrichs or Rusanov) fluxes, and the desired bounds are enforced by choosing the right amount of low-order fluxes. A similar blending strategy is used for the point value update. Several challenging benchmark tests are conducted to verify the accuracy, BP properties, and shock-capturing ability of the methods.

Key words. Hyperbolic conservation laws, finite volume method, active flux, flux vector splitting, bound-preserving, convex limiting, scaling limiter

MSC codes. 65M08, 65M12, 65M20, 35L65

1. Introduction.

This paper is concerned with solving systems of hyperbolic conservation laws

$$\frac{\partial \boldsymbol{U}(x,t)}{\partial t} + \frac{\partial \boldsymbol{F}(\boldsymbol{U})}{\partial x} = 0, \quad \boldsymbol{U}(x,0) = \boldsymbol{U}_0(x), \quad (x,t) \in \mathbb{R} \times \mathbb{R}^+, \tag{1.1}$$

where $\boldsymbol{U} \in \mathbb{R}^m$ is the vector of m conservative variables, $\boldsymbol{F} \in \mathbb{R}^m$ is the physical flux, and $\boldsymbol{U}_0(x)$ is assumed to be initial data of bounded variation. In this paper, we would like to consider two cases. The first is a scalar conservation law ($m = 1$)

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0, \quad u(x,0) = u_0(x). \tag{1.2}$$

The second case is that of compressible Euler equations of gas dynamics with $\boldsymbol{U} = (\rho, \rho v, E)^\top$ and $\boldsymbol{F} = (\rho v, \rho v^2 + p, (E+p)v)^\top$, i.e.,

$$\frac{\partial}{\partial t}\begin{pmatrix} \rho \\ \rho v \\ E \end{pmatrix} + \frac{\partial}{\partial x}\begin{pmatrix} \rho v \\ \rho v^2 + p \\ (E+p)v \end{pmatrix} = \mathbf{0}, \quad (\rho, v, p)(x,0) = (\rho_0, v_0, p_0). \tag{1.3}$$

Here ρ denotes the density, v the velocity, p the pressure, and $E = \frac{1}{2}\rho v^2 + \rho e$ the total energy with e the specific internal energy. The system (1.3) should be closed by

*Submitted to the editors DATE.

Funding: JD was supported by an Alexander von Humboldt Foundation Research fellowship CHN-1234352-HFST-P. CK and WB acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within *SPP 2410 Hyperbolic Balance Laws in Fluid Mechanics: Complexity, Scales, Randomness (CoScaRa)*, project number 525941602.

†Department of Mathematics, University of Würzburg, Germany (junming.duan@uni-wuerzburg.de, christian.klingenberg@uni-wuerzburg.de).

‡Institut de Mathématiques de Bordeaux (IMB), CNRS UMR 5251, University of Bordeaux, 33405 Talence, France (wasilij.barsukow@math.u-bordeaux.fr).

This manuscript is for review purposes only.

an equation of state (EOS). This paper considers the perfect gas EOS, $p = (\gamma - 1)\rho e$, with the adiabatic index $\gamma > 1$. Note that this paper uses bold symbols to refer to vectors and matrices, such that they are easier to distinguish from scalars.

The active flux (AF) method is a new finite volume method [13, 12, 14, 34], that Roe took inspiration by [39]. Apart from cell averages, it incorporates additional degrees of freedom as point values located at the cell interfaces, evolved independently from the cell average. The original AF method gives a global continuous representation of the numerical solution using a piecewise quadratic reconstruction, leading naturally to a third-order accurate method with a compact stencil. The introduction of point values at the cell interfaces avoids the usage of Riemann solvers as in usual Godunov methods, because the numerical solution is continuous across the cell interface and the numerical flux for the cell average update is available directly.

The independence of the point value update adds flexibility to the AF methods. Based on the evolution of the point value, there are generally two kinds of AF methods. The original one uses exact or approximate evolution operators and Simpson's rule for flux quadrature in time, i.e. it does not require time integration methods like Runge-Kutta methods. Exact evolution operators have been studied for linear equations in [7, 15, 14, 39]. Approximate evolution operators have been explored for Burgers' equation [13, 12, 34, 4], the compressible Euler equations in one spatial dimension [13, 25, 4], and hyperbolic balance laws [6, 5], etc. One of the advantages of the AF method over standard finite volume methods is its structure-preserving property. For instance, it preserves the vorticity and stationary states for multi-dimensional acoustic equations [7], and it is naturally well-balanced for acoustics with gravity [6].

Since it may not be convenient to derive exact or approximate evolution operators for nonlinear systems, especially in multi-dimensions, another kind of generalized AF method was presented in [1, 2]. A method of lines was used, where the cell average and point value updates are written in semi-discrete form and advanced in time with time integration methods. In the point values update, the Jacobian matrix is split based on the sign of the eigenvalues (Jacobian splitting (JS)), and upwind-biased stencils are used to compute the approximation of derivatives. There are some deficiencies of the JS when used for the AF methods, e.g., the transonic issue [25] for nonlinear problems, leading to spikes in the cell average. Some remedies are suggested in the literature, e.g., using discontinuous reconstruction [25] or evaluating the upwind direction using more information from the neighbors [4].

Solutions to hyperbolic systems (1.1) often stay in an *admissible state set* $\mathcal{G}$, also called the invariant domain. For instance, the solutions to initial value problems of scalar conservation laws (1.2) satisfy a strict maximum principle (MP) [11], i.e.,

$$\mathcal{G} = \{u \mid m_0 \leqslant u \leqslant M_0\}, \quad m_0 = \min_x u_0(x), \ M_0 = \max_x u_0(x). \tag{1.4}$$

Physically, both the density and pressure in the solutions to the compressible Euler equations (1.3) should stay positive, i.e.,

$$\mathcal{G} = \left\{ \boldsymbol{U} = (\rho, \rho v, E) \ \middle| \ \rho > 0, \ p = (\gamma - 1)\left(E - \frac{(\rho v)^2}{2\rho}\right) > 0 \right\}. \tag{1.5}$$

Throughout this paper, it is assumed that $\mathcal{G}$ is a *convex* set, which is obvious for the scalar case (1.4) and can be verified for the Euler equations (1.5), see e.g. [46]. It is desirable to conceive so-called bound-preserving (BP) methods, i.e., those guaranteeing that the numerical solutions at a later time will stay in $\mathcal{G}$, if the initial numerical solutions belong to $\mathcal{G}$. The BP property of numerical methods is very important for both

theoretical analysis and numerical stability. Many BP methods have been developed in the past few decades, e.g., a series of works by Shu and collaborators [45, 26, 43], a recent general framework on BP methods [42], and the convex limiting approach [17, 22, 29], which can be traced back to the flux-corrected transport (FCT) schemes for scalar conservation laws [10, 20, 32, 30]. The previous studies on the AF methods pay limited attention to high-speed flows, or problems containing strong discontinuities, with some efforts on the limiting for the point value update, see e.g. [4, 8]. However, those limitings are not enough to guarantee the BP property, as shown in our numerical tests. In a very recent paper, the MOOD [9] based stabilization was adopted to achieve the BP property [3] in an a posteriori fashion.

This paper presents a new way for the point value update to cure the transonic issue and develops suitable BP limiting strategies for the AF methods. The main contributions and findings in this work can be summarized as follows.

i). We propose to employ the flux vector splitting (FVS) methods for the point value update to cure the transonic issue, since it borrows information from the neighbors naturally and uniformly. The FVS was originally used to identify the upwind directions, which is simpler and somewhat more efficient than Godunov-type methods for solving hyperbolic systems [38]. In our numerical tests, the FVS is also shown to give better results than the JS, especially the local Lax-Friedrichs (LLF) or Rusanov FVS, in terms of the CFL number and shock-capturing ability. The FVS can also cure some defects in two dimensions observed in the JS, which will be shown in our future companion paper.

ii). We design BP limitings for both the update of the cell average and the point value by blending the high-order AF methods with the first-order LLF method in a convex combination. The convex limiting [17, 22, 29] and the scaling limiter [31] are applied to the cell average and point value updates, respectively. The main idea is to retain as much as possible of the high-order method while guaranteeing the numerical solutions to be BP, and the blending coefficients are computed by enforcing the bounds. We show that using a suitable time step size and BP limitings, the numerical solutions of the BP AF methods satisfy the MP for scalar conservation laws, and give positive density and pressure for the compressible Euler equations.

iii). Several challenging test cases such as the LeBlanc and double rarefaction Riemann problems, the Sedov point blast wave, and blast wave interaction problems are conducted to demonstrate the BP properties and the shock-capturing ability, which are rare in the literature for the AF methods.

The remainder of this paper is structured as follows. Section 2 introduces the AF methods based on the JS or FVS for the point value update, and the power law reconstruction for limiting the derivatives in the point value update. To design BP methods, Section 3 describes our convex limiting approach for the cell average, while Section 4 deals with the limiting for the point value. Some numerical tests are conducted in Section 5 to experimentally demonstrate the accuracy, BP properties, and shock-capturing ability of the methods. Section 6 concludes the paper with final remarks and future directions.

2. 1D active flux methods for hyperbolic conservation laws.

This section presents the 1D semi-discrete AF methods for the hyperbolic conservation laws (1.1), based on the JS [2] or FVS for the point value update. The fully-discrete methods are obtained using Runge-Kutta methods.

Assume that a 1D computational domain is divided into N cells $I_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ with cell centers $x_i = (x_{i-\frac{1}{2}} + x_{i+\frac{1}{2}})/2$ and cell sizes $\Delta x_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$, $i = 1, \cdots, N$.

The degrees of freedom of the AF methods are the approximations to cell averages of the conservative variable as well as point values at the cell interfaces, allowing some freedom in the choice of the point values, e.g. conservative variables, primitive variables, entropy variables, etc. This paper only considers using the conservative variables, and the degrees of freedom are denoted by

$$\overline{\boldsymbol{U}}_i(t) = \frac{1}{\Delta x_i}\int_{I_i} \boldsymbol{U}(x,t)\,\mathrm{d}x, \quad \boldsymbol{U}_{i+\frac{1}{2}}(t) = \boldsymbol{U}(x_{i+\frac{1}{2}},t). \tag{2.1}$$

The cell average is updated by integrating (1.1) over I_i in the following semi-discrete finite volume manner

$$\frac{\mathrm{d}\overline{\boldsymbol{U}}_i}{\mathrm{d}t} = -\frac{1}{\Delta x_i}\left[\boldsymbol{F}(\boldsymbol{U}_{i+\frac{1}{2}}) - \boldsymbol{F}(\boldsymbol{U}_{i-\frac{1}{2}})\right]. \tag{2.2}$$

Thus, the accuracy of (2.2) is determined by the approximation accuracy of the point values. It was so far (e.g. in [2]) considered sufficient to update the point values with any finite-difference-like formula

$$\frac{\mathrm{d}\boldsymbol{U}_{i+\frac{1}{2}}}{\mathrm{d}t} = -\boldsymbol{\mathcal{R}}\left(\boldsymbol{U}_{i+\frac{1}{2}-l_1}(t), \overline{\boldsymbol{U}}_{i+1-l_1}(t), \cdots, \overline{\boldsymbol{U}}_{i+l_2}(t), \boldsymbol{U}_{i+\frac{1}{2}+l_2}(t)\right),\ l_1, l_2 \geqslant 0, \tag{2.3}$$

with $\boldsymbol{\mathcal{R}}$ a consistent approximation of $\partial\boldsymbol{F}/\partial x$ at $x_{i+\frac{1}{2}}$, as long as it gave rise to a stable method. This paper explores further conditions on $\boldsymbol{\mathcal{R}}$ for nonlinear problems.

2.1. Point value update using Jacobian splitting. For smooth solutions, we have an equivalent formulation in the form

$$\frac{\partial \boldsymbol{U}}{\partial t} + \boldsymbol{J}(\boldsymbol{U})\frac{\partial \boldsymbol{U}}{\partial x} = 0, \quad \boldsymbol{J}(\boldsymbol{U}) = \frac{\partial \boldsymbol{F}(\boldsymbol{U})}{\partial \boldsymbol{U}}. \tag{2.4}$$

Inspired by the upwind scheme, (2.4) can be discretized by the JS [1, 2] as follows

$$\frac{\mathrm{d}\boldsymbol{U}_{i+\frac{1}{2}}}{\mathrm{d}t} = -\left[\boldsymbol{J}^+(\boldsymbol{U}_{i+\frac{1}{2}})\boldsymbol{D}^+_{i+\frac{1}{2}}(\boldsymbol{U}) + \boldsymbol{J}^-(\boldsymbol{U}_{i+\frac{1}{2}})\boldsymbol{D}^-_{i+\frac{1}{2}}(\boldsymbol{U})\right], \tag{2.5}$$

where the splitting of the Jacobian matrix $\boldsymbol{J} = \boldsymbol{J}^+ + \boldsymbol{J}^-$ is defined as

$$\begin{aligned}
&\boldsymbol{J}^+ = \boldsymbol{R}\boldsymbol{\Lambda}^+\boldsymbol{R}^{-1}, \quad \boldsymbol{J}^- = \boldsymbol{R}\boldsymbol{\Lambda}^-\boldsymbol{R}^{-1},\\
&\boldsymbol{\Lambda}^+ = \mathrm{diag}\{\max(\lambda_1,0),\ldots,\max(\lambda_m,0)\},\\
&\boldsymbol{\Lambda}^- = \mathrm{diag}\{\min(\lambda_1,0),\ldots,\min(\lambda_m,0)\},
\end{aligned}$$

based on the eigendecomposition $\partial\boldsymbol{F}/\partial\boldsymbol{U} = \boldsymbol{R}\boldsymbol{\Lambda}\boldsymbol{R}^{-1}$, $\boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_1,\ldots,\lambda_m\}$, where $\lambda_1,\cdots,\lambda_m$ are the eigenvalues, with the columns of $\boldsymbol{R}$ the corresponding eigenvectors.

To derive the approximation of the derivatives in (2.5), one can first obtain a high-order reconstruction for $\boldsymbol{U}$ in the upwind cell, and then differentiate the reconstructed polynomial. As an example, a parabolic reconstruction in cell i is

$$\begin{aligned}
\boldsymbol{U}_{\mathtt{para},1}(x) =& -3(2\overline{\boldsymbol{U}}_i - \boldsymbol{U}_{i-\frac{1}{2}} - \boldsymbol{U}_{i+\frac{1}{2}})\frac{x^2}{\Delta x_i^2} + (\boldsymbol{U}_{i+\frac{1}{2}} - \boldsymbol{U}_{i-\frac{1}{2}})\frac{x}{\Delta x_i}\\
&+ \frac{1}{4}(6\overline{\boldsymbol{U}}_i - \boldsymbol{U}_{i-\frac{1}{2}} - \boldsymbol{U}_{i+\frac{1}{2}})
\end{aligned} \tag{2.6}$$

satisfying $\boldsymbol{U}_{\mathtt{para},1}(\pm\Delta x_i/2) = \boldsymbol{U}_{i\pm\frac{1}{2}}$, $\frac{1}{\Delta x_i}\int_{-\Delta x_i/2}^{\Delta x_i/2} \boldsymbol{U}_{\mathtt{para},1}(x)\,\mathrm{d}x = \overline{\boldsymbol{U}}_i$. Then the derivatives are

$$\boldsymbol{D}^+_{i+\frac{1}{2}}(\boldsymbol{U}) = \boldsymbol{U}'_{\mathtt{para},1}(\Delta x_i/2) = \frac{1}{\Delta x_i}\left(2\boldsymbol{U}_{i-\frac{1}{2}} - 6\overline{\boldsymbol{U}}_i + 4\boldsymbol{U}_{i+\frac{1}{2}}\right), \tag{2.7a}$$

$$\boldsymbol{D}^-_{i+\frac{1}{2}}(\boldsymbol{U}) = \frac{1}{\Delta x_{i+1}}\left(-4\boldsymbol{U}_{i+\frac{1}{2}} + 6\overline{\boldsymbol{U}}_{i+1} - 2\boldsymbol{U}_{i+\frac{3}{2}}\right). \tag{2.7b}$$

They are third-order accurate. Higher-order extensions can be obtained by higher-order finite difference formulae using a larger spatial stencil, see [2] for examples.

2.2. Point value update using flux vector splitting. One of the deficiencies of using the JS is the transonic issue that appears for nonlinear problems, as observed in [25, 4] and described in more detail next. Consider Example 5.2, where we solve Burgers' equation with a square wave as the initial data. Figure 3 shows the cell averages and point values based on the JS with 200 cells, as well as the reference solution. The numerical solution based on the JS without limiting gives a spike at the initial discontinuity $x = 0.2$, which grows linearly in time. The reason for this behaviour is the inaccurate estimation of the upwind direction at the cell interface. In this example, there are two successive point values with different initial data near the initial discontinuity, denoted by $u_{i-\frac{1}{2}} = 2$, $u_{i+\frac{1}{2}} = -1$, respectively. At the cell interface $x_{i-\frac{1}{2}}$ or $x_{i+\frac{1}{2}}$, the upwind discretization in (2.7) only uses the data from the left or right, leading to zero derivatives, thus the point values $u_{i-\frac{1}{2}}$ and $u_{i+\frac{1}{2}}$ stay unchanged. However, according to the update of the cell average (2.2), $\bar{u}_i$ increases gradually (which is the observed spike). This deficiency cannot be eliminated by limitings, as one observes from Figure 3. Some remedies have been proposed, such as using discontinuous reconstruction [25] and an "entropy fix" that evaluates the upwind direction not only at the corresponding cell interface but also with values from its neighbors [4].

In this paper, we propose to use the FVS for the point value update, which borrows the information from the neighbors naturally, still based on the continuous reconstruction, and can eliminate the generation of the spike effectively, as shown in Figure 4. The FVS for the point value update reads

$$\frac{\mathrm{d}\boldsymbol{U}_{i+\frac{1}{2}}}{\mathrm{d}t} = -\left[\widetilde{\boldsymbol{D}}^+\boldsymbol{F}^+(\boldsymbol{U}) + \widetilde{\boldsymbol{D}}^-\boldsymbol{F}^-(\boldsymbol{U})\right]_{i+\frac{1}{2}}, \tag{2.8}$$

where the flux $\boldsymbol{F}$ is split into the positive and negative parts $\boldsymbol{F} = \boldsymbol{F}^+ + \boldsymbol{F}^-$ satisfying

$$\lambda\left(\frac{\partial \boldsymbol{F}^+}{\partial \boldsymbol{U}}\right) \geqslant 0, \quad \lambda\left(\frac{\partial \boldsymbol{F}^-}{\partial \boldsymbol{U}}\right) \leqslant 0, \tag{2.9}$$

i.e., all the eigenvalues of $\frac{\partial \boldsymbol{F}^+}{\partial \boldsymbol{U}}$ and $\frac{\partial \boldsymbol{F}^-}{\partial \boldsymbol{U}}$ are non-negative and non-positive, respectively. Different FVS can be adopted as long as they satisfy the constraint (2.9), to be discussed later. Finite difference formulae to approximate the flux derivatives are obtained similarly to the computation of the derivatives in the JS. A parabolic reconstruction of the flux can be obtained based on the three flux values as follows

$$\boldsymbol{F}_{\mathtt{para},2}(x) = 2(\boldsymbol{F}_{i-\frac{1}{2}} - 2\boldsymbol{F}_i + \boldsymbol{F}_{i+\frac{1}{2}})\frac{x^2}{\Delta x_i^2} + (\boldsymbol{F}_{i+\frac{1}{2}} - \boldsymbol{F}_{i-\frac{1}{2}})\frac{x}{\Delta x_i} + \boldsymbol{F}_i,$$

satisfying $\boldsymbol{F}_{\mathtt{para},2}(\pm\Delta x_i/2) = \boldsymbol{F}_{i\pm\frac{1}{2}}$, $\boldsymbol{F}_{\mathtt{para},2}(0) = \boldsymbol{F}_i$, with $\boldsymbol{F}_{i\pm\frac{1}{2}} = \boldsymbol{F}(\boldsymbol{U}_{i\pm\frac{1}{2}})$, and the cell-centered point value $\boldsymbol{F}_i = \boldsymbol{F}(\boldsymbol{U}_i)$ is obtained by evaluating the reconstruction of $\boldsymbol{U}$, i.e. according to Simpson's rule $\boldsymbol{U}_i = (-\boldsymbol{U}_{i-\frac{1}{2}} + 6\overline{\boldsymbol{U}}_i - \boldsymbol{U}_{i+\frac{1}{2}})/4$. Then the derivatives are

$$\left(\widetilde{\boldsymbol{D}}^+\boldsymbol{F}^+\right)_{i+\frac{1}{2}} = \boldsymbol{F}'_{\mathtt{para},2}(\Delta x_i/2) = \frac{1}{\Delta x_i}\left(\boldsymbol{F}_{i-\frac{1}{2}} - 4\boldsymbol{F}_i + 3\boldsymbol{F}_{i+\frac{1}{2}}\right), \tag{2.10a}$$

$$\left(\widetilde{\boldsymbol{D}}^-\boldsymbol{F}^-\right)_{i+\frac{1}{2}} = \frac{1}{\Delta x_{i+1}}\left(-3\boldsymbol{F}_{i+\frac{1}{2}} + 4\boldsymbol{F}_{i+1} - \boldsymbol{F}_{i+\frac{3}{2}}\right). \tag{2.10b}$$

These finite differences are third-order accurate. While the reconstructions of both $\boldsymbol{U}$ and $\boldsymbol{F}$ are parabolic, the coefficients in the formula (2.10) differ from (2.7) because (2.10) uses the cell-centered value rather than the cell average. Our numerical tests in Section 5 show that the AF methods based on the FVS generally give better results than the JS.

2.2.1. Local Lax-Friedrichs flux vector splitting. The first FVS we consider is the LLF FVS, defined as

$$\boldsymbol{F}^\pm = \frac{1}{2}(\boldsymbol{F}(\boldsymbol{U}) \pm \alpha\boldsymbol{U}),$$

where the choice of α should fulfill (2.9) across the spatial stencil. In our implementation, it is determined by

$$\alpha_{i+\frac{1}{2}} = \max_{r,\ell}\left\{|\lambda_\ell(\boldsymbol{U}_r)|\right\},\ r \in \left\{i-\frac{1}{2}, i, i+\frac{1}{2}, i+1, u+\frac{3}{2}\right\},\ \ell = 1,\cdots,m. \tag{2.11}$$

One can also choose α to be the maximal absolute value of the eigenvalues in the whole domain, corresponding to the (global) LF splitting. Note, however, that a larger α generally leads to a smaller time step size and more dissipation.

2.2.2. Upwind flux vector splitting. One can also split the Jacobian matrix based on each characteristic field,

$$\boldsymbol{F}^\pm = \frac{1}{2}(\boldsymbol{F}(\boldsymbol{U}) \pm |\boldsymbol{J}|\boldsymbol{U}), \quad |\boldsymbol{J}| = \boldsymbol{R}(\boldsymbol{\Lambda}^+ - \boldsymbol{\Lambda}^-)\boldsymbol{R}^{-1}. \tag{2.12}$$

For linear systems, one has $\boldsymbol{F} = \boldsymbol{J}\boldsymbol{U}$, so (2.12) reduces to the JS. To be specific,

$$\boldsymbol{F}^\pm = \frac{1}{2}(\boldsymbol{J} \pm |\boldsymbol{J}|)\boldsymbol{U} = \boldsymbol{R}\boldsymbol{\Lambda}^\pm\boldsymbol{R}^{-1}\boldsymbol{U} = \boldsymbol{J}^\pm\boldsymbol{U},$$

with $\boldsymbol{J}^\pm$ a constant matrix so that $\widetilde{\boldsymbol{D}}^\pm\boldsymbol{F}^\pm(\boldsymbol{U}) = \boldsymbol{J}^\pm\widetilde{\boldsymbol{D}}^\pm\boldsymbol{U}$, which is the same as $\boldsymbol{J}^\pm\boldsymbol{D}^\pm\boldsymbol{U}$ if $\boldsymbol{D}^+$ and $\widetilde{\boldsymbol{D}}^+$ are derived from the same reconstructed polynomial. In other words, the AF methods using this FVS enjoy the same properties as the original JS-based AF methods for linear systems.

Such an FVS is also known as the Steger-Warming (SW) FVS [36] for the Euler equations (1.3), since the "homogeneity property" holds [38], i.e., $\boldsymbol{F} = \boldsymbol{J}\boldsymbol{U}$. One can write down the SW FVS explicitly

$$\boldsymbol{F}^\pm = \begin{bmatrix} \frac{\rho}{2\gamma}\alpha^\pm \\ \frac{\rho}{2\gamma}\left(\alpha^\pm v + a(\lambda_2^\pm - \lambda_3^\pm)\right) \\ \frac{\rho}{2\gamma}\left(\frac{1}{2}\alpha^\pm v^2 + av(\lambda_2^\pm - \lambda_3^\pm) + \frac{a^2}{\gamma-1}(\lambda_2^\pm + \lambda_3^\pm)\right) \end{bmatrix},$$

where $\lambda_1 = v,\ \lambda_2 = v + a,\ \lambda_3 = v - a,\ \alpha^\pm = 2(\gamma-1)\lambda_1^\pm + \lambda_2^\pm + \lambda_3^\pm$, and $a = \sqrt{\gamma p/\rho}$ is the sound speed.

2.2.3. Van Leer-Hänel flux vector splitting for the Euler equations. Another popular FVS for the Euler equations was proposed by Van Leer [40], and improved by [23]. The flux can be split based on the Mach number $M = v/a$ as

$$\boldsymbol{F} = \begin{bmatrix} \rho a M \\ \rho a^2(M^2 + \frac{1}{\gamma}) \\ \rho a^3 M(\frac{1}{2}M^2 + \frac{1}{\gamma-1}) \end{bmatrix} = \boldsymbol{F}^+ + \boldsymbol{F}^-, \quad \boldsymbol{F}^\pm = \begin{bmatrix} \pm\frac{1}{4}\rho a(M \pm 1)^2 \\ \pm\frac{1}{4}\rho a(M \pm 1)^2 v + p^\pm \\ \pm\frac{1}{4}\rho a(M \pm 1)^2 H \end{bmatrix},$$

with the enthalpy $H = (E + p)/\rho$, and the pressure splitting $p^\pm = \frac{1}{2}(1 \pm \gamma M)p$. This FVS gives a quadratic differentiable splitting with respect to the Mach number.

2.3. 1D power law reconstruction for point value update. When the numerical solutions contain discontinuities, the computation of the derivatives (2.7) or (2.10) based on the parabolic reconstructions may cause oscillations. Thus, it is reasonable to seek finite difference approximations based on differentiating a modified reconstruction with improved monotonicity properties. This section only considers the scalar case and can be extended to systems of equations in a component-wise fashion.

The power law reconstruction proposed in [4] can be used to replace the original parabolic reconstruction to achieve monotonicity on some occasions. It is shown in Theorem 5 in [4] that the extremum is not avoidable in the cell $I_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ for continuous reconstructions if the cell average lies outside the range of the point values $(\bar{u}_i - u_{i-\frac{1}{2}})(u_{i+\frac{1}{2}} - \bar{u}_i) < 0$. The parabola is monotone, and thus no action is required when $(2u_{i-\frac{1}{2}} + u_{i+\frac{1}{2}})/3 < \bar{u}_i < (u_{i-\frac{1}{2}} + 2u_{i+\frac{1}{2}})/3$ or $(2u_{i-\frac{1}{2}} + u_{i+\frac{1}{2}})/3 > \bar{u}_i > (u_{i-\frac{1}{2}} + 2u_{i+\frac{1}{2}})/3$. Upon defining $r = \dfrac{u_{i+1/2} - \bar{u}_i}{\bar{u}_i - u_{i-1/2}}$, one can equivalently express that the parabola is monotone when $1/2 < r < 2$. In both these cases, the parabolic reconstruction is used, and the derivatives are obtained by (2.7) or (2.10). Otherwise, the following power law reconstruction is used.

PROPOSITION 2.1 (Barsukow [4]). The power law reconstruction

$$\begin{cases} u_{\mathtt{pwl},1}(x) = u_{i-\frac{1}{2}} + (u_{i+\frac{1}{2}} - u_{i-\frac{1}{2}}) \left(\dfrac{x - x_i}{\Delta x_i} + \dfrac{1}{2} \right)^r, & \text{if } \ r > 2 \\ u_{\mathtt{pwl},2}(x) = u_{i+\frac{1}{2}} - (u_{i+\frac{1}{2}} - u_{i-\frac{1}{2}}) \left(\dfrac{1}{2} - \dfrac{x - x_i}{\Delta x_i} \right)^{1/r}, & \text{if } \ 0 < r < 1/2 \end{cases} \tag{2.13}$$

is monotone and satisfies

$$u_{\mathtt{pwl},l}(x_{i-\frac{1}{2}}) = u_{i-\frac{1}{2}}, \ u_{\mathtt{pwl},l}(x_{i+\frac{1}{2}}) = u_{i+\frac{1}{2}}, \ \frac{1}{\Delta x_i} \int_{I_i} u_{\mathtt{pwl},l}(x) \, \mathrm{d}x = \bar{u}_i, \ l = 1, 2.$$

A comparison between the parabolic reconstruction (2.6) and power law reconstruction (2.13) is given in Figure 1 with point values fixed as -1 and 1 at the interfaces, and different cell averages $\{-1.1, -0.8, -1/3, 0.1, 1/3, 0.8, 1.1\}$. One can observe monotone profiles for the power law reconstruction when the cell average lies between the two point values. Based on (2.13), the derivatives can be computed directly

$$\begin{cases} u'_{\mathtt{pwl},1}(x) = \dfrac{u_{i+\frac{1}{2}} - u_{i-\frac{1}{2}}}{\Delta x_i} r \left(\dfrac{x - x_i}{\Delta x_i} + \dfrac{1}{2} \right)^{r-1}, & \text{if } \ r > 2, \\ u'_{\mathtt{pwl},2}(x) = \dfrac{u_{i+\frac{1}{2}} - u_{i-\frac{1}{2}}}{\Delta x_i} \dfrac{1}{r} \left(\dfrac{1}{2} - \dfrac{x - x_i}{\Delta x_i} \right)^{1/r-1}, & \text{if } \ 0 < r < 1/2. \end{cases}$$

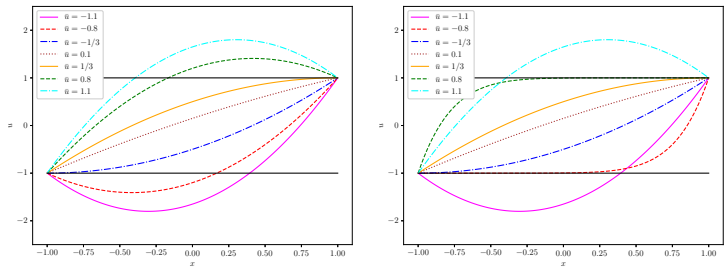

Fig. 1: The parabolic (2.6) and power law reconstruction (2.13) obtained with different cell averages $\{-1.1, -0.8, -1/3,\ 0.1,\ 1/3,\ 0.8,\ 1.1\}$, and fixed point values as -1 and 1 at the left and right interfaces.

At the left interface, the derivative is

$$\begin{cases} u'_{\texttt{pwl},1}(x^+_{i-\frac{1}{2}}) = 0, & \text{if } \ r > 2, \\ u'_{\texttt{pwl},2}(x^+_{i-\frac{1}{2}}) = \dfrac{u_{i+\frac{1}{2}} - u_{i-\frac{1}{2}}}{\Delta x_i}\dfrac{1}{r}, & \text{if } \ 0 < r < 1/2, \end{cases} \tag{2.14}$$

and at the right interface, the derivative is

$$\begin{cases} u'_{\texttt{pwl},1}(x^-_{i+\frac{1}{2}}) = \dfrac{u_{i+\frac{1}{2}} - u_{i-\frac{1}{2}}}{\Delta x_i} r, & \text{if } \ r > 2, \\ u'_{\texttt{pwl},2}(x^-_{i+\frac{1}{2}}) = 0, & \text{if } \ 0 < r < 1/2. \end{cases} \tag{2.15}$$

To avoid computational issues, when $r \notin [1/50, 50]$, the parabolic reconstruction is adopted directly.

For the FVS, as the cell average of the flux can be obtained through Simpson's rule, $\bar{f}_i = (f_{i-\frac{1}{2}} + 4f_i + f_{i+\frac{1}{2}})/6$, the flux derivatives can be computed by (2.14)-(2.15).

Remark 2.2. In [2], it is mentioned that if the signs of the derivatives of the parabolic reconstruction and the first-order reconstruction are the same, then the parabolic reconstruction is adopted. This strategy is not employed in this paper as the numerical results may be worse.

2.4. Time discretization. The fully-discrete scheme is obtained by using the SSP-RK3 method [16]

$$\begin{aligned} \boldsymbol{U}^* &= \boldsymbol{U}^n + \Delta t^n \boldsymbol{L}\left(\boldsymbol{U}^n\right), \\ \boldsymbol{U}^{**} &= \frac{3}{4}\boldsymbol{U}^n + \frac{1}{4}\left(\boldsymbol{U}^* + \Delta t^n \boldsymbol{L}\left(\boldsymbol{U}^*\right)\right), \\ \boldsymbol{U}^{n+1} &= \frac{1}{3}\boldsymbol{U}^n + \frac{2}{3}\left(\boldsymbol{U}^{**} + \Delta t^n \boldsymbol{L}\left(\boldsymbol{U}^{**}\right)\right), \end{aligned} \tag{2.16}$$

where $\boldsymbol{L}$ is the right-hand side of the semi-discrete schemes (2.2) or (2.3). The time step size is determined by the usual CFL condition

$$\Delta t^n = \frac{C_{\texttt{CFL}}}{\max\limits_{i,\ell}\{\lambda_\ell(\overline{\boldsymbol{U}}_i)/\Delta x_i\}}. \tag{2.17}$$

3. Convex limiting for the cell average. Although the power law reconstruction [4] has been shown to effectively reduce spurious oscillations, the numerical solutions may still violate certain bounds, e.g., the appearance of negative density or pressure, leading to unphysical solutions or even causing the simulations to blow up. Since the degrees of freedom in the AF methods include both cell averages and point values, it is necessary to design suitable BP limitings for both of them to achieve the BP property. The limiting for the cell average has not been addressed much in the literature, except for a very recent work [3].

DEFINITION 3.1. *An AF method is called* bound-preserving *(BP) if starting from cell averages and point values in the admissible state set $\mathcal{G}$, the cell averages and point values remain in $\mathcal{G}$ at the next time step.*

This section presents a convex limiting approach to achieve the BP property of the cell average update. The basic idea of the convex limiting approaches [17, 22, 29] is to enforce the preservation of local and global bounds by constraining individual numerical fluxes. The BP or invariant domain-preserving (IDP) properties of flux-limited approximations are shown using representations in terms of intermediate states that stay in convex admissible state sets [17, 21]. The low-order scheme is chosen as the first-order LLF scheme

$$\overline{\boldsymbol{U}}_i^{\mathrm{L}} = \overline{\boldsymbol{U}}_i^n - \mu_i \left(\widehat{\boldsymbol{F}}_{i+\frac{1}{2}}^{\mathrm{L}} - \widehat{\boldsymbol{F}}_{i-\frac{1}{2}}^{\mathrm{L}} \right), \quad \mu_i = \Delta t^n / \Delta x_i,$$

$$\widehat{\boldsymbol{F}}_{i+\frac{1}{2}}^{\mathrm{L}} = \frac{1}{2} \left(\boldsymbol{F}(\overline{\boldsymbol{U}}_i^n) + \boldsymbol{F}(\overline{\boldsymbol{U}}_{i+1}^n) \right) - \frac{\alpha_{i+\frac{1}{2}}}{2} \left(\overline{\boldsymbol{U}}_{i+1}^n - \overline{\boldsymbol{U}}_i^n \right),$$

where $\alpha_{i+\frac{1}{2}}$ is an *upper bound* for the maximum wave speed of the Riemann problem with the initial data $\boldsymbol{U}_i, \boldsymbol{U}_{i+1}$, whose estimation for scalar conservation laws and the Euler equations can be found in [19] and [18], respectively. Note that here $\alpha_{i+\frac{1}{2}}$ is not the same as the one in the LLF FVS (2.11). Following [19], the first-order LLF scheme can be rewritten as

$$\overline{\boldsymbol{U}}_i^{\mathrm{L}} = \left[1 - \mu_i \left(\alpha_{i-\frac{1}{2}} + \alpha_{i+\frac{1}{2}} \right) \right] \overline{\boldsymbol{U}}_i^n + \mu_i \alpha_{i-\frac{1}{2}} \widetilde{\boldsymbol{U}}_{i-\frac{1}{2}} + \mu_i \alpha_{i+\frac{1}{2}} \widetilde{\boldsymbol{U}}_{i+\frac{1}{2}}, \tag{3.1}$$

with the intermediate states defined as

$$\begin{aligned} \widetilde{\boldsymbol{U}}_{i-\frac{1}{2}} &:= \frac{1}{2} \left(\overline{\boldsymbol{U}}_{i-1}^n + \overline{\boldsymbol{U}}_i^n \right) + \frac{1}{2\alpha_{i-\frac{1}{2}}} \left[\boldsymbol{F}(\overline{\boldsymbol{U}}_{i-1}^n) - \boldsymbol{F}(\overline{\boldsymbol{U}}_i^n) \right], \\ \widetilde{\boldsymbol{U}}_{i+\frac{1}{2}} &:= \frac{1}{2} \left(\overline{\boldsymbol{U}}_i^n + \overline{\boldsymbol{U}}_{i+1}^n \right) + \frac{1}{2\alpha_{i+\frac{1}{2}}} \left[\boldsymbol{F}(\overline{\boldsymbol{U}}_i^n) - \boldsymbol{F}(\overline{\boldsymbol{U}}_{i+1}^n) \right]. \end{aligned} \tag{3.2}$$

Remark 3.2. As $\alpha_{i+\frac{1}{2}}$ is chosen to be larger than the leftmost and rightmost wave speed, the intermediate state defined in (3.2) is indeed an average of the exact Riemann solution [19], thus it belongs to $\mathcal{G}$. For systems, it is also the intermediate state of the HLL solver [24]. Moreover, the intermediate state (3.2) preserves all *convex invariants* (e.g., density and pressure positivity, and minimum entropy principle for the Euler equations) of initial value problems for hyperbolic systems [19].

LEMMA 3.3 (Guermond and Popov [19]). If the time step size Δt^n satisfies

$$\Delta t^n \leqslant \frac{\Delta x_i}{\alpha_{i-\frac{1}{2}} + \alpha_{i+\frac{1}{2}}}, \tag{3.3}$$

then (3.1) is a convex combination, and the first-order LLF scheme (3.1) is BP.

The proof relies on the fact that the intermediate state (3.2) stays in the admissible state set $\mathcal{G}$ and the convexity of $\mathcal{G}$.

Upon defining the anti-diffusive flux $\Delta\widehat{\boldsymbol{F}}_{i\pm\frac{1}{2}} := \widehat{\boldsymbol{F}}^{\mathtt{H}}_{i\pm\frac{1}{2}} - \widehat{\boldsymbol{F}}^{\mathtt{L}}_{i\pm\frac{1}{2}}$ with $\widehat{\boldsymbol{F}}^{\mathtt{H}}_{i\pm\frac{1}{2}} := \boldsymbol{F}(\boldsymbol{U}_{i\pm\frac{1}{2}})$, a forward-Euler step applied to the semi-discrete high-order scheme for the cell average (2.2) can be written as

$$\overline{\boldsymbol{U}}^{\mathtt{H}}_i = \overline{\boldsymbol{U}}^n_i - \mu_i(\widehat{\boldsymbol{F}}^{\mathtt{H}}_{i+\frac{1}{2}} - \widehat{\boldsymbol{F}}^{\mathtt{H}}_{i-\frac{1}{2}}) = \overline{\boldsymbol{U}}^n_i - \mu_i(\widehat{\boldsymbol{F}}^{\mathtt{L}}_{i+\frac{1}{2}} - \widehat{\boldsymbol{F}}^{\mathtt{L}}_{i-\frac{1}{2}}) - \mu_i(\Delta\widehat{\boldsymbol{F}}_{i+\frac{1}{2}} - \Delta\widehat{\boldsymbol{F}}_{i-\frac{1}{2}}) \tag{3.4}$$

$$=: \left[1 - \mu_i\left(\alpha_{i-\frac{1}{2}} + \alpha_{i+\frac{1}{2}}\right)\right]\overline{\boldsymbol{U}}^n_i + \mu_i\alpha_{i-\frac{1}{2}}\widetilde{\boldsymbol{U}}^{\mathtt{H}}_{i-\frac{1}{2}} + \mu_i\alpha_{i+\frac{1}{2}}\widetilde{\boldsymbol{U}}^{\mathtt{H}}_{i+\frac{1}{2}},$$

$$\widetilde{\boldsymbol{U}}^{\mathtt{H}}_{i-\frac{1}{2}} := \left(\widetilde{\boldsymbol{U}}_{i-\frac{1}{2}} + \frac{\Delta\widehat{\boldsymbol{F}}_{i-\frac{1}{2}}}{\alpha_{i-\frac{1}{2}}}\right), \quad \widetilde{\boldsymbol{U}}^{\mathtt{H}}_{i+\frac{1}{2}} := \left(\widetilde{\boldsymbol{U}}_{i+\frac{1}{2}} - \frac{\Delta\widehat{\boldsymbol{F}}_{i+\frac{1}{2}}}{\alpha_{i+\frac{1}{2}}}\right).$$

With the low-order scheme (3.1) and high-order scheme (3.4) having the same form one can now define the limited scheme for the cell average as

$$\overline{\boldsymbol{U}}^{\mathtt{Lim}}_i = \left[1 - \mu_i\left(\alpha_{i-\frac{1}{2}} + \alpha_{i+\frac{1}{2}}\right)\right]\overline{\boldsymbol{U}}^n_i + \mu_i\alpha_{i-\frac{1}{2}}\widetilde{\boldsymbol{U}}^{\mathtt{Lim},+}_{i-\frac{1}{2}} + \mu_i\alpha_{i+\frac{1}{2}}\widetilde{\boldsymbol{U}}^{\mathtt{Lim},-}_{i+\frac{1}{2}}, \tag{3.5}$$

with the limited intermediate states

$$\widetilde{\boldsymbol{U}}^{\mathtt{Lim},+}_{i-\frac{1}{2}} = \widetilde{\boldsymbol{U}}_{i-\frac{1}{2}} + \frac{\Delta\widehat{\boldsymbol{F}}^{\mathtt{Lim}}_{i-\frac{1}{2}}}{\alpha_{i-\frac{1}{2}}} := \widetilde{\boldsymbol{U}}_{i-\frac{1}{2}} + \frac{\theta_{i-\frac{1}{2}}\Delta\widehat{\boldsymbol{F}}_{i-\frac{1}{2}}}{\alpha_{i-\frac{1}{2}}},$$

$$\widetilde{\boldsymbol{U}}^{\mathtt{Lim},-}_{i+\frac{1}{2}} = \widetilde{\boldsymbol{U}}_{i+\frac{1}{2}} - \frac{\Delta\widehat{\boldsymbol{F}}^{\mathtt{Lim}}_{i+\frac{1}{2}}}{\alpha_{i+\frac{1}{2}}} := \widetilde{\boldsymbol{U}}_{i+\frac{1}{2}} - \frac{\theta_{i+\frac{1}{2}}\Delta\widehat{\boldsymbol{F}}_{i+\frac{1}{2}}}{\alpha_{i+\frac{1}{2}}},$$

where the coefficients $\theta_{i\pm\frac{1}{2}} \in [0,1]$.

PROPOSITION 3.4. If the cell average at the last time step $\overline{\boldsymbol{U}}^n_i$ and the limited intermediate states $\widetilde{\boldsymbol{U}}^{\mathtt{Lim},\mp}_{i\pm\frac{1}{2}}$ belong to the admissible state set $\mathcal{G}$, then the limited average update (3.5) is BP, i.e., $\overline{\boldsymbol{U}}^{\mathtt{Lim}}_i \in \mathcal{G}$, under the CFL condition (3.3). If the SSP-RK3 (2.16) is used for the time integration, the high-order scheme is also BP.

Proof. Under the constraint (3.3), the limited cell average update $\overline{\boldsymbol{U}}^{\mathtt{Lim}}_i$ is a convex combination of $\overline{\boldsymbol{U}}^n_i$ and $\widetilde{\boldsymbol{U}}^{\mathtt{Lim},\mp}_{i\pm\frac{1}{2}}$, thus it belongs to $\mathcal{G}$ due to the convexity of $\mathcal{G}$. Because the SSP-RK3 is a convex combination of forward-Euler stages, the high-order scheme equipped with the SSP-RK3 is also BP according to the convexity. ∎

Remark 3.5. The scheme (3.5) is conservative as it amounts to using the numerical flux $\widehat{\boldsymbol{F}}^{\mathtt{L}}_{i+\frac{1}{2}} + \theta_{i+\frac{1}{2}}\Delta\widehat{\boldsymbol{F}}_{i+\frac{1}{2}} = \theta_{i+\frac{1}{2}}\widehat{\boldsymbol{F}}^{\mathtt{H}}_{i+\frac{1}{2}} + (1-\theta_{i+\frac{1}{2}})\widehat{\boldsymbol{F}}^{\mathtt{L}}_{i+\frac{1}{2}}$, which is a convex combination of the high-order and low-order fluxes.

Remark 3.6. It should be noted that the time step size (3.3) is determined based on the solutions at t^n. If the constraint is not satisfied at the later stage of the SSP-RK3, the BP property may not be achieved because (3.5) is no longer a convex combination. In our implementation, we start from the usual CFL condition (2.17). Then, if the high-order AF states need BP limitings and (3.2) is not BP or (3.3) is not satisfied, the numerical solutions are set back to the last time step, and we rerun with a halved time step size until (3.2) is BP and the constraint (3.3) is satisfied. This is also a typical implementation to save computational costs in other BP methods.

The remaining task is to determine the coefficients at each interface $\theta_{i\pm\frac{1}{2}}$ such that $\widetilde{\boldsymbol{U}}^{\mathtt{Lim},\mp}_{i\pm\frac{1}{2}} \in \mathcal{G}$ and stay as close as possible to the high-order states $\widetilde{\boldsymbol{U}}^{\mathtt{H}}_{i\pm\frac{1}{2}}$, i.e., the goal is to find the largest $\theta_{i\pm\frac{1}{2}} \in [0,1]$ such that $\widetilde{\boldsymbol{U}}^{\mathtt{Lim},\mp}_{i\pm\frac{1}{2}} \in \mathcal{G}$.

3.1. Application to scalar conservation laws. This section is devoted to applying the convex limiting approach to scalar conservation laws (1.2), such that the numerical solutions satisfy the global or local MP. For the global MP, the blending coefficient $\theta_{i+\frac{1}{2}} \in [0,1]$ should be chosen such that $m_0 \leqslant \tilde{u}^{\mathtt{Lim},\pm}_{i+\frac{1}{2}} \leqslant M_0$, with m_0, M_0 defined in (1.4), which gives

$$\theta_{i+\frac{1}{2}} = \begin{cases} \min\left\{1, \frac{\alpha_{i+\frac{1}{2}}(\tilde{u}_{i+\frac{1}{2}}-m_0)}{\Delta\hat{f}_{i+\frac{1}{2}}}, \frac{\alpha_{i+\frac{1}{2}}(M_0-\tilde{u}_{i+\frac{1}{2}})}{\Delta\hat{f}_{i+\frac{1}{2}}}\right\}, & \text{if } \Delta\hat{f}_{i+\frac{1}{2}} > 0, \\ \min\left\{1, \frac{\alpha_{i+\frac{1}{2}}(m_0-\tilde{u}_{i+\frac{1}{2}})}{\Delta\hat{f}_{i+\frac{1}{2}}}, \frac{\alpha_{i+\frac{1}{2}}(\tilde{u}_{i+\frac{1}{2}}-M_0)}{\Delta\hat{f}_{i+\frac{1}{2}}}\right\}, & \text{if } \Delta\hat{f}_{i+\frac{1}{2}} < 0. \end{cases}$$

To avoid a small denominator, the limited anti-diffusive flux can be obtained directly,

$$\Delta\hat{f}^{\mathtt{Lim}}_{i+\frac{1}{2}} = \begin{cases} \min\left\{\Delta\hat{f}_{i+\frac{1}{2}},\ \alpha_{i+\frac{1}{2}}(\tilde{u}_{i+\frac{1}{2}}-m_0),\ \alpha_{i+\frac{1}{2}}(M_0-\tilde{u}_{i+\frac{1}{2}})\right\}, & \text{if } \Delta\hat{f}_{i+\frac{1}{2}} \geqslant 0, \\ \max\left\{\Delta\hat{f}_{i+\frac{1}{2}},\ \alpha_{i+\frac{1}{2}}(m_0-\tilde{u}_{i+\frac{1}{2}}),\ \alpha_{i+\frac{1}{2}}(\tilde{u}_{i+\frac{1}{2}}-M_0)\right\}, & \text{otherwise.} \end{cases}$$

On the other hand, one can also enforce the local MP $u_i^{\min} \leqslant \tilde{u}^{\mathtt{Lim},-}_{i+\frac{1}{2}} \leqslant u_i^{\max}$, $u_{i+1}^{\min} \leqslant \tilde{u}^{\mathtt{Lim},+}_{i+\frac{1}{2}} \leqslant u_{i+1}^{\max}$, which helps to suppress spurious oscillations and improve shock-capturing ability. The corresponding limited anti-diffusive flux is

$$\Delta\hat{f}^{\mathtt{Lim}}_{i+\frac{1}{2}} = \begin{cases} \min\left\{\Delta\hat{f}_{i+\frac{1}{2}},\ \alpha_{i+\frac{1}{2}}(\tilde{u}_{i+\frac{1}{2}}-u_i^{\min}),\ \alpha_{i+\frac{1}{2}}(u_{i+1}^{\max}-\tilde{u}_{i+\frac{1}{2}})\right\}, & \text{if } \Delta\hat{f}_{i+\frac{1}{2}} \geqslant 0, \\ \max\left\{\Delta\hat{f}_{i+\frac{1}{2}},\ \alpha_{i+\frac{1}{2}}(u_{i+1}^{\min}-\tilde{u}_{i+\frac{1}{2}}),\ \alpha_{i+\frac{1}{2}}(\tilde{u}_{i+\frac{1}{2}}-u_i^{\max})\right\}, & \text{otherwise.} \end{cases}$$

The choice of the local bounds can be based on the intermediate states

$$u_i^{\min} = \min\left\{\bar{u}_i^n,\ \tilde{u}_{i-\frac{1}{2}},\ \tilde{u}_{i+\frac{1}{2}}\right\},\ u_i^{\max} = \max\left\{\bar{u}_i^n,\ \tilde{u}_{i-\frac{1}{2}},\ \tilde{u}_{i+\frac{1}{2}}\right\}.$$

Finally, the numerical flux is

$$\hat{f}^{\mathtt{Lim}}_{i+\frac{1}{2}} = \hat{f}^{\mathtt{L}}_{i+\frac{1}{2}} + \Delta\hat{f}^{\mathtt{Lim}}_{i+\frac{1}{2}}. \tag{3.6}$$

3.2. Application to the compressible Euler equations. This section aims at enforcing the strict positivity of density and pressure, i.e., $\rho > \varepsilon,\ p > \varepsilon$, with ε a small positive number close to zero, chosen as 10^{-13} in our numerical tests.

3.2.1. Positivity of density. The first step is to impose the density positivity $\widetilde{\boldsymbol{U}}^{\mathtt{Lim},\pm,\rho}_{i+\frac{1}{2}} > \varepsilon$, where $\boldsymbol{U}^{*,\rho}$ denotes the density component of $\boldsymbol{U}^*$. The corresponding density component of the limited anti-diffusive flux is

$$\Delta\widehat{\boldsymbol{F}}^{\mathtt{Lim},*,\rho}_{i+\frac{1}{2}} = \begin{cases} \min\left\{\Delta\widehat{\boldsymbol{F}}^{\rho}_{i+\frac{1}{2}},\ \alpha_{i+\frac{1}{2}}\left(\widetilde{\boldsymbol{U}}^{\rho}_{i+\frac{1}{2}}-\varepsilon\right)\right\}, & \text{if } \Delta\widehat{\boldsymbol{F}}^{\rho}_{i+\frac{1}{2}} \geqslant 0, \\ \max\left\{\Delta\widehat{\boldsymbol{F}}^{\rho}_{i+\frac{1}{2}},\ \alpha_{i+\frac{1}{2}}\left(\varepsilon-\widetilde{\boldsymbol{U}}^{\rho}_{i+\frac{1}{2}}\right)\right\}, & \text{otherwise.} \end{cases}$$

Then the density component of the limited numerical flux is $\widehat{\boldsymbol{F}}^{\mathtt{Lim},*,\rho}_{i+\frac{1}{2}} = \widehat{\boldsymbol{F}}^{\mathtt{L},\rho}_{i+\frac{1}{2}} + \Delta\widehat{\boldsymbol{F}}^{\mathtt{Lim},*,\rho}_{i+\frac{1}{2}}$, with the other components remaining the same as $\widehat{\boldsymbol{F}}^{\mathtt{H}}_{i+\frac{1}{2}}$.

3.2.2. Positivity of pressure. The second step is to enforce pressure positivity $p(\widetilde{\boldsymbol{U}}^{\mathtt{Lim},\pm}_{i+\frac{1}{2}}) > \varepsilon$, where $p(\boldsymbol{U}^*)$ denotes the pressure recovered from $\boldsymbol{U}^*$, with

$$\widetilde{\boldsymbol{U}}^{\mathtt{Lim},\pm}_{i+\frac{1}{2}} = \widetilde{\boldsymbol{U}}_{i+\frac{1}{2}} \pm \frac{\theta_{i+\frac{1}{2}} \Delta \widehat{\boldsymbol{F}}^{\mathtt{Lim},*}_{i+\frac{1}{2}}}{\alpha_{i+\frac{1}{2}}}, \quad \Delta \widehat{\boldsymbol{F}}^{\mathtt{Lim},*}_{i+\frac{1}{2}} = \widehat{\boldsymbol{F}}^{\mathtt{Lim},*}_{i+\frac{1}{2}} - \widehat{\boldsymbol{F}}^{\mathtt{L}}_{i+\frac{1}{2}}.$$

Such constraints lead to two inequalities

$$\frac{A_{i+\frac{1}{2}}}{\alpha^2_{i+\frac{1}{2}}} \theta^2_{i+\frac{1}{2}} \pm \frac{B_{i+\frac{1}{2}}}{\alpha_{i+\frac{1}{2}}} \theta_{i+\frac{1}{2}} < C_{i+\frac{1}{2}}, \tag{3.7}$$

with the coefficients

$$A_{i+\frac{1}{2}} = \frac{1}{2}\left(\Delta \widehat{\boldsymbol{F}}^{\mathtt{Lim},*,\rho v}_{i+\frac{1}{2}}\right)^2 - \Delta \widehat{\boldsymbol{F}}^{\mathtt{Lim},*,\rho}_{i+\frac{1}{2}} \Delta \widehat{\boldsymbol{F}}^{\mathtt{Lim},*,E}_{i+\frac{1}{2}},$$

$$B_{i+\frac{1}{2}} = \alpha_{i+\frac{1}{2}}\left(\Delta \widehat{\boldsymbol{F}}^{\mathtt{Lim},*,\rho}_{i+\frac{1}{2}} \widetilde{\boldsymbol{U}}^{E}_{i+\frac{1}{2}} + \widetilde{\boldsymbol{U}}^{\rho}_{i+\frac{1}{2}} \Delta \widehat{\boldsymbol{F}}^{\mathtt{Lim},*,E}_{i+\frac{1}{2}} - \Delta \widehat{\boldsymbol{F}}^{\mathtt{Lim},*,\rho v}_{i+\frac{1}{2}} \widetilde{\boldsymbol{U}}^{\rho v}_{i+\frac{1}{2}} - \varepsilon \Delta \widehat{\boldsymbol{F}}^{\mathtt{Lim},*,\rho}_{i+\frac{1}{2}}\right),$$

$$C_{i+\frac{1}{2}} = \alpha^2_{i+\frac{1}{2}}\left(\widetilde{\boldsymbol{U}}^{\rho}_{i+\frac{1}{2}} \widetilde{\boldsymbol{U}}^{E}_{i+\frac{1}{2}} - \frac{1}{2}\left(\widetilde{\boldsymbol{U}}^{\rho v}_{i+\frac{1}{2}}\right)^2 - \varepsilon \widetilde{\boldsymbol{U}}^{\rho}_{i+\frac{1}{2}}\right).$$

Following [29], the inequalities (3.7) hold under the linear sufficient condition

$$\max\{0, A_{i+\frac{1}{2}}\} + |B_{i+\frac{1}{2}}| \leqslant C_{i+\frac{1}{2}},$$

if making use of $\theta^2_{i+\frac{1}{2}} \leqslant \theta_{i+\frac{1}{2}},\ \theta_{i+\frac{1}{2}} \in [0,1]$. Thus the coefficient can be chosen as

$$\theta_{i+\frac{1}{2}} = \min\left\{1,\ \frac{C_{i+\frac{1}{2}}}{\max\{0, A_{i+\frac{1}{2}}\} + |B_{i+\frac{1}{2}}|}\right\},$$

and the final limited numerical flux is

$$\widehat{\boldsymbol{F}}^{\mathtt{Lim}}_{i+\frac{1}{2}} = \widehat{\boldsymbol{F}}^{\mathtt{L}}_{i+\frac{1}{2}} + \theta_{i+\frac{1}{2}} \Delta \widehat{\boldsymbol{F}}^{\mathtt{Lim},*}_{i+\frac{1}{2}}. \tag{3.8}$$

4. Scaling limiter for point value. To achieve the BP property, it is also necessary to introduce BP limiting for the point value. As one will see in the numerical tests in Section 5, using power law reconstruction or BP limiting for cell average, individually or in combination, cannot guarantee the bounds. As there is no conservation requirement on the point value update, a simple scaling limiter [31] is directly performed on the high-order point values rather than on the flux for the cell average.

A first-order LLF scheme for the point value update can be

$$\boldsymbol{U}^{\mathtt{L}}_{i+\frac{1}{2}} = \boldsymbol{U}^{n}_{i+\frac{1}{2}} - \frac{2\Delta t^n}{\Delta x_i + \Delta x_{i+1}}\left(\widehat{\boldsymbol{F}}^{\mathtt{L}}_{i+1}(\boldsymbol{U}^{n}_{i+\frac{1}{2}}, \boldsymbol{U}^{n}_{i+\frac{3}{2}}) - \widehat{\boldsymbol{F}}^{\mathtt{L}}_{i}(\boldsymbol{U}^{n}_{i-\frac{1}{2}}, \boldsymbol{U}^{n}_{i+\frac{1}{2}})\right), \tag{4.1}$$

with the numerical flux

$$\widehat{\boldsymbol{F}}^{\mathtt{L}}_{i}(\boldsymbol{U}^{n}_{i-\frac{1}{2}}, \boldsymbol{U}^{n}_{i+\frac{1}{2}}) = \frac{1}{2}\left(\boldsymbol{F}(\boldsymbol{U}^{n}_{i-\frac{1}{2}}) + \boldsymbol{F}(\boldsymbol{U}^{n}_{i+\frac{1}{2}})\right) - \frac{\alpha_i}{2}\left(\boldsymbol{U}^{n}_{i+\frac{1}{2}} - \boldsymbol{U}^{n}_{i-\frac{1}{2}}\right),$$

$$\alpha_i = \max\{\lambda(\boldsymbol{U}^{n}_{i-\frac{1}{2}}),\ \lambda(\boldsymbol{U}^{n}_{i+\frac{1}{2}})\}.$$

Such an LLF scheme can be interpreted as a scheme on a staggered mesh if the point value is viewed as the cell average on the staggered mesh. Based on the proof in [33], it is straightforward to obtain the following Lemma.

LEMMA 4.1. The LLF scheme for the point value (4.1) is BP under the CFL condition

$$\Delta t^n \leqslant \frac{\Delta x_i + \Delta x_{i+1}}{4\alpha_i}. \tag{4.2}$$

The limited state is obtained by blending the high-order AF scheme (2.3) with the forward Euler scheme and the LLF scheme (4.1) as $\boldsymbol{U}^{\mathtt{Lim}}_{i+\frac{1}{2}} = \theta_{i+\frac{1}{2}}\boldsymbol{U}^{\mathtt{H}}_{i+\frac{1}{2}} + (1-\theta_{i+\frac{1}{2}})\boldsymbol{U}^{\mathtt{L}}_{i+\frac{1}{2}}$, such that $\boldsymbol{U}^{\mathtt{Lim}}_{i+\frac{1}{2}} \in \mathcal{G}$.

Remark 4.2. In the FVS for the point value update, the cell-centered value $\boldsymbol{U}_i$ is used. It is possible that $\boldsymbol{U}_i \notin \mathcal{G}$, then it is set as $\overline{\boldsymbol{U}}_i$ in such cases, which is a reasonable second-order approximation.

4.1. Application to scalar conservation laws. This section enforces the global MP $m_0 \leqslant u^{\mathtt{Lim}}_{i+\frac{1}{2}} \leqslant M_0$ by choosing the coefficient as

$$\theta_{i+\frac{1}{2}} = \begin{cases} \dfrac{u^{\mathtt{L}}_{i+\frac{1}{2}} - m_0}{u^{\mathtt{L}}_{i+\frac{1}{2}} - u^{\mathtt{H}}_{i+\frac{1}{2}}}, & \text{if } u^{\mathtt{H}}_{i+\frac{1}{2}} < m_0, \\ \dfrac{M_0 - u^{\mathtt{L}}_{i+\frac{1}{2}}}{u^{\mathtt{H}}_{i+\frac{1}{2}} - u^{\mathtt{L}}_{i+\frac{1}{2}}}, & \text{if } u^{\mathtt{H}}_{i+\frac{1}{2}} > M_0. \end{cases}$$

The final limited state is

$$u^{\mathtt{Lim}}_{i+\frac{1}{2}} = \theta_{i+\frac{1}{2}} u^{\mathtt{H}}_{i+\frac{1}{2}} + \left(1-\theta_{i+\frac{1}{2}}\right) u^{\mathtt{L}}_{i+\frac{1}{2}}. \tag{4.3}$$

4.2. Application to the compressible Euler equations. The limiting consists of two steps. First, the high-order state $\boldsymbol{U}^{\mathtt{H}}_{i+\frac{1}{2}}$ is modified as $\boldsymbol{U}^{\mathtt{Lim},*}_{i+\frac{1}{2}}$, such that its density component satisfies $\boldsymbol{U}^{\mathtt{Lim},*,\rho}_{i+\frac{1}{2}} > \varepsilon$. Solving this inequality gives the coefficient

$$\theta^*_{i+\frac{1}{2}} = \begin{cases} \dfrac{\boldsymbol{U}^{\mathtt{L},\rho}_{i+\frac{1}{2}} - \varepsilon}{\boldsymbol{U}^{\mathtt{L},\rho}_{i+\frac{1}{2}} - \boldsymbol{U}^{\mathtt{H},\rho}_{i+\frac{1}{2}}}, & \text{if } \boldsymbol{U}^{\mathtt{H},\rho}_{i+\frac{1}{2}} < \varepsilon, \\ 1, & \text{otherwise.} \end{cases}$$

Then the density component of the limited state is $\boldsymbol{U}^{\mathtt{Lim},*,\rho}_{i+\frac{1}{2}} = \theta^*_{i+\frac{1}{2}}\boldsymbol{U}^{\mathtt{H},\rho}_{i+\frac{1}{2}} + (1-\theta^*_{i+\frac{1}{2}})\boldsymbol{U}^{\mathtt{L},\rho}_{i+\frac{1}{2}}$, with the other components remaining the same as $\boldsymbol{U}^{\mathtt{H}}_{i+\frac{1}{2}}$.

Then the limited state $\boldsymbol{U}^{\mathtt{Lim},*}_{i+\frac{1}{2}}$ is modified as $\boldsymbol{U}^{\mathtt{Lim}}_{i+\frac{1}{2}}$, such that it gives positive pressure, i.e., $p\left(\boldsymbol{U}^{\mathtt{Lim}}_{i+\frac{1}{2}}\right) > \varepsilon$. Let $\boldsymbol{U}^{\mathtt{Lim}}_{i+\frac{1}{2}} = \theta^{**}_{i+\frac{1}{2}}\boldsymbol{U}^{\mathtt{Lim},*}_{i+\frac{1}{2}} + (1-\theta^{**}_{i+\frac{1}{2}})\boldsymbol{U}^{\mathtt{L}}_{i+\frac{1}{2}}$. Note that the pressure is a concave function (see e.g. [45]) of the conservative variables, such that

$$p\left(\boldsymbol{U}^{\mathtt{Lim}}_{i+\frac{1}{2}}\right) \geqslant \theta^{**}_{i+\frac{1}{2}} p\left(\boldsymbol{U}^{\mathtt{Lim},*}_{i+\frac{1}{2}}\right) + \left(1-\theta^{**}_{i+\frac{1}{2}}\right) p\left(\boldsymbol{U}^{\mathtt{L}}_{i+\frac{1}{2}}\right)$$

based on Jensen's inequality and $\boldsymbol{U}^{\mathtt{Lim},*,\rho}_{i+\frac{1}{2}} > 0$, $\boldsymbol{U}^{\mathtt{L},\rho}_{i+\frac{1}{2}} > 0$, $\theta^{**}_{i+\frac{1}{2}} \in [0,1]$. Thus a sufficient condition is

$$\theta^{**}_{i+\frac{1}{2}} = \begin{cases} \dfrac{p\left(\boldsymbol{U}^{\mathtt{L}}_{i+\frac{1}{2}}\right) - \varepsilon}{p\left(\boldsymbol{U}^{\mathtt{L}}_{i+\frac{1}{2}}\right) - p\left(\boldsymbol{U}^{\mathtt{Lim},*}_{i+\frac{1}{2}}\right)}, & \text{if } p\left(\boldsymbol{U}^{\mathtt{Lim},*}_{i+\frac{1}{2}}\right) < \varepsilon, \\ 1, & \text{otherwise.} \end{cases}$$

The final limited state is

$$\boldsymbol{U}^{\mathtt{Lim}}_{i+\frac{1}{2}} = \theta^{**}_{i+\frac{1}{2}} \boldsymbol{U}^{\mathtt{Lim},*}_{i+\frac{1}{2}} + \left(1-\theta^{**}_{i+\frac{1}{2}}\right) \boldsymbol{U}^{\mathtt{L}}_{i+\frac{1}{2}}. \tag{4.4}$$

Let us summarize the main results of the BP AF methods in this paper.

THEOREM 4.3. If the initial numerical solution $\overline{\boldsymbol{U}}^0_i, \boldsymbol{U}^0_{i+\frac{1}{2}} \in \mathcal{G}$ for all i, and the time step size satisfies (3.3) and (4.2), then the AF methods (2.2)-(2.3) equipped with the SSP-RK3 (2.16) and the BP limitings

- (3.6) and (4.3) preserve the maximum principle for scalar case;
- (3.8) and (4.4) preserve the density and pressure positivity for the Euler equations.

5. Numerical results. This section conducts some numerical tests to verify the accuracy of using the FVS for point value updates, the BP property, and the shock-capturing ability of our AF methods.

5.1. Scalar conservation laws. This section shows the results for the linear advection equation and the Burgers' equation, which demonstrate that the proposed limiting can preserve the MP and suppress oscillations well.

Example 5.1 (Advection equation). Consider the 1D advection equation $u_t + u_x = 0$, on the periodic domain $[-1, 1]$ with the initial data [27]

$$\begin{cases} \frac{1}{6}\left(G_1(x,\beta,z-\delta) + G_1(x,\beta,z+\delta) + 4G_1(x,\beta,z)\right), & \text{if } -0.8 \leqslant x \leqslant -0.6, \\ 1, & \text{if } -0.4 \leqslant x \leqslant -0.2, \\ 1 - |10(x-0.1)|, & \text{if } 0 \leqslant x \leqslant 0.2, \\ \frac{1}{6}\left(G_2(x,\alpha,a-\delta) + G_2(x,\alpha,a+\delta) + 4G_2(x,\alpha,a)\right), & \text{if } 0.4 \leqslant x \leqslant 0.6, \\ 0, & \text{otherwise}, \end{cases}$$

where $G_1(x,\beta,z) = \exp(-\beta(x-z)^2)$, $G_2(x,\alpha,a) = \sqrt{\max(1-\alpha^2(x-a)^2, 0)}$, and the constants are $a = -0.5, z = -0.7, \delta = 0.005, \alpha = 10, \beta = \ln 2/(36\delta^2)$. The problem is solved for one period, i.e., until $T = 2$.

For the advection equation, the JS and LLF FVS are equivalent. The maximal CFL number leading to a stable simulation is 0.41 without any limiting, and it reduces to 0.13 when only the power law reconstruction is activated, and it increases a little bit to 0.42 when only the BP limitings are used. When the power law reconstruction and the BP limitings are employed together, the maximal CFL number can be 0.4. The reduction of the CFL number with the power law reconstruction for semi-discrete AF has, in fact, not been noticed previously. Thus, in the following simulations we try not to use the power law reconstruction unless otherwise stated.

The results obtained with different limitings are shown in Figure 2, which are computed with 400 cells and the CFL number is 0.1. The ranges of the numerical solutions are listed in Table 1, considering both the cell averages and point values. One can observe that there are some oscillations near the discontinuities without any limiting, and that the power law reconstruction can eliminate the oscillations effectively but is still not BP. The activation of either the BP limiting for the cell average alone or the BP limiting for the point value alone also fails to preserve the bounds $[0, 1]$, as one can see from Table 1, as is the case when using both the BP limiting for the cell average and the power law reconstruction in the point value update. Only when a BP limiting is performed on both the cell average and the point value, the BP property is achieved, showing that using the two BP limitings

simultaneously is necessary for the preservation of the MP. Figure 2 also shows the results obtained by imposing the global or local MP for the cell average, and global MP for the point value (without power law reconstruction), indicating that the use of local MP tends to dissipate the numerical solutions near the discontinuities and clip maxima more than the global MP.

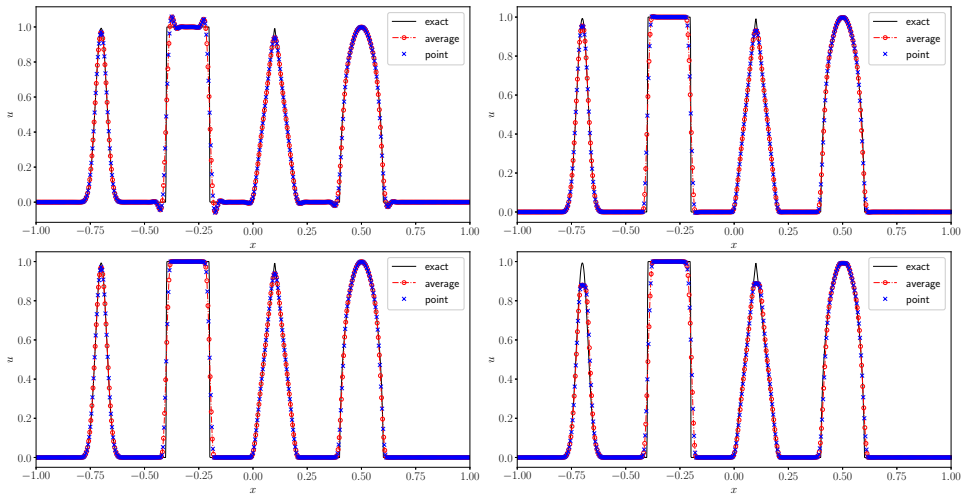

Fig. 2: Example 5.1, advection. The results are obtained without any limiting (upper left), with power law reconstruction (upper right), with BP limitings imposing global MP for the cell average and point value (lower left), with BP limitings imposing local and global MP for the cell average and point value (lower right).

none	$[-5.9 \times 10^{-2}, 1 + 5.9 \times 10^{-2}]$	✗
PLR	$[-2.7 \times 10^{-3}, 1 + 2.6 \times 10^{-3}]$	✗
global MP for average	$[-1.7 \times 10^{-3}, 1 + 1.7 \times 10^{-3}]$	✗
local MP for average	$[-1.3 \times 10^{-3}, 1 + 1.3 \times 10^{-3}]$	✗
global MP for point	$[-3.0 \times 10^{-4}, 1 + 2.6 \times 10^{-4}]$	✗
PLR + global MP for average	$[-9.8 \times 10^{-6}, 1 + 2.7 \times 10^{-6}]$	✗
PLR + local MP for average	$[-1.4 \times 10^{-5}, 1 + 1.9 \times 10^{-5}]$	✗
global MP for average + global MP for point	$[0.0, 1.0]$	✓
local MP for average + global MP for point	$[0.0, 1 - 9.4 \times 10^{-13}]$	✓
PLR + global MP for average + global MP for point	$[0.0, 1 - 1.1 \times 10^{-16}]$	✓
PLR + local MP for average + global MP for point	$[0.0, 1 - 7.3 \times 10^{-14}]$	✓

Table 1: Example 5.1, advection. The ranges of the numerical solutions (including both the cell averages and the point values) obtained with different limitings after one period. "PLR" denotes the power law reconstruction.

Example 5.2 (Self-steepening shock). Consider the 1D Burgers' equation $u_t + \left(\frac{1}{2}u^2\right)_x = 0$ on the domain $[-1, 1]$ with periodic boundary conditions. This test is solved until $T = 0.5$ with the initial condition as a square wave

$$u_0(x) = \begin{cases} 2, & \text{if } |x| < 0.2, \\ -1, & \text{otherwise.} \end{cases}$$

Figures 3 and 4 plot the cell averages and point values based on different point value updates with 200 cells, as well as the reference solution. The spike generation has been observed in [25], and the reason is also discussed in Subsection 2.2. Such spike generation cannot be eliminated by using the power law reconstruction, nor do both BP limitings help to eliminate artefacts, as can be seen from Figure 3. The numerical solutions based on the LLF or SW FVS are shown in Figure 4, in which no spike appears. There are some oscillations near the discontinuity without limitings, and the numerical solutions agree well with the reference solution when the limitings are activated.

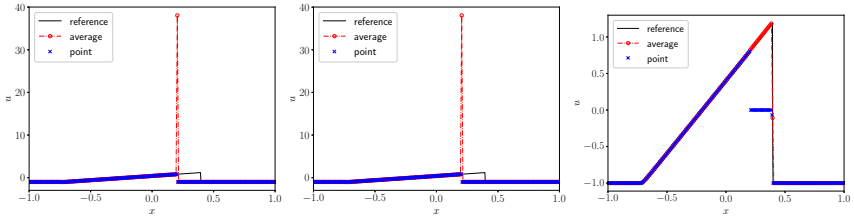

Fig. 3: Example 5.2, self-steepening shock for the Burgers' equation. The numerical solutions are based on the JS. From left to right: without limiting, with the power law reconstruction, with the BP limitings imposing local and global MP for the cell average and point value update, respectively.

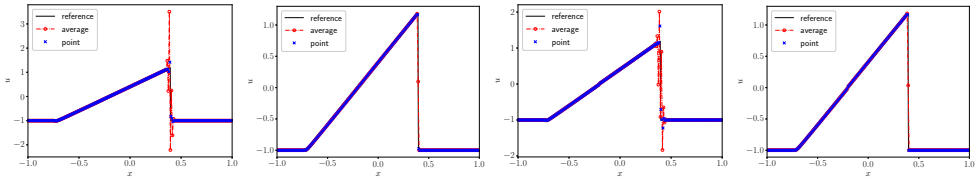

Fig. 4: Example 5.2, self-steepening shock for the Burgers' equation. From left to right: the LLF FVS without limiting, the LLF FVS with limitings, the SW FVS without limiting, the SW FVS with limitings. The limitings consider the local and global MP for the cell average and point value updates, respectively.

5.2. The compressible Euler equations. This section shows some challenging tests, which require the BP property of the numerical methods in order to prevent simulations from crashing at some time. The adiabatic index is $\gamma = 1.4$ unless otherwise stated. Note that the BP limiting naturally reduces some oscillations.

Example 5.3 (1D accuracy test for the Euler equations). This test is used to examine the accuracy of using different point value updates. The domain is $[0, 1]$ with periodic boundary conditions. Two manufactured solutions are constructed by adding additional source terms $\boldsymbol{S}$ to the Euler equations,

$$\rho = 4 + 0.1 s_1,\ v_1 = s_1,\ p = (6002 + 398 c_2 + 305 s_1 + 5 s_3)/1000, \tag{5.1}$$
$$\boldsymbol{S} = (\pi(39 c_1 + s_2)/5,\ -\pi(905 c_1 + 15 c_3 - 776 s_2)/125,$$
$$\pi c_1 (20421 + 1179 c_2 + 2160 s_1 + 20 s_3)/500),$$

and

$$\rho = 4 + 0.1s_1,\ v_1 = 2 + 0.5s_1,\ p = (12328 + 472c_2 - 5455s_1 + 5s_3)/4000, \tag{5.2}$$
$$\boldsymbol{S} = (\pi(42c_1 + s_2)/10,\ \pi(4855c_1 - 15c_3 + 914s_2)/500,$$
$$\pi c_1(14991 + 369c_2 - 2983s_1 + 5s_3)/1000),$$

with $s_k = \sin(2k\pi(x-t)), c_k = \cos(2k\pi(x-t)),\ k = 1, 2, 3$. The source terms are discretized by using Simpson's rule for the cell average update. The problem is solved until $T = 0.4$.

In this test, the maximal CFL number is around 0.18 for the VH FVS, while around 0.43 for the JS, LLF, and SW FVS, thus we run the test with the same CFL number as 0.18. Figure 5 shows the following errors and corresponding convergence rates for the conservative variables in the ℓ^1 norm. It is seen that for the first exact solution (5.1), the JS and all the FVS except for the SW FVS achieve the designed third-order accuracy, while the SW FVS only gives second-order accuracy. Figure 6 plots the density and velocity profiles obtained by the SW FVS with 80 cells. One can observe some defects in the density when the velocity is zero, similar to the "sonic point glitch" in the literature [37]. For the second exact solution (5.2), the velocity stays away from zero and no such issue appears. One possible reason is that the SW FVS is based on the absolute value of the eigenvalues, which is not smooth when the velocity is zero. Such an issue remains to be further explored in the future.

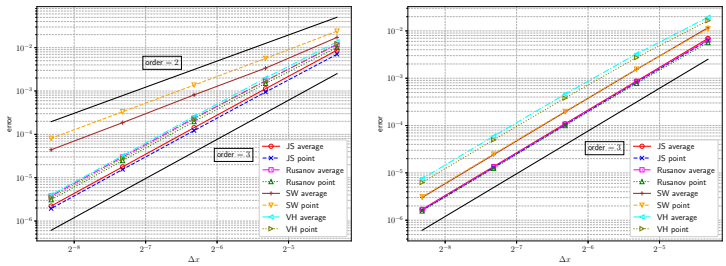

Fig. 5: Example 5.3, the accuracy tests for the 1D Euler equations based on the manufactured solutions (5.1) and (5.2) for the left and right plots, respectively.

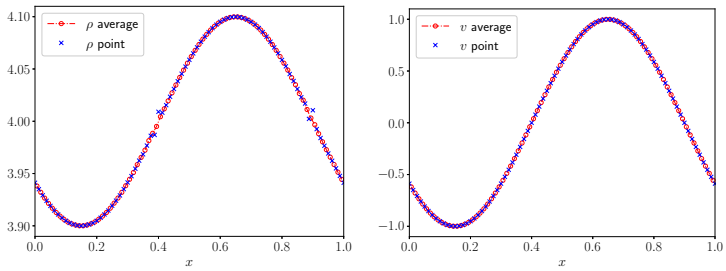

Fig. 6: Example 5.3, the density (left) and velocity (right) are obtained with the SW FVS and 80 cells for the 1D Euler equations based on the initial data (5.1).

Example 5.4 (Double rarefaction problem). The exact solution to this problem contains a vacuum, so that it is often used to verify the BP property of numerical methods. The test is solved on a domain $[0, 1]$ until $T = 0.3$ with the initial data

$$(\rho, v, p) = \begin{cases} (7, -1, 0.2), & \text{if } \ x < 0.5, \\ (7, 1, 0.2), & \text{otherwise.} \end{cases}$$

In this test, the AF method based on any kind of point value update mentioned in this paper gives negative density or pressure without the BP limitings. Figure 7 shows the density computed with 400 cells and the BP limitings for the cell average and point value updates. The power law reconstruction is not used in this test, and the CFL number is 0.4 for all kinds of point value updates, except for 0.1 for the VH FVS. One observes that the BP AF method gets good performance for this example.

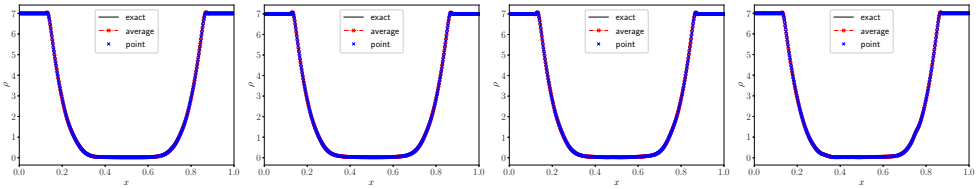

Fig. 7: Example 5.4, double rarefaction Riemann problem. The numerical solutions are computed with BP limitings for the cell average and point value updates on a uniform mesh of 400 cells. The power law reconstruction is not used. From left to right: JS, LLF, SW, and VH FVS.

Example 5.5 (LeBlanc shock tube). This is a Riemann problem with an extremely large initial pressure ratio. This test is solved until $T = 5 \times 10^{-6}$ on a domain $[0, 1]$ with the initial data

$$(\rho, v, p) = \begin{cases} (2, 0, 10^9), & \text{if } \ x < 0.5, \\ (10^{-3}, 0, 1), & \text{otherwise.} \end{cases}$$

Without the BP limitings, the simulation will stop due to negative density or pressure. Figure 8 shows the density computed on a uniform mesh of 400 and 6000 cells with the BP limitings for the cell average and point value updates. The CFL number is 0.4 for the LLF and SW FVS, and 0.15 for the JS and VH FVS for stability when the power law reconstruction is not used. It is seen that the numerical solutions on the coarse mesh deviate from the exact solutions, which has also been observed in other high-order BP methods, e.g., [44]. As the number of the mesh cells increases from 400 to 6000, one can observe from Figure 8 that the numerical solutions converge to the exact solutions with only a few overshoots/undershoots at the contact discontinuity. The LLF and SW FVS give better results.

To verify whether the power law reconstruction can suppress spurious oscillations and overshoots/undershoots, we rerun the test with the CFL number 0.1, and the density profiles are shown in Figure 9. It is obvious that only reducing the CFL number does not change the numerical solutions much except that the oscillations near the contact discontinuity based on the VH FVS are damped. When the power law reconstruction is activated, the overshoots/undershoots are reduced for the JS,

LLF, and SW FVS, while the VH FVS gives worse results even with a smaller CFL number (e.g. 0.02, not shown here), which needs further investigation.

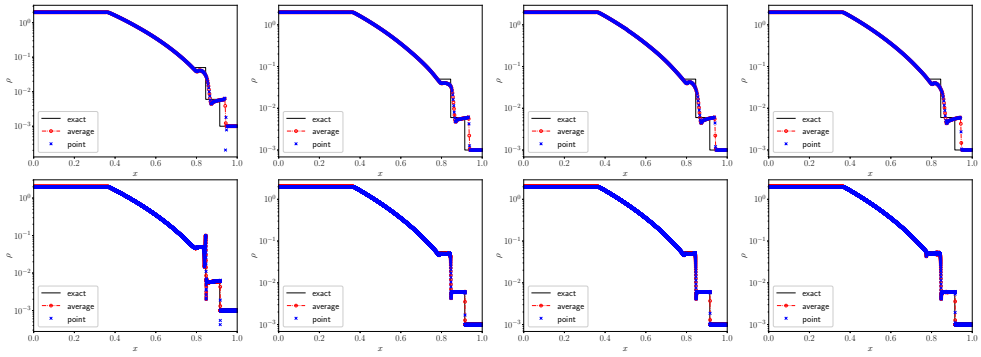

Fig. 8: Example 5.5, LeBlanc Riemann problem. The numerical solutions are computed with the BP limitings for the cell average and point value updates on a uniform mesh of 400 cells (top) and 6000 cells (bottom). The CFL number is 0.4 and the power law reconstruction is not used. From left to right: JS, LLF, SW, and VH FVS.

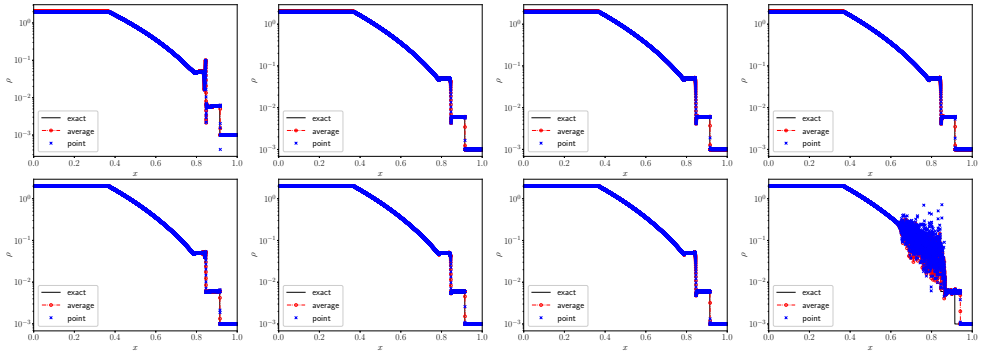

Fig. 9: Example 5.5, LeBlanc Riemann problem. The numerical solutions are computed with the BP limitings for the cell average and point value updates on a uniform mesh of 6000 cells. From left to right: JS, LLF, SW, and VH FVS. The CFL number is 0.1 and the power law reconstruction is not activated (top) and activated (bottom).

Example 5.6 (Sedov problem). In this problem, a volume of uniform density and temperature is initialized, and a large quantity of thermal energy is injected at the center, developing into a blast wave that evolves in time in a self-similar fashion [35]. An exact analytical solution based on self-similarity arguments is available [28], which contains very low density with strong shocks. The initial density is one, velocity is zero, and total energy is 10^{-12} everywhere except that in the center cell, the total energy of the cell average and point values at two cell interfaces are $3.2 \times 10^6/\Delta x$ with $\Delta x = 4/N$ with N the number of cells, which is used to emulate a δ-function at the center. The test is solved until $T = 5 \times 10^{-6}$.

This test is run with $N = 801$ cells, and the density plots in the right half domain are shown in Figure 10. The BP limitings are adopted for the cell average and point

value updates, while the power law reconstruction is not used. The maximal CFL numbers for different point value updates to be stable are also listed in the caption, i.e., 0.1 for the JS, 0.4, 0.3, and 0.25 for the LLF, SW, and VH FVS, respectively. The numerical solutions obtained by the three FVS are nearly the same, while there are some defects in the solution based on the JS. Thus the LLF FVS is superior to others regarding the time step size and the shock-capturing ability.

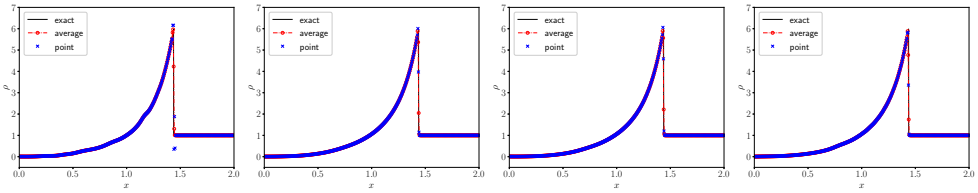

Fig. 10: Example 5.6, Sedov problem. The numerical solutions are computed with the BP limitings for the cell average and point value updates on a uniform mesh of 801 cells, without the power law reconstruction. The CFL number is (from left to right): 0.1 for the JS, 0.4 for the LLF FVS, 0.3 for the SW FVS, 0.25 for the VH FVS.

Example 5.7 (Blast wave interaction [41]). This test describes the interaction of two strong shocks in the domain $[0, 1]$ with reflective boundary conditions. The test is solved until $T = 0.038$.

Due to the low-pressure region, the schemes blow up without the BP limitings. Figure 11 shows the density profiles and corresponding enlarged views in $x \in [0.62, 0.82]$ obtained by using the BP limitings on a uniform mesh of 800 cells, in which the power law reconstruction is not activated. It is seen that the numerical solutions are close to the reference solution, although there are some oscillations in the enlarged views. Then the power law reconstruction is additionally adopted to see if it can suppress the oscillations. The results with the CFL number 0.1 and a refined mesh of 1600 cells are shown in Figure 12, from which one can observe that the oscillations reduce, and the LLF FVS gives the best result.

Remark 5.8. In the numerical tests, the maximal CFL numbers for stability are obtained by experiments. Note that the constraints (3.3) and (4.2) are used to guarantee the BP property, while the reduction of the CFL numbers is due to the stability issue for different FVS and power law reconstruction.

6. Conclusion. In the active flux (AF) methods, the way how point values at cell interfaces are updated is essential to achieve stability and high-order accuracy. The point value update based on Jacobian splitting (JS) may lead to the so-called transonic issue for nonlinear problems due to inaccurate estimation of the upwind direction. This paper proposed to use the flux vector splitting (FVS) for the point value update instead of the JS, which keeps the continuous reconstruction as the original AF methods, and offers a natural and uniform remedy to the transonic issue. To further improve the robustness of the AF methods, this paper developed bound-preserving (BP) AF methods for general one-dimensional hyperbolic conservation laws, achieved by blending the high-order AF methods with the first-order local Lax-Friedrichs (LLF) or Rusanov methods for both the cell average and point value updates, where the convex limiting and scaling limiter were employed, respectively. For scalar conservation laws, the blending coefficient was determined based on the global or local maximum

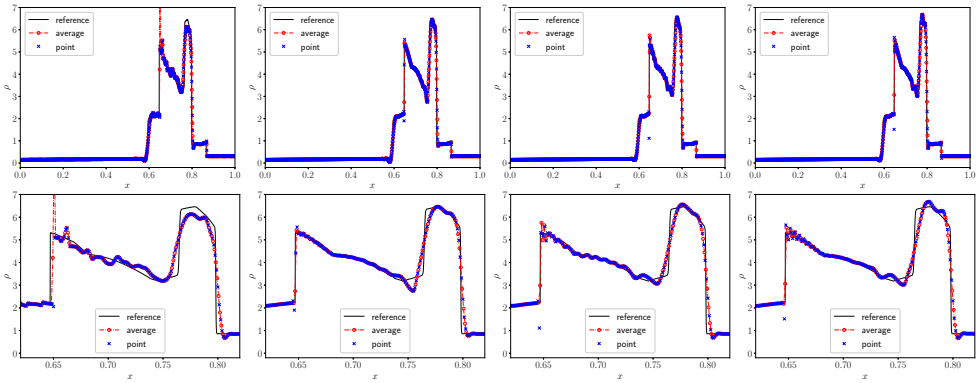

Fig. 11: Example 5.7, blast wave interaction. The numerical solutions are computed with the BP limitings for the cell average and point value updates on a uniform mesh of 800 cells. The power law reconstruction is not used, and from left to right: the CFL number is 0.4, 0.4, 0.4, 0.35 for the JS, LLF, SW, and VH FVS, respectively. The corresponding enlarged views in $[0.62, 0.82]$ are shown in the bottom row.

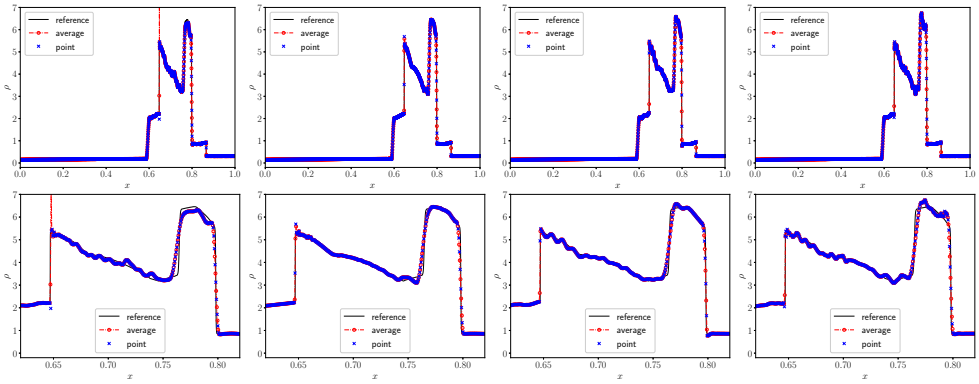

Fig. 12: Example 5.7, blast wave interaction. The numerical solutions are computed with the power law reconstruction and the BP limitings for the cell average and point values update on a uniform mesh of 1600 cells. The CFL number is 0.1 for all the point value updates, and the corresponding enlarged views in $[0.62, 0.82]$ are shown in the bottom row. From left to right: JS, LLF, SW, and VH FVS.

principle, while for the compressible Euler equations, it was obtained by enforcing the positivity of density and pressure. Some challenging benchmark tests were conducted based on different choices of the point value update, including the JS, LLF, Steger-Warming, and Van Leer-Hänel FVS. The numerical results confirmed the accuracy, BP property, and shock-capturing ability of our methods, and also showed that the LLF FVS is generally superior to others in terms of the CFL number and shock-capturing ability. Our future work will include, among others, extending the current BP limitings to two-dimensional cases. We may also explore other ways to further suppress oscillations for the Euler equations.

613 REFERENCES

614 [1] R. Abgrall, *A combination of residual distribution and the active flux formulations or a*
615 *new class of schemes that can combine several writings of the same hyperbolic problem:*
616 *Application to the 1D Euler equationss*, Commun. Appl. Math. Comput., 5 (2023), pp. 370–
617 402.
618 [2] R. Abgrall and W. Barsukow, *Extensions of active flux to arbitrary order of accuracy*,
619 ESAIM: Math. Model. Numer. Anal., 57 (2023), pp. 991–1027.
620 [3] R. Abgrall, J. Lin, and Y. Liu, *Active flux for triangular meshes for compressible flows*
621 *problems*, Dec. 2023, https://arxiv.org/abs/2312.11271.
622 [4] W. Barsukow, *The active flux scheme for nonlinear problems*, J. Sci. Comput., 86 (2021),
623 p. 3.
624 [5] W. Barsukow and J. P. Berberich, *A well-balanced active flux method for the shallow*
625 *water equations with wetting and drying*, Commun. Appl. Math. Comput., (2023).
626 [6] W. Barsukow, J. P. Berberich, and C. Klingenberg, *On the active flux scheme for*
627 *hyperbolic PDEs with source terms*, SIAM J. Sci. Comput., 43 (2021), pp. A4015–A4042.
628 [7] W. Barsukow, J. Hohm, C. Klingenberg, and P. L. Roe, *The active flux scheme on*
629 *Cartesian grids and its low Mach number limit*, J. Sci. Comput., 81 (2019), pp. 594–622.
630 [8] E. Chudzik, C. Helzel, and D. Kerkmann, *The Cartesian grid active flux method: Linear*
631 *stability and bound preserving limiting*, Appl. Math. Comput., 393 (2021), p. 125501.
632 [9] S. Clain, S. Diot, and R. Loubère, *A high-order finite volume method for systems of con-*
633 *servation laws–Multi-dimensional Optimal Order Detection (MOOD)*, J. Comput. Phys.,
634 230 (2011), pp. 4028–4050.
635 [10] C. J. Cotter and D. Kuzmin, *Embedded discontinuous Galerkin transport schemes with*
636 *localised limiters*, J. Comput. Phys., 311 (2016), pp. 363–373.
637 [11] C. M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Springer Berlin Hei-
638 delberg, 2000.
639 [12] T. Eymann and P. Roe, *Active flux schemes*, in 49th AIAA Aerospace Sciences Meeting
640 including the New Horizons Forum and Aerospace Exposition, Orlando, Florida, Jan. 2011,
641 American Institute of Aeronautics and Astronautics.
642 [13] T. Eymann and P. Roe, *Active flux schemes for systems*, in 20th AIAA Computational Fluid
643 Dynamics Conference, Fluid Dynamics and Co-located Conferences, American Institute of
644 Aeronautics and Astronautics, June 2011.
645 [14] T. A. Eymann and P. L. Roe, *Multidimensional active flux schemes*, in 21st AIAA Computa-
646 tional Fluid Dynamics Conference, Fluid Dynamics and Co-located Conferences, American
647 Institute of Aeronautics and Astronautics, June 2013.
648 [15] D. Fan and P. L. Roe, *Investigations of a new scheme for wave propagation*, in 22nd AIAA
649 Computational Fluid Dynamics Conference, American Institute of Aeronautics and Astro-
650 nautics, 2015.
651 [16] S. Gottlieb, C.-W. Shu, and E. Tadmor, *Strong Stability-Preserving High-Order Time*
652 *Discretization Methods*, SIAM Rev., 43 (2001), pp. 89–112.
653 [17] J.-L. Guermond, M. Nazarov, B. Popov, and I. Tomas, *Second-order invariant domain*
654 *preserving approximation of the Euler equations using convex limiting*, SIAM J. Sci. Com-
655 put., 40 (2018), pp. A3211–A3239.
656 [18] J.-L. Guermond and B. Popov, *Fast estimation from above of the maximum wave speed in*
657 *the Riemann problem for the Euler equations*, J. Comput. Phys., 321 (2016), pp. 908–926.
658 [19] J.-L. Guermond and B. Popov, *Invariant domains and first-order continuous finite element*
659 *approximation for hyperbolic systems*, SIAM J. Numer. Anal., 54 (2016), pp. 2466–2489.
660 [20] J.-L. Guermond and B. Popov, *Invariant domains and second-order continuous finite ele-*
661 *ment approximation for scalar conservation equations*, SIAM J. Numer. Anal., 55 (2017),
662 pp. 3120–3146.
663 [21] J.-L. Guermond, B. Popov, and I. Tomas, *Invariant domain preserving discretization-*
664 *independent schemes and convex limiting for hyperbolic systems*, Comput. Methods Appl.
665 Mech. Engrg., 347 (2019), pp. 143–175.
666 [22] H. Hajduk, *Monolithic convex limiting in discontinuous Galerkin discretizations of hyperbolic*
667 *conservation laws*, Comput. Math. Appl., 87 (2021), pp. 120–138.
668 [23] D. Hänel, R. Schwane, and G. Seider, *On the accuracy of upwind schemes for the solution*
669 *of the Navier-Stokes equations*, Fluid Dynamics and Co-located Conferences, American
670 Institute of Aeronautics and Astronautics, June 1987.
671 [24] A. Harten, P. D. Lax, and B. v. Leer, *On upstream differencing and Godunov-type*
672 *schemes for hyperbolic conservation laws*, SIAM Rev., 25 (1983), pp. 35–61.
673 [25] C. Helzel, D. Kerkmann, and L. Scandurra, *A new ADER method inspired by the active*

flux method, J. Sci. Comput., 80 (2019), pp. 1463–1497.
[26] 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.
[27] G. S. Jiang and C. W. Shu, *Efficient implementation of weighted ENO schemes*, J. Comput. Phys., 126 (1996), pp. 202–228.
[28] J. R. Kamm and F. X. Timmes, *On efficient generation of numerically robust Sedov solutions*, Tech. Report LA-UR-07-2849, 2007.
[29] D. Kuzmin, *Monolithic convex limiting for continuous finite element discretizations of hyperbolic conservation laws*, Comput. Methods Appl. Mech. Engrg., 361 (2020), p. 112804.
[30] D. Kuzmin, R. Löhner, and S. Turek, eds., *Flux-Corrected Transport: Principles, Algorithms, and Applications*, Scientific Computation, Springer Netherlands, Dordrecht, 2012.
[31] 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.
[32] 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*, J. Comput. Phys., 344 (2017), pp. 151–186.
[33] B. Perthame and C.-W. Shu, *On positivity preserving finite volume schemesfor Euler equations*, Numer. Math., 73 (1996), pp. 119–130.
[34] P. Roe, *Is discontinuous reconstruction really a good idea?*, J. Sci. Comput., 73 (2017), pp. 1094–1114.
[35] L. I. Sedov, *Similarity and Dimensional Methods in Mechanics*, Academic Press, New York,, 1959.
[36] J. L. Steger and R. F. Warming, *Flux vector splitting of the inviscid gasdynamic equations with application to finite-difference methods*, J. Comput. Phys., 40 (1981), pp. 263–293.
[37] H. Tang, *On the sonic point glitch*, J. Comput. Phys., 202 (2005), pp. 507–532.
[38] E. F. Toro, *Riemann Solvers and Numerical Methods for Fluid Dynamics*, Springer Berlin Heidelberg, 2009.
[39] B. Van Leer, *Towards the ultimate conservative difference scheme. IV. A new approach to numerical convection*, J. Comput. Phys., 23 (1977), pp. 276–299.
[40] B. van Leer, *Flux-vector splitting for the Euler equations*, in Eighth International Conference on Numerical Methods in Fluid Dynamics, E. Krause, ed., Lecture Notes in Physics, Berlin, Heidelberg, 1982, Springer, pp. 507–512.
[41] P. Woodward and P. Colella, *The numerical simulation of two-dimensional fluid flow with strong shocks*, J. Comput. Phys., 54 (1984), pp. 115–173.
[42] K. Wu and C.-W. Shu, *Geometric quasilinearization framework for analysis and design of bound-preserving schemes*, SIAM Rev., 65 (2023), pp. 1031–1073.
[43] Z. Xu, *Parametrized maximum principle preserving flux limiters for high order schemes solving hyperbolic conservation laws: one-dimensional scalar problem*, Math. Comput., 83 (2014), pp. 2213–2238.
[44] 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.
[45] 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.
[46] X. Zhang and C.-W. Shu, *Positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations with source terms*, J. Comput. Phys., 230 (2011), pp. 1238–1248.