ACTIVE FLUX METHODS FOR HYPERBOLIC CONSERVATION LAWS – FLUX VECTOR SPLITTING AND BOUND-PRESERVATION*

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

Abstract. The active flux (AF) method is a compact high-order finite volume method that simultaneously evolves cell averages and point values at cell interfaces. Within the method of lines framework, the existing Jacobian splitting-based point value update incorporates the upwind idea but suffers from a stagnation issue for nonlinear problems due to inaccurate estimation of the upwind direction and also from a mesh alignment issue partially resulting from decoupled point value updates. This paper proposes to use flux vector splitting for the point value update, offering a natural and uniform remedy to those two issues. To improve robustness, this paper also develops bound-preserving (BP) AF methods for 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 is rewritten as a convex combination of 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. In addition, a shock sensor-based limiting is proposed to enhance the convex limiting for the cell average, which can suppress oscillations well. Several challenging tests are conducted to verify the robustness and effectiveness of the BP AF methods, including flow past a forward-facing step and high Mach number jets.

Key words. Hyperbolic conservation laws, active flux, flux vector splitting, bound-preserving, convex limiting, shock sensor

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

1. Introduction.

This paper focuses on the development of robust active flux (AF) methods for hyperbolic conservation laws. The AF method is a new finite volume method [17, 16, 18, 36], that was inspired by [39]. Apart from cell averages, it incorporates additional degrees of freedom (DoFs) as point values located at the cell interfaces, evolved simultaneously with the cell average. The original AF method employs a globally 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 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 [8, 19, 18, 39]. Approximate evolution operators have been explored for Burgers'

*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.

†Corresponding author. Institute of Mathematics, University of Würzburg, Emil-Fischer-Straße 40, 97074 Würzburg, Germany (junming.duan@uni-wuerzburg.de).

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

§Institute of Mathematics, University of Würzburg, Emil-Fischer-Straße 40, 97074 Würzburg, Germany (christian.klingenberg@uni-wuerzburg.de).

equation [17, 16, 36, 5], the compressible Euler equations in one spatial dimension [17, 27, 5], and hyperbolic balance laws [7, 6], 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 [8], and it is naturally well-balanced for acoustics with gravity [7].

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, 3]. 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-based AF methods, e.g., the stagnation issue [27] for nonlinear problems, and mesh alignment issue in 2D to be introduced in Subsection 3.2. Some remedies are suggested for the stagnation issue, e.g., using discontinuous reconstruction [27] or evaluating the upwind direction using more neighboring information [5].

Solutions to hyperbolic conservation laws 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 satisfy a strict maximum principle (MP) [14]. Physically, both the density and pressure in the solutions to the compressible Euler equations should stay positive. It is desired 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 [46, 28, 43], a recent general framework on BP methods [42], and the convex limiting approach [21, 25, 30], which can be traced back to the flux-corrected transport (FCT) schemes for scalar conservation laws [13, 23, 33, 31]. The previous studies on the AF methods pay limited attention to high-speed flows, or problems involving strong discontinuities, with some efforts on the limiting for the point value update, see e.g. [5, 27, 10]. Although those limitings can reduce oscillations, the new cell average may violate the bound even for linear advection [5, 27], and it is not straightforward to extend them to the multi-dimensional case. In [10, 9], the authors proposed to adopt a discontinuous reconstruction based on the scaling limiter [46]. The flux is computed based on the limited point values, resulting in BP AF methods for scalar conservation laws. In a very recent paper, the MOOD [11] based stabilization was adopted to achieve the BP property [4] in an a posteriori fashion.

This paper presents a new way for the point value update to cure the stagnation and mesh alignment issues, develops suitable BP limitings for the AF methods, and also proposes a shock sensor-based limiting to further suppress oscillations. The main contributions and findings in this work can be summarized as follows.

i). We propose to employ the flux vector splitting (FVS) for the point value update, which can cure both the stagnation and the mesh alignment issues effectively, because the FVS couples the neighboring information in a uniform and natural way. The AF method based on the FVS is also shown to give better results than the JS, especially the local Lax-Friedrichs (LLF) FVS, in terms of the CFL number and shock-capturing ability.

ii). We develop BP limitings for both the cell average and point value by blending the high-order AF methods with the first-order LLF method in a convex combination. 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 BP AF methods satisfy the MP for scalar conservation laws, and preserve positive density and pressure for the compressible Euler equations.

iii). We design a shock sensor-based limiting, which helps to reduce oscillations by detecting shock strength. It is shown to strongly improve the shock-capturing ability in the numerical tests.

iv). Several challenging numerical tests are used to demonstrate the robustness and effectiveness of our BP AF methods. Moreover, for the forward-facing step problem, our BP AF method captures small-scale features better compared to the third-order DG method with the TVB limiter on the same mesh resolution, while using fewer DoFs, demonstrating its efficiency and potential for high Mach number flows.

The remainder of this paper is organized as follows. Section 2 introduces the 1D AF methods based on the FVS for the point value update. Section 3 extends our FVS-based AF methods to the 2D case. To design BP methods, Section 4 describes our convex limiting approach for the cell average, and the limiting for the point value. The shock sensor-based limiting is also proposed in Section 4 to suppress oscillations. The 1D limitings can be reduced from the 2D case, and more details are given in Section SM3 in the supplementary material. 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.

2. 1D active flux methods.

This section presents the generalized AF methods using the method of lines for the 1D hyperbolic conservation laws

$$\boldsymbol{U}_t + \boldsymbol{F}(\boldsymbol{U})_x = 0, \quad \boldsymbol{U}(x,0) = \boldsymbol{U}_0(x), \tag{2.1}$$

where $\boldsymbol{U} \in \mathbb{R}^m$ is the vector of m conservative variables, $\boldsymbol{F} \in \mathbb{R}^m$ is the flux function, and $\boldsymbol{U}_0(x)$ is assumed to be initial data of bounded variation. Two cases are of particular interest. The first is a scalar conservation law ($m = 1$)

$$u_t + f(u)_x = 0, \quad u(x,0) = u_0(x). \tag{2.2}$$

The second case is that of compressible Euler equations of gas dynamics with $\boldsymbol{U} = (\rho, \rho v, E)^\top$ and $\boldsymbol{F} = (\rho v, \rho v^2 + p, (E+p)v)^\top$, where 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 perfect gas equation of state (EOS) $p = (\gamma - 1)\rho e$ is used to close the system 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.

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 DoFs 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 DoFs are denoted by

$$\overline{\boldsymbol{U}}_i(t) = \frac{1}{\Delta x_i} \int_{I_i} \boldsymbol{U}_h(x,t)\, \mathrm{d}x, \quad \boldsymbol{U}_{i+\frac{1}{2}}(t) = \boldsymbol{U}_h(x_{i+\frac{1}{2}}, t), \tag{2.3}$$

where $\boldsymbol{U}_h(x,t)$ is the numerical solution. The cell average is updated by integrating

(2.1) over I_i in the following semi-discrete finite volume manner

$$\frac{\mathrm{d}\overline{\boldsymbol{U}}_i}{\mathrm{d}t} = -\frac{1}{\Delta x_i}\left[\boldsymbol{F}(\boldsymbol{U}_{i+\frac{1}{2}}) - \boldsymbol{F}(\boldsymbol{U}_{i-\frac{1}{2}})\right]. \tag{2.4}$$

Thus, the accuracy of (2.4) 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

$$\frac{\mathrm{d}\boldsymbol{U}_{i+\frac{1}{2}}}{\mathrm{d}t} = -\boldsymbol{\mathcal{R}}\left(\boldsymbol{U}_{i+\frac{1}{2}-l_1}(t), \overline{\boldsymbol{U}}_{i+1-l_1}(t), \cdots, \overline{\boldsymbol{U}}_{i+l_2}(t), \boldsymbol{U}_{i+\frac{1}{2}+l_2}(t)\right),\ l_1, l_2 \geqslant 0, \tag{2.5}$$

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. Stagnation issue when using Jacobian splitting. Let us first briefly describe the point value update based on the JS [2], which reads

$$\frac{\mathrm{d}\boldsymbol{U}_{i+\frac{1}{2}}}{\mathrm{d}t} = -\left[\boldsymbol{J}^+(\boldsymbol{U}_{i+\frac{1}{2}})\boldsymbol{D}^+_{i+\frac{1}{2}}(\boldsymbol{U}) + \boldsymbol{J}^-(\boldsymbol{U}_{i+\frac{1}{2}})\boldsymbol{D}^-_{i+\frac{1}{2}}(\boldsymbol{U})\right], \tag{2.6}$$

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

$$\boldsymbol{J}^\pm = \boldsymbol{R}\boldsymbol{\Lambda}^\pm\boldsymbol{R}^{-1},\ \boldsymbol{\Lambda}^\pm = \mathrm{diag}\{\lambda_1^\pm, \ldots, \lambda_m^\pm\},$$

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, and $a^+ = \max\{a, 0\}$, $a^- = \min\{a, 0\}$. To derive the approximation of the derivatives in (2.6), 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_i is

$$\begin{aligned}\boldsymbol{U}_{\mathtt{para},1}(x) = &-3(2\overline{\boldsymbol{U}}_i - \boldsymbol{U}_{i-\frac{1}{2}} - \boldsymbol{U}_{i+\frac{1}{2}})\frac{x^2}{\Delta x_i^2} + (\boldsymbol{U}_{i+\frac{1}{2}} - \boldsymbol{U}_{i-\frac{1}{2}})\frac{x}{\Delta x_i} \\ &+ \frac{1}{4}(6\overline{\boldsymbol{U}}_i - \boldsymbol{U}_{i-\frac{1}{2}} - \boldsymbol{U}_{i+\frac{1}{2}})\end{aligned} \tag{2.7}$$

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

$$\boldsymbol{D}^+_{i+\frac{1}{2}}(\boldsymbol{U}) = \boldsymbol{U}'_{\mathtt{para},1}(\Delta x_i/2) = \frac{1}{\Delta x_i}\left(2\boldsymbol{U}_{i-\frac{1}{2}} - 6\overline{\boldsymbol{U}}_i + 4\boldsymbol{U}_{i+\frac{1}{2}}\right), \tag{2.8a}$$

$$\boldsymbol{D}^-_{i+\frac{1}{2}}(\boldsymbol{U}) = \frac{1}{\Delta x_{i+1}}\left(-4\boldsymbol{U}_{i+\frac{1}{2}} + 6\overline{\boldsymbol{U}}_{i+1} - 2\boldsymbol{U}_{i+\frac{3}{2}}\right). \tag{2.8b}$$

One of the deficiencies of using the JS is the stagnation issue that appears in certain setups for nonlinear problems, as observed in [27, 5]. As shown in Example 5.1 for Burgers' equation, the numerical solution based on the JS without limiting gives a spike in the cell average at the initial discontinuity $x = 0.2$, which grows linearly in time. The reason for this behavior is the inaccurate estimation of the upwind direction at the cell interface, required to split the Jacobian in (2.6). In this example, there are two successive point values with different signs near the initial discontinuity: $u_{i-\frac{1}{2}} = 2$, $u_{i+\frac{1}{2}} = -1$. At the cell interface $x_{i-\frac{1}{2}}$ or $x_{i+\frac{1}{2}}$, depending on the details of

initialization, the upwind discretization in (2.8) 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 (2.4), $\bar{u}_i$ increases gradually (which is the observed spike). Proposed solutions to handle the stagnation issue involve estimating the Jacobian not only at the relevant cell interface, but also at the neighboring interfaces, and to select a better upwind direction (e.g. [5]), or achieve the same by blending (e.g. [9]). As will be shown below, using FVS instead of the JS naturally has a similar effect.

2.2. Point value update using flux vector splitting. In this paper, we propose to use the FVS for the point value update, which was originally used to identify the upwind directions, and is simpler and somewhat more efficient than Godunov-type methods for solving hyperbolic systems [38]. The FVS for the point value update reads

$$\frac{\mathrm{d}\boldsymbol{U}_{i+\frac{1}{2}}}{\mathrm{d}t} = -\left[\widetilde{\boldsymbol{D}}^{+}\boldsymbol{F}^{+}(\boldsymbol{U}) + \widetilde{\boldsymbol{D}}^{-}\boldsymbol{F}^{-}(\boldsymbol{U})\right]_{i+\frac{1}{2}}, \tag{2.9}$$

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

$$\lambda\left(\frac{\partial \boldsymbol{F}^{+}}{\partial \boldsymbol{U}}\right) \geqslant 0, \quad \lambda\left(\frac{\partial \boldsymbol{F}^{-}}{\partial \boldsymbol{U}}\right) \leqslant 0, \tag{2.10}$$

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 (2.10), to be discussed later. Finite difference formulae to approximate the flux derivatives are obtained as follows. From the reconstruction of $\boldsymbol{U}$ (2.7), one can evaluate the flux $\boldsymbol{F}$, and also the split fluxes $\boldsymbol{F}^{\pm}$ pointwise. We compute them at the endpoints of the cell and in the middle. Then a parabolic reconstruction for, say, $\boldsymbol{F}^{+}$ in the cell I_i is obtained 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 $\boldsymbol{F}^{+}_{\mathtt{para},2}(\pm\Delta x_i/2) = \boldsymbol{F}^{+}_{i\pm\frac{1}{2}} = \boldsymbol{F}^{+}(\boldsymbol{U}_{i\pm\frac{1}{2}})$, and $\boldsymbol{F}^{+}_{\mathtt{para},2}(0) = \boldsymbol{F}^{+}_{i} = \boldsymbol{F}^{+}(\boldsymbol{U}_i)$. The cell-centered point value is $\boldsymbol{U}_i = (-\boldsymbol{U}_{i-\frac{1}{2}} + 6\overline{\boldsymbol{U}}_i - \boldsymbol{U}_{i+\frac{1}{2}})/4$. Then the discrete derivatives are

$$\left(\widetilde{\boldsymbol{D}}^{+}\boldsymbol{F}^{+}\right)_{i+\frac{1}{2}} = \left(\boldsymbol{F}^{+}_{\mathtt{para},2}\right)'(\Delta x_i/2) = \frac{1}{\Delta x_i}\left(\boldsymbol{F}^{+}_{i-\frac{1}{2}} - 4\boldsymbol{F}^{+}_{i} + 3\boldsymbol{F}^{+}_{i+\frac{1}{2}}\right), \tag{2.11a}$$

$$\left(\widetilde{\boldsymbol{D}}^{-}\boldsymbol{F}^{-}\right)_{i+\frac{1}{2}} = \frac{1}{\Delta x_{i+1}}\left(-3\boldsymbol{F}^{-}_{i+\frac{1}{2}} + 4\boldsymbol{F}^{-}_{i+1} - \boldsymbol{F}^{-}_{i+\frac{3}{2}}\right). \tag{2.11b}$$

These finite differences are third-order accurate. While the reconstructions of both $\boldsymbol{U}$ and $\boldsymbol{F}$ are parabolic, the coefficients in the formula (2.11) differ from that in [2] because (2.11) uses the cell-centered value rather than the cell average.

The FVS-based point value update borrows the information from the neighbors naturally, and can eliminate the generation of the spike effectively, as shown in Figure 3, similar to the idea of the remedy in [5]. Note that we still use the original continuous reconstruction in the AF methods. We remark that, in AF methods, it is not clear how to define the point values at discontinuities, thus there may be other methods to fix the stagnation issue.

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 (2.10) across the spatial stencil. In our implementation, it is determined by

$$\alpha_{i+\frac{1}{2}} = \max_r \{\varrho(\boldsymbol{U}_r)\}, \ r \in \left\{i - \frac{1}{2}, i, i + \frac{1}{2}, i + 1, i + \frac{3}{2}\right\}, \tag{2.12}$$

where ϱ is the spectral radius of $\partial \boldsymbol{F}/\partial \boldsymbol{U}$. One can also choose α to be the maximal spectral radius 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,

$$\boldsymbol{F}^{\pm} = \frac{1}{2}(\boldsymbol{F}(\boldsymbol{U}) \pm |\boldsymbol{J}|\boldsymbol{U}), \quad |\boldsymbol{J}| = \boldsymbol{R}(\boldsymbol{\Lambda}^{+} - \boldsymbol{\Lambda}^{-})\boldsymbol{R}^{-1}. \tag{2.13}$$

Note that we evaluate the Jacobian at three locations in the cell I_i to get corresponding $\boldsymbol{F}^{+}$. For linear systems, one has $\boldsymbol{F} = \boldsymbol{J}\boldsymbol{U}$, so (2.13) reduces to the JS, because in this case

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

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, for linear systems, the AF methods using this FVS are the same as the original JS-based AF methods.

Such an FVS is also known as the Steger-Warming (SW) FVS [37] for the Euler equations, since the "homogeneity property" $\boldsymbol{F} = \boldsymbol{J}\boldsymbol{U}$ holds [38]. 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.

Remark 2.1. It should be noted that $\boldsymbol{F}^{\pm}$ in this FVS may not be differentiable with respect to $\boldsymbol{U}$ for nonlinear systems (e.g. Euler), as the splitting is based on the absolute value. In [40], the mass flux of $\boldsymbol{F}^{\pm}$ is shown to be not differentiable, which might explain the accuracy degradation in Example 5.7.

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 [26]. 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.

Remark 2.2. Different FVS may lead to different stability conditions but it is difficult to perform the analysis theoretically. We provide experimental CFL numbers for different FVS in some 1D tests. Our numerical tests in Section 5 show that the AF methods based on the FVS generally give better results than the JS, and the LLF FVS is the best among all the three FVS in terms of the CFL number and non-oscillatory property for high-speed flows involving strong discontinuities.

2.3. Time discretization. The fully-discrete scheme is obtained by using the SSP-RK3 method [20]

$$\begin{aligned} \boldsymbol{U}^{*} &= \boldsymbol{U}^{n} + \Delta t^{n} \boldsymbol{L}\left(\boldsymbol{U}^{n}\right), \\ \boldsymbol{U}^{**} &= \frac{3}{4}\boldsymbol{U}^{n} + \frac{1}{4}\left(\boldsymbol{U}^{*} + \Delta t^{n} \boldsymbol{L}\left(\boldsymbol{U}^{*}\right)\right), \\ \boldsymbol{U}^{n+1} &= \frac{1}{3}\boldsymbol{U}^{n} + \frac{2}{3}\left(\boldsymbol{U}^{**} + \Delta t^{n} \boldsymbol{L}\left(\boldsymbol{U}^{**}\right)\right), \end{aligned} \tag{2.14}$$

where $\boldsymbol{L}$ is the right-hand side of the semi-discrete schemes (2.4) or (2.5). The time step size is determined by the usual CFL condition

$$\Delta t^{n} = \frac{C_{\mathtt{CFL}}}{\max\limits_{i}\{\varrho(\overline{\boldsymbol{U}}_i)/\Delta x_i\}}. \tag{2.15}$$

3. 2D active flux methods on Cartesian meshes. This section presents the generalized AF methods using the method of lines for the 2D hyperbolic conservation laws

$$\boldsymbol{U}_t + \boldsymbol{F}_1(\boldsymbol{U})_x + \boldsymbol{F}_2(\boldsymbol{U})_y = 0, \quad \boldsymbol{U}(x,y,0) = \boldsymbol{U}_0(x,y). \tag{3.1}$$

We will consider the scalar conservation law

$$u_t + f_1(u)_x + f_2(u)_y = 0, \quad u(x,y,0) = u_0(x,y), \tag{3.2}$$

and the Euler equations with $\boldsymbol{U} = (\rho, \rho\boldsymbol{v}, E)^{\intercal}$, $\boldsymbol{F}_1 = (\rho v_1, \rho v_1^2 + p, \rho v_1 v_2, (E+p)v_1)^{\intercal}$, $\boldsymbol{F}_2 = (\rho v_2, \rho v_1 v_2, \rho v_2^2 + p, (E+p)v_2)^{\intercal}$, where $\boldsymbol{v} = (v_1, v_2)$ is the velocity vector, and the other notations are the same as for 1D in Section 2. The SSP-RK3 method is used to obtain the fully-discrete method.

Assume that a 2D computational domain is divided into $N_1 \times N_2$ cells, $I_{i,j} = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}]$ with the cell sizes $\Delta x_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}, \Delta y_j = y_{j+\frac{1}{2}} - y_{j-\frac{1}{2}}$, and cell centers $(x_i, y_j) = \left(\frac{1}{2}(x_{i-\frac{1}{2}} + x_{i+\frac{1}{2}}), \frac{1}{2}(y_{j-\frac{1}{2}} + y_{j+\frac{1}{2}})\right)$, $i = 1, \cdots, N_1$, $j = 1, \cdots, N_2$. The DoFs consist of the cell average $\overline{\boldsymbol{U}}_{i,j}(t) = \frac{1}{\Delta x_i \Delta y_j}\int_{I_{i,j}} \boldsymbol{U}_h(x,y,t)\,\mathrm{d}x\mathrm{d}y$, the face-centered values $\boldsymbol{U}_{i+\frac{1}{2},j}(t) = \boldsymbol{U}_h(x_{i+\frac{1}{2}}, y_j, t), \boldsymbol{U}_{i,j+\frac{1}{2}}(t) = \boldsymbol{U}_h(x_i, y_{j+\frac{1}{2}}, t)$, and the value at the corner $\boldsymbol{U}_{i+\frac{1}{2},j+\frac{1}{2}}(t) = \boldsymbol{U}_h(x_{i+\frac{1}{2}}, y_{j+\frac{1}{2}}, t)$, where $\boldsymbol{U}_h(x,y,t)$ is the numerical solution. A sketch of the DoFs for the third-order AF method (for the scalar case) is given in Figure 1.

The cell average is evolved as follows

$$\frac{\mathrm{d}\overline{\boldsymbol{U}}_{i,j}}{\mathrm{d}t} = -\frac{1}{\Delta x_i}\left(\widehat{\boldsymbol{F}}_{i+\frac{1}{2},j} - \widehat{\boldsymbol{F}}_{i-\frac{1}{2},j}\right) - \frac{1}{\Delta y_j}\left(\widehat{\boldsymbol{F}}_{i,j+\frac{1}{2}} - \widehat{\boldsymbol{F}}_{i,j-\frac{1}{2}}\right), \tag{3.3}$$

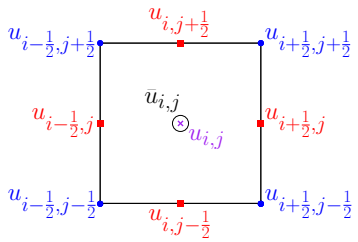

Fig. 1: The DoFs for the third-order AF method: cell average (circle), face-centered values (squares), values at corners (dots). Note that the cell-centered point value $u_{i,j}$ (cross) is used in constructing the scheme, but does not belong to the DoFs.

where $\widehat{\boldsymbol{F}}_{i+\frac{1}{2},j}$ and $\widehat{\boldsymbol{F}}_{i,j+\frac{1}{2}}$ are the numerical fluxes

(3.4)
$$\widehat{\boldsymbol{F}}_{i+\frac{1}{2},j} = \frac{1}{\Delta y_j}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \boldsymbol{F}_1(\boldsymbol{U}_h(x_{i+\frac{1}{2}},y))\,\mathrm{d}y,\ \widehat{\boldsymbol{F}}_{i,j+\frac{1}{2}} = \frac{1}{\Delta x_i}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \boldsymbol{F}_2(\boldsymbol{U}_h(x,y_{j+\frac{1}{2}}))\,\mathrm{d}x.$$

To achieve third-order accuracy, one can use Simpson's rule

$$\widehat{\boldsymbol{F}}_{i+\frac{1}{2},j} = \frac{1}{6}\Big(\boldsymbol{F}_1(\boldsymbol{U}_{i+\frac{1}{2},j-\frac{1}{2}}) + 4\boldsymbol{F}_1(\boldsymbol{U}_{i+\frac{1}{2},j}) + \boldsymbol{F}_1(\boldsymbol{U}_{i+\frac{1}{2},j+\frac{1}{2}})\Big). \tag{3.5}$$

3.1. Point value update using flux vector splitting. For the evolution of the point values, consider the following general form

$$\frac{\mathrm{d}\boldsymbol{U}_\sigma}{\mathrm{d}t} = -\boldsymbol{\mathcal{R}}\left(\overline{U}_c(t), \boldsymbol{U}_{\sigma'}(t)\right),\ c\in\mathcal{C}(\sigma), \sigma'\in\Sigma(\sigma), \tag{3.6}$$

where $\boldsymbol{\mathcal{R}}$ is a consistent approximation of $\partial\boldsymbol{F}_1/\partial x + \partial\boldsymbol{F}_2/\partial y$ at the point σ, $\mathcal{C}(\sigma)$ and $\Sigma(\sigma)$ are the spatial stencils containing the cell averages and point values, respectively. One can use the JS in [3], or employ the FVS for the point value update. E.g. for the point value at the corner $(x_{i+\frac{1}{2}}, y_{j+\frac{1}{2}})$ the FVS-based update reads

$$\frac{\mathrm{d}\boldsymbol{U}_{i+\frac{1}{2},j+\frac{1}{2}}}{\mathrm{d}t} = -\sum_{\ell=1}^{2}\left[\widetilde{\boldsymbol{D}}_\ell^+\boldsymbol{F}_\ell^+(\boldsymbol{U}) + \widetilde{\boldsymbol{D}}_\ell^-\boldsymbol{F}_\ell^-(\boldsymbol{U})\right]_{i+\frac{1}{2},j+\frac{1}{2}}, \tag{3.7}$$

where the fluxes are split as $\boldsymbol{F}_\ell = \boldsymbol{F}_\ell^+ + \boldsymbol{F}_\ell^-$, $\lambda\left(\frac{\partial\boldsymbol{F}_\ell^+}{\partial\boldsymbol{U}}\right)\geqslant 0$, $\lambda\left(\frac{\partial\boldsymbol{F}_\ell^-}{\partial\boldsymbol{U}}\right)\leqslant 0$. The explicit expressions of the 2D FVS used in this paper can be found in the supplementary material Section SM1. The finite difference operators $\widetilde{\boldsymbol{D}}_1^\pm$ and $\widetilde{\boldsymbol{D}}_2^\pm$ can be obtained similarly to Subsection 2.2. For third-order accuracy, starting with a bi-parabolic reconstruction of $\boldsymbol{U}$ and computing a bi-parabolic interpolation of $\boldsymbol{F}_\ell^\pm$, one thus obtains $\widetilde{\boldsymbol{D}}_1^\pm$ in the x-direction as

$$\left(\widetilde{\boldsymbol{D}}_1^+\boldsymbol{F}_1^+\right)_{i+\frac{1}{2},j+\frac{1}{2}} = \frac{1}{\Delta x_i}\left((\boldsymbol{F}_1)^+_{i-\frac{1}{2},j+\frac{1}{2}} - 4(\boldsymbol{F}_1)^+_{i,j+\frac{1}{2}} + 3(\boldsymbol{F}_1)^+_{i+\frac{1}{2},j+\frac{1}{2}}\right),$$

$$\left(\widetilde{\boldsymbol{D}}_1^-\boldsymbol{F}_1^-\right)_{i+\frac{1}{2},j+\frac{1}{2}} = \frac{1}{\Delta x_{i+1}}\left(-3(\boldsymbol{F}_1)^-_{i+\frac{1}{2},j+\frac{1}{2}} + 4(\boldsymbol{F}_1)^-_{i+1,j+\frac{1}{2}} - (\boldsymbol{F}_1)^-_{i+\frac{3}{2},j+\frac{1}{2}}\right).$$

For the face-centered point value at $(x_{i+\frac{1}{2}}, y_j)$, the FVS-based update reads

$$\frac{\mathrm{d}\boldsymbol{U}_{i+\frac{1}{2},j}}{\mathrm{d}t} = -\left[\widetilde{\boldsymbol{D}}_1^+\boldsymbol{F}_1^+(\boldsymbol{U}) + \widetilde{\boldsymbol{D}}_1^-\boldsymbol{F}_1^-(\boldsymbol{U})\right]_{i+\frac{1}{2},j} - \left(\widetilde{\boldsymbol{D}}_2\boldsymbol{F}_2(\boldsymbol{U})\right)_{i+\frac{1}{2},j}, \tag{3.9}$$

where

$$\left(\widetilde{\boldsymbol{D}}_1^+\boldsymbol{F}_1^+\right)_{i+\frac{1}{2},j} = \frac{1}{\Delta x_i}\left((\boldsymbol{F}_1)^+_{i-\frac{1}{2},j} - 4(\boldsymbol{F}_1)^+_{i,j} + 3(\boldsymbol{F}_1)^+_{i+\frac{1}{2},j}\right),$$

$$\left(\widetilde{\boldsymbol{D}}_1^-\boldsymbol{F}_1^-\right)_{i+\frac{1}{2},j} = \frac{1}{\Delta x_{i+1}}\left(-3(\boldsymbol{F}_1)^-_{i+\frac{1}{2},j} + 4(\boldsymbol{F}_1)^-_{i+1,j} - (\boldsymbol{F}_1)^-_{i+\frac{3}{2},j}\right),$$

$$\left(\widetilde{\boldsymbol{D}}_2\boldsymbol{F}_2\right)_{i+\frac{1}{2},j} = \frac{1}{\Delta y_j}\left((\boldsymbol{F}_2)_{i+\frac{1}{2},j+\frac{1}{2}} - (\boldsymbol{F}_2)_{i+\frac{1}{2},j-\frac{1}{2}}\right),$$

and the cell-centered point value is computed from the bi-parabolic reconstruction [3] as

$$\boldsymbol{U}_{i,j} = \frac{1}{16}\Big[36\overline{\boldsymbol{U}}_{i,j} - 4\left(\boldsymbol{U}_{i-\frac{1}{2},j} + \boldsymbol{U}_{i+\frac{1}{2},j} + \boldsymbol{U}_{i,j-\frac{1}{2}} + \boldsymbol{U}_{i,j+\frac{1}{2}}\right) \tag{3.11}$$

$$- \left(\boldsymbol{U}_{i-\frac{1}{2},j-\frac{1}{2}} + \boldsymbol{U}_{i+\frac{1}{2},j-\frac{1}{2}} + \boldsymbol{U}_{i-\frac{1}{2},j+\frac{1}{2}} + \boldsymbol{U}_{i+\frac{1}{2},j+\frac{1}{2}}\right)\Big]. \tag{3.12}$$

The update for the point value at $(x_i, y_{j+\frac{1}{2}})$ is omitted here, which is similar to (3.9).

3.2. Mesh alignment issue when using Jacobian splitting. The mesh alignment issue was observed for the fully-discrete AF methods in [34], where the convergence rate reduces to 2 for the linear advection problem, when the advection velocity is aligned with the grid. For the generalized AF methods based on the JS, such an issue is also observed. Consider Example 5.8, where we solve a quasi-2D Sod shock tube along the x-direction on a 100×2 uniform mesh. As shown in Figure 10, the density based on the JS shows large deviations between the contact discontinuity and shock wave. From Figure 11, it can be seen that the solutions of the DoFs at the corner $(x_{i+\frac{1}{2}}, y_{j+\frac{1}{2}})$ and horizontal face $(x_i, y_{j+\frac{1}{2}})$ are decoupled from that at the vertical face $(x_{i+\frac{1}{2}}, y_j)$ and cell averages. The reason is complicated because the mesh alignment issue seems to be caused by the decoupled point value update and its interaction with the JS.

3.3. Boundary treatment. The numerical boundary conditions can be implemented using ghost cells as usual finite volume methods. Take the reflective boundary for the Euler equations as an example. Let $x = x_{N_1-\frac{1}{2}}$ be the boundary, then the cell averages and point values in the ghost cell $I_{N_1,j}$ are given by

$$\overline{\boldsymbol{U}}_{N_1,j} = \mathcal{M}(\overline{\boldsymbol{U}}_{N_1-1,j}),\ \boldsymbol{U}_{N_1+\frac{1}{2},j} = \mathcal{M}(\boldsymbol{U}_{N_1-\frac{3}{2},j}),$$

$$\boldsymbol{U}_{N_1,j-\frac{1}{2}} = \mathcal{M}(\boldsymbol{U}_{N_1-1,j-\frac{1}{2}}),\ \boldsymbol{U}_{N_1+\frac{1}{2},j-\frac{1}{2}} = \mathcal{M}(\boldsymbol{U}_{N_1-\frac{3}{2},j-\frac{1}{2}}),$$

where $\mathcal{M}$ reverses the sign of the ρv_1 component while keeping others unchanged. Then the point value update at the boundary can be computed in the same way as the interior points, but the numerical flux on the boundary for the cell average is computed through the LLF flux as suggested in [4]. For instance, the flux $\boldsymbol{F}_1(\boldsymbol{U}_{N_1-\frac{1}{2},j-\frac{1}{2}})$ in the right-hand side of (3.5) is replaced by $\widehat{\boldsymbol{F}}_1^{\text{LLF}}(\boldsymbol{U}_{N_1-1,j-\frac{1}{2}}, \mathcal{M}(\boldsymbol{U}_{N_1-1,j-\frac{1}{2}}))$.

4. 2D bound-preserving active flux methods. In this paper, the admissible state set $\mathcal{G}$ is assumed to be convex. Two cases are considered. For the scalar conservation law (3.2), its solutions satisfy a strict maximum principle (MP) [14], i.e.,

$$\mathcal{G} = \{u \mid m_0 \leqslant u \leqslant M_0\}, \quad m_0 = \min_{x,y} u_0(x,y),\ M_0 = \max_{x,y} u_0(x,y). \tag{4.1}$$

For the compressible Euler equations, the admissible state set is

$$\mathcal{G} = \left\{ \boldsymbol{U} = (\rho, \rho\boldsymbol{v}, E) \;\middle|\; \rho > 0,\ p = (\gamma - 1)\left(E - \|\rho\boldsymbol{v}\|^2/(2\rho)\right) > 0 \right\}, \tag{4.2}$$

which is convex, see e.g. [47].

DEFINITION 4.1. *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.*

Note that to avoid the effect of the round-off error, we need to choose the desired lower bounds for the density and pressure. In the numerical tests, we will enforce $\rho \geqslant \varepsilon^\rho$, $p \geqslant \varepsilon^p$ with $\varepsilon^\rho, \varepsilon^p$ to be defined later. Since the DoFs 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 [4]. The 1D limitings can be reduced from this Section, given in Section SM3 in the supplementary material.

4.1. Convex limiting for the cell average. 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 [21, 25, 30] is to enforce the preservation of local or 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 [21, 24]. The low-order scheme is chosen as the first-order LLF scheme

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

where $\widehat{\boldsymbol{F}}^{\mathrm{L}}_{i+\frac{1}{2},j}$ and $\widehat{\boldsymbol{F}}^{\mathrm{L}}_{i,j+\frac{1}{2}}$ are the LLF fluxes. Take the x-direction as an example,

$$\begin{aligned}
\widehat{\boldsymbol{F}}^{\mathrm{L}}_{i+\frac{1}{2},j} &:= \widehat{\boldsymbol{F}}^{\mathrm{LLF}}_1(\overline{\boldsymbol{U}}^{n}_{i,j}, \overline{\boldsymbol{U}}^{n}_{i+1,j}) \\
&= \frac{1}{2}\left(\boldsymbol{F}_1(\overline{\boldsymbol{U}}^{n}_{i,j}) + \boldsymbol{F}_1(\overline{\boldsymbol{U}}^{n}_{i+1,j})\right) - \frac{(\alpha_1)_{i+\frac{1}{2},j}}{2}\left(\overline{\boldsymbol{U}}^{n}_{i+1,j} - \overline{\boldsymbol{U}}^{n}_{i,j}\right), \\
(\alpha_1)_{i+\frac{1}{2},j} &= \max\{\varrho_1(\overline{\boldsymbol{U}}^{n}_{i,j}),\ \varrho_1(\overline{\boldsymbol{U}}^{n}_{i+1,j})\}, \\
\mu_{1,i} &= \Delta t^n/\Delta x_i,
\end{aligned} \tag{4.3}$$

where ϱ_1 is the spectral radius of $\partial\boldsymbol{F}_1/\partial\boldsymbol{U}$. Note that here $\alpha_{i+\frac{1}{2},j}$ is not the same as the one in the LLF FVS. Following [22], the first-order LLF scheme can be rewritten as

$$\begin{aligned}
\overline{\boldsymbol{U}}^{\mathrm{L}}_{i,j} = &\left[1 - \mu_{1,i}\left((\alpha_1)_{i-\frac{1}{2},j} + (\alpha_1)_{i+\frac{1}{2},j}\right) - \mu_{2,j}\left((\alpha_2)_{i,j-\frac{1}{2}} + (\alpha_2)_{i,j+\frac{1}{2}}\right)\right]\overline{\boldsymbol{U}}^{n}_{i,j} \\
&+ \mu_{1,i}(\alpha_1)_{i-\frac{1}{2},j}\widetilde{\boldsymbol{U}}_{i-\frac{1}{2},j} + \mu_{1,i}(\alpha_1)_{i+\frac{1}{2},j}\widetilde{\boldsymbol{U}}_{i+\frac{1}{2},j} \\
&+ \mu_{2,j}(\alpha_2)_{i,j-\frac{1}{2}}\widetilde{\boldsymbol{U}}_{i,j-\frac{1}{2}} + \mu_{2,j}(\alpha_2)_{i,j+\frac{1}{2}}\widetilde{\boldsymbol{U}}_{i,j+\frac{1}{2}},
\end{aligned} \tag{4.4}$$

with four intermediate states, and the explicit expressions in the x-direction are

$$\widetilde{\boldsymbol{U}}_{i\pm\frac{1}{2},j} = \frac{1}{2}\left(\overline{\boldsymbol{U}}^{n}_{i,j} + \overline{\boldsymbol{U}}^{n}_{i\pm1,j}\right) \mp \frac{1}{2(\alpha_1)_{i\pm\frac{1}{2},j}}\left[\boldsymbol{F}_1(\overline{\boldsymbol{U}}^{n}_{i,j}) - \boldsymbol{F}_1(\overline{\boldsymbol{U}}^{n}_{i\pm1,j})\right]. \tag{4.5}$$

The proofs of $\widetilde{\boldsymbol{U}}_{i\pm\frac{1}{2},j}, \widetilde{\boldsymbol{U}}_{i,j\pm\frac{1}{2}} \in \mathcal{G}$ are given in the supplementary material Section SM2, for the scalar case and Euler equations.

LEMMA 4.2. *If the time step size Δt^n satisfies*

$$\Delta t^n \leqslant \frac{1}{2}\min\left\{\frac{\Delta x_i}{(\alpha_1)_{i-\frac{1}{2},j}+(\alpha_1)_{i+\frac{1}{2},j}}, \frac{\Delta y_j}{(\alpha_2)_{i,j-\frac{1}{2}}+(\alpha_2)_{i,j+\frac{1}{2}}}\right\}, \tag{4.6}$$

then (4.4) is a convex combination, and the first-order LLF scheme is BP.

The proof (see e.g. [22, 35]) relies on $\overline{\boldsymbol{U}}^n_{i,j}, \widetilde{\boldsymbol{U}}_{i\pm\frac{1}{2},j}, \widetilde{\boldsymbol{U}}_{i,j\pm\frac{1}{2}} \in \mathcal{G}$ and the convexity of $\mathcal{G}$.

Upon defining the anti-diffusive flux $\Delta\widehat{\boldsymbol{F}}_{i\pm\frac{1}{2},j} = \widehat{\boldsymbol{F}}^{\mathtt{H}}_{i\pm\frac{1}{2},j} - \widehat{\boldsymbol{F}}^{\mathtt{L}}_{i\pm\frac{1}{2},j}$, and $\widehat{\boldsymbol{F}}^{\mathtt{H}}_{i\pm\frac{1}{2},j}$ is given in (3.4), a forward-Euler step applied to the semi-discrete high-order scheme for the cell average (3.3) can be written as

$$\begin{aligned}
\overline{\boldsymbol{U}}^{\mathtt{H}}_{i,j} &= \overline{\boldsymbol{U}}^n_{i,j} - \mu_{1,i}\left(\widehat{\boldsymbol{F}}^{\mathtt{L}}_{i+\frac{1}{2},j} - \widehat{\boldsymbol{F}}^{\mathtt{L}}_{i-\frac{1}{2},j}\right) - \mu_{2,j}\left(\widehat{\boldsymbol{F}}^{\mathtt{L}}_{i,j+\frac{1}{2}} - \widehat{\boldsymbol{F}}^{\mathtt{L}}_{i,j-\frac{1}{2}}\right)\\
&\quad - \mu_{1,i}\left(\Delta\widehat{\boldsymbol{F}}_{i+\frac{1}{2},j} - \Delta\widehat{\boldsymbol{F}}_{i-\frac{1}{2},j}\right) - \mu_{2,j}\left(\Delta\widehat{\boldsymbol{F}}_{i,j+\frac{1}{2}} - \Delta\widehat{\boldsymbol{F}}_{i,j-\frac{1}{2}}\right)\\
&= \left[1 - \mu_{1,i}\left((\alpha_1)_{i-\frac{1}{2},j} + (\alpha_1)_{i+\frac{1}{2},j}\right) - \mu_{2,j}\left((\alpha_2)_{i,j-\frac{1}{2}} + (\alpha_2)_{i,j+\frac{1}{2}}\right)\right]\overline{\boldsymbol{U}}^n_{i,j}\\
&\quad + \mu_{1,i}(\alpha_1)_{i-\frac{1}{2},j}\widetilde{\boldsymbol{U}}^{\mathtt{H},+}_{i-\frac{1}{2},j} + \mu_{1,i}(\alpha_1)_{i+\frac{1}{2},j}\widetilde{\boldsymbol{U}}^{\mathtt{H},-}_{i+\frac{1}{2},j}\\
&\quad + \mu_{2,j}(\alpha_2)_{i,j-\frac{1}{2}}\widetilde{\boldsymbol{U}}^{\mathtt{H},+}_{i,j-\frac{1}{2}} + \mu_{2,j}(\alpha_2)_{i,j+\frac{1}{2}}\widetilde{\boldsymbol{U}}^{\mathtt{H},-}_{i,j+\frac{1}{2}},
\end{aligned} \tag{4.7}$$

with the high-order intermediate states

$$\widetilde{\boldsymbol{U}}^{\mathtt{H},\mp}_{i\pm\frac{1}{2},j} := \widetilde{\boldsymbol{U}}_{i\pm\frac{1}{2},j} \mp \frac{\Delta\widehat{\boldsymbol{F}}_{i\pm\frac{1}{2},j}}{(\alpha_1)_{i\pm\frac{1}{2},j}}, \quad \widetilde{\boldsymbol{U}}^{\mathtt{H},\mp}_{i,j\pm\frac{1}{2}} := \widetilde{\boldsymbol{U}}_{i,j\pm\frac{1}{2}} \mp \frac{\Delta\widehat{\boldsymbol{F}}_{i,j\pm\frac{1}{2}}}{(\alpha_2)_{i,j\pm\frac{1}{2}}}.$$

With the low-order scheme (4.4) and high-order scheme (4.7) having the same abstract form, one can blend them to define the limited scheme for the cell average as

$$\begin{aligned}
\overline{\boldsymbol{U}}^{\mathtt{Lim}}_{i,j} &= \left[1 - \mu_{1,i}\left((\alpha_1)_{i-\frac{1}{2},j} + (\alpha_1)_{i+\frac{1}{2},j}\right) - \mu_{2,j}\left((\alpha_2)_{i,j-\frac{1}{2}} + (\alpha_2)_{i,j+\frac{1}{2}}\right)\right]\overline{\boldsymbol{U}}^n_{i,j}\\
&\quad + \mu_{1,i}(\alpha_1)_{i-\frac{1}{2},j}\widetilde{\boldsymbol{U}}^{\mathtt{Lim},+}_{i-\frac{1}{2},j} + \mu_{1,i}(\alpha_1)_{i+\frac{1}{2},j}\widetilde{\boldsymbol{U}}^{\mathtt{Lim},-}_{i+\frac{1}{2},j}\\
&\quad + \mu_{2,j}(\alpha_2)_{i,j-\frac{1}{2}}\widetilde{\boldsymbol{U}}^{\mathtt{Lim},+}_{i,j-\frac{1}{2}} + \mu_{2,j}(\alpha_2)_{i,j+\frac{1}{2}}\widetilde{\boldsymbol{U}}^{\mathtt{Lim},-}_{i,j+\frac{1}{2}},
\end{aligned} \tag{4.8}$$

where the limited intermediate states are

$$\begin{aligned}
\widetilde{\boldsymbol{U}}^{\mathtt{Lim},\mp}_{i\pm\frac{1}{2},j} &= \widetilde{\boldsymbol{U}}_{i\pm\frac{1}{2},j} \mp \frac{\Delta\widehat{\boldsymbol{F}}^{\mathtt{Lim}}_{i\pm\frac{1}{2},j}}{(\alpha_1)_{i\pm\frac{1}{2},j}} := \widetilde{\boldsymbol{U}}_{i\pm\frac{1}{2},j} \mp \frac{\theta_{i\pm\frac{1}{2},j}\Delta\widehat{\boldsymbol{F}}_{i\pm\frac{1}{2},j}}{(\alpha_1)_{i\pm\frac{1}{2},j}},\\
\widetilde{\boldsymbol{U}}^{\mathtt{Lim},\mp}_{i,j\pm\frac{1}{2}} &= \widetilde{\boldsymbol{U}}_{i,j\pm\frac{1}{2}} \mp \frac{\Delta\widehat{\boldsymbol{F}}^{\mathtt{Lim}}_{i,j\pm\frac{1}{2}}}{(\alpha_2)_{i,j\pm\frac{1}{2}}} := \widetilde{\boldsymbol{U}}_{i,j\pm\frac{1}{2}} \mp \frac{\theta_{i,j\pm\frac{1}{2}}\Delta\widehat{\boldsymbol{F}}_{i,j\pm\frac{1}{2}}}{(\alpha_2)_{i,j\pm\frac{1}{2}}},
\end{aligned} \tag{4.9}$$

and $\theta_{i\pm\frac{1}{2},j}, \theta_{i,j\pm\frac{1}{2}} \in [0,1]$ are the blending coefficients. The limited scheme (4.8) reduces to the first-order LLF scheme if $\theta_{i\pm\frac{1}{2},j} = \theta_{i,j\pm\frac{1}{2}} = 0$, and recovers the high-order AF scheme (3.3) when $\theta_{i\pm\frac{1}{2},j} = \theta_{i,j\pm\frac{1}{2}} = 1$.

PROPOSITION 4.3. *If the cell average at the last time step $\overline{\boldsymbol{U}}^n_{i,j}$ and the limited intermediate states $\widetilde{\boldsymbol{U}}^{Lim,\mp}_{i\pm\frac{1}{2},j}$, $\widetilde{\boldsymbol{U}}^{Lim,\mp}_{i,j\pm\frac{1}{2}}$ belong to the admissible state set $\mathcal{G}$, then the limited average update (4.8) is BP, i.e., $\overline{\boldsymbol{U}}^{Lim}_{i,j} \in \mathcal{G}$, under the CFL condition (4.6). If the SSP-RK3 (2.14) is used for the time integration, the high-order scheme is also BP.*

Proof. Under the constraint (4.6), the limited cell average update $\overline{\boldsymbol{U}}^{\mathtt{Lim}}_{i,j}$ is a convex combination of $\overline{\boldsymbol{U}}^{n}_{i,j}$ $\widetilde{\boldsymbol{U}}^{\mathtt{Lim},\mp}_{i\pm\frac{1}{2},j}$, and $\widetilde{\boldsymbol{U}}^{\mathtt{Lim},\mp}_{i,j\pm\frac{1}{2}}$, 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 4.4. The scheme (4.8) is conservative as it amounts to using the x-directional numerical flux $\widehat{\boldsymbol{F}}^{\mathtt{L}}_{i+\frac{1}{2},j}+\theta_{i+\frac{1}{2},j}\Delta\widehat{\boldsymbol{F}}_{i+\frac{1}{2},j}=\theta_{i+\frac{1}{2},j}\widehat{\boldsymbol{F}}^{\mathtt{H}}_{i+\frac{1}{2},j}+(1-\theta_{i+\frac{1}{2},j})\widehat{\boldsymbol{F}}^{\mathtt{L}}_{i+\frac{1}{2},j}$, which is a convex combination of the high-order and low-order fluxes.

Remark 4.5. It should be noted that the time step size (4.6) 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 (4.8) is no longer a convex combination. In our implementation, we start from the usual CFL condition (2.15). Then, if the high-order AF solutions need BP limitings and (4.5) is not BP or (4.6) is not satisfied, the numerical solutions are set back to the last time step, and we rerun with a halved time step size until (4.5) is BP and the constraint (4.6) is satisfied. This is a typical implementation in other BP methods, e.g. [45].

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

4.1.1. Application to scalar conservation laws. This section is devoted to applying the convex limiting approach to scalar conservation laws (3.2), such that the limited cell averages (4.8) satisfy the MP $u^{\min}_{i,j}\leqslant\bar{u}^{\mathtt{Lim}}_{i,j}\leqslant u^{\max}_{i,j}$, where $u^{\min}_{i,j}=\min\mathcal{N}$, $u^{\max}_{i,j}=\max\mathcal{N}$, and $\mathcal{N}$ will be defined later. According to the convex decomposition, the blending coefficient $\theta_{i+\frac{1}{2},j}\in[0,1]$ or $\Delta\hat{f}^{\mathtt{Lim}}_{i+\frac{1}{2},j}$ should be chosen such that $u^{\min}_{i,j}\leqslant\tilde{u}^{\mathtt{Lim},-}_{i+\frac{1}{2},j}\leqslant u^{\max}_{i,j}$, $u^{\min}_{i+1,j}\leqslant\tilde{u}^{\mathtt{Lim},+}_{i+\frac{1}{2},j}\leqslant u^{\max}_{i+1,j}$. Solving the first condition, i.e. $u^{\min}_{i,j}\leqslant\tilde{u}_{i+\frac{1}{2},j}-\Delta\hat{f}^{\mathtt{Lim}}_{i+\frac{1}{2},j}/\alpha_{i+\frac{1}{2},j}\leqslant u^{\max}_{i,j}$, one has $\Delta\hat{f}^{\mathtt{Lim}}_{i+\frac{1}{2},j}\leqslant\alpha_{i+\frac{1}{2},j}(\tilde{u}_{i+\frac{1}{2},j}-u^{\min}_{i,j})$ if $\Delta\hat{f}_{i+\frac{1}{2},j}\geqslant 0$, or $\Delta\hat{f}^{\mathtt{Lim}}_{i+\frac{1}{2},j}\geqslant\alpha_{i+\frac{1}{2},j}(\tilde{u}_{i+\frac{1}{2},j}-u^{\max}_{i,j})$ if $\Delta\hat{f}_{i+\frac{1}{2},j}<0$. Solving the second condition $u^{\min}_{i+1,j}\leqslant\tilde{u}^{\mathtt{Lim},+}_{i+\frac{1}{2},j}\leqslant u^{\max}_{i+1,j}$ in the same way and combining the two sets of results yields

$$\Delta\hat{f}^{\mathtt{Lim}}_{i+\frac{1}{2},j}=\begin{cases}\min\left\{\Delta\hat{f}_{i+\frac{1}{2},j},\Delta\hat{f}^{+}_{i+\frac{1}{2},j}\right\}, & \text{if } \Delta\hat{f}_{i+\frac{1}{2},j}\geqslant 0,\\ \max\left\{\Delta\hat{f}_{i+\frac{1}{2},j},\Delta\hat{f}^{-}_{i+\frac{1}{2},j}\right\}, & \text{otherwise,}\end{cases}$$

$$\Delta\hat{f}^{+}_{i+\frac{1}{2},j}=(\alpha_1)_{i+\frac{1}{2},j}\min\left\{\tilde{u}_{i+\frac{1}{2},j}-u^{\min}_{i,j},u^{\max}_{i+1,j}-\tilde{u}_{i+\frac{1}{2},j}\right\},$$

$$\Delta\hat{f}^{-}_{i+\frac{1}{2},j}=(\alpha_1)_{i+\frac{1}{2},j}\max\left\{u^{\min}_{i+1,j}-\tilde{u}_{i+\frac{1}{2},j},\tilde{u}_{i+\frac{1}{2},j}-u^{\max}_{i,j}\right\}.$$

Finally, the limited numerical flux is

$$\hat{f}^{\mathtt{Lim}}_{i+\frac{1}{2},j}=\hat{f}^{\mathtt{L}}_{i+\frac{1}{2},j}+\Delta\hat{f}^{\mathtt{Lim}}_{i+\frac{1}{2},j}. \tag{4.10}$$

If considering the global MP, $\mathcal{N}=\bigcup_{i,j,\sigma}\{\bar{u}^n_{i,j},u^n_\sigma\}$. One can also enforce the local MP, which helps to suppress spurious oscillations [21, 31, 22], by choosing

$$\mathcal{N}=\left\{\bar{u}^n_{i,j},\ \tilde{u}_{i-\frac{1}{2},j},\ \tilde{u}_{i+\frac{1}{2},j},\ \tilde{u}_{i,j-\frac{1}{2}},\ \tilde{u}_{i,j+\frac{1}{2}},\bar{u}^n_{i-1,j},\bar{u}^n_{i+1,j},\bar{u}^n_{i,j-1},\bar{u}^n_{i,j+1}\right\},$$

which includes the intermediate states and neighboring cell averages.

4.1.2. Application to the compressible Euler equations. This section aims at enforcing the positivity of density and pressure. To avoid the effect of the round-off error, we need to choose the desired lower bounds. Denote the lowest density and pressure in the domain by

$$\varepsilon^{\rho} := \min_{i,j,\sigma}\{\overline{\boldsymbol{U}}^{n,\rho}_{i,j}, \boldsymbol{U}^{n,\rho}_{\sigma}\},\ \ \varepsilon^{p} := \min_{i,j,\sigma}\{p(\overline{\boldsymbol{U}}^{n}_{i,j}), p(\boldsymbol{U}^{n}_{\sigma})\}, \tag{4.11}$$

where $\boldsymbol{U}^{*,\rho}$ and $p(\boldsymbol{U}^{*})$ denote the density component and pressure recovered from $\boldsymbol{U}^{*}$, respectively, and σ denotes the locations of point values in the DoFs. The limiting (4.9) is feasible if the constraints are satisfied by the first-order LLF intermediate states (4.5), thus the lower bounds can be defined as

$$\begin{aligned}
\varepsilon^{\rho}_{i,j} &:= \min\{10^{-13}, \varepsilon^{\rho}, \widetilde{\boldsymbol{U}}^{\rho}_{i-\frac{1}{2},j}, \widetilde{\boldsymbol{U}}^{\rho}_{i+\frac{1}{2},j}, \widetilde{\boldsymbol{U}}^{\rho}_{i,j-\frac{1}{2}}, \widetilde{\boldsymbol{U}}^{\rho}_{i,j+\frac{1}{2}}\},\\
\varepsilon^{p}_{i,j} &:= \min\{10^{-13}, \varepsilon^{p}, p(\widetilde{\boldsymbol{U}}_{i-\frac{1}{2},j}), p(\widetilde{\boldsymbol{U}}_{i+\frac{1}{2},j}), p(\widetilde{\boldsymbol{U}}_{i,j-\frac{1}{2}}), p(\widetilde{\boldsymbol{U}}_{i,j+\frac{1}{2}})\}.
\end{aligned}$$

i) **Positivity of density.** The first step is to impose the density positivity $\widetilde{\boldsymbol{U}}^{\mathtt{Lim},\pm,\rho}_{i+\frac{1}{2},j} \geqslant \bar{\varepsilon}^{\rho}_{i+\frac{1}{2},j} := \min\{\varepsilon^{\rho}_{i,j}, \varepsilon^{\rho}_{i+1,j}\}$. Similarly to the derivation of the scalar case, the corresponding density component of the limited anti-diffusive flux is

$$\Delta\widehat{\boldsymbol{F}}^{\mathtt{Lim},\rho}_{i+\frac{1}{2},j} = \begin{cases} \min\left\{\Delta\widehat{\boldsymbol{F}}^{\rho}_{i+\frac{1}{2},j},\ (\alpha_1)_{i+\frac{1}{2},j}\left(\widetilde{\boldsymbol{U}}^{\rho}_{i+\frac{1}{2},j} - \bar{\varepsilon}^{\rho}_{i+\frac{1}{2},j}\right)\right\}, & \text{if } \Delta\widehat{\boldsymbol{F}}^{\rho}_{i+\frac{1}{2},j} \geqslant 0,\\ \max\left\{\Delta\widehat{\boldsymbol{F}}^{\rho}_{i+\frac{1}{2},j},\ (\alpha_1)_{i+\frac{1}{2},j}\left(\bar{\varepsilon}^{\rho}_{i+\frac{1}{2},j} - \widetilde{\boldsymbol{U}}^{\rho}_{i+\frac{1}{2},j}\right)\right\}, & \text{otherwise.}\end{cases}$$

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

ii) **Positivity of pressure.** The second step is to enforce pressure positivity $p(\widetilde{\boldsymbol{U}}^{\mathtt{Lim},\pm}_{i+\frac{1}{2},j}) \geqslant \bar{\varepsilon}^{p}_{i+\frac{1}{2},j} := \min\{\varepsilon^{p}_{i,j}, \varepsilon^{p}_{i+1,j}\}$. Since

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

the constraints lead to two inequalities after some algebraic operations

$$A_{i+\frac{1}{2},j}\theta^{2}_{i+\frac{1}{2},j} \pm B_{i+\frac{1}{2},j}\theta_{i+\frac{1}{2},j} \leqslant C_{i+\frac{1}{2},j}, \tag{4.12}$$

with the coefficients (the subscript $(\cdot)_{i+\frac{1}{2},j}$ is omitted in the right-hand side)

$$\begin{aligned}
A_{i+\frac{1}{2},j} &= \frac{1}{2}\left\|\Delta\widehat{\boldsymbol{F}}^{\mathtt{Lim},*,\rho\boldsymbol{v}}\right\|_2^2 - \Delta\widehat{\boldsymbol{F}}^{\mathtt{Lim},*,\rho}\Delta\widehat{\boldsymbol{F}}^{\mathtt{Lim},*,E},\\
B_{i+\frac{1}{2},j} &= \alpha_1\left(\Delta\widehat{\boldsymbol{F}}^{\mathtt{Lim},*,\rho}\widetilde{\boldsymbol{U}}^{E} + \widetilde{\boldsymbol{U}}^{\rho}\Delta\widehat{\boldsymbol{F}}^{\mathtt{Lim},*,E} - \Delta\widehat{\boldsymbol{F}}^{\mathtt{Lim},*,\rho\boldsymbol{v}}\cdot\widetilde{\boldsymbol{U}}^{\rho\boldsymbol{v}} - \tilde{\varepsilon}\Delta\widehat{\boldsymbol{F}}^{\mathtt{Lim},*,\rho}\right),\\
C_{i+\frac{1}{2},j} &= \alpha_1^2\left(\widetilde{\boldsymbol{U}}^{\rho}\widetilde{\boldsymbol{U}}^{E} - \frac{1}{2}\left\|\widetilde{\boldsymbol{U}}^{\rho\boldsymbol{v}}\right\|_2^2 - \tilde{\varepsilon}\widetilde{\boldsymbol{U}}^{\rho}\right), \quad \tilde{\varepsilon} = \bar{\varepsilon}^{p}_{i+\frac{1}{2},j}/(\gamma-1).
\end{aligned}$$

Following [30], the inequalities (4.12) hold under the linear sufficient condition

$$\left(\max\{0, A_{i+\frac{1}{2},j}\} + |B_{i+\frac{1}{2},j}|\right)\theta_{i+\frac{1}{2},j} \leqslant C_{i+\frac{1}{2},j},$$

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

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

and the final limited numerical flux is

$$\widehat{\boldsymbol{F}}^{\mathtt{Lim},**}_{i+\frac{1}{2},j} = \widehat{\boldsymbol{F}}^{\mathtt{L}}_{i+\frac{1}{2},j} + \theta_{i+\frac{1}{2},j}\Delta\widehat{\boldsymbol{F}}^{\mathtt{Lim},*}_{i+\frac{1}{2},j}. \tag{4.13}$$

4.1.3. Shock sensor-based limiting. Spurious oscillations are observed, especially near strong shock waves, if only the BP limitings are employed, see Example 5.10. We propose to further limit the numerical fluxes using another parameter $\theta^s_{i+\frac{1}{2},j}$ based on shock sensors. Consider the Jameson's shock sensor in [29],

$$(\varphi_1)_{i,j} = \frac{|\bar{p}_{i+1,j} - 2\bar{p}_{i,j} + \bar{p}_{i-1,j}|}{|\bar{p}_{i+1,j} + 2\bar{p}_{i,j} + \bar{p}_{i-1,j}|},$$

and a modified Ducros' shock sensor [15]

$$(\varphi_2)_{i,j} = \max\left\{\frac{-(\nabla\cdot\bar{\boldsymbol{v}})_{i,j}}{\sqrt{(\nabla\cdot\bar{\boldsymbol{v}})^2_{i,j} + (\nabla\times\bar{\boldsymbol{v}})^2_{i,j} + 10^{-40}}},\ 0\right\},$$

where

$$(\nabla\cdot\bar{\boldsymbol{v}})_{i,j} \approx \frac{2\left((\bar{v}_1)_{i+1,j} - (\bar{v}_1)_{i-1,j}\right)}{\Delta x_i + \Delta x_{i+1}} + \frac{2\left((\bar{v}_2)_{i,j+1} - (\bar{v}_2)_{i,j-1}\right)}{\Delta y_j + \Delta y_{j+1}},$$

$$(\nabla\times\bar{\boldsymbol{v}})_{i,j} \approx \frac{2\left((\bar{v}_2)_{i+1,j} - (\bar{v}_2)_{i-1,j}\right)}{\Delta x_i + \Delta x_{i+1}} - \frac{2\left((\bar{v}_1)_{i,j+1} - (\bar{v}_1)_{i,j-1}\right)}{\Delta y_j + \Delta y_{j+1}},$$

with $\bar{v}_{i,j}$ and $\bar{p}_{i,j}$ the velocity and pressure recovered from the cell average $\overline{\boldsymbol{U}}_{i,j}$. We consider the sign of the velocity divergence, such that the shock waves can be located better. The blending coefficient is designed as

$$\theta^s_{i+\frac{1}{2},j} = \exp(-\kappa(\varphi_1)_{i+\frac{1}{2},j}(\varphi_2)_{i+\frac{1}{2},j}) \in (0,1],$$

$$(\varphi_s)_{i+\frac{1}{2},j} = \max\left\{(\varphi_s)_{i,j}, (\varphi_s)_{i+1,j}\right\},\ s = 1,2,$$

where the problem-dependent parameter κ adjusts the strength of the limiting, and its optimal choice needs further investigation. The final limited numerical flux is

$$\widehat{\boldsymbol{F}}^{\mathtt{Lim}}_{i+\frac{1}{2},j} = \widehat{\boldsymbol{F}}^{\mathtt{L}}_{i+\frac{1}{2},j} + \theta^s_{i+\frac{1}{2},j}\Delta\widehat{\boldsymbol{F}}^{\mathtt{Lim},**}_{i+\frac{1}{2},j}, \tag{4.14}$$

with $\Delta\widehat{\boldsymbol{F}}^{\mathtt{Lim},**}_{i+\frac{1}{2},j} = \widehat{\boldsymbol{F}}^{\mathtt{Lim},**}_{i+\frac{1}{2},j} - \widehat{\boldsymbol{F}}^{\mathtt{L}}_{i+\frac{1}{2},j}$, and $\widehat{\boldsymbol{F}}^{\mathtt{Lim},**}_{i+\frac{1}{2},j}$ given in (4.13).

4.2. Scaling limiter for point value. To achieve the BP property, it is also necessary to introduce BP limiting for the point value, because using the BP limiting for cell average alone cannot guarantee the bounds, see Example 5.5. As there is no conservation requirement on the point value update, a simple scaling limiter [32] is directly performed on the high-order solution rather than on the flux for the cell average.

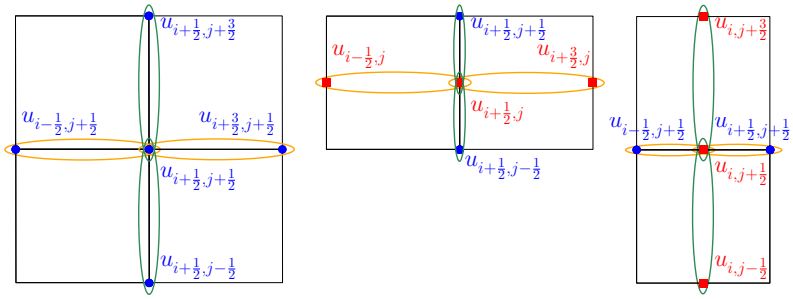

Fig. 2: The stencils for the first-order LLF schemes.

The first step is to define suitable first-order LLF schemes. The stencils are shown in Figure 2.

For the point value at the corner, one can choose

$$\begin{aligned} \boldsymbol{U}^{\mathrm{L}}_{i+\frac{1}{2},j+\frac{1}{2}} = \boldsymbol{U}^{n}_{i+\frac{1}{2},j+\frac{1}{2}} &- \frac{2\Delta t^n}{\Delta x_i + \Delta x_{i+1}} \left(\widehat{\boldsymbol{F}}^{\mathrm{L}}_{i+1,j+\frac{1}{2}} - \widehat{\boldsymbol{F}}^{\mathrm{L}}_{i,j+\frac{1}{2}} \right) \\ &- \frac{2\Delta t^n}{\Delta y_j + \Delta y_{j+1}} \left(\widehat{\boldsymbol{F}}^{\mathrm{L}}_{i+\frac{1}{2},j+1} - \widehat{\boldsymbol{F}}^{\mathrm{L}}_{i+\frac{1}{2},j} \right), \end{aligned} \tag{4.15}$$

with the LLF numerical fluxes

$$\widehat{\boldsymbol{F}}^{\mathrm{L}}_{i+1,j+\frac{1}{2}} := \widehat{\boldsymbol{F}}^{\mathrm{LLF}}_1(\boldsymbol{U}^{n}_{i+\frac{1}{2},j+\frac{1}{2}}, \boldsymbol{U}^{n}_{i+\frac{3}{2},j+\frac{1}{2}}), \ \widehat{\boldsymbol{F}}^{\mathrm{L}}_{i+\frac{1}{2},j+1} := \widehat{\boldsymbol{F}}^{\mathrm{LLF}}_2(\boldsymbol{U}^{n}_{i+\frac{1}{2},j+\frac{1}{2}}, \boldsymbol{U}^{n}_{i+\frac{1}{2},j+\frac{3}{2}}).$$

Note that the x-directional LLF flux has been used in (4.3). For the vertical face-centered point value, we choose the first-order LLF scheme as

$$\boldsymbol{U}^{\mathrm{L}}_{i+\frac{1}{2},j} = \boldsymbol{U}^{n}_{i+\frac{1}{2},j} - \frac{2\Delta t^n}{\Delta x_i + \Delta x_{i+1}} \left(\widehat{\boldsymbol{F}}^{\mathrm{L}}_{i+1,j} - \widehat{\boldsymbol{F}}^{\mathrm{L}}_{i,j} \right) - \frac{\Delta t^n}{\Delta y_j} \left(\widehat{\boldsymbol{F}}^{\mathrm{L}}_{i+\frac{1}{2},j+\frac{1}{2}} - \widehat{\boldsymbol{F}}^{\mathrm{L}}_{i+\frac{1}{2},j-\frac{1}{2}} \right), \tag{4.16}$$

with the LLF numerical fluxes

$$\widehat{\boldsymbol{F}}^{\mathrm{L}}_{i+1,j} := \widehat{\boldsymbol{F}}^{\mathrm{LLF}}_1(\boldsymbol{U}^{n}_{i+\frac{1}{2},j}, \boldsymbol{U}^{n}_{i+\frac{3}{2},j}), \ \widehat{\boldsymbol{F}}^{\mathrm{L}}_{i+\frac{1}{2},j+\frac{1}{2}} := \widehat{\boldsymbol{F}}^{\mathrm{LLF}}_2(\boldsymbol{U}^{n}_{i+\frac{1}{2},j}, \boldsymbol{U}^{n}_{i+\frac{1}{2},j+1}).$$

The LLF scheme for the face-centered value on the horizontal face can be chosen as

$$\boldsymbol{U}^{\mathrm{L}}_{i,j+\frac{1}{2}} = \boldsymbol{U}^{n}_{i,j+\frac{1}{2}} - \frac{\Delta t^n}{\Delta x_i} \left(\widehat{\boldsymbol{F}}^{\mathrm{L}}_{i+\frac{1}{2},j+\frac{1}{2}} - \widehat{\boldsymbol{F}}^{\mathrm{L}}_{i-\frac{1}{2},j+\frac{1}{2}} \right) - \frac{2\Delta t^n}{\Delta y_j + \Delta y_{j+1}} \left(\widehat{\boldsymbol{F}}^{\mathrm{L}}_{i,j+1} - \widehat{\boldsymbol{F}}^{\mathrm{L}}_{i,j} \right), \tag{4.17}$$

with similarly defined LLF numerical fluxes as for the vertical face.

Similarly to Lemma 4.2, it is straightforward to obtain the following Lemma.

LEMMA 4.6. *The LLF schemes (4.15)-(4.17) for the point value update are BP under the following time step size constraint*

$$\begin{aligned} \Delta t^n \leqslant \frac{1}{2} \min \Bigg\{ & \frac{\Delta x_i + \Delta x_{i+1}}{2\left((\alpha_1)_{i,j+\frac{1}{2}} + (\alpha_1)_{i+1,j+\frac{1}{2}} \right)}, \frac{\Delta y_j + \Delta y_{j+1}}{2\left((\alpha_2)_{i+\frac{1}{2},j} + (\alpha_2)_{i+\frac{1}{2},j+1} \right)}, \\ & \frac{\Delta x_i + \Delta x_{i+1}}{2\left((\alpha_1)_{i,j} + (\alpha_1)_{i+1,j} \right)}, \frac{\Delta y_j}{(\alpha_2)_{i+\frac{1}{2},j+\frac{1}{2}} + (\alpha_2)_{i+\frac{1}{2},j-\frac{1}{2}}}, \\ & \frac{\Delta x_i}{(\alpha_1)_{i+\frac{1}{2},j+\frac{1}{2}} + (\alpha_1)_{i-\frac{1}{2},j+\frac{1}{2}}}, \frac{\Delta y_j + \Delta y_{j+1}}{2\left((\alpha_2)_{i,j} + (\alpha_2)_{i,j+1} \right)} \Bigg\}, \end{aligned} \tag{4.18}$$

where $(\alpha_1)_*$ *and* $(\alpha_2)_*$ *are the viscosity coefficients in the LLF schemes.*

The limited solution is obtained by blending the high-order AF scheme (3.6) with the forward-Euler scheme and the LLF schemes (4.15)-(4.17) as $\boldsymbol{U}_\sigma^{\mathtt{Lim}} = \theta_\sigma \boldsymbol{U}_\sigma^{\mathtt{H}} + (1-\theta_\sigma)\boldsymbol{U}_\sigma^{\mathtt{L}}$, such that $\boldsymbol{U}_\sigma^{\mathtt{Lim}} \in \mathcal{G}$.

Remark 4.7. In the FVS, the cell-centered value obtained based on Simpson's rule $\boldsymbol{U}_i = (-\boldsymbol{U}_{i-\frac{1}{2}} + 6\overline{\boldsymbol{U}}_i - \boldsymbol{U}_{i+\frac{1}{2}})/4$ in 1D or (3.11) in 2D is not a convex combination, thus it is possible that $\boldsymbol{U}_i, \boldsymbol{U}_{i,j} \notin \mathcal{G}$. For the scalar case, it does not affect the BP property. However, for the Euler equations, the computation of $\boldsymbol{F}_i$ (resp. $(\boldsymbol{F}_\ell)_{i,j}$) requires that $\boldsymbol{U}_i \in \mathcal{G}$ (resp. $\boldsymbol{U}_{i,j} \in \mathcal{G}$), thus the scaling limiter [45] is applied in the cell I_i (resp. $I_{i,j}$), a procedure also mentioned in [10]. See more details in Remark 4.8.

4.2.1. Application to scalar conservation laws. This section enforces the MP $u_\sigma^{\min} \leqslant u_\sigma^{\mathtt{Lim}} \leqslant u_\sigma^{\max}$ using the scaling limiter [44]. The limited solution is

$$u_\sigma^{\mathtt{Lim}} = \theta_\sigma u_\sigma^{\mathtt{H}} + (1-\theta_\sigma)\, u_\sigma^{\mathtt{L}}, \tag{4.19}$$

with the coefficient

$$\theta_\sigma = \min\left\{1,\ \left|\frac{u_\sigma^{\mathtt{L}} - m_0}{u_\sigma^{\mathtt{L}} - u_\sigma^{\mathtt{H}}}\right|,\ \left|\frac{M_0 - u_\sigma^{\mathtt{L}}}{u_\sigma^{\mathtt{H}} - u_\sigma^{\mathtt{L}}}\right|\right\}.$$

The bounds are determined by $u_\sigma^{\min} = \min\mathcal{N}$, $u_\sigma^{\max} = \max\mathcal{N}$, where the set $\mathcal{N}$ consists of all the DoFs in the domain, i.e., $\mathcal{N} = \bigcup_{i,j,\sigma}\{\bar{u}_{i,j}^n, u_\sigma^n\}$ for the global MP. One can also consider the neighboring DoFs for the local MP. For the point value at the corner $(x_{i+\frac{1}{2}}, y_{j+\frac{1}{2}})$, we choose

$$\mathcal{N} = \left\{u_{i+\frac{1}{2},j+\frac{1}{2}}^n, u_{i-\frac{1}{2},j+\frac{1}{2}}^n, u_{i+\frac{3}{2},j+\frac{1}{2}}^n, u_{i+\frac{1}{2},j-\frac{1}{2}}^n, u_{i+\frac{1}{2},j+\frac{3}{2}}^n\right\},$$

which should at least include all the DoFs that appeared in the first-order LLF scheme (4.15). For the point value at the vertical face center $(x_{i+\frac{1}{2}}, y_j)$, similarly we choose

$$\mathcal{N} = \left\{u_{i+\frac{1}{2},j}^n, u_{i-\frac{1}{2},j}^n, u_{i+\frac{3}{2},j}^n, u_{i+\frac{1}{2},j-\frac{1}{2}}^n, u_{i+\frac{1}{2},j+\frac{1}{2}}^n\right\}.$$

For the point value at the horizontal face center $(x_i, y_{j+\frac{1}{2}})$, we choose

$$\mathcal{N} = \left\{u_{i,j+\frac{1}{2}}^n, u_{i,j-\frac{1}{2}}^n, u_{i,j+\frac{3}{2}}^n, u_{i-\frac{1}{2},j+\frac{1}{2}}^n, u_{i+\frac{1}{2},j+\frac{1}{2}}^n\right\}.$$

4.2.2. Application to the compressible Euler equations. The limiting consists of two steps.

i) Positivity of density. First, the high-order solution $\boldsymbol{U}_\sigma^{\mathtt{H}}$ is modified as $\boldsymbol{U}_\sigma^{\mathtt{Lim},*}$, such that its density component satisfies $\boldsymbol{U}_\sigma^{\mathtt{Lim},*,\rho} \geqslant \varepsilon_\sigma^\rho := \min\{10^{-13}, \varepsilon^\rho, \boldsymbol{U}_\sigma^{\mathtt{L},\rho}\}$ with ε^ρ given in (4.11). Solving the inequality yields

$$\theta_\sigma^* = \begin{cases} \dfrac{\boldsymbol{U}_\sigma^{\mathtt{L},\rho} - \varepsilon_\sigma^\rho}{\boldsymbol{U}_\sigma^{\mathtt{L},\rho} - \boldsymbol{U}_\sigma^{\mathtt{H},\rho}}, & \text{if } \ \boldsymbol{U}_\sigma^{\mathtt{H},\rho} < \varepsilon_\sigma^\rho, \\ 1, & \text{otherwise.} \end{cases}$$

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

ii) Positivity of pressure. Then the limited solution $\boldsymbol{U}_\sigma^{\mathtt{Lim},*}$ is modified as $\boldsymbol{U}_\sigma^{\mathtt{Lim}}$, such that it gives positive pressure, i.e., $p(\boldsymbol{U}_\sigma^{\mathtt{Lim}}) \geqslant \varepsilon_\sigma^p := \min\{10^{-13}, \varepsilon^p, p(\boldsymbol{U}_\sigma^{\mathtt{L}})\}$, with ε^p given in (4.11). Let the final limited solution be

$$\boldsymbol{U}_\sigma^{\mathtt{Lim}} = \theta_\sigma^{**}\boldsymbol{U}_\sigma^{\mathtt{Lim},*} + (1-\theta_\sigma^{**})\,\boldsymbol{U}_\sigma^{\mathtt{L}}. \tag{4.20}$$

The pressure is a concave function of the conservative variables (see e.g. [46]), so that $p(\boldsymbol{U}_\sigma^{\mathtt{Lim}}) \geqslant \theta_\sigma^{**}p(\boldsymbol{U}_\sigma^{\mathtt{Lim},*}) + (1-\theta_\sigma^{**})\,p(\boldsymbol{U}_\sigma^{\mathtt{L}})$ based on Jensen's inequality and $\boldsymbol{U}_\sigma^{\mathtt{Lim},*,\rho} > 0$, $\boldsymbol{U}_\sigma^{\mathtt{L},\rho} > 0$, $\theta_\sigma^{**} \in [0,1]$. Thus the coefficient can be chosen as

$$\theta_\sigma^{**} = \begin{cases} \dfrac{p(\boldsymbol{U}_\sigma^{\mathtt{L}}) - \varepsilon_\sigma^p}{p(\boldsymbol{U}_\sigma^{\mathtt{L}}) - p(\boldsymbol{U}_\sigma^{\mathtt{Lim},*})}, & \text{if } \ p(\boldsymbol{U}_\sigma^{\mathtt{Lim},*}) < \varepsilon_\sigma^p, \\ 1, & \text{otherwise.} \end{cases}$$

Remark 4.8. To compute the high-order FVS-based point value update, we should limit the cell-centered value $\boldsymbol{U}_i$ in 1D (resp. $\boldsymbol{U}_{i,j}$ in 2D) at the beginning of each Runge-Kutta stage. For example, in 2D, we modify $\boldsymbol{U}_{i,j}$ as $\boldsymbol{U}_{i,j}^{\mathtt{Lim}} = \theta_{i,j}\boldsymbol{U}_{i,j} + (1-\theta_{i,j})\overline{\boldsymbol{U}}_{i,j}$ such that

$$\boldsymbol{U}_{i,j}^{\mathtt{Lim},\rho} \geqslant \min\{10^{-13}, \overline{\boldsymbol{U}}_{i,j}^{\rho}\},\ \ p(\boldsymbol{U}_{i,j}^{\mathtt{Lim}}) \geqslant \min\{10^{-13}, p(\overline{\boldsymbol{U}}_{i,j})\}.$$

The computation of $\theta_{i,j}$ is similar to the procedure in this section.

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

PROPOSITION 4.9. *If the initial numerical solution $\overline{\boldsymbol{U}}_{i,j}^0, \boldsymbol{U}_\sigma^0 \in \mathcal{G}$ for all i,j,σ, and the time step size satisfies* (4.6) *and* (4.18)*, then the AF methods* (3.3)-(3.6) *equipped with the SSP-RK3* (2.14) *and the BP limitings*

- (4.10) *and* (4.19) *preserve the maximum principle for scalar case;*
- (4.13) *and* (4.20) *preserve positive density and pressure for the Euler equations.*

Remark 4.10. For uniform meshes, and if taking the maximal spectral radius of $\partial\boldsymbol{F}_1/\partial\boldsymbol{U}$ and $\partial\boldsymbol{F}_2/\partial\boldsymbol{U}$ in the domain as $\|\varrho_1\|_\infty$ and $\|\varrho_2\|_\infty$, the following CFL condition

$$\Delta t^n \leqslant \frac{1}{4}\min\left\{\frac{\Delta x}{\|\varrho_1\|_\infty}, \frac{\Delta y}{\|\varrho_2\|_\infty}\right\}$$

fulfills the time step size constraints (4.6) and (4.18).

5. Numerical results. This section presents some numerical tests to verify the accuracy, BP property, and shock-capturing ability of the proposed BP AF methods. The adiabatic index is $\gamma = 1.4$ for the Euler equations except for Example 5.11, where it is 5/3. In the 2D plots, the numerical solutions are visualized on a refined mesh with half the mesh size, where the values at the grid points are the cell averages or point values on the original mesh. Note that the BP limitings naturally reduce some oscillations. Some additional tests are provided in Section SM4 in the supplementary material, including a 1D accuracy test for the Euler equations, double rarefaction problem, blast wave interaction problem using the power law reconstruction [5], and double Mach reflection problem.

Example 5.1 (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. The initial condition is a square wave, $u_0(x) = 2$ if $|x| < 0.2$, otherwise $u_0(x) = -1$.

Figure 3 shows the cell averages and point values at $T = 0.5$ based on different point value updates with 200 cells. The CFL number is 0.2. The spike generation when using the JS has been observed in [27], and the reason is also discussed in Subsection 2.2. Such an issue cannot be eliminated by our BP limitings alone, but can be cured by additionally using the FVS for the point value update. The numerical solutions based on the FVS agree well with the reference solution when the limitings are activated.

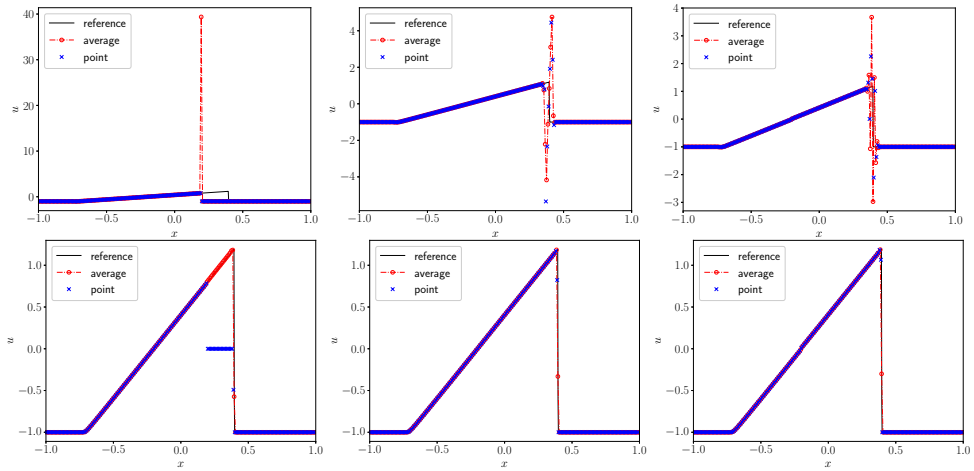

Fig. 3: Example 5.1, self-steepening shock for the Burgers' equation. The numerical solutions computed without limiting (top row) and with the BP limitings imposing the local MP (bottom row). From left to right: JS, LLF, and upwind FVS.

Example 5.2 (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 the domain $[0, 1]$ with the initial data $(\rho, v, p) = (2, 0, 10^9)$ if $x < 0.5$, otherwise $(\rho, v, p) = (10^{-3}, 0, 1)$.

Without the BP limitings, the simulation will stop due to negative density or pressure. Figure 4 shows the density computed on a uniform mesh of 6000 cells with the BP limitings and shock sensor-based limiting. Note that, the numerical methods typically need a small mesh size to accurately obtain the right location of the shock wave. The CFL number is 0.4 for the JS, LLF, and SW FVS, and 0.1 for the VH FVS for stability. The numerical solutions agree well with the exact solution with only a few undershoots at the discontinuities.

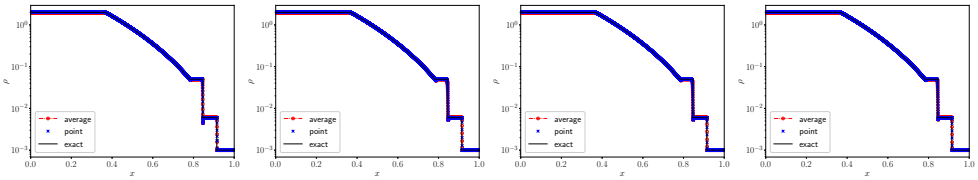

Fig. 4: Example 5.2, LeBlanc Riemann problem. The density computed with the BP limitings and the shock sensor-based limiting ($\kappa = 10$) on a uniform mesh of 6000 cells. From left to right: JS, LLF, SW, and VH FVS.

Example 5.3 (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 5 shows the density plots obtained by using the BP limitings and shock sensor-based limiting on a uniform mesh of 800 cells. The CFL number is 0.4 for the JS, LLF, and SW FVS, and 0.36 for the VH FVS. The numerical solutions agree well the reference solution with a few overshoots/undershoots.

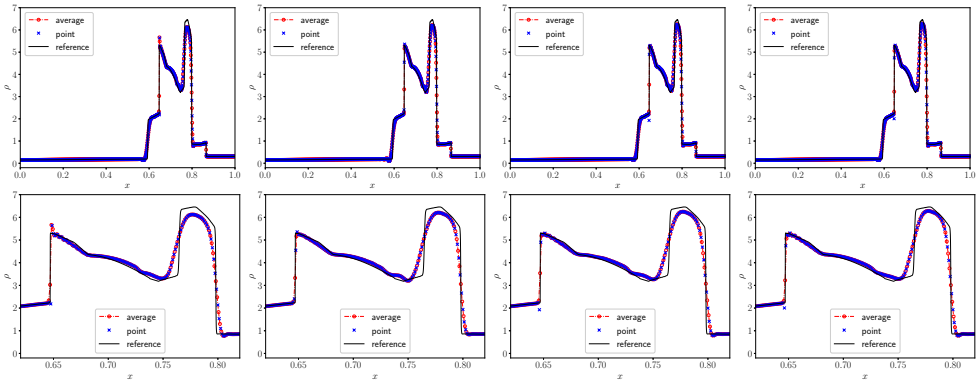

Fig. 5: Example 5.3, blast wave interaction. The density computed with the BP limitings and the shock sensor-based limiting ($\kappa = 1$). The corresponding enlarged views in $x \in [0.62, 0.82]$ are shown in the bottom row.

Remark 5.4. In the numerical tests, the maximal CFL numbers for stability are obtained experimentally. Note that the reduction of the CFL numbers is due to different stability bounds for different point value updates, and is not related to the BP property. The study of such an issue is beyond the scope of this paper, which will be explored in the future.

Example 5.5 (2D advection equation). This test solves $u_t + u_x + u_y = 0$, on the periodic domain $[0,1]\times[0,1]$ with the following initial data

$$u_0(x,y) = \begin{cases} 1 - |5r|, & \text{if} \quad r = \sqrt{(x-0.3)^2 + (y-0.3)^2} < 0.2, \\ 1, & \text{if} \quad \max\{|x-0.7|,\ |y-0.7|\} < 0.2, \\ 0, & \text{otherwise.} \end{cases}$$

For the advection equation, the JS and LLF FVS are equivalent. The results on the uniform 100×100 mesh obtained without and with BP limitings at $T = 2$ are presented in Figure 6. The BP limitings suppress the overshoots and undershoots well near the discontinuities. Table 1 lists whether the numerical solutions stay in the bound $[0,1]$. The bound is only preserved when both the BP limitings for the cell average and point value are activated, demonstrating that it is necessary to use the two kinds of BP limitings simultaneously.

Example 5.6 (2D Burgers' equation). We solve $u_t + \left(\frac{1}{2}u^2\right)_x + \left(\frac{1}{2}u^2\right)_y = 0$ on the periodic domain $[0,1]\times[0,1]$, with the initial condition $u_0(x,y) = 0.5 + \sin(2\pi(x+y))$. This test is solved until $T = 0.3$, when the shock waves have emerged.

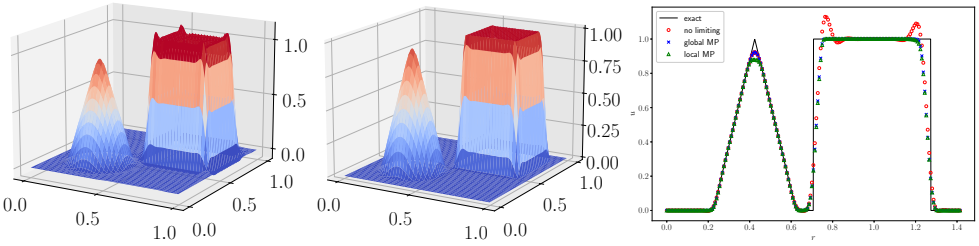

Fig. 6: Example 5.5, 2D advection equation. From left to right: without any limiting, with BP limitings imposing the global MP, cut-line along $y = x$.

cell average \ point value	no limiting	global MP	local MP
no limiting	✗	✗	✗
global MP	✗	✓	✓
local MP	✗	✓	✓

Table 1: Example 5.5, 2D advection equation. We list whether the numerical solutions stay in the bound $[0, 1]$ with different limitings.

Figure 7 plots the solutions using the LLF FVS on the uniform 100×100 mesh without and with limitings. The oscillations near the shock waves are suppressed well when the limitings are activated, and the numerical solutions agree well with the reference solution. The blending coefficients $\theta_{i+\frac{1}{2},j}, \theta_{i,j+\frac{1}{2}}$ for the cell average and θ_σ for the point value when using the global MP are also presented in Figure 8, verifying that the limitings are only locally activated near the shock waves.

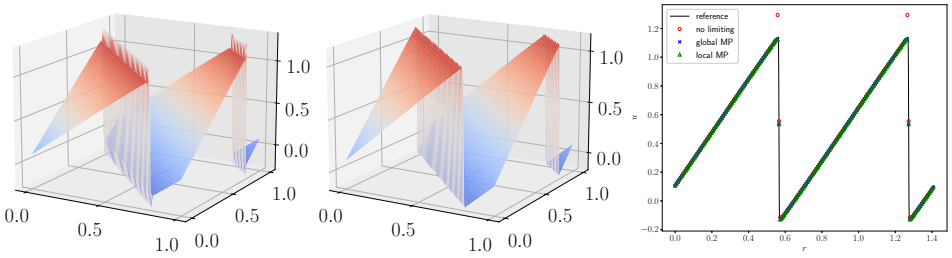

Fig. 7: Example 5.6, 2D Burgers' equation. From left to right: without limiting, with BP limitings imposing the global MP, cut-line along $y = x$.

Example 5.7 (2D isentropic vortex). The domain is $[-5, 5] \times [-5, 5]$ with periodic boundary conditions and the initial condition is

$$\rho = T_0^{\frac{1}{\gamma-1}}, \ (v_1, v_2) = (1, 1) + k_0(y, -x), \ p = T_0\rho, \ k_0 = \frac{\epsilon}{2\pi} e^{0.5(1-r^2)}, \ T_0 = 1 - \frac{\gamma - 1}{2\gamma} k_0^2,$$

where $r^2 = x^2 + y^2$, and $\epsilon = 10.0828$ is the vortex strength. The lowest initial density and pressure are around 7.83×10^{-15} and 1.78×10^{-20}, respectively, so that the BP

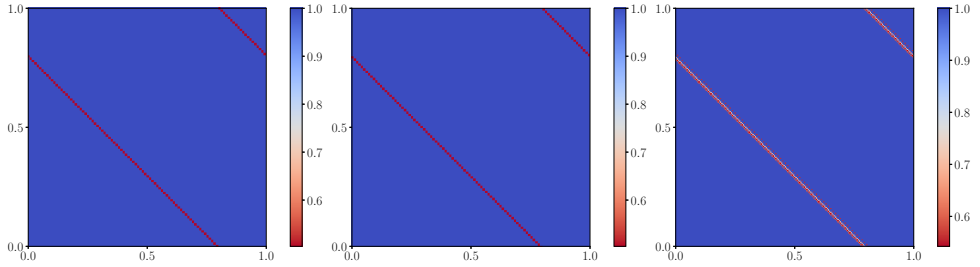

Fig. 8: Example 5.6, 2D Burgers' equation. The blending coefficients in the limitings. From left to right: $\theta_{i+\frac{1}{2},j}$ and $\theta_{i,j+\frac{1}{2}}$ for the cell average, θ_σ for the point value.

limitings are necessary to run this test case. The problem is solved until $T = 1$.

Figure 9 shows the errors and corresponding convergence rates of the conservative variables in the ℓ^1 norm with the CFL number 0.2. The BP AF methods based on the JS, LLF, and VH FVS achieve the third-order accuracy, which is not affected by the BP limitings. The convergence rate based on the SW FVS reduces to around 2, due to the non-smoothness of the SW FVS as mentioned in Remark 2.1.

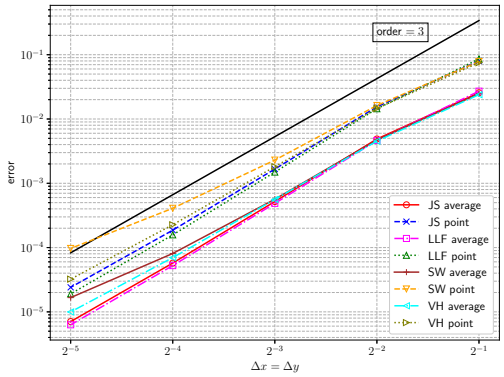

Fig. 9: Example 5.7, 2D isentropic vortex problem. The errors and convergence rates.

Example 5.8 (Quasi-2D Sod shock tube). This test solves the Sod shock tube problem along the x-direction on the domain $[0,1]\times[0,1]$ with a 100×2 uniform mesh until $T = 0.2$. The initial condition is $(\rho, v_1, v_2, p) = (1, 0, 0, 1)$ if $x < 0.5$, otherwise $(\rho, v_1, v_2, p) = (0.125, 0, 0, 0.1)$.

The density plots obtained by using different ways for the point value update without and with the shock sensor ($\kappa = 1$) are shown in Figure 10. The density based on the JS shows large deviations between the contact discontinuity and shock wave, which cannot be reduced by the limiting. Seen from Figure 11, the solutions belonging to the DoFs for different point values are decoupled, known as the mesh alignment issue, and has been explained in Subsection 3.2. The results of all the FVS-based methods agree well with the exact solution when the limiting is activated. The FVS-based AF methods are more advantageous in simulations since they can cure both the stagnation and mesh alignment issues. To save space, in the following tests, we only show the results obtained using the LLF FVS.

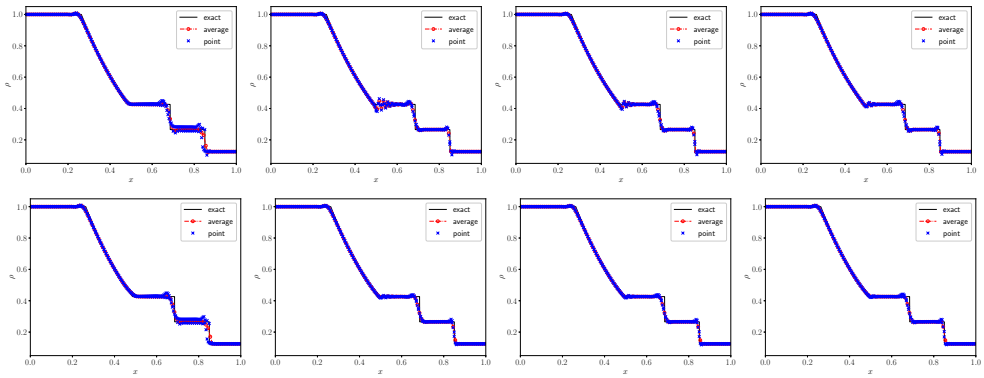

Fig. 10: Example 5.8, quasi-2D Sod shock tube. The density are computed without (top row) and with the shock sensor-based limiting ($\kappa = 1$, bottom row). From left to tight: JS, LLF, SW, and VH FVS.

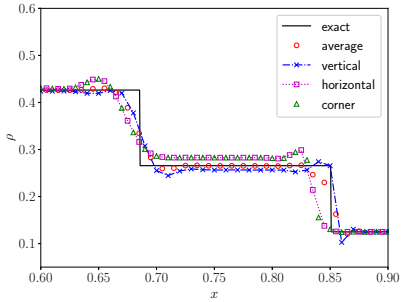

Fig. 11: Example 5.8, quasi-2D Sod shock tube. Based on the JS, the solutions that belong to different kinds of DoFs are decoupled.

Example 5.9 (Sedov blast wave). The domain is $[-1.1, 1.1] \times [-1.1, 1.1]$ with outflow boundary conditions. The initial density is one, velocity is zero, and the total energy is 10^{-12} everywhere except that for the centered cell, the total energy of the cell average and the point values on its faces are $\frac{0.979264}{\Delta x \Delta y}$ with $\Delta x = 2.2/N_1, \Delta y = 2.2/N_2$, which is used to emulate a δ-function at the center.

This test is solved until $T = 1$ and the BP limitings are necessary, otherwise, the simulation fails due to negative pressure. The density plots obtained with the shock sensor ($\kappa = 0.5$) are shown in Figure 12. The circular shock wave is well-captured and the numerical solutions converge to the exact solution without spurious oscillations. The blending coefficients based on the shock sensor are presented in Figure 13, indicating that the limiting is locally activated.

Example 5.10 (A Mach 3 wind tunnel with a forward-facing step). The initial condition is a Mach 3 flow $(\rho, v_1, v_2, p) = (1.4, 3, 0, 1)$. The computational domain is $[0, 3] \times [0, 1]$ and the step is of height 0.2 located from $x = 0.6$ to $x = 3$. The inflow and outflow boundary conditions are applied at the entrance ($x = 0$) and exit ($x = 3$), respectively, and the reflective boundary conditions are imposed at other boundaries.

The density computed by the BP AF method without and with the shock sensor-

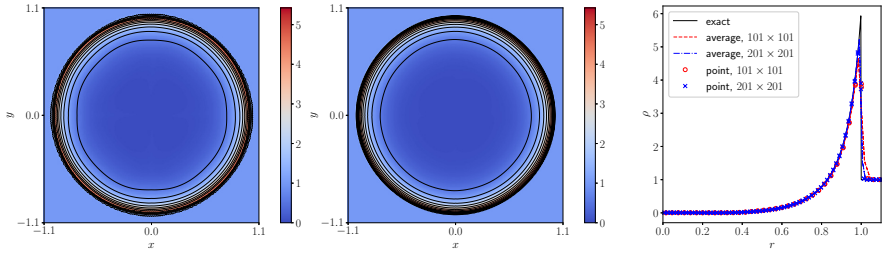

Fig. 12: Example 5.9, 2D Sedov blast wave. The density plots computed by the BP AF method. From left to right: 10 equally spaced contour lines from 0 to 5.423 on the uniform 101×101 and 201×201 meshes, respectively, cut-line along $y = x$.

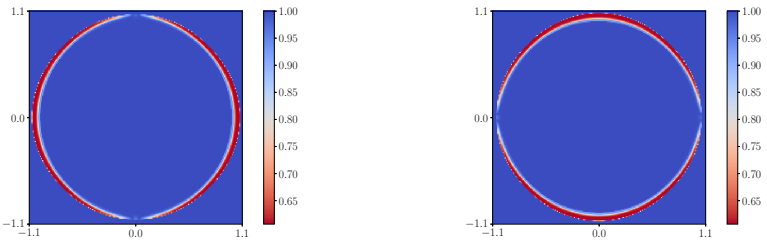

Fig. 13: Example 5.9, 2D Sedov blast wave. The shock sensor-based blending coefficients $\theta^s_{i+\frac{1}{2},j}$ (left) and $\theta^s_{i,j+\frac{1}{2}}$ (right) on the 201×201 uniform mesh.

based limiting at $T = 4$ are shown in Figure 14, and the blending coefficients $\theta^s_{i+\frac{1}{2},j}$, $\theta^s_{i,j+\frac{1}{2}}$ are presented in Figure 15. If only the BP limitings are used, there are oscillations in the numerical solutions, but the BP property is not violated. The numerical solutions can be improved by our shock sensor-based limiting. Our BP AF method can capture the main features and well-developed Kelvin–Helmholtz roll-ups that originate from the triple point. The noise after the shock waves is reduced by the shock sensor-based limiting, while the roll-ups are preserved well. Compared to the results obtained by the third-order P^2 DG method with the TVB limiter [12], the vortices are better captured with the same mesh size $\Delta x = \Delta y = 1/160, 1/320$. Note that the AF method uses fewer DoFs, showing its efficiency and potential for high Mach number flows.

Example 5.11 (High Mach number astrophysical jets). This test follows the setup in [45]. The first case considers a Mach 80 jet on a computational domain $[0, 2] \times [-0.5, 0.5]$, initially filled with ambient gas with $(\rho, v_1, v_2, p) = (0.5, 0, 0, 0.4127)$. A jet is injected into the domain with $(\rho, v_1, v_2, p) = (5, 30, 0, 0.4127)$ at the left boundary when $|y| < 0.05$. The free boundary conditions are applied on other boundaries. The second case considers a Mach 2000 jet on a computational domain $[0, 1] \times [-0.25, 0.25]$. The initial condition and boundary conditions are the same as the first case except that the state of the jet is $(\rho, v_1, v_2, p) = (5, 800, 0, 0.4127)$. The adiabatic index is $\gamma = 5/3$, and the output time is 0.07 and 0.001 for the two cases, respectively.

The numerical solutions obtained by the BP AF methods with the shock sensor on the uniform 400×200 mesh are shown in Figure 16. The main flow structures and small-scale features are captured well, comparable to those in [45].

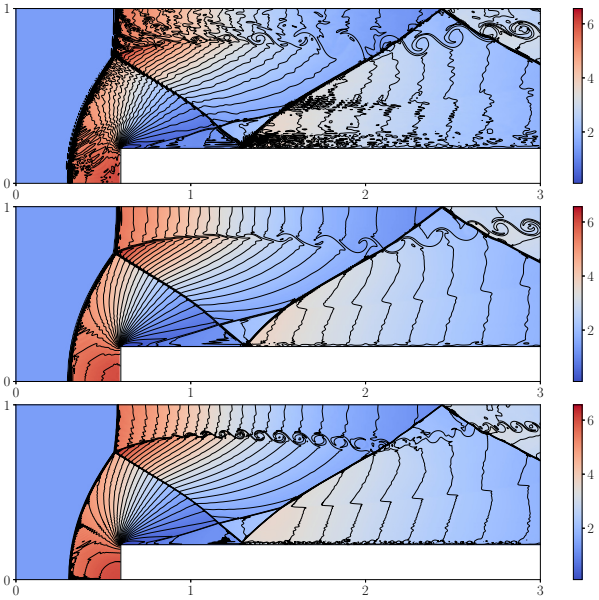

Fig. 14: Example 5.10, forward-facing step problem. 30 equally spaced contour lines of the density from 0.098 to 6.566. From top to bottom: 480×160 mesh without shock sensor, 480×160 mesh with $\kappa=1$, 960×320 mesh with $\kappa=1$.

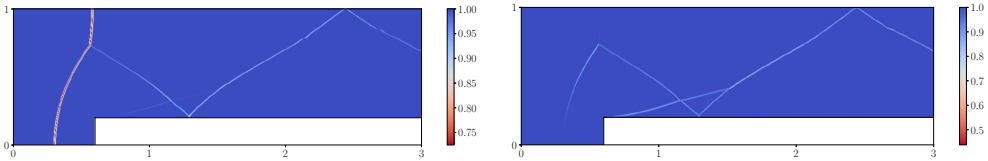

Fig. 15: Example 5.10, forward-facing step problem. The blending coefficients $\theta^s_{i+\frac{1}{2},j}$ (left) and $\theta^s_{i,j+\frac{1}{2}}$ (right) based on the shock sensor with $\kappa=1$ on the 960×320 mesh.

6. Conclusion. In the active flux (AF) methods, it is pivotal to design suitable point values update at cell interfaces, to achieve stability and high-order accuracy. The point value update based on the Jacobian splitting (JS) may lead to the stagnation and mesh alignment issues. 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 those two issues. To further improve the robustness of the AF methods, this paper developed bound-preserving (BP) AF methods for 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. The shock sensor-based limiting was proposed to further improve the shock-capturing ability. The challenging numerical tests verified the robustness and effectiveness of our BP AF methods, and also showed that the LLF FVS is generally superior to others in terms of the CFL number and non-oscillatory property. Moreover, for the forward-facing step problem, the present FVS-

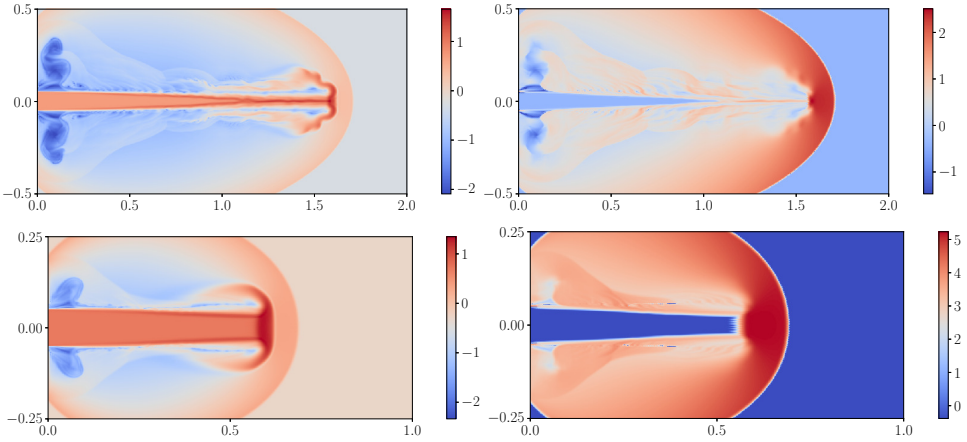

Fig. 16: Example 5.11, the Mach 80 jet (top row) and Mach 2000 jet (bottom row). $\log_{10}\rho$ (left) and $\log_{10} p$ (right) obtained with the BP limitings and shock sensor-based limiting ($\kappa = 1$ for Mach 80 and 10 for Mach 2000, respectively).

based BP AF method was able to capture small-scale features better compared to the third-order discontinuous Galerkin method with the TVB limiter on the same mesh resolution [12], while using fewer degrees of freedom, demonstrating the efficiency and potential of our BP AF method for high Mach number flows.

Acknowledgments. We acknowledge helpful discussions with Praveen Chandrashekar at TIFR-CAM Bangalore on the Ducros' shock sensor.

REFERENCES

[1] 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 equationss*, Commun. Appl. Math. Comput., 5 (2023), pp. 370–402.

[2] R. Abgrall and W. Barsukow, *Extensions of active flux to arbitrary order of accuracy*, ESAIM: Math. Model. Numer. Anal., 57 (2023), pp. 991–1027.

[3] R. Abgrall, W. Barsukow, and C. Klingenberg, *The active flux method for the Euler equations on Cartesian grids*, Oct. 2023. arXiv:2310.00683.

[4] R. Abgrall, J. Lin, and Y. Liu, *Active flux for triangular meshes for compressible flows problems*, Dec. 2023.

[5] W. Barsukow, *The active flux scheme for nonlinear problems*, J. Sci. Comput., 86 (2021), p. 3.

[6] W. Barsukow and J. P. Berberich, *A well-balanced active flux method for the shallow water equations with wetting and drying*, Commun. Appl. Math. Comput., (2023).

[7] W. Barsukow, J. P. Berberich, and C. Klingenberg, *On the active flux scheme for hyperbolic PDEs with source terms*, SIAM J. Sci. Comput., 43 (2021), pp. A4015–A4042.

[8] W. Barsukow, J. Hohm, C. Klingenberg, and P. L. Roe, *The active flux scheme on Cartesian grids and its low Mach number limit*, J. Sci. Comput., 81 (2019), pp. 594–622.

[9] E. Chudzik and C. Helzel, *A Review of Cartesian Grid Active Flux Methods for Hyperbolic Conservation Laws*, in Finite Volumes for Complex Applications X—Volume 1, Elliptic and Parabolic Problems, E. Franck, J. Fuhrmann, V. Michel-Dansac, and L. Navoret, eds., Cham, 2023, Springer Nature Switzerland, pp. 93–109.

[10] E. Chudzik, C. Helzel, and D. Kerkmann, *The Cartesian grid active flux method: Linear stability and bound preserving limiting*, Appl. Math. Comput., 393 (2021), p. 125501.

[11] S. Clain, S. Diot, and R. Loubère, *A high-order finite volume method for systems of con-*

servation laws–Multi-dimensional Optimal Order Detection (MOOD), J. Comput. Phys., 230 (2011), pp. 4028–4050.
[12] B. Cockburn and C. W. Shu, *Runge-Kutta discontinuous Galerkin methods for convection-dominated problems*, J. Sci. Comput., 16 (2001), pp. 173–261.
[13] C. J. Cotter and D. Kuzmin, *Embedded discontinuous Galerkin transport schemes with localised limiters*, J. Comput. Phys., 311 (2016), pp. 363–373.
[14] C. M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Springer Berlin Heidelberg, 2000.
[15] F. Ducros, V. Ferrand, F. Nicoud, C. Weber, D. Darracq, C. Gacherieu, and T. Poinsot, *Large-eddy simulation of the shock/turbulence interaction*, Journal of Computational Physics, 152 (1999), pp. 517–549.
[16] T. Eymann and P. Roe, *Active flux schemes*, in 49th AIAA Aerospace Sciences Meeting including the New Horizons Forum and Aerospace Exposition, Orlando, Florida, Jan. 2011, American Institute of Aeronautics and Astronautics.
[17] T. Eymann and P. Roe, *Active flux schemes for systems*, in 20th AIAA Computational Fluid Dynamics Conference, Fluid Dynamics and Co-located Conferences, American Institute of Aeronautics and Astronautics, June 2011.
[18] T. A. Eymann and P. L. Roe, *Multidimensional active flux schemes*, in 21st AIAA Computational Fluid Dynamics Conference, Fluid Dynamics and Co-located Conferences, American Institute of Aeronautics and Astronautics, June 2013.
[19] D. Fan and P. L. Roe, *Investigations of a new scheme for wave propagation*, in 22nd AIAA Computational Fluid Dynamics Conference, American Institute of Aeronautics and Astronautics, 2015.
[20] S. Gottlieb, C.-W. Shu, and E. Tadmor, *Strong Stability-Preserving High-Order Time Discretization Methods*, SIAM Rev., 43 (2001), pp. 89–112.
[21] 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.
[22] 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.
[23] 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.
[24] J.-L. Guermond, B. Popov, and I. Tomas, *Invariant domain preserving discretization-independent schemes and convex limiting for hyperbolic systems*, Comput. Method. Appl. M., 347 (2019), pp. 143–175.
[25] H. Hajduk, *Monolithic convex limiting in discontinuous Galerkin discretizations of hyperbolic conservation laws*, Comput. Math. Appl., 87 (2021), pp. 120–138.
[26] D. Hänel, R. Schwane, and G. Seider, *On the accuracy of upwind schemes for the solution of the Navier-Stokes equations*, Fluid Dynamics and Co-located Conferences, American Institute of Aeronautics and Astronautics, June 1987.
[27] 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.
[28] 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.
[29] A. Jameson, W. Schmidt, and E. Turkel, *Solutions of the Euler equations by finite volume methods using Runge-Kutta time-stepping schemes*, AIAA J., 1259 (1981).
[30] D. Kuzmin, *Monolithic convex limiting for continuous finite element discretizations of hyperbolic conservation laws*, Computer Methods in Applied Mechanics and Engineering, 361 (2020), p. 112804.
[31] D. Kuzmin, R. Löhner, and S. Turek, eds., *Flux-Corrected Transport: Principles, Algorithms, and Applications*, Scientific Computation, Springer Netherlands, Dordrecht, 2012.
[32] 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.
[33] 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.
[34] J. Maeng, *On the Advective Component of Active Flux Schemes for Nonlinear Hyperbolic Conservation Laws*, PhD thesis, 2017.
[35] B. Perthame and C.-W. Shu, *On positivity preserving finite volume schemes for Euler equations*, Numer. Math., 73 (1996), pp. 119–130.

[36] P. Roe, *Is discontinuous reconstruction really a good idea?*, J. Sci. Comput., 73 (2017), pp. 1094–1114.
[37] 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.
[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 maximum-principle-satisfying high order schemes for scalar conservation laws*, J. Comput. Phys., 229 (2010), pp. 3091–3120.
[45] 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.
[46] 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.
[47] 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.