High accurate monotonicity-preserving Semi-Lagrangian scheme for Vlasov-Poisson Simulations

C Yang, Michel Mehrenberger

High accurate monotonicity-preserving Semi-Lagrangian scheme for Vlasov-Poisson Simulations.

C. Yang[†], M. Mehrenberger[‡]

[†]School of Mathematics,
Harbin Institute of Technology
92 West Dazhi Street, Nan Gang District,
Harbin, 150001, China
yangchang@hit.edu.cn

[‡]Institut de Mathématiques de Marseille (I2M)
Aix-Marseille Université,
39, rue F. Joliot-Curie,
F-13 453 Marseille, France,
michel.mehrenberger@univ-amu.fr

October 4, 2020

Abstract

In this paper, we study a high accurate monotonicity-preserving (MP) Semi-Lagrangian scheme for Vlasov-Poisson simulations. The classical Semi-Lagrangian scheme is known to be high accurate and free from CFL condition, but it does not satisfy local maximum principle. To remedy this drawback, using the conservative form of the Semi-Lagrangian scheme, we recast existing MP schemes for the numerical flux in a common framework, and then substitute the local minimum/maximum by some "better" guess, in order to avoid as much as possible loss of accuracy and clipping near extrema, while keeping the monotonicity on monotone portions. With the limiter, on the one hand, the scheme keeps the good properties of the unlimited scheme: it is conservative, free from CFL condition and high accurate. On the other hand, for locally monotonic data, the monotonicity of the solution is preserved. Numerical tests are made on free transport equation and Vlasov-Poisson system to illustrate the robustness of our method.

1 Introduction

The Vlasov-Poisson system describes the evolution of charged particles under a self-consistent electric field. One important application of the Vlasov-Poisson system is in the study of the controlled fusion. The Vlasov-Poisson system has a number of conservation properties that need special attention when developing numerical methods. Ideally, we want numerical method to retain the exact invariants in numerical methods. However, when it is not possible to keep them all, they can be used to monitor the validity of the simulation by checking accuracy of these invariants. Many attempts have been made for solving Vlasov-Poisson system, including classical discretizations as finite difference methods [2], finite element method [49, 36, 1], finite volume method [24, 25, 13, 43] , spectral method [31], discontinuous Galerkin methods [29, 17, 10, 9, 48, 39, 30], statistical based method as particle-in-cell method [11, 42, 19, 26, 18, 14]. There is also another important category named Semi-Lagrangian methods [35, 3, 15, 33, 32, 41, 8, 44, 4, 22, 23], which is popular thanks to its good precision and as it is free from time step limitation. In this paper, we will focus on designing a Semi-Lagrangian method for the Vlasov-Poisson system.

The Semi-Lagrangian methods use the characteristics of the scalar hyperbolic equation, along with an interpolation method, to update the unknown from one time step to the next. The classical Semi-Lagrangian method with high order interpolation can generate new extrema, violate the monotonicity and develop numerical oscillations.

Some remedies have been proposed. Conservative Semi-Lagrangian methods were introduced by using flux formulation [15], permitting to add filters to impose monotonicity or positivity, while keeping the conservativity, which turns out to be satisfied for the current applications in this paper (constant advection equation with periodic boundary conditions). Such design was essentially developed by changing the derivatives at the cell edges and a full monotonic or bounded preserving solution was obtained, in the framework of Hermite representation of the solution, that is locally a polynomial of degree ≤ 3. The case of higher order reconstructions is more complex (see [20]), as one has to have conditions for a polynomial to be monotone or positive. A nearby solution is the PFC scheme that was earlier developed [27], and a local variant has been proposed by Umeda [40], and then a generalization has been performed for polynomials of degree ≤ 4 [41] instead of degree ≤ 3. Another strategy is based on a WENO type reconstruction [33, 44], typically with polynomials of degree ≤ 5. However, WENO type reconstruction is too much dissipative for long term simulation as shown in section 4. The key point of WENO type method is to find "optimal" non-linear combination weights. In contrast, we will position our design in the framework of Monotonicity-Preserving (MP) constraint, that is applied directly on classical interpolations and the modification is directly on the numerical flux [37], which removes the problem of dealing with high order polynomials and corresponding criteria to get monotonicity or positivity.

Suresh *et al.* [37] proposed such a limiter for numerical flux, which can retain high accuracy and in the meantime preserves monotonicity, so that the scheme can efficiently remove spurious oscillations. The idea is to distinguish automatically monotone portion and extrema portion of solution. On the one hand, for monotone solution, the limiter preserves monotonicity of the solution, on the other hand, it provides enough relaxation space to retain high accuracy of solution. However, this limiter has limitation of CFL condition. It has been applied recently to the Semi-Lagrangian method and for the Vlasov-Poisson system [38]. Note also that an extension work of [37] has been made [16] to give more relaxation space. We refer also to the recent paper [34] for some complementary references on the numerical resolution of the Vlasov equation by Semi-Lagrangian type methods focussing precisely on removing spurious oscillations, which is the concern of our paper.

In this paper, we focus on developing a new limiter for Semi-Lagrangian method. We first reformulate the existing MP limiters in a common framework, that is the Monotonicity-Preserving constraint mentioned

in [21], then we substitute local maximum/minimum by some "better" guess maximum/minimum. In this framework, we can easily compare among the existing limiters and propose our best choice. In numerical tests, we use L^1 norm invariant to estimate growing of spurious oscillation and L^2 norm invariant for dissipation of the solution.
The outline of this paper is following. The high accurate Semi-Lagrangian scheme is revised in the section 2. The construction for left flux and right flux is presented respectively. Some symmetry argument between left flux and right flux is also explained here. In the section 3, the new limiter is introduced. We first define the Monotonicity-Preserving constraints for both left flux and right flux. Then the proposed relaxations to the MP constraints are explained in detail and a MP property is proven. Finally, a short analysis of comparison is given for different limiters. In the section 4, we give a portion of C code for the limiters. Then numerical results for linear advection equation and Vlasov-Poisson system are collected. At the end, we give a conclusion and perspectives.

2 High accurate Semi-Lagrangian scheme

In this section, we will introduce the high accurate Semi-Lagrangian scheme for the free transport equation

$$\partial_t f + v\partial_x f = 0. \tag{1}$$

Let us first introduce uniform mesh in space as $x_i = ih$, $i \in \mathbb{Z}$, where h is a fixed mesh size. Similarly, we give discrete time as $t^n = n\Delta t$, $n \in \mathbb{Z}^+$, where Δt is time step. Hence the numerical solution at discrete grid is denoted by f_i^n.
The classical Semi-Lagrangian scheme is divided in two steps. The first step is devoted to find foot of characteristic curve. For the free transport equation (1), the foot is simply $x_i - v\Delta t$, if we start from x_i. The second step consists to approximate the solution by Lagrangian interpolation thanks to the following relation

$$f(t^{n+1}, x_i) = f(t^n, x_i - v\Delta t).$$

Next, we will recast the Lagrangian interpolation into a flux formulation, which is a preparation step for MP limiter. Two types of flux formulation are considered, left flux or right flux, corresponding to wind direction, as we have in a classical finite volume scheme.

2.1 Construction of left flux

We can always suppose the foot of characteristic curve locates between x_{j^L-1} and x_{j^L}, and can write $x_{j^L} - (x_i - v\Delta t) = \nu_L h$, thus the numerical solution by interpolation of degree one is given by $f_i^{n+1} = \nu_L f_{j^L-1}^n + (1-\nu_L) f_{j^L}^n$, and more generally of odd degree $2d+1$, with $d \in \mathbb{N}$, is given by

$$f_i^{n+1} = \sum_{\ell=-d}^{d+1} L_\ell(1-\nu_L) f_{j^L-1+\ell}^n = \sum_{\ell=-d}^{d+1} L_{1-\ell}(\nu_L) f_{j^L-1+\ell}^n = \sum_{\ell=-d}^{d+1} L_\ell(\nu_L) f_{j^L-\ell}^n, \tag{2}$$

where L_ℓ is Lagrangian basis function defined by $L_\ell(x) = \prod_{k=-d, k\neq\ell}^{d+1} \frac{x-k}{\ell-k}$.
The Semi-Lagrangian scheme (2) can be written under a conservative form. For $d=0$, we have $f_i^{n+1} = f_{j^L}^n - \nu_L\left(f_{j^L}^n - f_{j^L-1}^n\right)$. We shall more generally write the Semi-Lagrangian scheme (2) with left flux (symbolized by L), expressed as

$$f_i^{n+1} = f_{j^L}^n - \nu_L\left(f_{j^L+1/2}^L - f_{j^L-1/2}^L\right), \tag{3}$$

defining the linear formula

$$f_{j+1/2}^L = \sum_{\ell=-d}^{d} c_\ell^L f_{j+\ell}^n, \ j \in \mathbb{Z}, \tag{4}$$

where c_ℓ^L, $\ell \in \{-d, \dots, d\}$ are the coefficients to be determined. We will omit superscript "L" if there is no confusion. Thanks to this linear definition, the Semi-Lagrangian scheme can be recast as

$$\begin{aligned} f_i^{n+1} &= f_j^n - \nu \left(\sum_{\ell=-d}^{d} c_\ell^L f_{j+\ell}^n - \sum_{\ell=-d}^{d} c_\ell^L f_{j+\ell-1}^n \right) \\ &= -\nu c_d^L f_{j+d}^n + \nu c_{-d}^L f_{j-d-1}^n - \nu \sum_{\ell=-d, \ell\neq 0}^{d-1} (c_\ell^L - c_{\ell+1}^L) f_{j+\ell}^n + (1 - \nu(c_0^L - c_1^L)) f_j^n. \end{aligned}$$

Comparing the corresponding term in (2), we get

$$\begin{cases} \nu c_{-d}^L &= L_{d+1}(\nu), \\ -\nu(c_\ell^L - c_{\ell+1}^L) &= L_{-\ell}(\nu), \quad \ell = -d, \dots, -1, \\ -\nu c_d^L &= L_{-d}(\nu), \\ -\nu(c_\ell^L - c_{\ell+1}^L) &= L_{-\ell}(\nu), \quad \ell = 1, \dots, d-1, \\ 1 - \nu(c_0^L - c_1^L) &= L_0(\nu). \end{cases}$$

Solving the linear system yields, for $\nu \neq 0$, the coefficients

$$c_\ell^L = \begin{cases} \frac{1}{\nu} \sum_{k=-d-1}^{\ell-1} L_{-k}(\nu), & \ell = -d, \dots, 0, \\ -\frac{1}{\nu} \sum_{k=\ell}^{d} L_{-k}(\nu), & \ell = 1, \dots, d, \end{cases} \tag{5}$$

since $\nu c_{-d}^L + \sum_{\ell=-d}^{d-1} -\nu(c_\ell^L - c_{\ell+1}^L) - \nu c_d^L + \sum_{\ell=1}^{d-1} -\nu(c_\ell^L - c_{\ell+1}^L) + 1 - \nu(c_0^L - c_1^L) = 1$, and $L_{d+1}(\nu) + \sum_{\ell=-d}^{d-1} L_{-\ell}(\nu) + L_{-d}(\nu) + \sum_{\ell=1}^{d-1} L_{-\ell}(\nu) + L_0(\nu) = \sum_{\ell=-d}^{d+1} L_\ell(\nu) = 1$. For $\nu = 0$, the c_ℓ^L can be arbitrary from the system; we define them by taking the limit of (5), as $\nu \to 0$, which is well defined, since 0 is a root of L_k, for $k \neq 0$. For implementing issues, we can use the following formulae which have no evaluation problem:

$$c_\ell^L = \begin{cases} \sum_{k=-d-1}^{\ell-1} \tilde{L}_{-k}(\nu), & \ell = -d, \dots, 0, \\ -\sum_{k=\ell}^{d} \tilde{L}_{-k}(\nu), & \ell = 1, \dots, d, \end{cases} \tag{6}$$

with

$$\tilde{L}_\ell(x) = \frac{1}{\ell} \prod_{k=-d, k\notin\{\ell,0\}}^{d+1} \frac{x-k}{\ell-k},$$

for $\ell = -d, \dots, d+1$, $\ell \neq 0$. We report on Table 1, the expressions for some values of d.

c_0^L	1

c_{-1}^L	$\frac{1}{6}\nu^2 - \frac{1}{6}$
c_0^L	$-\frac{1}{3}\nu^2 + \frac{1}{2}\nu + \frac{5}{6}$
c_1^L	$\frac{1}{6}\nu^2 - \frac{1}{2}\nu + \frac{1}{3}$

c_{-2}^L	$\frac{1}{120}\nu^4 - \frac{1}{24}\nu^2 + \frac{1}{30}$
c_{-1}^L	$-\frac{1}{30}\nu^4 + \frac{1}{24}\nu^3 + \frac{1}{4}\nu^2 - \frac{1}{24}\nu - \frac{13}{60}$
c_0^L	$\frac{1}{20}\nu^4 - \frac{1}{8}\nu^3 - \frac{1}{3}\nu^2 + \frac{5}{8}\nu + \frac{47}{60}$
c_1^L	$-\frac{1}{30}\nu^4 + \frac{1}{8}\nu^3 + \frac{1}{12}\nu^2 - \frac{5}{8}\nu + \frac{9}{20}$
c_2^L	$\frac{1}{120}\nu^4 - \frac{1}{24}\nu^3 + \frac{1}{24}\nu^2 + \frac{1}{24}\nu - \frac{1}{20}$

Table 1: Values of $c_k^L = c_{-k}^R,\ k = -d, \dots, d$, for $d = 0, 1, 2$, using (5).

Remark 2.1. *We can propose another formula to define the left flux. Using (4) and (5), we have for* $j \in \mathbb{Z}$,

$$\begin{aligned}
\nu f_{j+1/2}^L &= \sum_{\ell=-d}^{0} \sum_{k=-d-1}^{\ell-1} L_{-k}(\nu) f_{j+\ell}^n - \sum_{\ell=1}^{d} \sum_{k=\ell}^{d} L_{-k}(\nu) f_{j+\ell}^n, \\
&= \sum_{k=-d-1}^{-1} L_{-k}(\nu) \sum_{\ell=k+1}^{0} f_{j+\ell}^n - \sum_{k=1}^{d} L_{-k}(\nu) \sum_{\ell=1}^{k} f_{j+\ell}^n \\
&= \sum_{k=-d}^{-1} \left(-\sum_{\ell=1}^{-k} f_{j+\ell}^n \right) L_k(\nu) + \sum_{k=1}^{d+1} \left(\sum_{\ell=-k+1}^{0} f_{j+\ell}^n \right) L_k(\nu).
\end{aligned}$$

Defining $S_k = -\sum_{\ell=1}^{-k} f_{j+\ell}^n,\ k = -d, \dots, -1$ *and* $S_k = \sum_{\ell=-k+1}^{0} f_{j+\ell}^n,\ k = 1, \dots, d+1$, *we have* $S_{k+1} - S_k = f_{j-k}^n$, *for* $k = -d, \dots, -2$ *and* $S_{k+1} - S_k = f_{j-k}^n$, *for* $k = 1, \dots, d$. *Denoting* $S_0 = 0$, *we have* $S_0 - S_{-1} = -S_{-1} = f_{j-1}^n$, *and* $S_1 - S_0 = S_1 = f_j^n$, *so that we have* $S_{k+1} - S_k = f_{j-k}^n$, *for* $k = -d, \dots, d$. *Thus, the* S_k *are uniquely determined, by* $S_{k+1} - S_k = f_{j-k}^n,\ k = -d, \dots, d$ *and* $S_0 = 0$ *and finally,* $\nu f_{j+1/2}^L = \sum_{k=-d}^{d+1} S_k L_k(\nu)$.

2.2 Construction of right flux

In order now, to use the right flux, we can also always suppose that the foot of characteristic curve locates between x_{j^R} and x_{j^R+1}, and can write $(x_i - v\Delta t) - x_{j^R} = \nu_R h$, thus the numerical solution by interpolation of degree one is now given by $f_i^{n+1} = (1 - \nu_R) f_{j^R}^n + \nu_R f_{j^R+1}^n$, and more generally of odd degree $2d+1$, with $d \in \mathbb{N}$, is given by

$$f_i^{n+1} = \sum_{\ell=-d}^{d+1} L_\ell(\nu_R) f_{j^R+\ell}^n. \tag{7}$$

We get the same solution as before, but we now will make a link with the conservative form using the right flux (symbolized by the letter R) instead of the left flux. We can write for $d = 0$, $f_i^{n+1} = f_{j^R}^n + \nu_R(f_{j^R+1}^n - f_{j^R}^n)$, and more generally

$$f_i^{n+1} = f_{j^R}^n + \nu_R \left(f_{j^R+1/2}^R - f_{j^R-1/2}^R \right), \tag{8}$$

defining the linear formula

$$f_{j-1/2}^R = \sum_{\ell=-d}^{d} c_\ell^R f_{j+\ell}^n,\ j \in \mathbb{Z}, \tag{9}$$

where c_ℓ^R, $\ell \in \{-d, \dots, d\}$ are the coefficients to be determined. We will also omit superscript "R" if there is no confusion.
Now, the Semi-Lagrangian scheme can be recast as

$$\begin{aligned} f_i^{n+1} &= f_j^n + \nu \left(\sum_{\ell=-d}^{d} c_\ell^R f_{j+\ell+1}^n - \sum_{\ell=-d}^{d} c_\ell^R f_{j+\ell}^n \right) \\ &= \nu c_d^R f_{j+d+1} - \nu c_{-d}^R f_{j-d} + \nu \sum_{\ell=-d, \ell\neq -1}^{d-1} (c_\ell^R - c_{\ell+1}^R) f_{j+\ell+1} + (1 + \nu(c_{-1}^R - c_0^R)) f_j \\ &= \nu c_d^R f_{j+d+1} - \nu c_{-d}^R f_{j-d} + \nu \sum_{\ell=-d+1, \ell\neq 0}^{d} (c_{\ell-1}^R - c_\ell^R) f_{j+\ell} + (1 + \nu(c_{-1}^R - c_0^R)) f_j. \end{aligned}$$

Comparing the corresponding term in (7), we get

$$\begin{cases} -\nu c_{-d}^R &= L_{-d}(\nu), \\ \nu(c_{\ell-1}^R - c_\ell^R) &= L_\ell(\nu), \quad \ell = -d+1, \dots, -1, \\ \nu c_d^R &= L_{d+1}(\nu), \\ \nu(c_{\ell-1}^R - c_\ell^R) &= L_\ell(\nu), \quad \ell = 1, \dots, d, \\ 1 + \nu(c_{-1}^R - c_0^R) &= L_0(\nu). \end{cases}$$

We find similarly the coefficients as follows

$$c_\ell^R = \begin{cases} -\sum_{k=-d}^{\ell} \tilde{L}_k(\nu), & \ell = -d, \dots, -1, \\ \sum_{k=\ell+1}^{d+1} \tilde{L}_k(\nu), & \ell = 0, \dots, d, \end{cases} \tag{10}$$

We have, for $\ell = -d, \dots, -1$, $c_\ell^R = -\sum_{k=-d}^{\ell} \tilde{L}_k(\nu)$ and $c_{-\ell}^L = -\sum_{k=-\ell}^{d} \tilde{L}_{-k}(\nu) = -\sum_{k=-d}^{\ell} \tilde{L}_k(\nu) = c_\ell^R$. On the other hand, we also have for $\ell = 0, \dots, d$, $c_\ell^R = \sum_{k=\ell+1}^{d+1} \tilde{L}_k(\nu)$ and $c_{-\ell}^L = \sum_{k=-d-1}^{-\ell-1} \tilde{L}_{-k}(\nu) = \sum_{k=\ell+1}^{d+1} \tilde{L}_k(\nu) = c_\ell^R$. We conclude that $c_\ell^R = c_{-\ell}^L$, for $\ell = -d, \dots, d$.

Remark 2.2. *By following a classical finite volume scheme (not semi-Lagrangian), we may think of using the left flux formula for $v > 0$ and the right flux formula for $v < 0$, but it is not necessary.*
On the other hand, one natural choice is to always consider the solution for which $0 \leq \nu \leq \frac{1}{2}$. Let $J \in \mathbb{Z}$ (which can be arbitrary), and consider $g(t,x) = f(t, -x + 2x_J)$ (putting $y = -x + 2x_J$, we have $x = -y + 2x_J$, and also $f(t,x) = g(t, -x + 2x_J)$). We have $g(t^{n+1}, x_i) = g(t^n, x_i + v\Delta t)$. Now, using the formula with the left flux, we write $x_j^L - (x_i + v\Delta t) = \nu_R h$, if $x_i + v\Delta t$ is between x_{j^L-1} and x_{j^L}. We then have $g_i^{n+1} = g_{j^L}^n - \nu_R \left(g_{j^L+1/2}^L - g_{j^L-1/2}^L \right)$, with $g_{j+1/2}^L = \sum_{\ell=-d}^{d} c_\ell^L g_{j+\ell}^n$, $j \in \mathbb{Z}$. Finally, we have $f_i^{n+1} = g_{2J-i}^{n+1}, i \in \mathbb{Z}$.

Remark 2.3. *The right flux formula can also be deduced by the symmetry argument. We have $f(t, x_{J+i}) = f(t, x_0 + (J+i)h) = g(t, -x_0 - (J+i)h + 2x_0 + 2Jh) = g(t, x_0 + (J-i)h) = g(t, x_{J-i})$. So, from the*

scheme for g, we can deduce a scheme for f, taking $f^n_{J+i} = g^n_{J-i}$. Let us denote $(\cdot)^{*,L}/(\cdot)^{*,R}$ for the nearest grid point of characteristic foot for left/right flux. So we have $f^{n+1}_{J+i} = g^{n+1}_{J-i} = g^n_{(J-i)^{*,L}} - \nu_R \left(g^L_{(J-i)^{*,L}+1/2} - g^L_{(J-i)^{*,L}-1/2}\right)$, with $g^L_{j+1/2} = \sum_{\ell=-d}^{d} c^L_\ell g^n_{j+\ell}$, $j \in \mathbb{Z}$. We have $x_{(J-i)^{*,L}} - (x_{J-i} + v\Delta t) = \nu_R h$, giving $(J-i)^* - \nu = J - i + v\Delta t/h$, that is $(J-i)^{*,L} - J + i - \nu_R = v\Delta t/h$ and thus $2J - (J-i)^{*,L} + \nu_R = i + J - v\Delta t/h$, that is $x_{i+J} - v\Delta t - x_{2J-(J-i)^{*,L}} = \nu_R h$, and we get $(J+i)^{*,R} = 2J-(J-i)^{*,L}$. Now, we have $g^n_{(J-i)^{*,L}} = g^n_{J+(J-i)^{*,L}-J} = f^n_{2J-(J-i)^{*,L}} = f^n_{(J+i)^{*,R}}$. We also have $g^L_{j+1/2} = \sum_{\ell=-d}^{d} c^L_\ell f^n_{2J-j-\ell}$, $j \in \mathbb{Z}$, leading to $g^L_{(J-i)^{*,L}+1/2} = \sum_{\ell=-d}^{d} c^L_\ell f^n_{2J-(J-i)^{*,L}-\ell} = \sum_{\ell=-d}^{d} c^L_\ell f^n_{(J+i)^{*,R}-\ell} = \sum_{\ell=-d}^{d} c^L_{-\ell} f^n_{(J+i)^{*,R}+\ell} = f^R_{(J+i)^{*,R}-1/2}$, and $g^L_{(J-i)^{*,L}-1/2} = \sum_{\ell=-d}^{d} c^L_\ell f^n_{(J+i)^{*,R}-\ell+1} = \sum_{\ell=-d}^{d} c^L_{-\ell} f^R_{(J+i)^{*,R}+1+\ell} = f^n_{(J+i)^{*,R}+1/2}$, where we have defined $f^R_{j-1/2} = \sum_{\ell=-d}^{d} c^L_{-\ell} f^n_{j+\ell}$, $j \in \mathbb{Z}$, and we have finally the formula $f^{n+1}_{J+i} = f^n_{(J+i)^{*,R}} - \nu_R \left(f^R_{(J+i)^{*,R}-1/2} - f^R_{(J+i)^{*,R}+1/2}\right)$, which is exactly (8), as $c^L_{-\ell} = c^R_\ell$.

3 A new limiter

The proposed scheme in the previous section is simple and very accurate for regular solutions. However, it may provoke spurious oscillations for irregular solutions. In this section, we focus on proposing a flux limiter, such that on the one hand, the scheme has a monotonicity-preserving property, and on the other hand, the scheme can preserve high accuracy.
The strategy for deriving flux limiter consists of two steps: the first is to identify the monotonicity-preserving constraints, the second is to relax the MP constraints near extrema.

3.1 Monotonicity-preserving constraints

The monotonicity-preserving constraint consists of two parts, as mentioned in [21]. The first one is that the flux $f^n_{j+1/2}$ should locate between f^n_j and f^n_{j+1}, *i.e.*

$$m_{j+1/2} \le f^n_{j+1/2} \le M_{j+1/2}, \tag{11}$$

where $m_{j+1/2} = \min\left(f^n_j, f^n_{j+1}\right)$ and $M_{j+1/2} = \max\left(f^n_j, f^n_{j+1}\right)$, for $j \in \mathbb{Z}$. The second one is TVD (Total Variation Diminishing) condition, *i.e.*

$$m_{j-1/2} \le f^{n+1}_i \le M_{j-1/2}.$$

Let us first present the scheme for left flux. Using the conservative form $f^{n+1}_i = f^n_j - \nu\left(f^n_{j+1/2} - f^n_{j-1/2}\right)$, the above TVD condition is equivalent to

$$f^n_{j-1/2} + \frac{1}{\nu}\left(f^n_j - M_{j-1/2}\right) \le f^n_{j+1/2} \le f^n_{j-1/2} + \frac{1}{\nu}\left(f^n_j - m_{j-1/2}\right).$$

Consider that $m_{j-1/2} \le f^n_{j-1/2} \le M_{j-1/2}$, then a sufficient condition of TVD scheme is

$$M_{j-1/2} + \frac{1}{\nu}\left(f^n_j - M_{j-1/2}\right) \le f^n_{j+1/2} \le m_{j-1/2} + \frac{1}{\nu}\left(f^n_j - m_{j-1/2}\right). \tag{12}$$

Hence the inequalities (11) and (12) are monotonicity-preserving constraints.

In the sequel, we will use same notations as in [37]:

$$\begin{aligned}\text{minmod}(x,y) &= \frac{1}{2}(\text{sgn}(x)+\text{sgn}(y))\min(|x|,|y|),\\ \text{median}(x,y,z) &= x+\text{minmod}(y-x,z-x),\\ I[x_1,\ldots,x_k] &= [\min(x_1,\ldots,x_k),\max(x_1,\ldots,x_k)],\end{aligned}$$

where sgn is the sign function.
As mentioned in literatures [45, 47], these monotonicity-preserving constraints will limit numerical solution to first order, the so called clipping near extrema, thus they alter high accuracy of scheme.
For instance, in [37], two cases near extrema are identified:

Case 1: When $f_j^n = f_{j+1}^n$, the constraint (11) leads that the numerical flux is limited as $f_{j+1/2}^n = f_j^n$.

Case 2: When $f_{j-1}^n = f_j^n$, the constraint (12) leads also that $f_{j+1/2}^n = f_j^n$.

It is clear that when the solution is not constant near extrema, we will lose accuracy.
To remedy this drawback of monotonicity-preserving constraints, we should relax the constraints (11) and (12) near extrema. The rule of relaxation is on the one hand to provide relaxed space as much as possible near extrema, and on the other hand to preserve monotonicity for monotone portion.
We notice that the constraints (11) and (12) are defined by the local maximum $M_{j+1/2}$ and local minimum $m_{j+1/2}$. Thus, one possible way to relax the monotonicity-preserving constraints near extrema is to replace the local maximum/minimum by some "better" guess of maximum/minimum. More precisely, we denote the "better" guess maximum/minimum by $M^{(1)}/m^{(1)}$ and $M^{(2)}/m^{(2)}$. Then injecting them into the constraints (11) and (12) yields

$$m_{j+1/2}^{(1)} \le f_{j+1/2}^n \le M_{j+1/2}^{(1)}, \tag{13}$$

and

$$M_{j-1/2}^{(2)} + \frac{1}{\nu}\left(f_j^n - M_{j-1/2}^{(2)}\right) \le f_{j+1/2}^n \le m_{j-1/2}^{(2)} + \frac{1}{\nu}\left(f_j^n - m_{j-1/2}^{(2)}\right). \tag{14}$$

Thus the definition of new monotonicity-preserving constraint is deduced as follows

Definition 3.1. *For monotonic data, that is $f_{j-2} \le f_{j-1} \le f_j \le f_{j+1} \le f_{j+2}$ or $f_{j-2} \ge f_{j-1} \ge f_j \ge f_{j+1} \ge f_{j+2}$, if the following constraints are verified*

1. $m_{j+1/2} \le m_{j+1/2}^{(1)}$,
2. $M_{j+1/2} \ge M_{j+1/2}^{(1)}$,
3. $m_{j-1/2}^{(2)} + \frac{1}{\nu}(f_j - m_{j-1/2}^{(2)}) \le m_{j-1/2} + \frac{1}{\nu}(f_j - m_{j-1/2})$,
4. $M_{j-1/2}^{(2)} + \frac{1}{\nu}(f_j - M_{j-1/2}^{(2)}) \ge M_{j-1/2} + \frac{1}{\nu}(f_j - M_{j-1/2})$,

then, the constraint defined in (13)-(14) is Monotonicity-Preserving (MP for short).

Remark 3.1. *For increasing data, the MP constraint can also be recast as*

$$\max\left(m_{j+1/2}^{(1)}, M_{j-1/2}^{(2)} + \frac{1}{\nu}(f_j - M_{j-1/2}^{(2)})\right) \ge \max\left(m_{j+1/2}, M_{j-1/2} + \frac{1}{\nu}(f_j - M_{j-1/2})\right) = f_j,$$

and

$$\min\left(M^{(1)}_{j+1/2}, m^{(2)}_{j-1/2}+\frac{1}{\nu}(f_j-m^{(2)}_{j-1/2})\right)\leq \min\left(M_{j+1/2}, m_{j-1/2}+\frac{1}{\nu}(f_j-m_{j-1/2})\right)=\min(f_{j+1},\Phi_{\nu,j}(f_{j-1})).$$

For decreasing data, the MP constraint is equivalent to

$$\max\left(m^{(1)}_{j+1/2}, M^{(2)}_{j-1/2}+\frac{1}{\nu}(f_j-M^{(2)}_{j-1/2})\right)\geq \max\left(f_{j+1},\Phi_{\nu,j}(f_{j-1})\right),$$

and

$$\min\left(M^{(1)}_{j+1/2}, m^{(2)}_{j-1/2}+\frac{1}{\nu}(f_j-m^{(2)}_{j-1/2})\right)\leq f_j.$$

3.2 Limiter corresponding to right flux

Similarly, one can write down corresponding MP constraints for right flux. However, we shall show a little bit more the relation between MP constraints for left flux and right flux. For the left formula, we have

$$f_i^{n+1}=f^n_{j^L}-\nu_L\left(f^L_{j^L+1/2}-f^L_{j^L-1/2}\right).$$

When we use the right formula, we have

$$f_i^{n+1}=f^n_{j^R}+\nu_R\left(f^R_{j^R+1/2}-f^R_{j^R-1/2}\right)$$

We have $x_i-v\Delta t=x_{j^L}-\nu_L h$ and $x_i-v\Delta t=x_{j^R}+\nu_R h$, with $j^L,j^R\in\mathbb{Z}$ and $\nu_L,\nu_R\in]0,1[$. So we get $j^L-\nu_L=j^R+\nu_R=j^L-1+1-\nu_L$. So, we get $j^R=j^L-1$ and $\nu_R=1-\nu_L$. Thus for the right formula, we have

$$f_i^{n+1}=f^n_{j^L-1}+(1-\nu_L)\left(f^R_{j^R+1/2}-f^R_{j^R-1/2}\right)=f^n_{j^L}+f^n_{j^L-1}-f^n_{j^L}+(1-\nu_L)\left(f^R_{j^R+1/2}-f^R_{j^R-1/2}\right),$$

that is

$$f_i^{n+1}=f^n_{j^L}-(f^n_{j^L}-(1-\nu_L)f^R_{j^R+1/2})+f^n_{j^L-1}-(1-\nu_L)f^R_{j^R-1/2}.$$

From the unicity of the formula, we have

$$\nu_L f^L_{j^L+1/2}=f^n_{j^L}-(1-\nu_L)f^R_{j^R+1/2}=f^n_{j^L}-\nu_R f^R_{j^R+1/2}.$$

For the limiting, we have $m^{(1,L)}_{j^L+1/2}\leq f^L_{j^L+1/2}\leq M^{(1,L)}_{j^L+1/2}$ together with

$$M^{(2,L)}_{j^L-1/2}+\frac{1}{\nu_L}\left(f^n_{j^L}-M^{(2,L)}_{j^L-1/2}\right)\leq f^L_{j^L+1/2}\leq m^{(2,L)}_{j^L-1/2}+\frac{1}{\nu_L}\left(f^n_{j^L}-m^{(2,L)}_{j^L-1/2}\right),\tag{15}$$

leading to the limiting

$$f^{n,L}_{j^L+1/2}=\text{median}\left(f_{\min,j^L},f^L_{j^L+1/2},f_{\max,j^L}\right)$$

with

$$\begin{aligned}
f_{\min,j^L} &= \max\left(m^{(1,L)}_{j^L+1/2}, M^{(2,L)}_{j^L-1/2}+\frac{1}{\nu_L}\left(f^n_{j^L}-M^{(2,L)}_{j^L-1/2}\right)\right),\\
f_{\max,j^L} &= \min\left(M^{(1,L)}_{j^L+1/2}, m^{(2,L)}_{j^L-1/2}+\frac{1}{\nu_L}\left(f^n_{j^L}-m^{(2,L)}_{j^L-1/2}\right)\right).
\end{aligned}$$

The conditions for the right flux is on the other hand $m^{(1,R)}_{j^R-1/2} \le f^R_{j^R-1/2} \le M^{(1,R)}_{j^R-1/2}$ together with

$$M^{(2,R)}_{j^R+1/2} + \frac{1}{\nu_R}\left(f^n_{j^R} - M^{(2,R)}_{j^R+1/2}\right) \le f^R_{j^R-1/2} \le m^{(2,R)}_{j^R+1/2} + \frac{1}{\nu_R}\left(f^n_{j^R} - m^{(2,R)}_{j^R+1/2}\right), \tag{16}$$

leading to the limiting

$$f^{n,R}_{j^R-1/2} = \text{median}\left(f_{\min,j^R}, f^R_{j^R-1/2}, f_{\max,j^R}\right)$$

with

$$\begin{aligned} f_{\min,j^R} &= \max\left(m^{(1,R)}_{j^R-1/2}, M^{(2,R)}_{j^R+1/2} + \frac{1}{\nu_R}\left(f^n_{j^R} - M^{(2,R)}_{j^R+1/2}\right)\right), \\ f_{\max,j^R} &= \min\left(M^{(1,R)}_{j^R-1/2}, m^{(2,R)}_{j^R+1/2} + \frac{1}{\nu_R}\left(f^n_{j^R} - m^{(2,R)}_{j^R+1/2}\right)\right). \end{aligned}$$

From (15), we have

$$\nu_L M^{(2,L)}_{j^L-1/2} + \left(f^n_{j^L} - M^{(2,L)}_{j^L-1/2}\right) \le \nu_L f^L_{j^L+1/2} \le \nu_L m^{(2,L)}_{j^L-1/2} + \left(f^n_{j^L} - m^{(2,L)}_{j^L-1/2}\right),$$

and so

$$(1-\nu_L) m^{(2,L)}_{j^L-1/2} \le f^n_{j^L} - \nu_L f^L_{j^L+1/2} \le (1-\nu_L) M^{(2,L)}_{j^L-1/2},$$

which leads to

$$m^{(2,L)}_{j^L-1/2} \le f^R_{j^R+1/2} \le M^{(2,L)}_{j^L-1/2}.$$

We have also from $m^{(1,L)}_{j^L+1/2} \le f^L_{j^L+1/2} \le M^{(1,L)}_{j^L+1/2}$, that

$$\nu_L m^{(1,L)}_{j^L+1/2} \le \nu_L f^L_{j^L+1/2} \le \nu_L M^{(1,L)}_{j^L+1/2},$$

leading to

$$f^n_{j^L} - \nu_L M^{(1,L)}_{j^L+1/2} \le f^n_{j^L} - \nu_L f^L_{j^L+1/2} \le f^n_{j^L} - \nu_L m^{(1,L)}_{j^L+1/2},$$

that is

$$f^n_{j^L} - (1-\nu_R) M^{(1,L)}_{j^L+1/2} \le \nu_R f^R_{j^R+1/2} \le f^n_{j^L} - (1-\nu_R) m^{(1,L)}_{j^L+1/2},$$

which leads to

$$M^{(1,L)}_{j^L+1/2} + \frac{1}{\nu_R}(f^n_{j^L} - M^{(1,L)}_{j^L+1/2}) \le f^R_{j^R+1/2} \le m^{(1,L)}_{j^L+1/2} + \frac{1}{\nu_R}(f^n_{j^L} - m^{(1,L)}_{j^L+1/2}).$$

We deduce that when we have $m^{(1,R)}_{j^R+1/2} = m^{(2,L)}_{j^L-1/2}$, $M^{(1,R)}_{j^R+1/2} = M^{(2,L)}_{j^L-1/2}$ and $m^{(2,R)}_{j^R+1/2} = m^{(1,L)}_{j^L-1/2}$, $M^{(2,R)}_{j^R+1/2} = M^{(1,L)}_{j^L-1/2}$, then the limiter for right flux is equivalent to the one for left flux.

3.3 Relax the MP constraint (11)

To relax the constraint (11), there are several ways. In the sequel, we will only consider limiter for left flux.

3.3.1 Linear extrapolation

We can find guess of maximum/minimum by using linear extrapolation. That is to search maximum/minimum by extrapolating from f_{j-1}^n and f_j^n (or f_j^n and f_{j+1}^n) to interval $]x_j, x_{j+1}[$. Let us define the following notations:

$$\begin{aligned} f^{FL}(\alpha) &= f_j^n + \alpha\left(f_j^n - f_{j-1}^n\right), \\ f^{FR}(\alpha) &= f_{j+1}^n + (1-\alpha)\left(f_{j+1}^n - f_{j+2}^n\right), \\ f^{AV}(\alpha) &= (1-\alpha)f_j^n + \alpha f_{j+1}^n, \end{aligned}$$

where $\alpha \in [0,1]$. Let us set

$$f_{j+1/2}^{MD}(\alpha) = \text{median}(f^{AV}(\alpha), f^{FL}(\alpha), f^{FR}(\alpha)) = (1-\alpha)f_j + \alpha f_{j+1} - \text{minmod}(\alpha d_j, (1-\alpha)d_{j+1}).$$

We write here $f^{MD}(\alpha)$ instead of $f_{j+1/2}^{MD}(\alpha)$ for brevity. We have $f^{MD}(0) = f_j$, $f^{MD}(1) = f_{j+1}$, $f^{MD}(1/2) = \frac{f_j+f_{j+1}}{2} - \frac{1}{2}\text{minmod}(d_j, d_{j+1})$.
Using the above notations, we can define the following guess of maximum/minimum:

- Median (MD for short)

$$m_{j+1/2}^{MD} = \min_{0\le\alpha\le 1} f^{MD}(\alpha),\ M_{j+1/2}^{MD} = \max_{0\le\alpha\le 1} f^{MD}(\alpha).$$

  In practice, we take

$$\alpha = \text{median}\left(0, 1, \frac{2f_{j+1}^n - f_j^n - f_{j+2}^n}{f_j^n - f_{j-1}^n + f_{j+1}^n - f_{j+2}^n + \varepsilon}\right), \tag{17}$$

  with $\varepsilon = 10^{-10}$.

- Daru-Tenaud (DaTe for short)

$$m_{j+1/2}^{DaTe} = \min_{\alpha\in\{0,1/2,1\}} f^{MD}(\alpha),\ M_{j+1/2}^{DaTe} = \max_{\alpha\in\{0,1/2,1\}} f^{MD}(\alpha),$$

- Total variation diminishing (TVD for short)

$$m_{j+1/2}^{TVD} = \min_{\alpha\in\{0,1\}} f^{MD}(\alpha),\ M_{j+1/2}^{TVD} = \max_{\alpha\in\{0,1\}} f^{MD}(\alpha).$$

Proposition 3.1. *For monotonic data, the constraint defined in (13)-(14) is Monotonicity-Preserving when substituting $m_{j+1/2}^{(1)}/M_{j+1/2}^{(1)}$ by $m_{j+1/2}^{X}/M_{j+1/2}^{X}$ and $m_{j-1/2}^{(2)}/M_{j-1/2}^{(2)}$ by $m_{j-1/2}^{X}/M_{j-1/2}^{X}$, with $X \in \{TVD, DaTe, MD\}$.*

Proof. We suppose that the data are increasing or decreasing. More precisely, we suppose that we have $f_{j-2} \le f_{j-1} \le f_j \le f_{j+1} \le f_{j+2}$ (or $f_{j-2} \ge f_{j-1} \ge f_j \ge f_{j+1} \ge f_{j+2}$). For the first case, f^{MD} is derivable, except maybe on a finite number of points, and the derivative is $f_j - f_{j-1}$ or $f_{j+2} - f_{j+1}$ or $f_{j+1} - f_j$, which is nonnegative, if as soon as $f_{j-1} \le f_j \le f_{j+1} \le f_{j+2}$. So, we get under this condition (and also under the condition $f_{j-1} \ge f_j \ge f_{j+1} \ge f_{j+2}$), that $m_{j+1/2}^{TVD} \le m_{j+1/2}^{X}$ and $M_{j+1/2}^{X} \le M_{j+1/2}^{TVD}$, for $X \in \{TVD, DaTe, MD\}$.

We define $\Phi_{j,\nu}(x) = x + \frac{1}{\nu}(f_j - x)$, which is non increasing, as we suppose $0 < \nu < 1$. We thus get $\Phi_{j,\nu}(m^X_{j-1/2}) \leq \Phi_{j,\nu}(m^{TVD}_{j-1/2})$, for $X \in \{TVD, DaTe, MD\}$, as soon as $f_{j-2} \leq f_{j-1} \leq f_j \leq f_{j+1}$ (or $f_{j-2} \geq f_{j-1} \geq f_j \geq f_{j+1}$), and similarly $\Phi_{j,\nu}(M^{TVD}_{j-1/2}) \leq \Phi_{j,\nu}(M^X_{j-1/2})$.
On the other hand, by the definition, we always have $m^{TVD}_{j+1/2} \geq m^X_{j+1/2}$ and $M^X_{j+1/2} \geq M^{TVD}_{j+1/2}$, for $X \in \{TVD, DaTe, MD\}$.
We deduce that, if we have $f_{j-2} \leq f_{j-1} \leq f_j \leq f_{j+1} \leq f_{j+2}$ (or $f_{j-2} \geq f_{j-1} \geq f_j \geq f_{j+1} \geq f_{j+2}$), we get $m^X_{j+1/2} = m^{TVD}_{j+1/2}$, $M^X_{j+1/2} = M^{TVD}_{j+1/2}$, $m^X_{j-1/2} = m^{TVD}_{j-1/2}$, $M^X_{j-1/2} = M^{TVD}_{j-1/2}$, together with $\Phi_{j,\nu}(m^X_{j-1/2}) = \Phi_{j,\nu}(m^{TVD}_{j-1/2})$, $\Phi_{j,\nu}(M^X_{j-1/2}) = \Phi_{j,\nu}(M^{TVD}_{j-1/2})$, $\Phi_{j,\nu}(m^X_{j+1/2}) = \Phi_{j,\nu}(m^{TVD}_{j+1/2})$, $\Phi_{j,\nu}(M^X_{j+1/2}) = \Phi_{j,\nu}(M^{TVD}_{j+1/2})$. □

Remark 3.2. *For monotonic data, we also have for $X, Y \in \{TVD, DaTe, MD\}$*

$$m^X_{j+1/2} = m^{TVD}_{j+1/2}, M^X_{j+1/2} = M^{TVD}_{j+1/2}, m^Y_{j-1/2} = m^{TVD}_{j-1/2}, M^Y_{j-1/2} = M^{TVD}_{j-1/2},$$

and

$$\Phi_{j,\nu}(m^Y_{j-1/2}) = \Phi_{j,\nu}(m^{TVD}_{j-1/2}), \Phi_{j,\nu}(M^Y_{j-1/2}) = \Phi_{j,\nu}(M^{TVD}_{j-1/2}),$$

$$\Phi_{j,\nu}(m^X_{j+1/2}) = \Phi_{j,\nu}(m^{TVD}_{j+1/2}), \Phi_{j,\nu}(M^X_{j+1/2}) = \Phi_{j,\nu}(M^{TVD}_{j+1/2}).$$

Thus the constraint (13)-(14) is still Monotonicity-Preserving.

3.3.2 Umeda's method

Umeda [40] (Um for short) has proposed another way to define guess maximum/minimum by linear extrapolation, that is

$$m^{Um}_{j+1/2} = \min(\min(f^{FL}(0), f^{FR}(1)), \max(f^{FL}(1), f^{FR}(0))).$$

and

$$M^{Um}_{j+1/2} = \max(\max(f^{FL}(0), f^{FR}(1)), \min(f^{FL}(1), f^{FR}(0))).$$

Proposition 3.2. *For monotonic data, the constraint defined in (13)-(14) is Monotonicity-Preserving when substituting $m^{(1)}_{j+1/2}/M^{(1)}_{j+1/2}$ by $m^{Um}_{j+1/2}/M^{Um}_{j+1/2}$ and $m^{(1)}_{j-1/2}/M^{(1)}_{j-1/2}$ by $m^{Um}_{j-1/2}/M^{Um}_{j-1/2}$.*

Proof. For increasing data, we have

$$\begin{aligned} m^{Um}_{j+1/2} &= \min\left(\min(f_j f_{j+1}), \max(2f_j - f_{j-1}, 2f_{j+1} - f_{j+2})\right) \\ &= \min\left(f_j, \max(f_j + (f_j - f_{j-1}), 2f_{j+1} - f_{j+2})\right). \end{aligned}$$

It is clear that $f_j + (f_j - f_{j-1}) \geq f_j$, so that $\max(f_j + (f_j - f_{j-1}), 2f_{j+1} - f_{j+2}) \geq f_j$. Thus $m^{Um}_{j+1/2} = f_j$.
On the other hand, we have

$$\begin{aligned} M^{Um}_{j+1/2} &= \max\left(\max(f_j f_{j+1}), \min(2f_j - f_{j-1}, 2f_{j+1} - f_{j+2})\right) \\ &= \max\left(f_{j+1}, \min(2f_j - f_{j-1}, f_{j+1} + (f_{j+1} - f_{j+2}))\right) \end{aligned}$$

It is clear that $f_{j+1} + (f_{j+1} - f_{j+2}) \leq f_{j+1}$, so that $\min(2f_j - f_{j-1}, f_{j+1} + (f_{j+1} - f_{j+2})) \geq f_{j+1}$. Thus $M^{Um}_{j+1/2} = f_{j+1}$.
Similarly, we have $m^{Um}_{j-1/2} = f_{j-1}$ and $M^{Um}_{j-1/2} = f_j$. Therefore, we have the result. □

3.4 Relax the MP constraint (12)

Let us consider the Case 2 of the subsection 3.1. Without loss of generality, we suppose that the solution is monotone increasing, so the constraint (12) becomes

$$f_j^n \le f_{j+1/2}^n \le f_{j-1}^n + \frac{1}{\nu}(f_j^n - f_{j-1}^n) = f_j^n + \left(\frac{1}{\nu} - 1\right)(f_j^n - f_{j-1}^n).$$

In the case $f_j^n = f_{j-1}^n$, it reduces that $f_{j+1/2}^n = f_j^n$. One way to relax the constraint (12) is to replace $f_j^n - f_{j-1}^n$ by $d_{j-1/2} = \text{minmod}(d_{j-1}, d_j)$, that is $f_j^n + \left(\frac{1}{\nu} - 1\right) d_{j-1/2}$. For positive $d_{j-1/2}$, we obviously gain more space. On the other hand, for increasing data, we always have $d_{j-1/2} \le f_j^n - f_{j-1}^n$, thus the monotonicity is preserved.
Now we can reformulate $f_j^n + \left(\frac{1}{\nu} - 1\right) d_{j-1/2}$ as

$$f_j^n + \left(\frac{1}{\nu} - 1\right) d_{j-1/2} = f_{j^-}^{LC} + \frac{1}{\nu}(f_j^n - f_{j^-}^{LC}) = \Phi_{\nu,j}(f_{j^-}^{LC}),$$

with $f_{j^-}^{LC} = f_j^n - d_{j-1/2}$. Combining with the limiter introduced in section 3.3, we get

$$\max\left(m_{j+1/2}^X, \Phi_{\nu,j}(\max(M_{j-1/2}^Y, f_{j^-}^{LC}))\right) \le f_{j+1/2} \le \min\left(M_{j+1/2}^X, \Phi_{\nu,j}(\min(m_{j-1/2}^Y, f_{j^-}^{LC}))\right),$$

for $X, Y \in \{MD, DaTe, TVD, Um\}$. The following proposition shows that the above limiter is monotonicity-preserving.

Proposition 3.3. *For increasing data, we have*

$$\max\left(m_{j+1/2}^X, \Phi_{\nu,j}(\max(M_{j-1/2}^Y, f_{j^-}^{LC}))\right) = f_j,$$

and

$$\min\left(M_{j+1/2}^X, \Phi_{\nu,j}(\min(m_{j-1/2}^Y, f_{j^-}^{LC}))\right) = \min(f_{j+1}, \Phi_{\nu,j}(f_{j-1})).$$

For decreasing data, we have

$$\max\left(m_{j+1/2}^X, \Phi_{\nu,j}(\max(M_{j-1/2}^Y, f_{j^-}^{LC}))\right) = \max(f_{j+1}, f_{j^-}^{UL}),$$

and

$$\min\left(M_{j+1/2}^X, \Phi_{\nu,j}(\min(m_{j-1/2}^Y, f_{j^-}^{LC}))\right) = f_j.$$

Proof. For increasing data, we obviously have

$$\begin{aligned}\max\left(m_{j+1/2}^X, \Phi_{\nu,j}(\max(M_{j-1/2}^Y, f_{j^-}^{LC}))\right) &= \max\left(f_j, \Phi_{\nu,j}(\max(f_j, f_{j^-}^{LC}))\right) \\ &= \max\left(f_j, \min(f_j, \Phi_{\nu,j}(f_{j^-}^{LC})\right) = f_j.\end{aligned}$$

On the other hand,

$$\begin{aligned}&\min\left(M_{j+1/2}^X, \Phi_{\nu,j}(\min(m_{j-1/2}^Y, f_{j^-}^{LC}))\right) = \min\left(f_{j+1}, \Phi_{\nu,j}(\min(f_{j-1}, f_{j^-}^{LC}))\right) \\ &= \min\left(f_{j+1}, \max(\Phi_{\nu,j}(f_{j-1}), \Phi_{\nu,j}(f_{j^-}^{LC}))\right).\end{aligned}$$

We want to prove that, if the data are increasing, then $\Phi_{\nu,j}(f_{j-1}) \geq \Phi_{\nu,j}(f_{j^-}^{LC})$, that is $f_{j-1} \leq f_{j^-}^{LC}$, that is

$$\text{minmod}(d_j, d_{j-1}) \leq f_j - f_{j-1}$$

If $d_j d_{j-1} \leq 0$ or ($d_j < 0$ and $d_{j-1} < 0$), this is true, since $f_j \geq f_{j-1}$. Otherwise $d_j > 0$ and $d_{j-1} > 0$, that is $f_{j-1} - 2f_j + f_{j+1} > 0$ and $f_{j-2} - 2f_{j-1} + f_j > 0$. Do we have:

$$f_{j-2} - 2f_{j-1} + f_j \leq f_j - f_{j-1}?$$

that is

$$f_{j-2} - f_{j-1} \leq 0,$$

which is true.
Now we consider that the date are decreasing, we obviously have

$$\begin{aligned}\min\left(M_{j+1/2}^X, \Phi_{\nu,j}(\min(m_{j-1/2}^Y, f_{j^-}^{LC}))\right) &= \min\left(f_j, \Phi_{\nu,j}(\min(f_j, f_{j^-}^{LC}))\right) \\ &= \min\left(f_j, \max(f_j, \Phi_{\nu,j}(f_{j^-}^{LC}))\right) = f_j.\end{aligned}$$

On the other hand,

$$\begin{aligned}\max\left(m_{j+1/2}^X, \Phi_{\nu,j}(\max(M_{j-1/2}^Y, f_{j^-}^{LC}))\right) &= \max\left(f_{j+1}, \Phi_{\nu,j}(\max(f_{j-1}, f_{j^-}^{LC}))\right) \\ &= \max\left(f_{j+1}, \min(\Phi_{\nu,j}(f_{j-1}), \Phi_{\nu,j}(f_{j^-}^{LC}))\right)\end{aligned}$$

We want to prove that, if the data is decreasing, then $\Phi_{\nu,j}(f_{j-1}) \leq \Phi_{\nu,j}(f_{j^-}^{LC})$, which rewrites $f_{j^-}^{LC} \leq f_{j-1}$ that is

$$f_j - f_{j-1} \leq \text{minmod}(d_j, d_{j-1}).$$

If $d_j d_{j-1} \leq 0$ or ($d_j > 0$ and $d_{j-1} > 0$), then it is true, since $f_j \leq f_{j-1}$. Otherwise, we have $d_{j-1} < 0$ and $d_j < 0$, and $\text{minmod}(d_j, d_{j-1}) = -\min(-d_j, -d_{j-1}) = \max(d_j, d_{j-1})$. Do we have

$$f_j - f_{j-1} \leq f_j - 2f_{j-1} + f_{j-2}?$$

that is $f_{j-1} \leq f_{j-2}$, which is true. □

We now consider the limiter for right flux, by symmetric argument, we have the constraint

$$\max\left(m_{j^R+1/2}^X, \Phi_{\nu_R,j^R+1}(\max(M_{j^R+3/2}^Y, f_{j^+}^{LC}))\right) \leq f_{j^R+1/2}^R \leq \min\left(M_{j^R+1/2}^X, \Phi_{\nu_R,j^R+1}(\min(m_{j^R+3/2}^Y, f_{j^+}^{LC}))\right),$$

with $f_{j^+}^{LC} := f_{j^R+1} - \text{minmod}(d_{j^R+1}, d_{j^R+2}) = f_{j^L} - \text{minmod}(d_{j^L}, d_{j^L+1})$. As shown in section 3.2, we have $m_{j^R+3/2}^{(2,R)} = m_{j^L+1/2}^{(1,L)}$ and $M_{j^R+3/2}^{(2,R)} = M_{j^L+1/2}^{(1,L)}$, which can also act in left flux as (omitting superscript "L")

$$\max\left(\min(m_{j+1/2}^X, f_{j^+}^{LC}), \Phi_{\nu,j}(M_{j-1/2}^Y)\right) \leq f_{j+1/2} \leq \min\left(\max(M_{j+1/2}^X, f_{j^+}^{LC}), \Phi_{\nu,j}(m_{j-1/2}^Y)\right),$$

for $X, Y \in \{MD, DaTe, TVD, Um\}$. The next proposition shows that this limiter is also monotonicity-preserving.

Proposition 3.4. *For increasing data, we have*

$$\max\left(\min(m^X_{j+1/2}, f^{LC}_{j^+}), \Phi_{\nu,j}(M^Y_{j-1/2})\right) = f_j,$$

and

$$\min\left(\max(M^X_{j+1/2}, f^{LC}_{j^+}), \Phi_{\nu,j}(m^Y_{j-1/2})\right) = \min(f_{j+1}, \Phi_{\nu,j}(f_{j-1})).$$

For decreasing data, we have

$$\max\left(\min(f_{j+1}, f^{LC}_{j^+}), \Phi_{\nu,j}(f_{j-1}))\right) = \max(f_{j+1}, \Phi_{\nu,j}(f_{j-1})),$$

and

$$\min\left(\max(f_j, f^{LC}_{j^+}), f_j\right) = f_j.$$

Proof. For increasing data, we first have

$$\max\left(\min(m^X_{j+1/2}, f^{LC}_{j^+}), \Phi_{\nu,j}(M^Y_{j-1/2})\right) = \max\left(\min(f_j, f^{LC}_{j^+}), \Phi_{\nu,j}(f_j)\right) = \max\left(\min(f_j, f^{LC}_{j^+}), f_j\right) = f_j.$$

We also have

$$\min\left(\max(M^X_{j+1/2}, f^{LC}_{j^+}), \Phi_{\nu,j}(m^Y_{j-1/2})\right) = \min\left(\max(f_{j+1}, f^{LC}_{j^+}), \Phi_{\nu,j}(f_{j-1})\right).$$

We want to prove that, if the data are increasing, then $f^{LC}_{j^+} \leq f_{j+1}$, that is $f_j - \text{minmod}(d_j, d_{j+1}) \leq f_{j+1}$, that is $f_j - f_{j+1} \leq \text{minmod}(d_j, d_{j+1})$; this is true, if $\text{minmod}(d_j, d_{j+1}) \geq 0$. Now, if $\text{minmod}(d_j, d_{j+1}) < 0$, we have $d_j < 0$ and $d_{j+1} < 0$, that is $f_{j-1} - 2f_j + f_{j+1} < 0$ and $f_j - 2f_{j+1} + f_{j+2} < 0$. We also have $d_{j+1} \leq \text{minmod}(d_j, d_{j+1}) < 0$. So the question is: do we have $f_j - f_{j+1} \leq f_j - 2f_{j+1} + f_{j+2}$? this is equivalent to $f_{j+1} \leq f_{j+2}$ which is true, and we get the result.
For decreasing data, the equality $\min\left(\max(f_j, f^{LC}_{j^+}), f_j\right) = f_j$ is always true. So, it remains to prove that, for decreasing data, we have $f_{j+1} \leq f^{LC}_{j^+}$, which writes $f_{j+1} \leq f_j - \text{minmod}(d_j, d_{j+1})$, that is $f_{j+1} - f_j \leq -\text{minmod}(d_j, d_{j+1})$ this is true, if $\text{minmod}(d_j, d_{j+1}) \leq 0$, since we have then $f_{j+1} - f_j \leq 0 \leq -\text{minmod}(d_j, d_{j+1})$. Now, if $\text{minmod}(d_j, d_{j+1}) < 0$, we have $d_j < 0$ and $d_{j+1} < 0$, that is $f_{j-1} - 2f_j + f_{j+1} < 0$ and $f_j - 2f_{j+1} + f_{j+2} < 0$. We also have $d_j \leq \text{minmod}(d_j, d_{j-1}) < 0$. So the question is: do we have $f_{j+1} - f_j \leq f_{j-1} - 2f_j + f_{j+1}$? this is equivalent to $f_j \leq f_{j-1}$ which is true, and we get the result. □

Finally, we conclude that we can propose the following limiting:

$$f^n_{j+1/2} = \text{median}\left(f_{\min,j}, f^L_{j+1/2}, f_{\max,j}\right),$$

with

$$f_{\min,j} = \min\left(\max\left(m^X_{j+1/2}, \Phi_{\nu,j}(\max(M^Y_{j-1/2}, f^{LC}_{j^-}))\right), \max\left(\min(m^X_{j+1/2}, f^{LC}_{j^+}), \Phi_{\nu,j}(M^Y_{j-1/2})\right)\right)$$

and

$$f_{\max,j} = \max\left(\min\left(M^X_{j+1/2}, \Phi_{\nu,j}(\min(m^Y_{j-1/2}, f^{LC}_{j^-}))\right), \min\left(\max(M^X_{j+1/2}, f^{LC}_{j^+}), \Phi_{\nu,j}(m^Y_{j-1/2})\right)\right),$$

and such limiting has the property that

$$f_{\min,j} = \max\left(m^{TVD}_{j+1/2}, \Phi_{\nu,j}(M^{TVD}_{j-1/2})\right),\ f_{\max,j} = \min\left(M^{TVD}_{j+1/2}, \Phi_{\nu,j}(m^{TVD}_{j-1/2})\right),$$

for monotonic data, that is, if $f_{j-2} \leq f_{j-1} \leq f_j \leq f_{j+1} \leq f_{j+2}$ or $f_{j-2} \geq f_{j-1} \geq f_j \geq f_{j+1} \geq f_{j+2}$, as we have just proven it.

3.5 Comparison among monotonicity-preserving limiters

The objective of this subsection is to compare among the proposed MP limiters. More specially, we will look at which one provide more relaxation space. In this subsection, without loss of generality, we suppose that $f_{j-1}^n \geq f_j^n$ and $f_{j+1}^n \leq f_{j+2}^n$.
We first study guess minimum/maximum provided by the methods $\{MD, DaTe, TVD, Um\}$. Um gives the most relaxation space. Indeed, we assume $f^{FL}(1/2) \geq f^{FR}(1/2)$, that is

$$f_j^n + \frac{1}{2}\left(f_j^n - f_{j-1}^n\right) \geq f_{j+1}^n + \frac{1}{2}\left(f_{j+1}^n - f_{j+2}^n\right).$$

It is also equivalent to

$$3(f_{j+1}^n - f_j^n) - (f_{j+2}^n - f_{j-1}^n) \leq 0.$$

So we have

$$\frac{2f_{j+1}^n - f_j^n - f_{j+2}^n}{f_j^n - f_{j-1}^n + f_{j+1}^n - f_{j+2}^n} - \frac{1}{2} = \frac{3(f_{j+1}^n - f_j^n) - (f_{j+2}^n - f_{j-1}^n)}{2(f_j^n - f_{j-1}^n + f_{j+1}^n - f_{j+2}^n)} \geq 0,$$

that is

$$\alpha = \frac{2f_{j+1}^n - f_j^n - f_{j+2}^n}{f_j^n - f_{j-1}^n + f_{j+1}^n - f_{j+2}^n} \geq \frac{1}{2},$$

Therefore,

$$f^{FL}(\alpha) \leq f^{FL}(1/2).$$

This yields also

$$f^{FL}(1) \leq f^{FL}(\alpha) \leq f^{FL}(1/2) \leq f^{FL}(0).$$

Finally, by definition, we have

$$m_{j+1/2}^{Um} \leq m_{j+1/2}^{MD} \leq m_{j+1/2}^{DaTe} \leq m_{j+1/2}^{TVD}.$$

Second, we make an assumption that the extrema can be approximately by a convex parabolic curve, that means the curvature is a constant $d_{j+1/2} = d$ and this parabolic curve can be expressed as

$$f(\alpha) = f_j^n + \alpha(f_{j+1}^n - f_j^n) + \frac{\alpha(\alpha-1)}{2}d.$$

We can thus express f_{j-1}^n and f_{j+2}^n by f_j^n, f_{j+1}^n and d, *i.e.*

$$f_{j-1}^n = 2f_j^n - f_{j+1}^n + d,$$

$$f_{j+2}^n = -f_j^n + 2f_{j+1}^n + d.$$

Injecting them into (17), we obtain $\alpha = \frac{1}{2}$, that means $m_{j+1/2}^{MD} = m_{j+1/2}^{DaTe} = \frac{1}{2}(f_j + f_{j+1}) - \frac{1}{2}d$. On the other hand, for the method Um, we have

$$\begin{aligned} m_{j+1/2}^{Um} &= \min(\min(f^{FL}(0), f^{FR}(1)), \max(f^{FL}(1), f^{FR}(0)) \\ &= \min(\min(f_j, f_{j+1}), \max(f_{j+1} - d, f_j - d). \end{aligned}$$

Without loss generality, we assume that $f_j \le f_{j+1}$, thus

$$m^{Um}_{j+1/2} = \min(f_j, f_{j+1} - d) = \min(f_j, f_j + (f_j - f_{j-1})) = f_{j+1} - d.$$

Taking difference between $m^{MD}_{j+1/2}$ and $m^{Um}_{j+1/2}$ yields

$$m^{MD}_{j+1/2} - m^{Um}_{j+1/2} = \frac{1}{2}(f_j - f_{j+1} + d) = \frac{1}{2}(f_{j-1} - f_j) \ge 0.$$

Again, we have

$$m^{Um}_{j+1/2} \le m^{MD}_{j+1/2} = m^{DaTe}_{j+1/2} \le m^{TVD}_{j+1/2}.$$

Third, we will focus on the relaxed MP constraint (12). We assume again that the curvature is a constant $d_{j+1/2} = d$. We only consider the guess maximum $\Phi_{\nu,j+1}(m^{(2)}_{j+1/2}) = m^{(2)}_{j+1/2} + \frac{1}{\nu}(f_{j+1} - m^{(2)}_{j+1/2})$, which should be large enough, for the case $f_{j-1} \ge f_j$ and $f_j \le f_{j+1} \le f_{j+2}$, to provide enough relaxation space for flux $f_{j+3/2}$. As $\Phi_{\nu,j+1}$ is a decreasing function, we search the minimum for $m^{(2)}_{j+1/2}$. According the last subsection, $m^{(2)}_{j+1/2}$ takes form as

$$m^{(2)}_{j+1/2} = \min(m^{Y}_{j+1/2}, f^{LC}_{j+1^-}), \text{ for } Y \in \{MD, DaTe, TVD, Um\}.$$

It is easy to find that $f^{LC}_{j+1^-} = f_{j+1} - d$. On the other hand, $f_{j-1} \ge f_j$ gives $d = f_{j+1} - 2f_j + f_{j-1} \ge f_{j+1} - f_j$, thus $\frac{f^n_j + f^n_{j+1}}{2} - \frac{1}{2}d \ge f^n_{j+1} - d$, so that $m^{MD}_{j+1/2} \ge f^{LC}_{j+1^-}$. We have previously obtained that $m^{Um}_{j+1/2} = \max(f_j - d, f_{j+1} - d)$, thus $m^{Um}_{j+1/2} \ge f^{LC}_{j+1^-}$. As a consequence, we have $m^{(2)}_{j+1/2} = f^{LC}_{j+1^-}$, which means the f_{LC} acts the most to provide relaxation space.

3.6 Summary of MP limiter

In this part, we give a summary of MP limiter for left flux, the one corresponding to right flux can be obtained similarly. We consider a conservative form

$$f^{n+1}_i = f^n_j - \nu\left(f^n_{j+1/2} - f^n_{j-1/2}\right).$$

We write f_j instead of f^n_j for better readability. We first define $d_j = f_{j+1} - 2f_j + f_{j-1}$.

1. we define

$$\begin{aligned} f^{FL}(\alpha) &= f_j + \alpha\,(f_j - f_{j-1})\,, \\ f^{FR}(\alpha) &= f_{j+1} + (1-\alpha)\,(f_{j+1} - f_{j+2})\,, \\ f^{AV}(\alpha) &= (1-\alpha) f_j + \alpha f_{j+1}, \end{aligned}$$

   where $\alpha \in [0, 1]$. Let us set

$$f^{MD}(\alpha) = \text{median}(f^{AV}(\alpha), f^{FL}(\alpha), f^{FR}(\alpha)) = (1-\alpha)f_j + \alpha f^n_{j+1} - \text{minmod}(\alpha d_j, (1-\alpha)d_{j+1}).$$

   We have $f^{MD}(0) = f_j$, $f^{MD}(1) = f_{j+1}$, $f^{MD}(1/2) = \frac{f_j + f_{j+1}}{2} - \frac{1}{2}\text{minmod}(d_j, d_{j+1})$.

We define

$$m_{j+1/2}^{MD} = \min_{0\le\alpha\le 1} f^{MD}(\alpha),\ M_{j+1/2}^{MD} = \max_{0\le\alpha\le 1} f^{MD}(\alpha).$$

We also define

$$m_{j+1/2}^{DaTe} = \min_{\alpha\in\{0,1/2,1\}} f^{MD}(\alpha),\ M_{j+1/2}^{DaTe} = \max_{\alpha\in\{0,1/2,1\}} f^{MD}(\alpha),$$

and

$$m_{j+1/2}^{TVD} = \min_{\alpha\in\{0,1\}} f^{MD}(\alpha),\ M_{j+1/2}^{TVD} = \max_{\alpha\in\{0,1\}} f^{MD}(\alpha).$$

$DaTe$ stands for Daru-Tenaud, and TVD for total variation diminishing.

2. We define

$$m_{j+1/2}^{Um} = \min(\min(f^{FL}(0), f^{FR}(1)), \max(f^{FL}(1), f^{FR}(0))).$$

and

$$M_{j+1/2}^{Um} = \max(\max(f^{FL}(0), f^{FR}(1)), \min(f^{FL}(1), f^{FR}(0))).$$

Um stands for Umeda.

Now we define $X, Y \in \{MD, DaTe, TVD, Um\}$, depending on the choice of method we make. We define also $f_{j^-}^{LC} = f_j - \text{minmod}(d_j, d_{j-1})$ and $f_{j^+}^{LC} = f_j - \text{minmod}(d_j, d_{j+1})$.
Finally we define the limiter as follows. We start with $f_{j+1/2}^{L}$, obtained with a given d (high order unlimited flux value), and define

$$f_{j+1/2}^{n} = \text{median}\left(f_{\min,j}, f_{j+1/2}^{L}, f_{\max,j}\right),$$

with

$$f_{\min,j} = \min\left(\max\left(m_{j+1/2}^{X}, \Phi_{\nu,j}(\max(M_{j-1/2}^{Y}, f_{j^-}^{LC}))\right), \max\left(\min(m_{j+1/2}^{X}, f_{j^+}^{LC}), \Phi_{\nu,j}(M_{j-1/2}^{Y})\right)\right)$$

and

$$f_{\max,j} = \max\left(\min\left(M_{j+1/2}^{X}, \Phi_{\nu,j}(\min(m_{j-1/2}^{Y}, f_{j^-}^{LC}))\right), \min\left(\max(M_{j+1/2}^{X}, f_{j^+}^{LC}), \Phi_{\nu,j}(m_{j-1/2}^{Y})\right)\right).$$

4 Numerical results

4.1 Implementation issues

We first give part of an implementation in the langage `C` of the limiter in Figure 1; we have chosen here the right flux and take the example of the limiter with Um and LC. We have not tried to fully optimize the code, but we remark that the limiter has a moderate overhead of computation, thanks to some reuse of computation, in this flux form. Note that some implementations can lead to large floating point errors and unsymmetric results; we have tried to limit this, in particular, by avoiding to have $1/\nu$ factors; we also choose to switch to first order limiter if the difference with it is very small.

4.2 Free transport equation

We first consider the classical constant advection equation $\partial_t f + \partial_x f = 0$ with initial condition $f(t = 0, x) = f_0(x)$ on the periodic domain $[-1, 1]$ and for $t \in [0, T]$, with $T \in \mathbb{R}^+$, the final time. Spatial mesh is $x_i = -1 + ih,\ i = 0, \ldots, N$, with spatial step $h = \frac{2}{N}$, and $N \in \mathbb{N}^*$ is the number of cells. The time step is $\Delta t = T/M \geq 0$, with $M \in \mathbb{N}^*$, the number of time steps.

```
flux2 = 0.;
for (ii=-d;ii<=d;ii++) flux2+=w[ii]*q[ii];
flux2 = nu*flux2;

//TVD
fmin[1] = DMIN(q[0],q[1]);
fmax[1] = DMAX(q[0],q[1]);
fmin[0] = DMIN(q[-1],q[0]);
fmax[0] = DMAX(q[-1],q[0]);
//Um
fmin[1] = DMIN(fmin[1],DMAX(2.*q[0]-q[-1],2.*q[1]-q[2]));
fmax[1] = DMAX(fmax[1],DMIN(2.*q[0]-q[-1],2.*q[1]-q[2]));
fmin[0] = DMIN(fmin[0],DMAX(2.*q[-1]-q[-2],2.*q[0]-q[1]));
fmax[0] = DMAX(fmax[0],DMIN(2.*q[-1]-q[-2],2.*q[0]-q[1]));
//fLC
dj = minmod(q[-1]-2.*q[0]+q[1],q[0]-2.*q[1]+q[2]);
fmin2[1] = DMIN(fmin[1],q[0]-dj);
fmax2[1] = DMAX(fmax[1],q[0]-dj);
fmin3[0] = DMIN(fmin[0],q[0]-minmod(q[-1]-2*q[0]+q[1],q[-2]-2*q[-1]+q[0]));
fmax3[0] = DMAX(fmax[0],q[0]-minmod(q[-1]-2*q[0]+q[1],q[-2]-2*q[-1]+q[0]));

bound0 = DMAX(nu*fmin3[0],q[0]-(1.-nu)*fmax[1]);
bound0 = DMIN(bound0,DMAX(nu*fmin[0],q[0]-(1.-nu)*fmax2[1]));
bound1 = DMIN(nu*fmax3[0],q[0]-(1.-nu)*fmin[1]);
bound1 = DMAX(bound1,DMIN(nu*fmax[0],q[0]-(1.-nu)*fmin2[1]));

flux2 = DMIN(flux2,bound1);
flux2 = DMAX(bound0,flux2);
for(i=0;i<N;i++){
        flux1 = flux2;
        q++;
        flux2 = 0.;
        for (ii=-d;ii<=d;ii++)flux2+=w[ii]*q[ii];
        flux2 = nu*flux2;

        //TVD
        fmin[0] = fmin[1];
        fmax[0] = fmax[1];
        fmin[1] = DMIN(q[0],q[1]);
        fmax[1] = DMAX(q[0],q[1]);
        //Um
        fmin[1] = DMIN(fmin[1],DMAX(2.*q[0]-q[-1],2.*q[1]-q[2]));
        fmax[1] = DMAX(fmax[1],DMIN(2.*q[0]-q[-1],2.*q[1]-q[2]));
        //fLC
        fmin3[0] = DMIN(fmin[0],q[0]-dj);
        fmax3[0] = DMAX(fmax[0],q[0]-dj);
        dj = minmod(q[-1]-2.*q[0]+q[1],q[0]-2.*q[1]+q[2]);
        fmin2[1] = DMIN(fmin[1],q[0]-dj);
        fmax2[1] = DMAX(fmax[1],q[0]-dj);

        bound0 = DMAX(nu*fmin3[0],q[0]-(1.-nu)*fmax[1]);
        bound0 = DMIN(bound0,DMAX(nu*fmin[0],q[0]-(1.-nu)*fmax2[1]));
        bound1 = DMIN(nu*fmax3[0],q[0]-(1.-nu)*fmin[1]);
        bound1 = DMAX(bound1,DMIN(nu*fmax[0],q[0]-(1.-nu)*fmin2[1]));

        flux2 = DMIN(flux2,bound1);
        flux2 = DMAX(bound0,flux2);
        if(fabs(flux2-nu*q[0])<1.e-16) flux2 = nu*q[0];
        p[i] =  q[-1]-(flux1-flux2);
}
```


Figure 1: Implementation of the limiter

4.2.1 Square wave

The initial condition is $f_0(x) = 1$, if $x \in [-0.75, 0.25]$ and $f_0(x) = 0$, if $x \in [-1, -0.75[\cup]0.25, 1]$. We represent here on Table 2, the error in L^1, L^2 norm, its order (in L^1 and L^2 norm), and the error in total variation (TV) (defined here as the total variation minus 2), for $CFL = 2.5$, 0.25 and 0.025, for SLWENO5 [33], cubic splines, $d = 2$ and $d = 8$ with and without limiter. We remark that the schemes with and without limiter converge in L^1 and L^2 norm, with the same order. This is also the case for $d = 1, 3, \ldots, 7$, where we have found that the order in L^1 is around 0.75, 0.86, 0.89, 0.90, 0.91, 0.92 and half for the L^2 norm. Note that the theoretical order for the unlimited scheme is $\frac{2d+1}{2d+2}$ in L^1 and the half value in L^2, and the numerical results reproduce rather faithfully the theoretical ones. The L^1 error is almost always better for the limited scheme. For the L^2 norm, the situation changes: for $d = 2$, it is slightly better, but for $d = 8$ it is worser. The total variation error (TV) is increasing (in absolute value) with the degree for the unlimited scheme. On the contrary, for the limited scheme, the total variation is very well preserved, but we can observe some little degradation for the small value of $CFL = 0.025$ and when N is small. This can be due to diffusion (negative value of TV), or it can be that the diffusion smooths the solution and then the scheme finds smooth extremas and relax the limiter at such places, and this is not in contraction with the theoretical property proven. We did in fact sometimes encounter situations where new extrema were found, but this was due to the propagation of round off errors, which has lead us to do a dedicated modification previously explained. For the SLWENO5 scheme, we observe that the total variation error is also small, even if it is not as well preserved, but the L^1 and L^2 errors are quite big, and there is no clear order of convergence. For cubic splines, the order of accuracy is 3/4, for $CFL = 2.5$ and improves for small CFL, which is linked to the fact that the cubic splines derivatives have a higher order of accuracy. Concerning the total variation error, it is at the level of $d = 1$ (without limiter) which is around 0.47, but goes higher when the CFL is small (for comparison, the total variation error is around 1 for $d = 3$ and around 1.2 for $d = 4$).
On Table 3 (top), we have compared the limiter (lim=Um+LC), with other combinations: DaTe,Um and Date+LC, when $N = 25$, $CFL = 0.25$ and $d = 3, 6$ and 8. We have also tested MD instead of Um which generally gives very similar results to Um. For this test, for $d = 3$, Um seems to act favorably w.r.t DaTe; for $d = 6$, we see that LC is useful; for $d = 8$, DaTe seems however to behave better (note that the total variation error is positive for lim, as shown on the previous Table). For other tests, in particular with more points, the limiters are less distinguishable.

4.2.2 Sinusoïdal wave

The initial condition is $f_0(x) = \sin(\pi x)$. We represent here on Table 3 (middle), the error in L^1, L^2 norm, its order r (in L^1 and L^2 norm) for $CFL = 2.5$. for SLWENO5, cubic splines, $d = 1, \ldots, 4$ with limiter. In fact for this test, the solution with and without limiter is the same. In order to see differences, we have to take N smaller (typically less than 10). We note that this is not true for SLWENO5 which shows effectively the 5th order of convergence, but not with the full stencil corresponding to $d = 2$.

4.2.3 Quartic sine function

The initial condition is $f_0(x) = \sin^4(4\pi x)$. Numerical results are shown on Table 3 (bottom). For $N = 25$, the mesh is coarse to discretize a function with 8 sin-like maxima and 8 more flat minima. In that case, the method with and without limiter differ only for the L^1 norm whose value is 0.654 without limiter (instead of 0.655 or 0.656), but then on refined meshes it is no more the case. Note that the maximum principle is here typically not satisfied and a further study enforcing global maximum principle would be worth to be added, but is not tackled here, in order to already see the sole influence of the monotonicity preserving

CFL=2.5

SLWENO5

N	L^1	order	L^2	order	TV
25	0.339		0.321		0.129
50	0.216	0.65	0.245	0.39	0.000164
100	0.123	0.81	0.183	0.42	0.00065
200	0.0744	0.73	0.138	0.4	0.000256
400	0.065	0.19	0.138	0.0063	0.00499
800	0.0575	0.18	0.132	0.061	0.00117

cubic splines

N	L^1	order	L^2	order	TV
25	0.329		0.302		0.415
50	0.201	0.71	0.231	0.38	0.417
100	0.124	0.69	0.179	0.37	0.486
200	0.0738	0.75	0.138	0.37	0.486
400	0.0437	0.75	0.107	0.37	0.483
800	0.026	0.75	0.0821	0.37	0.479

$d = 2$ with limiter

N	L^1	order	L^2	order	TV
25	0.26		0.279		0
50	0.146	0.83	0.21	0.41	0
100	0.0825	0.83	0.158	0.41	0
200	0.0464	0.83	0.119	0.41	0
400	0.0261	0.83	0.0891	0.41	0
800	0.0147	0.83	0.0669	0.41	4.44e-16

$d = 2$ without limiter

N	L^1	order	L^2	order	TV
25	0.326		0.298		0.493
50	0.194	0.75	0.214	0.48	0.766
100	0.108	0.84	0.159	0.43	0.784
200	0.0614	0.82	0.12	0.41	0.799
400	0.0344	0.84	0.0899	0.41	0.801
800	0.0194	0.83	0.0674	0.41	0.8

$d = 8$ with limiter

N	L^1	order	L^2	order	TV
25	0.133		0.191		0
50	0.07	0.92	0.14	0.45	0
100	0.0371	0.92	0.102	0.45	0
200	0.0196	0.92	0.0743	0.46	0
400	0.0103	0.92	0.0541	0.46	4.44e-16
800	0.00543	0.93	0.0393	0.46	0

$d = 8$ without limiter

N	L^1	order	L^2	order	TV
25	0.159		0.163		1.19
50	0.083	0.94	0.119	0.45	1.41
100	0.0476	0.8	0.0878	0.44	1.63
200	0.0246	0.95	0.064	0.46	1.64
400	0.0129	0.93	0.0466	0.46	1.66
800	0.0069	0.91	0.0339	0.46	1.66

CFL=0.25

SLWENO5

N	L^1	order	L^2	order	TV
25	0.436		0.364		-0.233
50	0.331	0.4	0.296	0.3	-0.201
100	0.437	-0.4	0.375	-0.34	-0.0354
200	0.323	0.43	0.315	0.25	0.00369
400	0.234	0.46	0.265	0.25	0.00564
800	0.161	0.55	0.22	0.27	0.00806

cubic splines

N	L^1	order	L^2	order	TV
25	0.364		0.334		0.00555
50	0.318	0.19	0.297	0.17	0.41
100	0.189	0.75	0.223	0.41	0.482
200	0.114	0.73	0.172	0.38	0.483
400	0.0674	0.75	0.132	0.38	0.484
800	0.04	0.75	0.102	0.38	0.481

$d = 2$ with limiter

N	L^1	order	L^2	order	TV
25	0.338		0.32		0.146
50	0.205	0.72	0.25	0.36	4.44e-16
100	0.116	0.83	0.188	0.41	4.44e-16
200	0.0649	0.83	0.141	0.41	0
400	0.0365	0.83	0.106	0.42	4.44e-16
800	0.0205	0.83	0.0791	0.42	-2.22e-16

$d = 2$ without limiter

N	L^1	order	L^2	order	TV
25	0.337		0.306		0.451
50	0.237	0.51	0.25	0.29	0.256
100	0.146	0.7	0.189	0.4	0.721
200	0.0858	0.77	0.142	0.41	0.802
400	0.0482	0.83	0.106	0.42	0.803
800	0.0271	0.83	0.0798	0.42	0.805

$d = 8$ with limiter

N	L^1	order	L^2	order	TV
25	0.165		0.213		0.0389
50	0.0824	1	0.152	0.48	0
100	0.0433	0.93	0.111	0.46	4.44e-16
200	0.0227	0.93	0.0805	0.46	4.44e-16
400	0.0119	0.93	0.0584	0.46	0
800	0.00624	0.93	0.0424	0.46	-2.22e-16

$d = 8$ without limiter

N	L^1	order	L^2	order	TV
25	0.145		0.173		0.95
50	0.101	0.52	0.132	0.39	1.62
100	0.0544	0.89	0.0956	0.47	1.64
200	0.0292	0.9	0.0695	0.46	1.69
400	0.0152	0.94	0.0505	0.46	1.65
800	0.00807	0.91	0.0366	0.46	1.56

CFL=0.025

SLWENO5

N	L^1	order	L^2	order	TV
25	0.457		0.373		-0.296
50	0.351	0.38	0.307	0.28	-0.193
100	0.326	0.11	0.306	0.0029	-0.0316
200	0.279	0.23	0.303	0.016	0.00122
400	0.204	0.45	0.253	0.26	0.00482
800	0.145	0.49	0.212	0.26	0.00805

cubic splines

N	L^1	order	L^2	order	TV
25	0.341		0.311		0.456
50	0.285	0.26	0.272	0.19	0.635
100	0.183	0.64	0.208	0.39	1.23
200	0.111	0.72	0.158	0.4	1.31
400	0.0625	0.83	0.119	0.4	1.22
800	0.035	0.84	0.0903	0.4	1.09

$d = 2$ with limiter

N	L^1	order	L^2	order	TV
25	0.342		0.323		0.12
50	0.211	0.69	0.253	0.35	-1.73e-06
100	0.119	0.83	0.19	0.41	-9.9e-13
200	0.0667	0.83	0.143	0.41	-4e-14
400	0.0375	0.83	0.107	0.42	-4.06e-14
800	0.021	0.83	0.0802	0.42	-4.15e-14

$d = 2$ without limiter

N	L^1	order	L^2	order	TV
25	0.337		0.306		0.437
50	0.249	0.44	0.256	0.26	0.266
100	0.149	0.74	0.191	0.42	0.681
200	0.0883	0.76	0.144	0.41	0.806
400	0.0496	0.83	0.108	0.42	0.81
800	0.0278	0.83	0.0809	0.42	0.805

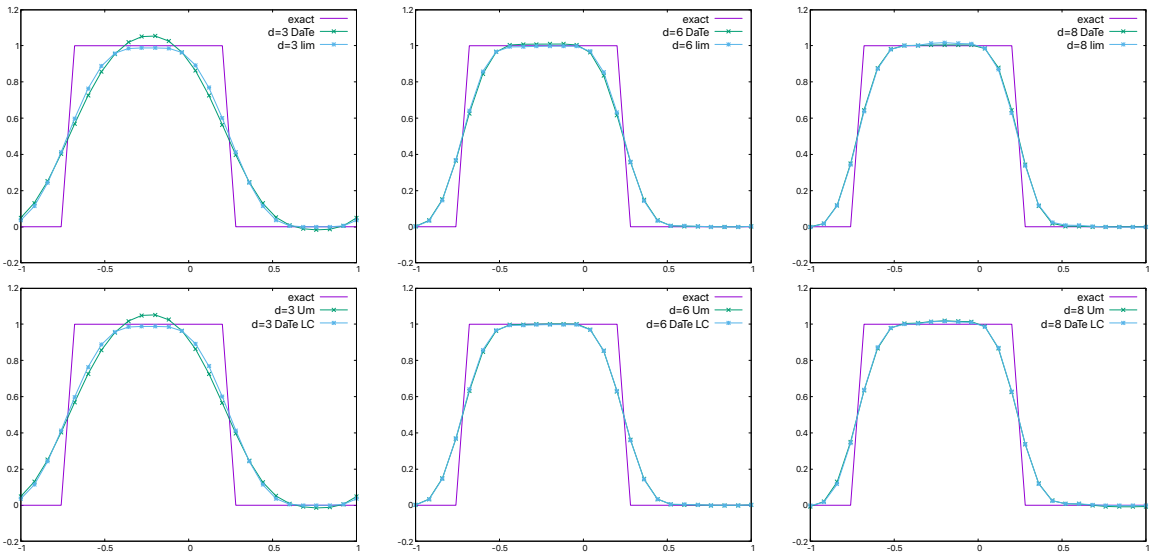
$d = 8$ with limiter

N	L^1	order	L^2	order	TV
25	0.216		0.22		-0.165
50	0.107	1	0.165	0.42	-0.000688
100	0.0565	0.92	0.121	0.45	6.8e-10
200	0.0294	0.94	0.0872	0.47	1.27e-06
400	0.0154	0.94	0.0628	0.47	5.68e-09
800	0.00809	0.92	0.0458	0.46	3.41e-09

$d = 8$ without limiter

N	L^1	order	L^2	order	TV
25	0.149		0.175		0.958
50	0.103	0.53	0.133	0.39	1.66
100	0.0562	0.88	0.0965	0.47	1.72
200	0.0303	0.89	0.0701	0.46	1.78
400	0.0158	0.94	0.0509	0.46	1.72
800	0.00839	0.91	0.0369	0.46	1.68

Table 2: Error for square wave at $T = 800$



sinusoïdal wave

SLWENO5

N	L^1	r	L^2	r
25	0.0393		0.0301	
50	0.00138	4.8	0.00106	4.8
100	4.29e-05	5	3.36e-05	5
200	1.34e-06	5	1.05e-06	5
400	4.17e-08	5	3.29e-08	5
800	1.29e-09	5	1.03e-09	5

cubic splines

L^1	r	L^2	r
0.0526		0.0413	
0.00661	3	0.0052	3
0.000827	3	0.00065	3
0.000103	3	8.12e-05	3
1.29e-05	3	1.01e-05	3
1.61e-06	3	1.27e-06	3

$d = 1$

L^1	r	L^2	r
0.395		0.311	
0.058	2.8	0.0456	2.8
0.00741	3	0.00583	3
0.00093	3	0.00073	3
0.000116	3	9.13e-05	3
1.45e-05	3	1.14e-05	3

$d = 2$

L^1	r	L^2	r
0.00618		0.00486	
0.000195	5	0.000153	5
6.11e-06	5	4.8e-06	5
1.91e-07	5	1.5e-07	5
5.98e-09	5	4.69e-09	5
1.87e-10	5	1.47e-10	5

$d = 3$

L^1	r	L^2	r
8.52e-05		6.7e-05	
6.73e-07	7	5.29e-07	7
5.28e-09	7	4.15e-09	7
4.13e-11	7	3.24e-11	7
1.14e-13	8.5	1.03e-13	8.3
1.14e-13	0.0056	1.03e-13	0.00083

$d = 4$

L^1	r	L^2	r
1.21e-06		9.48e-07	
2.39e-09	9	1.88e-09	9
4.67e-12	9	3.69e-12	9
1.14e-13	5.4	1.03e-13	5.2
1.14e-13	0.00058	1.03e-13	0.00018
1.14e-13	0.0055	1.03e-13	0.00064

quartic sine function

SLWENO5

N	L^1	r	L^2	r
25	0.657		0.515	
50	0.654	0.0052	0.515	0.00016
100	0.652	0.0046	0.514	0.0046
200	0.219	1.6	0.184	1.5
400	0.208	0.075	0.169	0.12
800	0.0625	1.7	0.0645	1.4

cubic splines

L^1	r	L^2	r
0.654		0.515	
0.654	2.1e-12	0.515	2.2e-12
0.619	0.081	0.487	0.082
0.227	1.4	0.19	1.4
0.0802	1.5	0.0647	1.6
0.013	2.6	0.0103	2.6

$d = 1$ with limiter

L^1	order	L^2	order
0.655		0.515	
0.654	0.00051	0.515	2.2e-06
0.654	8.7e-11	0.515	8.8e-11
0.623	0.07	0.491	0.071
0.241	1.4	0.2	1.3
0.0863	1.5	0.0696	1.5

$d = 2$ with limiter

L^1	r	L^2	r
0.655		0.515	
0.654	0.0013	0.515	1e-05
0.474	0.47	0.373	0.47
0.145	1.7	0.116	1.7
0.0119	3.6	0.0094	3.6
0.000391	4.9	0.000307	4.9

$d = 4$ with limiter

L^1	r	L^2	r
0.654		0.515	
0.516	0.34	0.406	0.35
0.151	1.8	0.118	1.8
0.00116	7	0.000911	7
2.41e-06	8.9	1.9e-06	8.9
4.79e-09	9	3.76e-09	9

$d = 5$ with limiter

L^1	r	L^2	r
0.656		0.515	
0.218	1.6	0.184	1.5
0.0745	1.6	0.0585	1.7
6.61e-05	10	5.19e-05	10
3.48e-08	11	2.73e-08	11
1.7e-11	11	1.35e-11	11

Table 3: On top: square wave, $T = 800$, $N = 25$, $CFL = 0.25$ (left, $d = 3$, center, $d = 6$ and right $d = 8$); then, error for sinusoïdal wave and quartic sine function, $CFL = 2.5$, $T = 800$

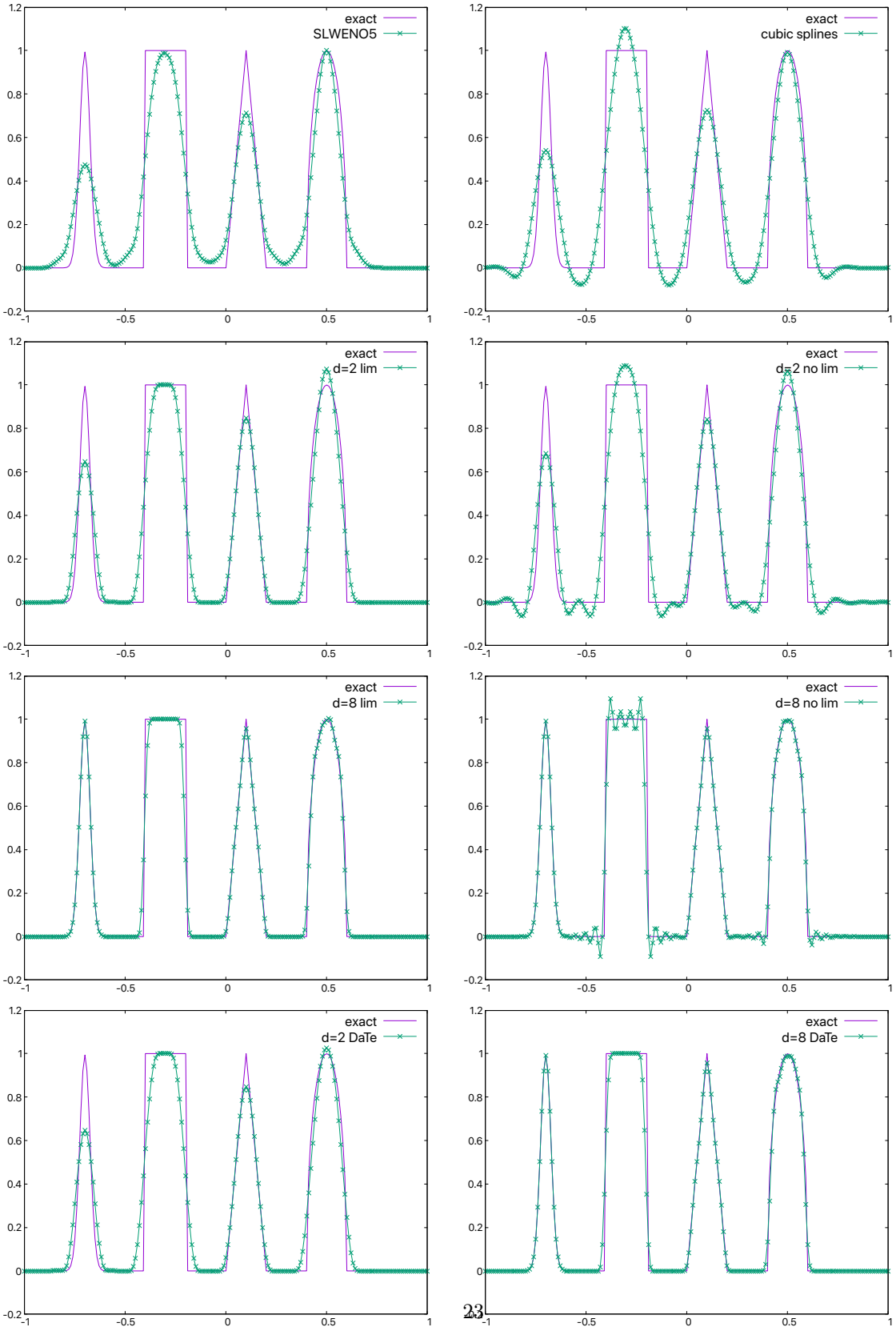


Figure 2: Shu test case, $T = 800,\ N = 200,\ CFL = 2.5$

property. We also note that, in contrary to the previous case, we do not attain machine precision error, and the right order of convergence is seen only when the grid is fine enough, with enhanced results when the order of the method gets higher. For the SLWENO5 method, the mesh is not fine enough to get the right order of convergence and is not as accurate as cubic splines. On the other hand, the method with $d = 2$ shows better results than cubic splines. The method with $d = 1$ is less accurate than SLWENO5, especially for $N = 200$, but for $N = 400$ (or even 800) it is not so different; in fact, it is almost of the accuracy of cubic splines (little less accurate) with twice less points.

4.2.4 Shu test case

The initial condition is given by

$$f_0(x) = \begin{cases} \frac{1}{6}\left(G(x,\beta,z-\delta) + G(x,\beta,z+\delta) + 4G(x,\beta,z)\right), & x \in [-0.8,-0.6] \\ 1, & x \in [-0.4,-0.2] \\ 1 - |10(x-0.1)|, & x \in [0,0.2] \\ \frac{1}{6}\left(F(x,\alpha,a-\delta) + F(x,\alpha,a+\delta) + 4F(x,\alpha,a)\right), & x \in [0.4,0.6] \\ 0, & x \in [-1,-0.8] \cup [-0.2,0] \cup [0.2,0.4] \cup [0.6,1], \end{cases}$$

with $G(x,\beta,z) = e^{-\beta(x-z)^2}$ and $F(x,\alpha,a) = \sqrt{\max(1-\alpha^2(x-a)^2, 0)}$, and $a = 0.5$, $z = -0.7$, $\delta = 0.005$, $\alpha = 10$ and $\beta = \frac{\ln(2)}{36\delta^2}$. Numerical results are given on Figure 2. We observe that SLWENO5 gives a quite diffusive behavior, except for the last bump. The cubic splines method is also diffusive, and has more oscillations; the last bump is also well captured. We see oscillations for the scheme without limiter which are amplified for $d = 8$. On the other hand, the monotonicity is well preserved for the scheme with limiter. The results are less diffusive and we see that the high order interpolation is useful; the last bump is resolved with lower accuracy for $d = 2$, but the results are improved going to $d = 8$; it seems that here the DaTe limiter behaves better for the last bump. We also remark that for the rectangular function, the result is not symmetric, but this could be changed by just shifting the grid of $h/2$.

4.3 Vlasov-Poisson system

We solve the Vlasov equation $\partial_t f + v\partial_x f + E\partial_v f = 0$, coupled with the Poisson equation; $E = -\partial_x \phi$ and $-\partial_x^2 \phi = \rho - 1$, using Strang splitting. We also have an initial condition $f(t = 0, x, v) = f_0(x, v)$. We can either first do the advection in x: $\partial_t f + v\partial_x f$ or the advection in v: Poisson equation and then $\partial_t f + E\partial_v f = 0$. In the numerical results we consider to do at each time step the advection in x for $\Delta t/2$, the advection in v for Δt and then again the advection in v for $\Delta t/2$. Then unknowns are $f_{i,j}^n \simeq f(t_n, x_i, v_j)$, with $t_n = n\Delta t$. The phase-space domain is $[0, L] \times [-v_{\max}, v_{\max}]$, and we have $x_i = i\Delta x$, $i = 0, \ldots, N_x$, with $\Delta x = L/N_x$ and $v_j = -v_{\max} + j\Delta v$, $j = 0, \ldots, N_v$, with $\Delta v = 2v_{\max}/N_v$, with $N_x, N_v \in \mathbb{N}^*$. We also have $t \in [0, T]$, with $T \in \mathbb{R}^+$, the final time, and the time step is $\Delta t = T/M \geq 0$, with $M \in \mathbb{N}^*$, the number of time steps.

4.3.1 Bump on tail [15]

Initial condition is $f_0(x, v) = (1 + \varepsilon \cos(kx))(\frac{0.9}{\sqrt{2\pi}} e^{-v^2/2} + \frac{0.1}{v_{th}\sqrt{2\pi}} e^{-(v-u)^2/(2v_{th}^2)})$, with $u = 4.5$, $\varepsilon = 0.04$, $v_{th} = 0.5$ and $k = 0.3$. The domain is $[0, 3\frac{2\pi}{k}] \times [-v_{\max}, v_{\max}]$, with $v_{\max} = 9$. We use $\Delta t = 0.1$ and final time $T = 400$. Numerical results are given on Figure 3 and 4.
In this test, we have three small vortices. In order to discretize them well in space, we use a number of points in x that is a multiple of 3; otherwise, the method with limiter is not able to keep the three vortices and some merging appears leading also to a break in electric energy, as observed in [15] for some methods.

On Figure 3, we represent the time evolution of the electric energy. Time step is fixed to $\Delta t = 0.1$. As reference solution, we use $d = 2$ with limiter on 1023×1024 grid (similar results are obtained without limiter), and we compare it to SLWENO5, cubic splines, $d = 2$, $d = 6$ with limiter, without limiter and with DaTe limiter, on 63×64 and 129×128 grid. We see that the limiter damps the electric energy and this makes the convergence slower to the reference solution, and on the other hand the solution has less oscillations in time. Increasing the degree permits to damp less on 63×64 grid, but this seems not to remain true on the finer grid 129×128, which might be surprising at first sight. This is confirmed looking at time evolution of the L^2 norm on Figure 4. First it is better preserved using the higher order, but then it is no more the case, maybe because the solution has more details and this can lead to more diffusion in later times. Concerning the L^1 norm, we clearly see the effect of the limiter which does the job of better conserving it, and this is improved when using a higher degree. The total energy is however generally better conserved without limiter. The DaTe limiter is more diffusive, as we can see it on the electric energy and the L^2 norm, and on the other hand the L^1 norm can be better preserved. For the total energy, there is an effect, but no clear winner. The comparison with SLWENO5 confirms that the monotonicity preserving schemes are much less diffusive; the L^1 norm is better preserved than for $d = 2$, but worser than for $d = 6$. Even if the convergence is not as fast as the method without limiter and high degree, the method with limiter is quite comparable to the cubic splines method. In particular, the L^2 norm is in the same range, or sometimes even better.

4.3.2 Two stream instability [15]

Initial condition is $f_0(x,v) = (1 + \varepsilon\cos(kx))(\frac{1}{2v_{th}\sqrt{2\pi}}e^{-(v-u)^2/(2v_{th}^2)} + \frac{1}{2v_{th}\sqrt{2\pi}}e^{-(v+u)^2/(2v_{th}^2)})$, with $\varepsilon = 0.05$, $u = 0.99, v_{th} = 0.3$ and $k = 1$. The domain is $[0, 26\pi] \times [-5, 5]$. Numerical results are given on Figure 5, 6, 7 and 8. We use $\Delta t = 0.1$.
We first look at the solution at time $T = 1000$, on Figure 5. We see that on a 128×128 grid, two remaining holes as in [15]. One exception is for the DaTe limiter where the centered hole has disappeared, for $d = 2$. We note also that the solution with limiter is less oscillatory and that the SLWENO5 method is more diffusive, almost as the method with $d = 1$. Looking for finer grid leads however to different behaviors: the holes are not always at the same place and the number of holes can differ. We remark also that the limiter removes oscillations. Comparing the solution with 6-th order scheme in time (o6) [7] leads to different results, indicating that the solution is not converged at that time (both in phase space and time). On the other hand, we see the convergence on 128×128 grid until $T \simeq 30$ and on 1024×1024 until $T \simeq 55$, looking at the plots of the electric energy (bottom of Figure 6). At time $T = 70$, the solution seems to be converged, as the o6 scheme or Strang scheme with or without limiter give very similar results; note that there is no merging at that time for the converged solution, and this might be true also at time $T = 1000$, but the convergence seems to be unreachable; if the way to converge is the same, we would need more than 2^{200} points per direction to converge (computed on the basis that we can gain 8 in time by multiplying the grid per direction by 2). Note that the current biggest simulations are on $128^6 = 2^{42}$ grid which is still far from 2^{200}. We see on such test case, that more oscillations appear when there is no limiter (see Figure 7). For $d = 2$ with limiter, there is also more oscillations, but we see that the L^1 norm is also worse conserved, which is not true for the DaTe limiter, that acts more. We see that the L^1 norm is better preserved with the limiter, and the situation even improves when using a higher degree, which is less the case for the solution without limiter.
Concerning the total energy, on a 128×128 grid, it is conserved up to 0.5% for cubic splines, 2% for $d = 2$ and 3% for $d = 4$ (with limiter). Without limiter, the energy can be better preserved (0.3% for $d = 2$ and 0.1% for $d = 4$).
Note also that the results can even be improved with the order 6 scheme (see [7]).

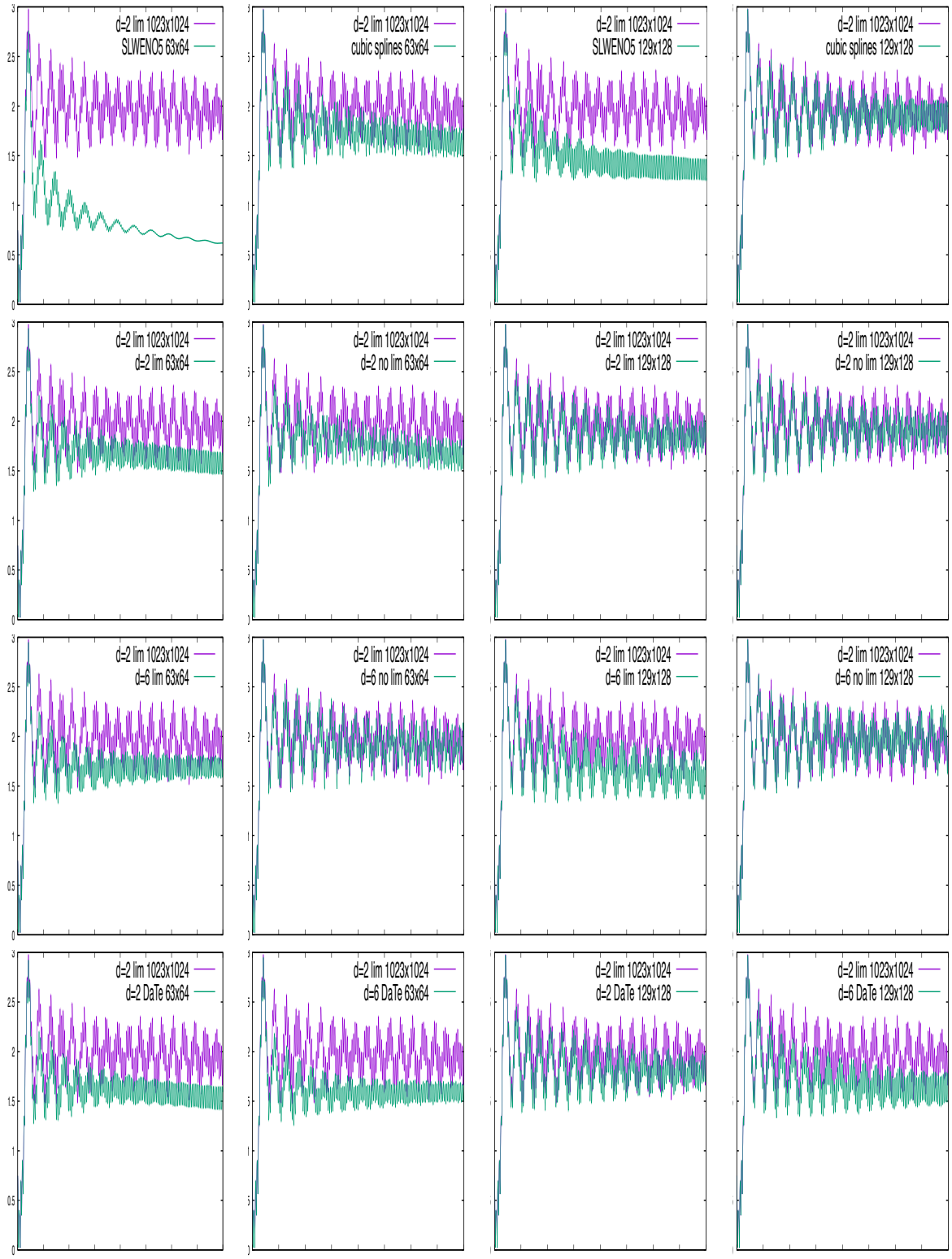


Figure 3: Electric energy for bump on tail test case 63×64 (left) and 129×128 (right) grid

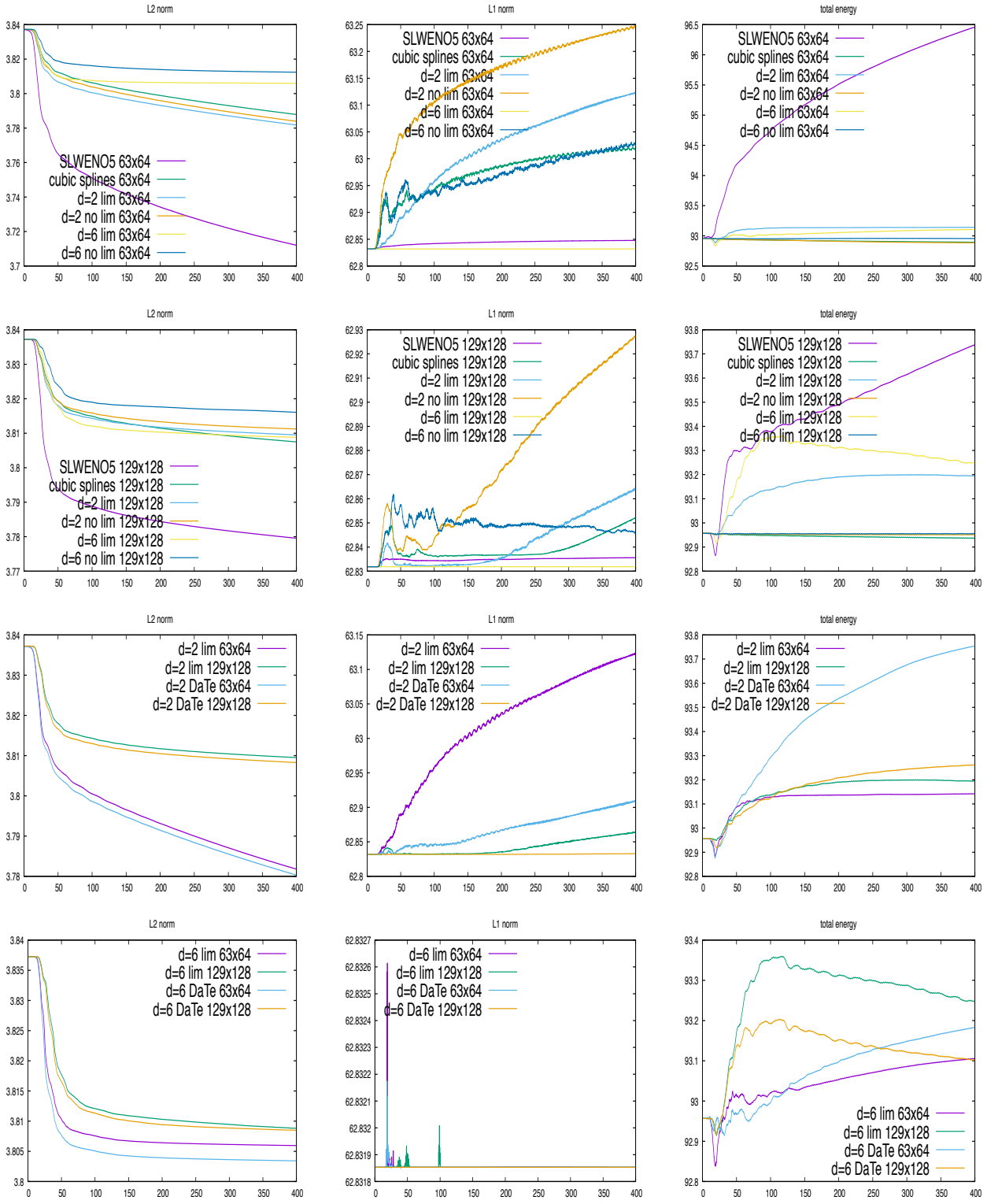


Figure 4: Time evolution of L^2 norm, L^1 norm and total energy for bump on tail test case

On Figure 8, we observe convergence of the methods at time $T = 70$: the methods with or without limiter and with Strang or o6 scheme on 4096×4096 grids are almost undistinguishable, which confirms the result seen on the electric energy (Figure 6).

4.3.3 Beam test case [46]

Initial condition is $f_0(x,v) = \frac{4}{\sqrt{2\pi\alpha}}\chi(x)e^{-v^2/(2\alpha)}$, with $\chi(x) = \frac{1}{2}\text{erf}(\frac{x+1.2}{0.3}) - \frac{1}{2}\text{erf}(\frac{x-1.2}{0.3})$ and $\alpha = 0.2$. Also we do not solve exactly the previous Vlasov-Poisson system, but $\partial_t f + \frac{v}{\varepsilon}\partial_x f + (E - \frac{x}{\varepsilon})\partial_v f = 0$, together with electric field E satisfying $\frac{1}{x}\partial_x(xE) = \int f dv$, on $[0, L/2]$ and imposing that E is odd function. E is given by $E = \frac{1}{x}\int_0^x s\rho(t,s)dx$ and we use a trapezoidal formula for the approximation of the integral. We choose $\varepsilon = 0.7$ and the domain is $[-L/2, L/2] \times [-v_{\max}, v_{\max}]$, with $L = 8$ and $v_{\max} = 4$. Numerical results are given on Figure 9, 10, 11 and 12. As reference solution, we use $d = 4$ with limiter on 4096×4096 grid; we take $\Delta t = 0.1$. We see that the limiter permits again to better preserve the L^1 norm. The L^2 norm is better preserved increasing the degree on the coarse mesh 64×64 and for short time on 256×256 grid, but then it is the contrary. On the other hand, we see that the solution with $d = 8$ is little less diffused than with cubic splines; the most diffused solution is again SLWENO5.

Conclusion and perspectives

We have revisited the monotonicity preserving schemes for semi-Lagrangian schemes based on odd order Lagrange interpolation. A detailed numerical study is performed for $1d$ constant advection and Vlasov-Poisson simulations. The new scheme has a proven monotonicity preserving property and controls in particular the L^1 norm, with some limited degradation of the L^2 norm. Comparison with cubic splines and SLWENO5 is made to show the accuracy of our method. One natural extension of this work is to add global maximum principle (in particular positivity), that we have here not added, permitting to measure the L^1 norm conservation as indicator of the well behavior of the scheme (if positivity is ensured, in our conservative setting, the L^1 norm automatically exactly satisfied). For this, we can follow the works [44, 38]. One other more demanding extension is to consider the non constant advection. For this, we can work as in [15] on the splitted conservative form, but this has the disadvantage of breaking of the conservation of constant states at the level of the equations [28]. Another more common and popular method is to work on the $2D$ unsplit advective form [35, 46, 28]. However, the conservation of mass is lost, and this is generally amplified when using the limiters [46, 28]. A dual way is to work with the more involved and technical $2D$ unsplit conservative form [12], which has been later developed in a Semi-Lagrangian Discontinuous Galerkin context [5, 6].

Acknowledgements

This work has been supported by Heilongjiang Natural Science Foundation (LH2019A013).
MM acknowledges invitations as a scholar professor by Harbin Institute of Technology in 2018, 2019 and 2020. This work has been carried out within the framework of the EUROfusion consortium and has received funding from the Euratom research and training programme 2014-2018 and 2019-2020 under grant agreement No 633053. The views and opinions expressed herein do not necessarily reflect those of the European Commission. Centre de Calcul Intensif d'Aix-Marseille is acknowledged for granting access to its high performance computing resource

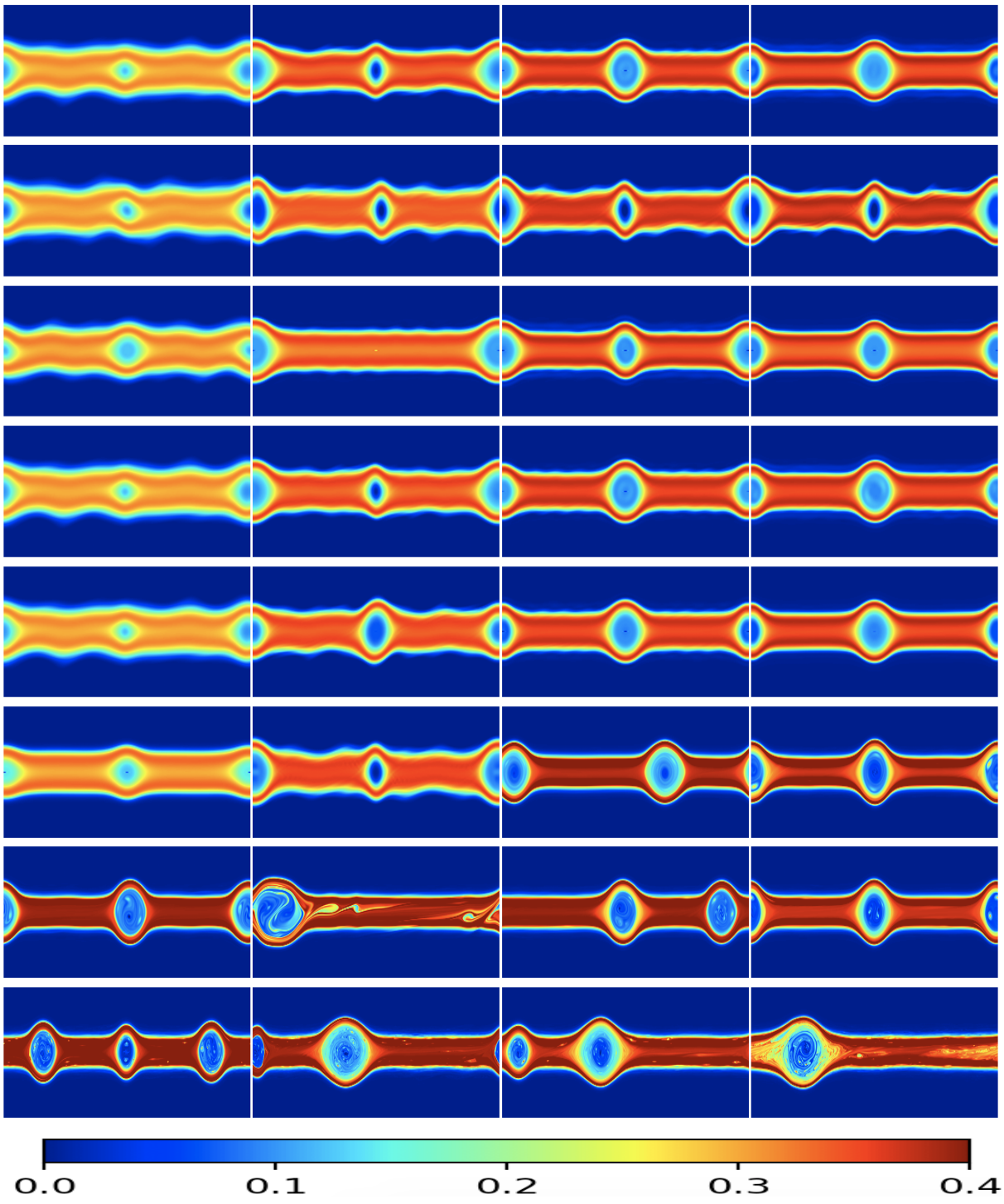


Figure 5: Two stream instability: from left to right: $d = 1, 2, 3, 4$; from top to bottom: lim, no lim, DaTe, Um, DaTe+LC, on 128×128 grid; 6th row: SLWENO5 and splines: 128×128 grid; then 1024×1024 grid; 7th row: $d = 2$ lim and no lim; then $d = 4$ lim and no lim, all on 1024×1024 grid; last row: $d = 4$ lim, $d = 4$ no lim; $d = 4$ lim (o6), $d = 4$ no lim (o6) on 4096×4096 grid.

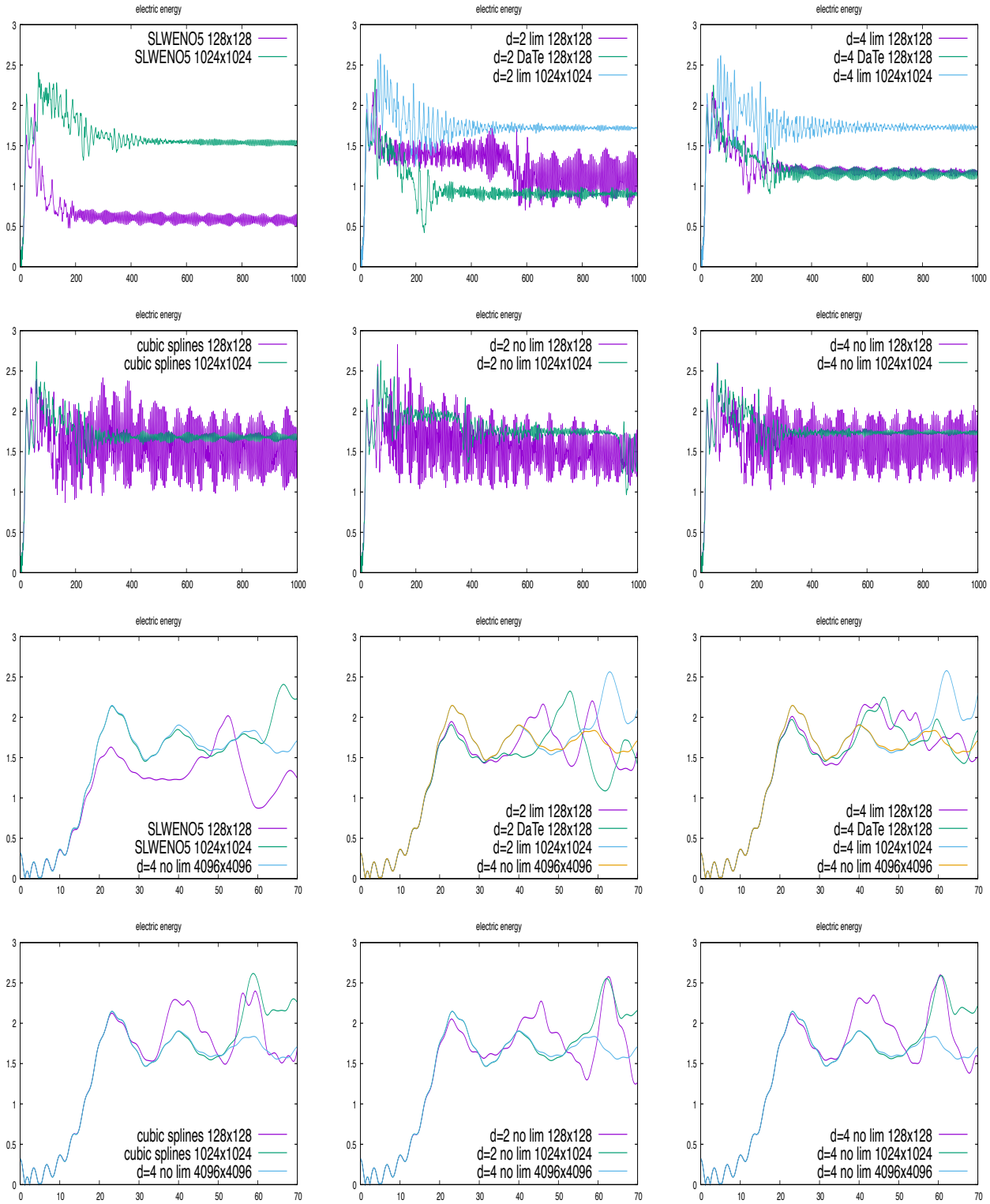


Figure 6: Time evolution of electric energy for two stream instability test case (top: until $T = 1000$, bottom until $T = 70$)

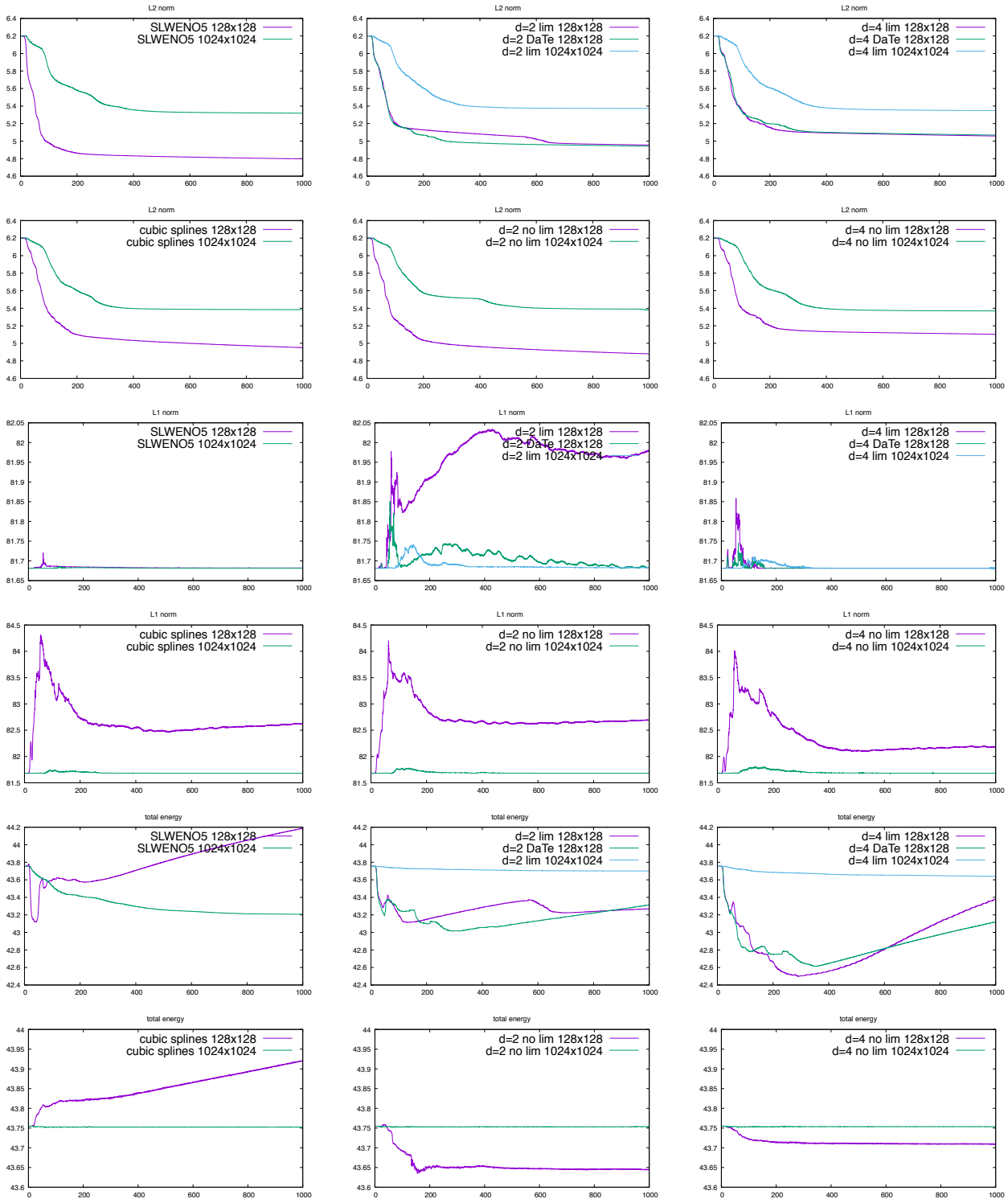


Figure 7: Time evolution of L^2 norm (top), L^1 norm (middle) and total energy (bottom) for two stream instability test case

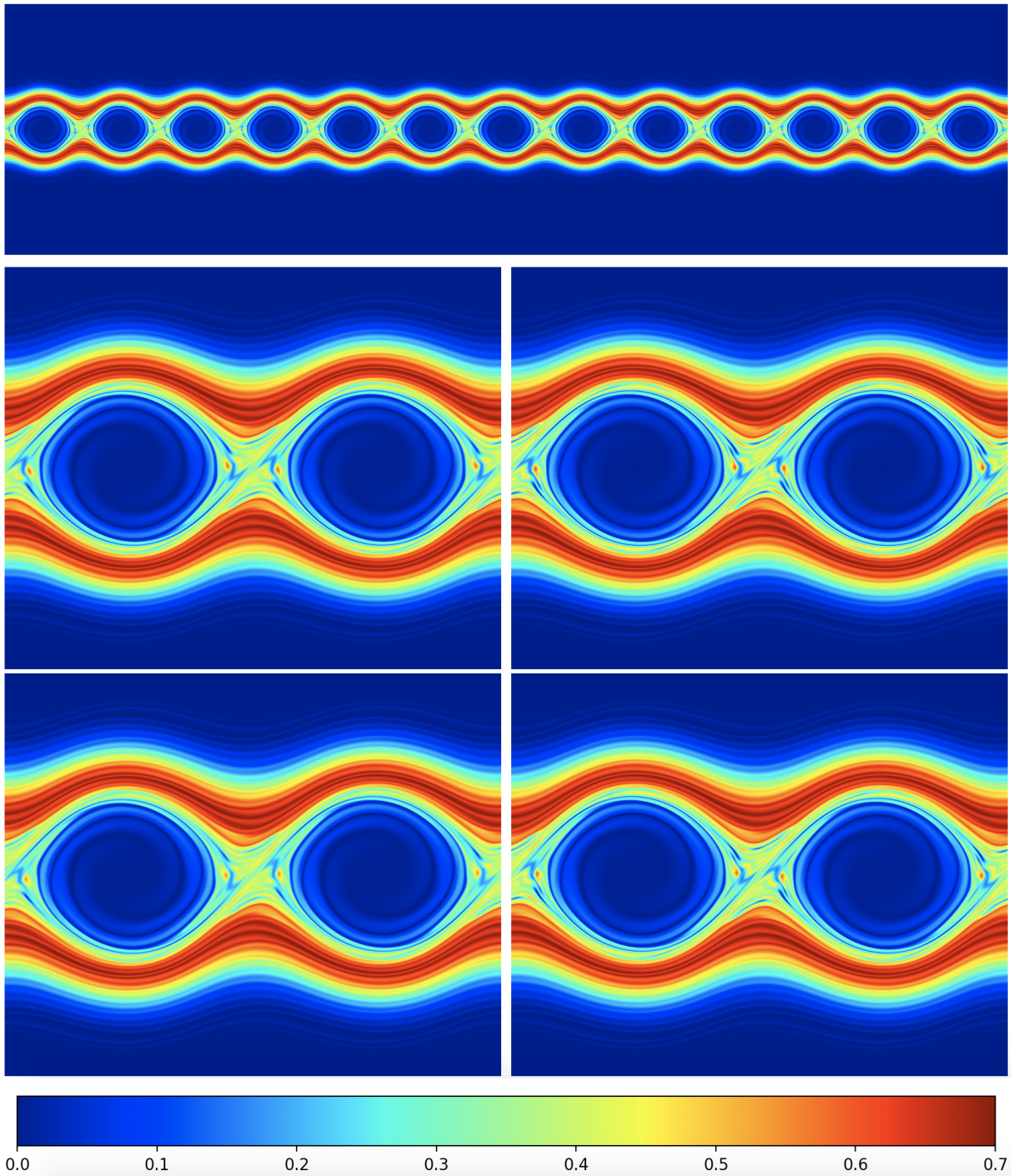


Figure 8: Two stream instability:at time $T = 70$ with $d = 4$, $\Delta t = 0.1$ on 4096×4096 grid; top: lim with order 6 splitting; then zoom of the two left holes: top/bottom: no lim/lim with order 6/Strang splitting (left/right)

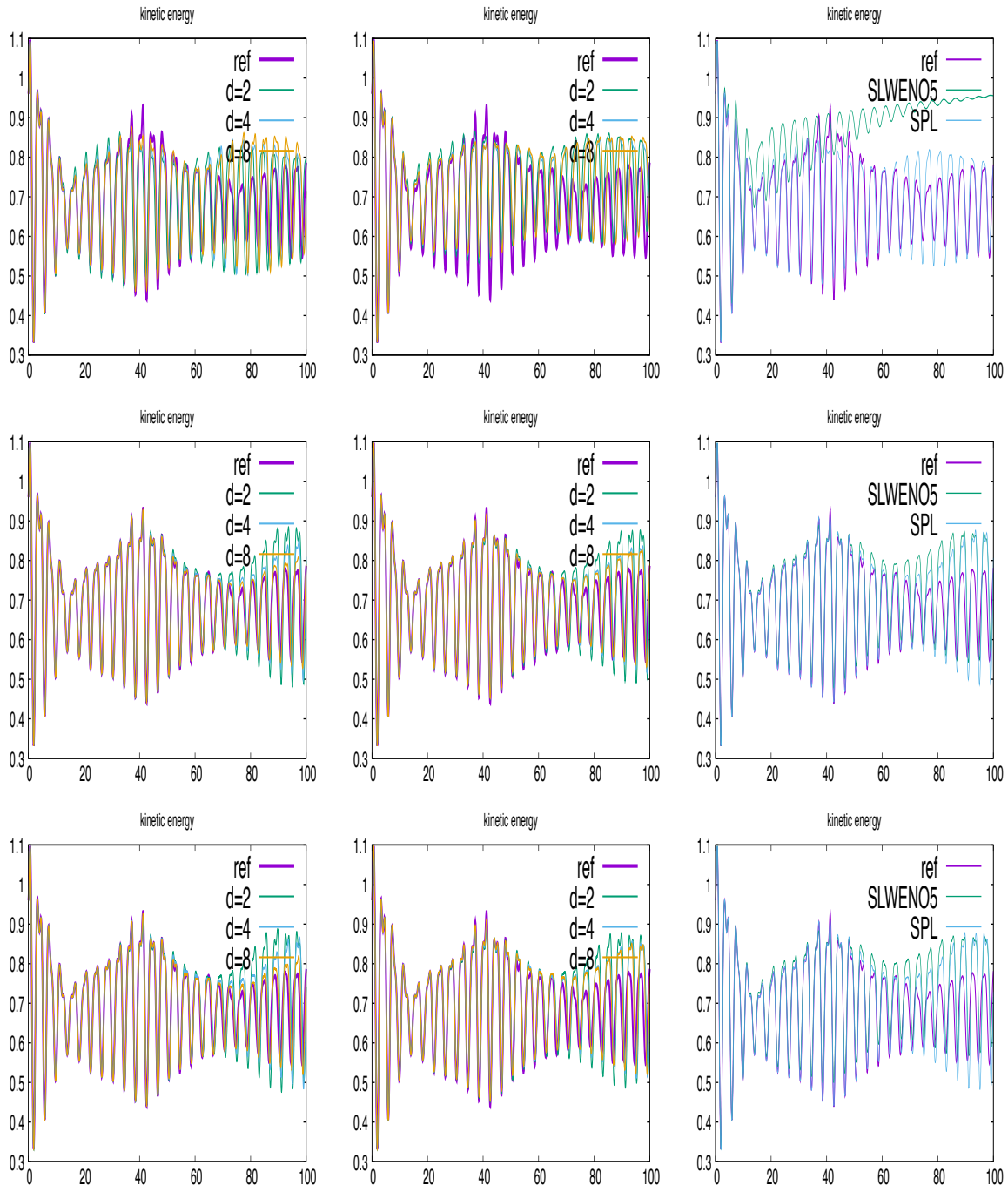


Figure 9: Time evolution of kinetic energy for beam test case; top, 64×64 grid and $\Delta t = 0.1$, middle 256×256 grid and $\Delta t = 0.1$ bottom 256×256 grid and $\Delta t = 0.01$; from left to right: without limiter, with limiter and cubic splines/SLWENO5

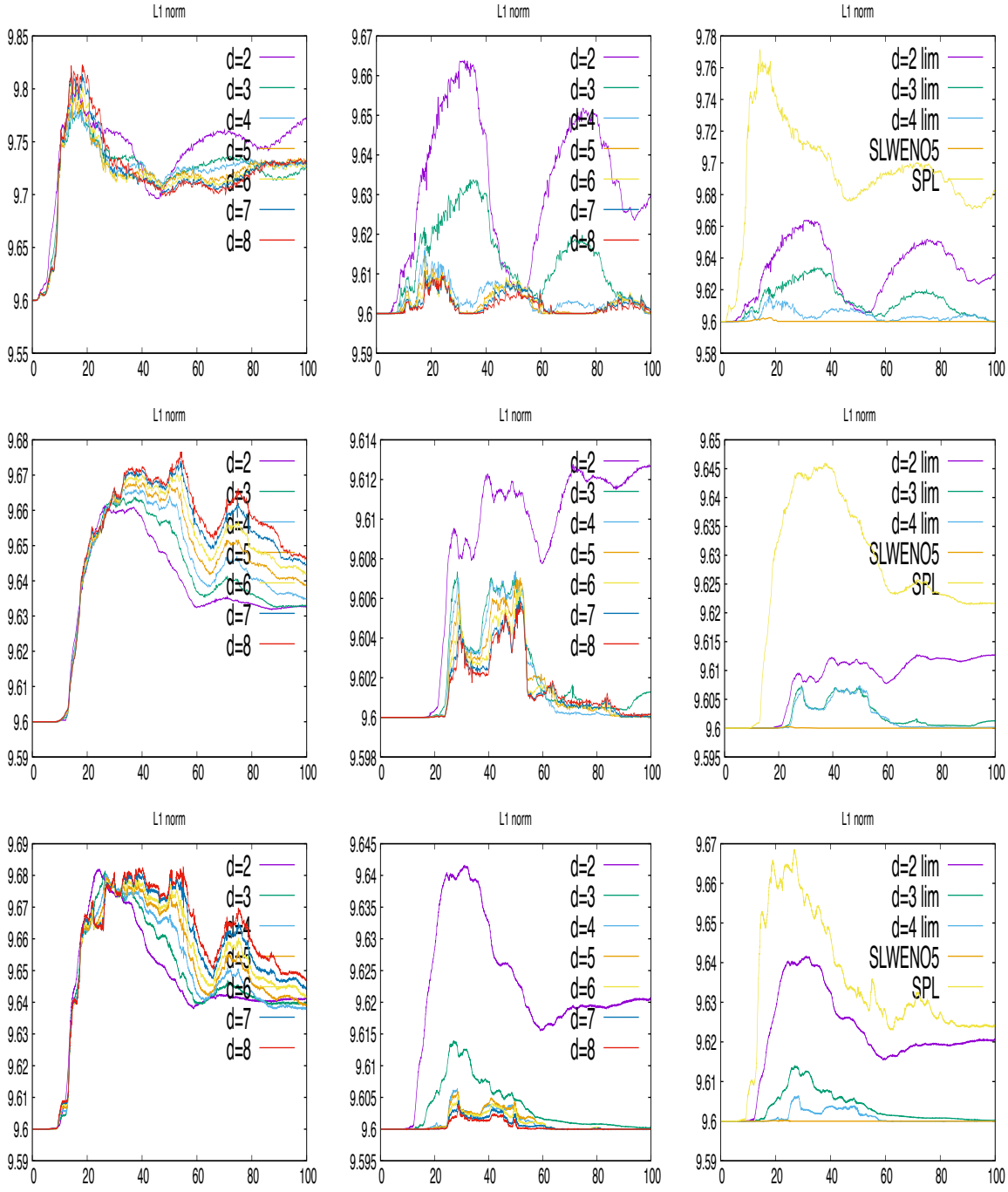


Figure 10: Time evolution of L^1 norm for beam test case; top, 64×64 grid and $\Delta t = 0.1$, middle 256×256 grid and $\Delta t = 0.1$ bottom 256×256 grid and $\Delta t = 0.01$; left without limiter; right with limiter; reference is $d = 4$ with limiter on 4096×4096 grid

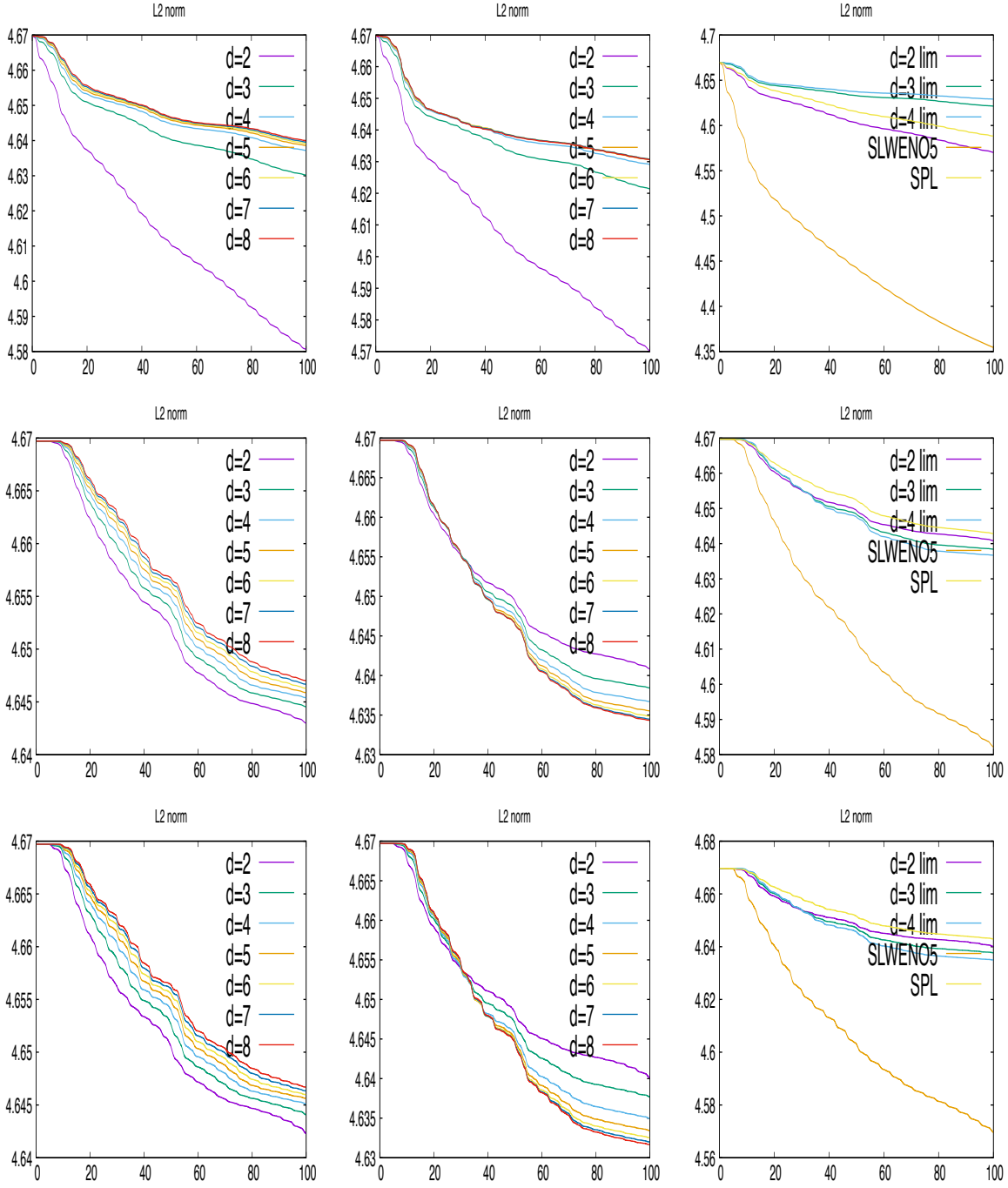


Figure 11: Time evolution of L^2 norm for beam test case; top, 64×64 grid and $\Delta t = 0.1$, middle 256×256 grid and $\Delta t = 0.1$ bottom 256×256 grid and $\Delta t = 0.01$; left without limiter; right with limiter

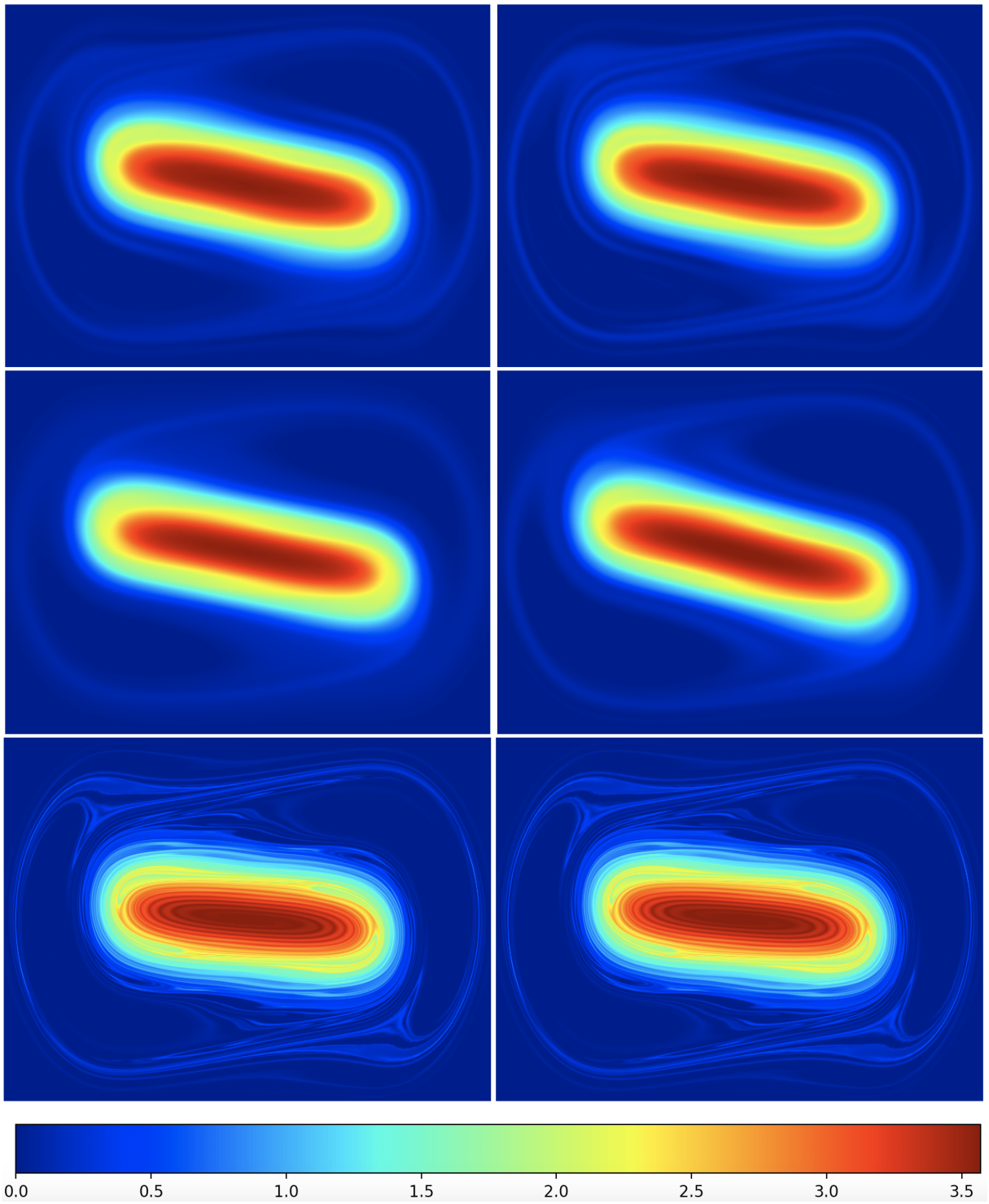


Figure 12: Beam test case: distribution fonction at time $T = 100$ with $d = 8$, $\Delta t = 0.1$. From top to bottom and left to right: $d = 8$ lim, $d = 8$ no lim, SLWENO5, cubic splines all on 256×256 grid; no lim on 4096×4096 grid, lim on 4096×4096 grid

References

[1] A. Back and E. Sonnendrucker. Finite element hodge for spline discrete differential forms. application to the vlasov-poisson system. *Applied Numerical Mathematics*, 79:124–136, 2014.

[2] J. W. Banks, A. G. Odu, R. L. Berger, T. Chapman, W. Arrighi, and S. Brunner. High-order accurate conservative finite difference methods for vlasov equations in 2d+2v. *SIAM Journal on Scientific Computing*, 41(5), 2019.

[3] N. Besse and M. Mehrenberger. Convergence of classes of high-order semi-lagrangian schemes for the vlasov-poisson system. *Mathematics of Computation*, 77(261):93–123, 2008.

[4] X. Cai, W. Guo, and J. Qiu. A high order semi-lagrangian discontinuous galerkin method for vlasov-poisson simulations without operator splitting. *Journal of Computational Physics*, 354:529–551, 2018.

[5] X. Cai, W. Guo, and J.-M. Qiu. A high order semi-lagrangian discontinuous galerkin method for the two-dimensional incompressible euler equations and the guiding center vlasov model without operator splitting. *Journal of Scientific Computing*, 79(2):1111–1134, 2019.

[6] X. Cai, J.-M. Qiu, and Y. Yang. An eulerian-lagrangian discontinuous galerkin method for transport problems and its application to nonlinear dynamics. *arXiv preprint arXiv:2002.02930*, 2020.

[7] F. Casas, N. Crouseilles, E. Faou, and M. Mehrenberger. High-order hamiltonian splitting for the vlasov–poisson equations. *Numerische Mathematik*, 135(3):769–801, 2017.

[8] F. Charles, B. Despres, and M. Mehrenberger. Enhanced convergence estimates for semi-lagrangian schemes application to the vlasov-poisson equation. *SIAM Journal on Numerical Analysis*, 51(2):840–863, 2013.

[9] Y. Cheng, A. J. Christlieb, and X. Zhong. Energy-conserving discontinuous galerkin methods for the vlasov-maxwell system. *Journal of Computational Physics*, 256(1):630–655, 2014.

[10] Y. Cheng, I. M. Gamba, and P. J. Morrison. Study of conservation and recurrence of runge-kutta discontinuous galerkin schemes for vlasov-poisson systems. *Journal of Scientific Computing*, 56(2):319–349, 2013.

[11] G. H. Cottet and P. A. Raviart. On particle-in-cell methods for the vlasov-poisson equations. *Transport Theory and Statistical Physics*, 15:1–31, 1986.

[12] N. Crouseilles, P. Glanc, S. A. Hirstoaga, E. Madaule, M. Mehrenberger, and J. Pétri. A new fully two-dimensional conservative semi-lagrangian method: applications on polar grids, from diocotron instability to itg turbulence. *The European Physical Journal D*, 68(9):252, 2014.

[13] N. Crouseilles, P. Glanc, M. Mehrenberger, and C. Steiner. Finite volume schemes for vlasov. cemracs 2011: Multiscale coupling of complex models in scientific computing. In *ESAIM Proc.*, volume 38, pages 275–297, 2012.

[14] N. Crouseilles, M. Lemou, F. Mehats, and X. Zhao. Uniformly accurate particle-in-cell method for the long time solution of the two-dimensional vlasov-poisson equation with uniform strong magnetic field. *Journal of Computational Physics*, 346:172–190, 2017.

[15] N. Crouseilles, M. Mehrenberger, and E. Sonnendrücker. Conservative semi-lagrangian schemes for vlasov equations. *Journal of Computational Physics*, 229(6):1927–1953, 2010.

[16] V. Daru and C. Tenaud. High order one-step monotonicity-preserving schemes for unsteady compressible flow calculations. *Journal of Computational Physics*, 193(2):563–594, 2004.

[17] B. A. De Dios, J. A. Carrillo, and C. Shu. Discontinuous galerkin methods for the multi-dimensional vlasov–poisson problem. *Mathematical Models and Methods in Applied Sciences*, 22(12):1250042, 2012.

[18] P. Degond, F. Deluzet, and D. Doyen. Asymptotic-preserving particle-in-cell methods for the vlasov-maxwell system in the quasi-neutral limit. *Journal of Computational Physics*, 330:467 – 492, 2017.

[19] P. Degond, F. Deluzet, L. Navoret, A. Sun, and M. Vignal. Asymptotic-preserving particle-in-cell method for the vlasov-poisson system near quasineutrality. *Journal of Computational Physics*, 229(16):5630–5652, 2010.

[20] B. Després. Polynomials with bounds and numerical approximation. *Numerical Algorithms*, 76(3):829–859, 2017.

[21] B. Després and F. Lagoutière. Contact discontinuity capturing schemes for linear advection and compressible gas dynamics. *Journal of Scientific Computing*, 16(4):479–524, 2002.

[22] L. Einkemmer. A performance comparison of semi-lagrangian discontinuous galerkin and spline based vlasov solvers in four dimensions. *Journal of Computational Physics*, 376:937–951, 2019.

[23] L. Fatone, D. Funaro, and G. Manzini. A semi-lagrangian spectral method for the vlasov-poisson system based on fourier, legendre and hermite polynomials. *arXiv: Numerical Analysis*, 2018.

[24] F. Filbet. Convergence d'un schéma de type volumes finis pour la résolution numérique du système de vlasov–poisson en dimension un. *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics*, 330(11):979 – 984, 2000.

[25] F. Filbet. Convergence of a finite volume scheme for the vlasov-poisson system. *SIAM Journal on Numerical Analysis*, 39(4):1146–1169, 2001.

[26] F. Filbet and L. M. Rodrigues. Asymptotically stable particle-in-cell methods for the vlasov-poisson system with a strong external magnetic field. *SIAM Journal on Numerical Analysis*, 54(2):1120–1146, 2016.

[27] F. Filbet, E. Sonnendrücker, and P. Bertrand. Conservative numerical schemes for the vlasov equation. *Journal of Computational Physics*, 172(1):166–187, 2001.

[28] A. Hamiaz, M. Mehrenberger, H. Sellama, and E. Sonnendrücker. The semi-lagrangian method on curvilinear grids. *Communications in Applied and Industrial Mathematics*, 7(3):99–137, 2016.

[29] R. E. Heath, I. M. Gamba, P. J. Morrison, and C. Michler. A discontinuous galerkin method for the vlasov-poisson system. *Journal of Computational Physics*, 231(4):1140–1174, 2012.

[30] E. Madaule, M. Restelli, and E. Sonnendrucker. Energy conserving discontinuous galerkin spectral element method for the vlasov-poisson system. *Journal of Computational Physics*, 279:261–288, 2014.

[31] G. Manzini, G. L. Delzanno, J. Vencels, and S. Markidis. A legendre-fourier spectral method with exact conservation laws for the vlasov-poisson system. *Journal of Computational Physics*, 317:82–107, 2016.

[32] J. Qiu and C. Shu. Positivity preserving semi-lagrangian discontinuous galerkin formulation: Theoretical analysis and application to the vlasov-poisson system. *Journal of Computational Physics*, 230(23):8386–8409, 2011.

[33] J.-M. Qiu and C.-W. Shu. Conservative semi-lagrangian finite difference weno formulations with applications to the vlasov equation. *Communications in Computational Physics*, 10(4):979, 2011.

[34] D. Sirajuddin and W. N. Hitchon. A truly forward semi-lagrangian weno scheme for the vlasov-poisson system. *Journal of Computational Physics*, 392:619–665, 2019.

[35] E. Sonnendrucker, J. Roche, P. Bertrand, and A. Ghizzo. The semi-lagrangian method for the numerical resolution of the vlasov equation. *Journal of Computational Physics*, 149(2):201–220, 1999.

[36] C. Standar. *On Finite Element Schemes for Vlasov-Maxwell System and Schrodinger Equation*. PhD thesis, Chalmers Tekniska Hogskola, 2017.

[37] A. Suresh and H. T. Huynh. Accurate monotonicity-preserving schemes with runge-kutta time stepping. *Journal of Computational Physics*, 136(1):83–99, 1997.

[38] S. Tanaka, K. Yoshikawa, T. Minoshima, and N. Yoshida. Multidimensional vlasov–poisson simulations with high-order monotonicity- and positivity-preserving schemes. *The Astrophysical Journal*, 849(2):76, nov 2017.

[39] Z. Tao, W. Guo, and Y. Cheng. Sparse grid discontinuous galerkin methods for the vlasov-maxwell system. *arXiv: Numerical Analysis*, 2018.

[40] T. Umeda. A conservative and non-oscillatory scheme for vlasov code simulations. *Earth, planets and space*, 60(7):773–779, 2008.

[41] T. Umeda, Y. Nariyuki, and D. Kariya. A non-oscillatory and conservative semi-lagrangian scheme with fourth-degree polynomial interpolation for solving the vlasov equation. *Computer Physics Communications*, 183(5):1094–1100, 2012.

[42] H. D. Victory and E. J. Allen. The convergence theory of particle-in-cell methods for multidimensional vlasov-poisson systems. *SIAM Journal on Numerical Analysis*, 28(5):1207–1241, 1991.

[43] G. Vogman, U. Shumlak, and P. Colella. Conservative fourth-order finite-volume vlasov–poisson solver for axisymmetric plasmas in cylindrical $(r, vr, v\theta)$ phase space coordinates. *Journal of Computational Physics*, 373:877 – 899, 2018.

[44] T. Xiong, J.-M. Qiu, Z. Xu, and A. Christlieb. High order maximum principle preserving semi-lagrangian finite difference weno schemes for the vlasov equation. *Journal of Computational Physics*, 273:618–639, 2014.

[45] Z. Xu and C. Shu. Anti-diffusive flux corrections for high order finite difference weno schemes. *Journal of Computational Physics*, 205(2):458–485, 2005.

[46] C. Yang and F. Filbet. Conservative and non-conservative methods based on hermite weighted essentially non-oscillatory reconstruction for vlasov equations. *Journal of Computational Physics*, 279:18–36, 2014.

[47] C. Yang and L. M. Tine. A hybrid finite volume method for advection equations and its applications in population dynamics. *Numerical Methods for Partial Differential Equations*, 33(4):1114–1142, 2017.

[48] H. Yang and F. Li. Discontinuous galerkin methods for relativistic vlasov-maxwell system. *Journal of Scientific Computing*, 73(2):1216–1248, 2017.

[49] S. Zaki, L. Gardner, and T. Boyd. A finite element code for the simulation of one-dimensional vlasov plasmas. i. theory. *Journal of Computational Physics*, 79(1):184–199, 1988.