

# The Lax–Wendroff theorem for Patankar-type methods applied to hyperbolic conservation laws

Janina Bender<sup>a,\*</sup>, Thomas Izgin<sup>a</sup>, Philipp Öffner<sup>b</sup>, Davide Torlo<sup>c</sup>

<sup>a</sup>*Institute of Mathematics, University of Kassel, Heinrich-Plett-Str. 40, 34132, Kassel, Germany*

<sup>b</sup>*Institute of Mathematics, Clausthal University of Technology, Clausthal-Zellerfeld, Germany*

<sup>c</sup>*Dipartimento di Matematica Guido Castelnuovo, Università di Roma La Sapienza, piazzale Aldo Moro, 5, Roma, 00185, Italy*

---

## Abstract

For hyperbolic conservation laws, the famous Lax–Wendroff theorem delivers sufficient conditions for the limit of a convergent numerical method to be a weak (entropy) solution. This theorem is a fundamental result, and many investigations have been done to verify its validity for finite difference, finite volume, and finite element schemes, using either explicit or implicit linear time-integration methods.

Recently, the use of modified Patankar (MP) schemes as time-integration methods for the discretization of hyperbolic conservation laws has gained increasing interest. These schemes are unconditionally conservative and positivity-preserving and only require the solution of a linear system. However, MP schemes are by construction nonlinear, which is why the theoretical investigation of these schemes is more involved. We prove an extension of the Lax–Wendroff theorem for the class of MP methods. This is the first extension of the Lax–Wendroff theorem to nonlinear time integration methods with just an additional hypothesis on the total time variation boundedness of the numerical solutions. We provide some numerical simulations that validate the theoretical observations.

**Keywords:** Local conservation, Implicit Lax–Wendroff theorem, Total time variation, Patankar-type schemes

**2000 MSC:** 65M06, 65M20, 65L06

---

## 1. Introduction

In the context of hyperbolic conservation laws, the development of high-order, structure-preserving methods has attracted a broad attention. These methods are essential to ensure physically reliable solutions satisfying suitable constraints. One such constraint is the positivity of certain quantities, such as density in the Euler equations of gas dynamics or the water height in the context of shallow water flows where positivity preservation is usually obtained within a numerical method by a positivity-preserving reconstruction/limiting technique and an additional time step constraint, see for instance [26, 39, 3]. To overcome the limitation of time step constraints without solving a completely nonlinear system and to guarantee the preservation of positivity, modified Patankar (MP) schemes have recently been applied, see e.g. [12, 19, 11]. These schemes, designed for production-destruction systems (PDS), are nonlinear, they preserve positivity unconditionally, and are conservative [9, 35]. Moreover, due to their specific structure, only linear equations need to be solved. Applying them as time-integration schemes for hyperbolic problems has yielded promising results, as outlined in recent works [34, 36, 37]. For instance in [12, 11], the MP Deferred Correction (MPDeC)

---

\*Corresponding author:

Email address: jmbender@mathematik.uni-kassel.de (Janina Bender)

method from [35] is used to ensure the positivity of the water height in the shallow water equations discretized with finite volumes in space. In [34, 36], it is also ensured with a Patankar-type method, where a discontinuous Galerkin (DG) method is used for space discretization. However, the theoretical background of these schemes has not yet been sufficiently investigated, especially concerning their integration into the existing convergence theory for numerical methods of hyperbolic systems. A cornerstone in the numerical treatment of hyperbolic conservation laws within this framework is the well-known Lax–Wendroff theorem [29]. This theorem establishes sufficient conditions for the convergence of numerical approximations to weak (entropy) solutions of the underlying nonlinear hyperbolic conservation law. Much of the literature, see e. g. [1, 14, 27, 40, 6, 28, 16, 41, 26], has been focusing on spatial discretization (even with unstructured grids and finite element (FE) methods), mostly using linear explicit time integration methods. One exception is given in [8, 32], where a Lax–Wendroff theorem was proven for implicit Runge–Kutta methods using explicit pseudo-time Runge–Kutta schemes for approximating the solution to the occurring (non-)linear system. However, here only finitely many pseudo-time steps are allowed to approximate the update of the difference scheme. Given a fixed spatial discretization of  $N$  points, this is not a problem for linear implicit systems, as suitable Krylov subspace methods converge to the exact solution in  $N$  iterations. Problems arise when we are interested in the convergence towards the exact solution for  $N \rightarrow \infty$ . In this case, the previous result does not directly apply, leaving some gaps in the proof of the Lax–Wendroff theorem for the implicit setting, either linear or nonlinear. Nevertheless, a first step towards a deeper understanding is given in [31].

In this paper, we propose a different ansatz introducing the notion of total time variation and we shift the focus to *nonlinear* time-integration schemes extending the convergence theory of Lax and Wendroff to Patankar-type methods. Altogether, the extension of the Lax–Wendroff theorem additionally assumes that the total variation in the time variable (see Definition 6) is uniformly bounded. This extension allows us, for the first time, to establish a theory of their convergence in the context of partial differential equations (PDEs). The theoretical results are validated through numerical tests for different conservation laws.

The rest of the paper is organized as follows: In Section 2, we introduce the concept of consistency and convergence for difference schemes for conservation laws and repeat the basic formulation of the Lax–Wendroff theorem which can be found nowadays in many excellent lecture books like [30, 25, 17]. Afterwards, in Section 3, MP schemes will be introduced for production-destruction systems in context of ordinary differential equations (ODE). Here, the recent interpretation as non-standard additive Runge–Kutta (NSARK) methods [21] will also be used. Then, we explain how to employ the MP ideas within difference schemes (finite volume (FV), finite differences, for example with Weighted Essentially Non-Oscillatory (WENO) reconstruction) to guarantee positivity preservation. In Section 4, we demonstrate the extension of the classical Lax–Wendroff Theorem for a large class of NSARK methods, including MP difference schemes. Here, we focus on the first order version (Theorem 3) but we give also a sketch of the proof for the extension to higher-order. In Section 5, we validate our theoretical findings by some numerical experiments. In particular, we demonstrate that we are unconditionally positivity-preserving and we converge to a weak solution. In Section 6, we summarize the work, we derive conclusions and give perspectives for future works. Finally, in Appendix A, we demonstrate how the total variation, which is an essential aspect of the Lax–Wendroff theorem, can in some specific frameworks be proven to be diminishing for the studied MP schemes.

## 2. Consistency and convergence for difference schemes for conservation laws

To transfer the convergence theory of the Lax–Wendroff theorem to Patankar-type methods, we first summarize the theoretical background of the classical theory. We consider a one-dimensional, scalar Cauchy problem

$$\partial_t u(x, t) + \partial_x f(u(x, t)) = 0, \quad u(x, 0) = \psi(x), \quad (1)$$

with  $u: \mathbb{R} \times \mathbb{R}_0^+ \rightarrow \mathbb{R}$  and  $f, \psi: \mathbb{R} \rightarrow \mathbb{R}$ , where  $f$  denotes the physical flux function.

We examine difference schemes of the form

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

to find approximations  $U_i^n \in \mathbb{R}$  to the solution  $u(x_i, t_n)$  at  $N$  discrete grid points  $\{x_i\}_{i=1}^N$  and at time points  $\{t_n\}_{n=0}^{N_t}$ . Here, the time step of a given grid is denoted by  $\Delta t = t_{n+1} - t_n$  for all  $n$  and the spatial mesh size by  $\Delta x = x_{i+1} - x_i$  for all  $i$ . Furthermore, we have

$$g_{i+\frac{1}{2}}^n = g(U_{i-p}^n, U_{i-p+1}^n, \dots, U_{i+q}^n), \quad (3)$$

where  $g$  is a continuous function, depending on  $p+q+1$  arguments, called the *numerical flux*. The following results and definitions hold for both explicit and implicit numerical fluxes [25, p. 44], but for simplicity, we only consider explicit schemes. First, we note the main properties of these fluxes.

**Definition 1 (Consistency).** The numerical flux function  $g$  from (3) is *consistent* to the physical flux function  $f$ , if

- i.  $g(\bar{u}, \dots, \bar{u}) = f(\bar{u})$  for any  $\bar{u}$  and
- ii.  $g$  is locally Lipschitz continuous in every argument, that is if it exists a  $K \in \mathbb{R}$  such that  $|g(U_{i-p}^n, U_{i-p+1}^n, \dots, U_{i+q}^n) - f(\bar{u})| \leq K \max_{-p \leq j \leq q} |U_{i+j}^n - \bar{u}|$  holds for  $U_{i+j}^n$  sufficiently close to  $\bar{u}$ .

If  $g$  is consistent, then we also say that the difference scheme (2) is consistent with the PDE (1).

To ensure that the limit of a converging difference scheme is a weak solution of the system, *Lax* and *Wendroff* [29] proved the following theorem. Here, we present a formulation with slightly different requirements.

**Theorem 1 (Lax–Wendroff [30]).** Consider a sequence of step sizes  $\Delta x_{\{l\}} > 0$  with  $\lim_{l \rightarrow \infty} \Delta x_{\{l\}} = 0$ . Let  $(U^{(l)})_l$  be a sequence of piece-wise constant functions constructed from numerical solutions obtained by a difference scheme consistent with (1) with time step size  $\Delta t_{\{l\}} = \lambda \Delta x_{\{l\}}, \lambda \equiv \text{const}$ . If for any domain  $[a, b] \times [0, T]$  the total variation  $TV(U^{(l)}(\cdot, t))$  is uniformly bounded for all  $l$  and  $t \in [0, T]$  and if  $U^{(l)} \xrightarrow{l \rightarrow \infty} u$  in  $L_{\text{loc}}^1(\mathbb{R} \times [0, T])$ , then  $u$  is a weak solution of the conservation law (1).

**Remark 1.** The Lax–Wendroff theorem 1 guarantees that the limit of a convergent conservative numerical scheme is a weak solution of the conservation law. Furthermore, if the approximated solutions also satisfy a discrete form of the entropy inequality, then the limit corresponds to the weak entropy solution. Therefore, the theorem remains applicable in identifying the entropy solution, provided the entropy condition is met. Therefore, applying entropy stable schemes which fulfill the Lax–Wendroff theorem yield weak entropy solutions [43, 10, 15, 17].

The main idea of this paper is to extend Theorem 1 if MP schemes (therefore nonlinear time-integration) schemes are used inside the discretization. We shortly introduce now the concept of MP schemes in the following section.

### 3. Patankar-type schemes

#### 3.1. Modified Patankar for Production-Destruction Systems

Patankar-type schemes are introduced to solve systems of ODE problems

$$\partial_t \mathbf{u}(t) = \Phi(\mathbf{u}(t)), \quad \mathbf{u}(t_0) = \mathbf{u}^0, \quad (4)$$

with  $\mathbf{u} : \mathbb{R}_0^+ \rightarrow \mathbb{R}^N$  and  $\Phi : \mathbb{R}^N \rightarrow \mathbb{R}^N$ . Similar to PDE problems, we denote the  $i$ -th component of an approximate solution obtained by a numerical scheme at time step  $t_n$  as  $U_i^n$ , while  $\mathbf{U}^n$  denotes the whole solution vector. As mentioned in the introduction, MP schemes are constructed to preserve positivity and conservativity, properties that are defined as follows.

**Definition 2.** An ODE scheme is called *unconditionally positive* if  $\mathbf{U}^n > 0$  implies  $\mathbf{U}^{n+1} > 0$  for all  $n \in \mathbb{N}_0$  and  $\Delta t > 0$ . Furthermore it is *unconditionally conservative* if  $\sum_{i=1}^N U_i^{n+1} = \sum_{i=1}^N U_i^n$  for all  $n \in \mathbb{N}_0$  and  $\Delta t > 0$ .

Throughout the paper, inequalities involving vectors, such as  $\mathbf{U}^n > 0$ , are meant componentwise for all components. Consider an ODE written as a PDS

$$\partial_t u_i(t) = \sum_{j=1}^N p_{ij}(\mathbf{u}(t)) - \sum_{j=1}^N d_{ij}(\mathbf{u}(t)), \quad i = 1, \dots, N \quad (5)$$

with  $p_{ij}, d_{ij}$  non-negative production and destruction terms for  $\mathbf{u} \geq 0$ . In the past, such PDSs were commonly used to describe *inter alia* biological reactions as in [9]. In later works, PDSs were also used to describe the semi-discretization of hyperbolic balance laws as in [35, 34]. For both contexts it is important to have conservative and unconditionally positive schemes. Therefore, it is necessary to denote these properties also for PDSs.

**Definition 3.** A PDS (5) is said to be *non-negative* if, for an initial value  $u_i(0) \geq 0$ ,  $u_i(t) \geq 0$  follows for all  $t > 0$  and  $i = 1, \dots, N$ . If the strict inequalities hold, the PDS is called *positive*.

**Definition 4.** The PDS (5) is called *conservative* if  $p_{ij}(\mathbf{u}) = d_{ji}(\mathbf{u})$  for all  $i, j = 1, \dots, N$  and *fully conservative* if additionally  $p_{ii}(\mathbf{u}) = d_{ii}(\mathbf{u}) = 0$  for all  $i = 1, \dots, N$ .

We remark that a conservative and a fully conservative PDS are analytically the same model, but treating the production and destruction matrices numerically, it is safer to choose the fully conservative one. Hereafter, all methods are supposed to be applied to a fully conservative and positive PDS. Based on such a conservative and positive PDS we can apply an explicit Runge–Kutta (RK) method and use the modified Patankar-trick to obtain a modified Patankar–Runge–Kutta (MPRK) scheme.

**Definition 5.** Given an explicit  $s$ -stage RK scheme with non-negative Butcher tableau  $\begin{array}{c|c} \mathbf{c} & \mathbf{A} \\ \hline & \mathbf{b}^T \end{array}$ , i.e.,  $\mathbf{A} \geq 0$  and  $\mathbf{b} \geq 0$ , applied to a fully conservative PDS (5), the method

$$U_i^{(k)} = U_i^n + \Delta t \sum_{v=1}^{k-1} a_{kv} \sum_{j=1}^N \left( p_{ij}(\mathbf{U}^{(v)}) \frac{U_j^{(k)}}{\pi_j^{(k)}} - d_{ij}(\mathbf{U}^{(v)}) \frac{U_i^{(k)}}{\pi_i^{(k)}} \right), \quad k = 1, \dots, s \quad (6)$$

$$U_i^{n+1} = U_i^n + \Delta t \sum_{k=1}^s b_k \sum_{j=1}^N \left( p_{ij}(\mathbf{U}^{(k)}) \frac{U_j^{n+1}}{\sigma_j} - d_{ij}(\mathbf{U}^{(k)}) \frac{U_i^{n+1}}{\sigma_i} \right) \quad (7)$$

is called *MPRK scheme* if the Patankar weight denominators (PWDs)  $\pi_i^{(k)}$  and  $\sigma_i$  are unconditionally positive,  $\pi_i^{(k)} = \pi_i^{(k)}(\mathbf{U}^n, \mathbf{U}^{(1)}, \dots, \mathbf{U}^{(k-1)})$  is independent of  $\mathbf{U}^{(k)}$ , and  $\sigma_i = \sigma_i(\mathbf{U}^n, \mathbf{U}^{(1)}, \dots, \mathbf{U}^{(s)})$  is independent of  $\mathbf{U}^{n+1}$ .

The huge advantage of such methods is that, given PWDs, MPRK schemes are unconditionally positive and conservative scheme, see for instance [23], while they merely require the solution of  $s$  linear systems

$$\mathbf{U}^{(1)} = \mathbf{U}^n,$$

$$\begin{aligned}\mathbf{M}^{(k)} \mathbf{U}^{(k)} &= \mathbf{U}^n, \quad k = 2, \dots, s, \\ \mathbf{M} \mathbf{U}^{n+1} &= \mathbf{U}^n,\end{aligned}$$

all of which are well-posed having a unique and positive solution. We note that MPRK schemes can also be devised for non-conservative PDS, see e.g. [44, 20], which is relevant when applying these schemes in the context of PDEs with non-periodic boundary conditions.

**Example 1** (Modified Patankar Euler). *As an example, we illustrate here the first order modified Patankar Euler (MPE) time discretization of a conservative PDS. In this case, there is only the final update (7) where  $\sigma_i = U_i^n$  for all  $i = 1, \dots, N$ . The scheme reads*

$$U_i^{n+1} = U_i^n + \Delta t \sum_{j=1}^N \left( p_{ij}(\mathbf{U}^n) \frac{U_j^{n+1}}{U_j^n} - d_{ij}(\mathbf{U}^n) \frac{U_i^{n+1}}{U_i^n} \right).$$

Clearly, each time step includes a linear system to be solved, with a matrix defined by

$$M_{ij} = \begin{cases} -\Delta t \frac{p_{ij}(\mathbf{U}^n)}{U_j^n} & \text{if } i \neq j, \\ 1 + \sum_{k \neq i} \Delta t \frac{d_{ik}(\mathbf{U}^n)}{U_k^n} & \text{if } i = j. \end{cases}$$

This scheme will be the prototype for high order methods and we will use it to define the successive examples.

After the first introduction of MPRK methods based on explicit first- and second-order RK schemes in [9], the Patankar-trick has since been successfully extended to a variety of RK frameworks, including state-of-the-art strong-stability-preserving Runge–Kutta (SSPRK) methods [42] resulting in MPSSPRK schemes [18], and arbitrary high-order Deferred Correction (DeC) schemes [13] yielding MPDeC methods [12]. It has also been applied to multistep methods in a recent work [22]. Further positivity-preserving schemes are Geometric Conservative (GeCo) [33] and generalized BBKS (gBBKS) [5]. To give the main ideas for the extension of the Lax–Wendroff theorem to all these schemes and many more, we use the environment of NSARK methods as introduced in [20] to obtain results that are as universal as possible.

Therefore, first recall the definition of additive Runge–Kutta (ARK) methods. These are constructed based on an additive splitting

$$\frac{d}{dt} \mathbf{u}(t) = \sum_{v=1}^N \mathbf{F}^{[v]}(\mathbf{u}(t)), \quad \mathbf{u}(0) = \mathbf{u}^0 \in \mathbb{R}^d$$

of the right hand side of the ODE. Thus, several schemes given by an extended Butcher tableau

$$\begin{array}{c|c|c|c|c} \mathbf{c} & \mathbf{A}^{[1]} & \mathbf{A}^{[2]} & \dots & \mathbf{A}^{[N]} \\ \hline & \mathbf{b}^{[1]} & \mathbf{b}^{[2]} & \dots & \mathbf{b}^{[N]} \end{array},$$

can be applied to the different summands leading to a scheme of the form

$$\begin{aligned}\mathbf{U}^{(k)} &= \mathbf{U}^n + \Delta t \sum_{r=1}^s \sum_{v=1}^N a_{kr}^{[v]} \mathbf{F}^{[v]}(\mathbf{U}^{(r)}), \quad k = 1, \dots, s, \\ \mathbf{U}^{n+1} &= \mathbf{U}^n + \Delta t \sum_{r=1}^s \sum_{v=1}^N b_r^{[v]} \mathbf{F}^{[v]}(\mathbf{U}^{(r)}).\end{aligned}$$

Notice that these Runge–Kutta coefficients are fixed for every summand. Now, the idea of non-standard additive Runge–Kutta (NSARK) methods is to allow a dependency on both solution and stepsize in the coefficients, which read as

$$\frac{\mathbf{c} \begin{vmatrix} A^{[1]}(\hat{\mathbf{U}}^n, t_n, \Delta t) \\ A^{[2]}(\hat{\mathbf{U}}^n, t_n, \Delta t) \\ \dots \\ A^{[N]}(\hat{\mathbf{U}}^n, t_n, \Delta t) \end{vmatrix}}{\mathbf{b} \begin{vmatrix} b^{[1]}(\hat{\mathbf{U}}^n, t_n, \Delta t) \\ b^{[2]}(\hat{\mathbf{U}}^n, t_n, \Delta t) \\ \dots \\ b^{[N]}(\hat{\mathbf{U}}^n, t_n, \Delta t) \end{vmatrix}},$$

where  $\hat{\mathbf{U}}^n = \left( \mathbf{U}^n \mid \mathbf{U}^{(1)} \mid \dots \mid \mathbf{U}^{(s)} \mid \mathbf{U}^{n+1} \right)$ , resulting in

$$\begin{aligned} \mathbf{U}^{(k)} &= \mathbf{U}^n + \Delta t \sum_{r=1}^s \sum_{v=1}^N a_{kr}^{[v]}(\hat{\mathbf{U}}^n, t_n, \Delta t) \mathbf{F}^{[v]}(\mathbf{U}^{(r)}), \quad k = 1, \dots, s, \\ \mathbf{U}^{n+1} &= \mathbf{U}^n + \Delta t \sum_{r=1}^s \sum_{v=1}^N b_r^{[v]}(\hat{\mathbf{U}}^n, t_n, \Delta t) \mathbf{F}^{[v]}(\mathbf{U}^{(r)}). \end{aligned}$$

As introduced in equation (5), we focus on ODEs which can be split into a PDS. So we can understand such a right-hand side (RHS) as

$$F_i^{[v]}(\mathbf{U}(t)) = \begin{cases} p_{iv}(\mathbf{U}(t)), & v \neq i, \\ -\sum_{j=1}^N d_{ij}(\mathbf{U}(t)), & v = i \end{cases}$$

and interpret a Patankar-type method as NSARK method with

$$\begin{aligned} a_{kr}^{[v]}(\hat{\mathbf{U}}^n, t_n, \Delta t) &= a_{kr} \gamma_v^{(k)}(\hat{\mathbf{U}}^n, t_n, \Delta t), \\ b_r^{[v]}(\hat{\mathbf{U}}^n, t_n, \Delta t) &= b_r \delta_v(\hat{\mathbf{U}}^n, t_n, \Delta t), \end{aligned} \tag{8}$$

where the Patankar-trick is incorporated into the *non-standard weights (NSWs)*  $\gamma_v^{(k)}, \delta_v$ . In particular, we obtain the MPRK methods introduced in Definition 5 by choosing the NSWs

$$\gamma_v^{(k)} = \frac{U_v^{(k)}}{\pi_v^{(k)}}, \quad \delta_v = \frac{U_v^{n+1}}{\sigma_v}.$$

Note that we use “( )” in the upper indices whenever we refer to a stage while “[ ]” is used to refer to a different RK scheme used within the ARK framework.

### 3.2. Modified Patankar for difference schemes

The key properties of Patankar-type methods, being unconditionally positivity-preserving and conservative, can be a huge advantage when used as the time integration scheme of the semidiscretization of a hyperbolic PDE. For instance, they can guarantee to produce nonnegative solutions when approaching zero. This is in contrast with almost all classical time discretizations where this bound can either be obtained with constraints on the time-step or it cannot be obtained at all. Negative solutions in such contexts can lead not only to unphysical solutions, but also to NaN in the code and stop the execution. The only major exception is the implicit Euler method, which is unfortunately only first order accurate and very dissipative for hyperbolic conservation laws.

To apply Patankar-type time integration methods to semidiscrete hyperbolic PDEs, it is necessary to find an adequate splitting of the obtained ODE into a PDS. After that, the production and destruction terms are properly weighted. Note that such a splitting is not unique, see Remark 2.

The transfer to Patankar-type difference schemes is done along the difference scheme formulation as originally introduced in [29]. That is, a flux formulation of the RHS treated by an explicit Euler scheme

in time, which in our context is associated with the MPE scheme. Thus, the modification of the numerical fluxes is the key point of this extension. Therefore, we first introduce an explicit conservative PDS based on a classical numerical flux  $g$  as in (2) and add the corresponding weights. These are kept in a general notation to clarify the later extension to general Patankar-type methods with more stages.

Starting with the PDS, we interpret the fluxes at the cell boundaries as production or destruction terms according to their signs, i.e.,

$$\begin{aligned} p_{i,i+1}(\mathbf{U}^n) &= d_{i+1,i}(\mathbf{U}^n) := -\min\{g_{i+\frac{1}{2}}^n, 0\}, \\ d_{i,i+1}(\mathbf{U}^n) &= p_{i+1,i}(\mathbf{U}^n) := \max\{g_{i+\frac{1}{2}}^n, 0\}, \end{aligned} \quad (9)$$

where  $\mathbf{U}^n$  denotes the vector containing  $U_i^n$  for  $i = 1, \dots, N$ . With that, we can set  $p_{i,j}, d_{i,j}$  to zero for  $j \notin \{i-1, i+1\} \subseteq \{1, \dots, N\}$ . Now, we consider a one-stage Patankar-type method interpreted as a non-standard additive Runge–Kutta (NSARK) method with non-standard weights (NSWs)  $\delta_v = \delta_v(\hat{\mathbf{U}}^n, t_n, \Delta t)$  [20]. We write it in terms of this PDS in difference form:

$$\begin{aligned} U_i^{n+1} &= U_i^n + \Delta t \sum_{j=1}^N \frac{1}{\Delta x} \left( p_{i,j}(\mathbf{U}^n) \delta_j - d_{i,j}(\mathbf{U}^n) \delta_i \right) \\ &= U_i^n + \frac{\Delta t}{\Delta x} \left( p_{i,i+1}(\mathbf{U}^n) \delta_{i+1} - d_{i,i+1}(\mathbf{U}^n) \delta_i + p_{i,i-1}(\mathbf{U}^n) \delta_{i-1} - d_{i,i-1}(\mathbf{U}^n) \delta_i \right), \end{aligned} \quad (10)$$

where we use the conservation property of the PDS, i.e.,  $p_{i,j} = d_{j,i}$ , leading to the difference scheme formulation

$$U_i^{n+1} = U_i^n - \frac{\Delta t}{\Delta x} \left( F_{g_{i+\frac{1}{2}}}^\delta(\mathbf{U}^n) - F_{g_{i-\frac{1}{2}}}^\delta(\mathbf{U}^n) \right)$$

with the new numerical flux

$$F_{g_{i+\frac{1}{2}}}^\delta(\mathbf{U}^n) := d_{i,i+1}(\mathbf{U}^n) \delta_i - p_{i,i+1}(\mathbf{U}^n) \delta_{i+1} = \max\{g_{i+\frac{1}{2}}^n, 0\} \delta_i + \min\{g_{i+\frac{1}{2}}^n, 0\} \delta_{i+1} \quad (11)$$

and the short notation  $F_{g_{i+\frac{1}{2}}}^{n,\delta} = F_{g_{i+\frac{1}{2}}}^\delta(\mathbf{U}^n)$ . As mentioned earlier, the NSWs  $\delta_v$  depend on the particular method. For example, considering an MPRK scheme as in [23],  $\delta_v$  represent the corresponding Patankar weights, which are  $\delta_i = \frac{U_i^{n+1}}{U_i^n}$  for the MPE.

**Remark 2 (Non uniqueness of the splitting).** First of all, with the construction (9), there always exists a splitting of the conservation law into a conservative PDS. However, it should be noted that the splitting in production and destruction terms is not unique as, in general, one could add a positive constant to all production and destructions  $p_{ij}^{\text{new}} := p_{ij}^{\text{old}} + c$  for all  $i \neq j$ , and  $d_{ij}^{\text{new}} := p_{ji}^{\text{new}}$ , leading to another valid splitting of the original PDS for any  $c \geq 0$ .

Furthermore, there might be even multiple non-trivial splittings, e. g., for the inviscid Burgers' equation using an upwind FV discretization, where  $U_i \geq 0$ ,  $f(U_i) = \frac{1}{2} U_i^2 \geq 0$ , we obtain

$$\begin{aligned} \frac{dU_i}{dt} &= -\frac{1}{\Delta x} (f(U_i) - f(U_{i-1})) = \frac{1}{\Delta x} f(U_{i-1}) - \frac{1}{\Delta x} f(U_i) \\ &= \frac{1}{\Delta x} \left( -(f(U_i) - f(U_{i-1}))^- + f(U_i) \right) - \frac{1}{\Delta x} (f(U_i) - f(U_{i-1}))^+ \\ &\quad + \frac{1}{\Delta x} \left( (f(U_{i+1}) - f(U_i))^- - f(U_{i+1}) \right) + (f(U_{i+1}) - f(U_i))^+, \end{aligned}$$

where we introduced  $\tau^+ := \max(\tau, 0)$ ,  $\tau^- := \min(\tau, 0)$  for a scalar function  $\tau$ , i. e.,  $\tau^+ + \tau^- = \tau$ . Here, both re-formulations may be interpreted as different conservative PDS splittings satisfying  $p_{i,j} = 0$  for

189  $j \notin \{i-1, i+1\} \subseteq \{1, \dots, N\}$ ,  $d_{i,j} = p_{j,i}$  for all  $i, j$ , and

$$p_{i,i-1}(\mathbf{U}) = \frac{1}{\Delta x} f(U_{i-1}), \quad d_{i,i+1} = -\frac{1}{\Delta x} f(U_i), \quad p_{i,i+1}(\mathbf{U}) = d_{i,i-1} = 0 \quad \text{or}$$

190

$$\begin{aligned} p_{i,i-1}(\mathbf{U}) &= \frac{1}{\Delta x} \left( -(f(U_i) - f(U_{i-1}))^- + f(U_i) \right), & d_{i,i-1}(\mathbf{U}) &= \frac{1}{\Delta x} (f(U_i) - f(U_{i-1}))^+, \\ d_{i,i+1}(\mathbf{U}) &= -\frac{1}{\Delta x} \left( (f(U_{i+1}) - f(U_i))^- - f(U_{i+1}) \right), & p_{i,i+1}(\mathbf{U}) &= \frac{1}{\Delta x} (f(U_{i+1}) - f(U_i))^+, \end{aligned}$$

191 respectively. It is worth noting that the first splitting corresponds to (9) and possesses differentiable produc-  
 192 tion and destruction terms in this example, while the second splitting does not. Also, by comparing  $p_{i,i+1}$   
 193 it becomes evident that the two splittings differ. In what follows, we require that any splitting of the PDE  
 194 results in a conservative PDS consisting of production and destruction terms which are non-negative on their  
 195 domain.

196 Although, we first introduced one stage methods, also an  $s$ -stage Patankar-type difference scheme based  
 197 on an explicit RK scheme with  $\mathbf{A} = (a_{kr})_{k,r=1,\dots,s} \geq \mathbf{0}$  and weights  $b_1, \dots, b_s \geq 0$  can be devised (using  
 198  $p_{i,j} = d_{j,i}$ ) by

$$\begin{aligned} U_i^{(k)} &= U_i^n - \frac{\Delta t}{\Delta x} \sum_{r=1}^{k-1} a_{kr} \left( d_{i,i+1}(\mathbf{U}^{(r)}) \gamma_i^{(k)} - p_{i,i+1}(\mathbf{U}^{(r)}) \gamma_{i+1}^{(k)} - (d_{i-1,i}(\mathbf{U}^{(r)}) \gamma_{i-1}^{(k)} - p_{i-1,i}(\mathbf{U}^{(r)}) \gamma_i^{(k)}) \right), \\ U_i^{n+1} &= U_i^n - \frac{\Delta t}{\Delta x} \sum_{r=1}^s b_r \left( d_{i,i+1}(\mathbf{U}^{(r)}) \delta_i - p_{i,i+1}(\mathbf{U}^{(r)}) \delta_{i+1} - (d_{i-1,i}(\mathbf{U}^{(r)}) \delta_{i-1} - p_{i-1,i}(\mathbf{U}^{(r)}) \delta_i) \right), \end{aligned} \quad (12)$$

199 for  $k = 1, \dots, s$ , where new NSWs  $\gamma_v^{(k)} = \gamma_v^{(k)}(\hat{\mathbf{U}}^n, t_n, \Delta t)$  are introduced for the stages  $U_i^{(k)}$ . We define the  
 200 new numerical flux based on  $F_{g_{i+\frac{1}{2}}}^\delta(\mathbf{U}^n)$  from (11) as

$$\bar{F}_{g_{i+\frac{1}{2}}}^{n,\delta} := \sum_{r=1}^s b_r F_{g_{i+\frac{1}{2}}}^\delta(\mathbf{U}^{(r)}). \quad (13)$$

201 **Remark 3 (Systems of conservation laws).** This approach can also be extended to systems of conservation  
 202 laws. There, we apply the MP trick on every component of the vector  $\mathbf{u}$  of the initial value problem that  
 203 requires positivity, and we consider the numerical flux for every such component as a scalar function  $g$ . With  
 204 that it becomes evident that this approach can be applied also for different space-discretizations as long as  
 205 they can be written in flux formulation [1].

206 **Remark 4 (Negative RK weights).** The MPRK can be designed also when some of the  $\mathbf{A}$  and  $\mathbf{b}$  coefficients  
 207 are negative, see [23], by simply changing the role of production and destruction for that particular term in  
 208 that stage. This has been applied systematically to design arbitrary high order MP methods as the MPDeC  
 209 [35] that will be used in the numerical section.

210 **Remark 5 (High order discretizations).** The difference schemes framework (2) is very general and can  
 211 encompass various spatial discretizations. This includes also various high order discretizations as Essentially  
 212 Non-Oscillatory (ENO), WENO, FV [12, 11] and finite difference [18] reconstructions schemes. A slightly  
 213 different formulation should be introduced to include also other high order discretizations as continuous and  
 214 DG FE methods [34, 46], where the PDS is either written only on the interfaces of the cells or there is more  
 215 freedom in the choice of the PD terms between the degrees of freedom inside each cell.



#### 4. Lax–Wendroff theorem for MP difference schemes

The aim of this section is to adapt the proof of Theorem 1 to Patankar-type schemes in the form (12). To achieve this, we need to control the nonlinear weights, resulting in the extra assumption that the numerical solution has a bounded total variation also in time (classical Lax–Wendroff requires only bounded total variation in space). So, let us introduce the concept of total time variation (TTV).

**Definition 6 (Total time variation).** The TTV over a domain  $[a, b] \times [0, T]$  of a function  $U \in L^1_{\text{loc}}(\mathbb{R} \times \mathbb{R}^+)$  is

$$\text{TTV}(U(x, [0, T])) := \sup \sum_{n=1}^{n_T} |U(x, t_n) - U(x, t_{n-1})|, \quad (14)$$

where the supremum is taken over all possible subdivisions  $t_0 = 0 < \dots < t_{n_T} = T$  for any  $n_T \in \mathbb{N}$ .

**Remark 6 (Discrete total time variation).** On a domain  $[a, b] \times [0, T]$  for a discrete function  $\{U^n_i\}_{i,n} = \{U(x_i, t_n)\}_{i,n}$  for  $i = 1, \dots, N$  and  $n = 0, \dots, n_T$  with  $t_0 = 0$  and  $t_{n_T} = T$ , the TTV can be equivalently defined using a piece-wise constant reconstruction of the discrete function on the space-time grid, obtaining the formula

$$\text{TTV}(U(x, [0, T])) := \sum_{n=1}^{n_T} |U^n(x) - U^{n-1}(x)|, \quad (15)$$

with  $U^n(x)$  being the piece-wise constant reconstruction in space of  $\{U^n_i\}_i$ .

It should be noticed that this definition differs from the definition of total of time variation (TOTV) given in [45], and it is the equivalent of the TV definition in space, but switching the roles of time and space.

**Remark 7 (TTV uniform boundedness).** The TTV uniform boundedness, i.e., having a bound for the TTV of an approximate solution while the mesh is refined, will be used as a requirement in the next sections in many theorems. As for the TV boundedness in space, it is related to the non-oscillatory behavior of the numerical solution in time and, hence, their stability. Moreover, we believe that there is a link between the TV boundedness and the TTV boundedness for hyperbolic problems, though it is not straightforward to prove it. During numerical results, we have seen that TTV boundedness seems naturally achieved (with  $\text{TTV}(U^{\text{num}}(x))$  converging to the exact  $\text{TTV}(U^{\text{ex}}(x))$ ) when a stable CFL number is used. On the contrary, when oscillations arise in the space discretization, they also propagate in time, leading to not converging  $\text{TTV}(U^{\text{num}}(x))$ .

**Lemma 2 (MPE weights convergence).** Let  $U^{(l)}(x_i, t_n)$  be the approximation of problem (1) at the space  $x_i^{(l)}$  and time  $t_n^{(l)}$  for a mesh discretization  $\Delta x_{(l)} \sim \Delta t_{(l)} \rightarrow 0$  for  $l \rightarrow \infty$ . For every  $l$ , consider  $U^{(l)}(x, t)$  as a piece-wise constant reconstruction of the numerical solution  $U^{(l)}(x_i, t_n)$ . Let  $\text{TTV}(U^{(l)}(x, [0, T]))$  be uniformly bounded for all  $x \in [a, b]$  and for all  $l \in \mathbb{N}$ . Suppose that  $U^{(l)} \rightarrow u$  in  $L^1_{\text{loc}}([a, b] \times [0, T])$  with  $u$  positive. Defining  $\delta^{(l)}(x, t)$  as the piece-wise constant reconstruction of the MPE weights  $\delta^{(l)}(x_i, t_n) = U^{(l)}(x_i, t_{n+1})/U^{(l)}(x_i, t_n)$  for all  $i$  and  $n$ . Then,  $\delta^{(l)}(x, t) \rightarrow 1$  almost everywhere (a.e.) in  $[a, b] \times [0, T]$ .

*Proof.* Consider the MPE weights

$$\delta^{(l)}(x, t) := \frac{U^{(l)}(x, t + \Delta t_{(l)})}{U^{(l)}(x, t)} = 1 + \frac{U^{(l)}(x, t + \Delta t_{(l)}) - U^{(l)}(x, t)}{U^{(l)}(x, t)}.$$

Now, we can say that  $\lim_{l \rightarrow \infty} U^{(l)}(x, t + \Delta t_{(l)}) - U^{(l)}(x, t) = 0$  a. e. because the numerical solutions are TTV bounded uniformly for all  $l$  and  $x$  and the denominator,  $U^{(l)}$  is converging to a weak solution  $u$  in  $L^1_{\text{loc}}$ .  $\square$

The arguments of this proof are the same used in space for the Lax–Wendroff theorem in the proof of [30]. These follow from some classical analysis results using the compactness of bounded total variation functions in  $L^1_{loc}$ , the fact that in 1D functions of bounded variation have at most countably many discontinuities and the Helly’s selection theorem to deduce that bounded sequences of 1D bounded total variation functions are precompact for pointwise convergence almost everywhere [2].

**Theorem 3** (Lax–Wendroff for MPE difference schemes). *Consider a sequence of step sizes  $\Delta x_{[l]} > 0$  with  $\lim_{l \rightarrow \infty} \Delta x_{[l]} = 0$ . Let  $(U^{[l]})_l$  be a sequence of piece-wise constant functions constructed of numerical solutions obtained by a MPE difference scheme (10) consistent with (1) with time step size  $\Delta t_{[l]} = \lambda \Delta x_{[l]}$ ,  $\lambda \equiv \text{const}$ . If for any domain  $[a, b] \times [0, T]$  the total variation  $TV(U^{[l]}([a, b], t))$  from (15) is uniformly bounded for all  $l$  and  $t \in [0, T]$ , the total time variation  $TTV(U^{[l]}(x, [0, T]))$  is uniformly bounded for all  $l$  and  $x \in [a, b]$  and if  $U^{[l]} \xrightarrow{l \rightarrow \infty} u$  in  $L^1_{loc}(\mathbb{R} \times [0, T])$ , then  $u$  is a weak solution of the conservation law (1).*

*Proof.* We follow the original Lax–Wendroff proof from [29, 30] and adapt it for our case where needed. The main goal is to show that the limit function  $u(x, t)$  of the considered method satisfies

$$\int_0^\infty \int_{-\infty}^\infty [\varphi_t u + \varphi_x f(u)] dx dt = - \int_{-\infty}^\infty \varphi(x, 0) u(x, 0) dx. \quad (16)$$

for all  $\varphi \in C^1_0([a, b] \times [0, T])$ . In what follows, we omit to write the dependency on the refinement represented by the index  $l$  if it is clear from the context.

Multiplying the numerical method (10) by any test function  $\varphi \in C^1_0$  and summing over all  $i$  and  $n \geq 0$  yields

$$\sum_{n=0}^\infty \sum_{i=-\infty}^\infty \varphi(x_i, t_n) (U_i^{n+1} - U_i^n) = \frac{\Delta t}{\Delta x} \sum_{n=0}^\infty \sum_{i=-\infty}^\infty \varphi(x_i, t_n) \left[ F_{g_{i+\frac{1}{2}}}^\delta(U^n) - F_{g_{i-\frac{1}{2}}}^\delta(U^n) \right].$$

We note that the value of, e. g.,  $U_i^n$  does not have an effect for  $(x_i, t_n) \notin \text{supp}(\varphi) \subseteq [a, b] \times [0, T]$ . Using summation by parts, the fact  $\varphi$  has a compact support, multiplying by  $\Delta x$  and rearranging leads to

$$\begin{aligned} & \Delta x \Delta t \left[ \sum_{n=1}^\infty \sum_{i=-\infty}^\infty \left( \frac{\varphi(x_i, t_n) - \varphi(x_i, t_{n-1})}{\Delta t} \right) U_i^n + \sum_{n=0}^\infty \sum_{i=-\infty}^\infty \left( \frac{\varphi(x_{i+1}, t_n) - \varphi(x_i, t_n)}{\Delta x} \right) F_{g_{i+\frac{1}{2}}}^\delta(U^n) \right] \\ &= - \Delta x \sum_{i=-\infty}^\infty \varphi(x_i, 0) U_i^0. \end{aligned} \quad (17)$$

Now we show that this equation coincides with the desired assertion (16) as we pass to the limit  $\lim_{l \rightarrow \infty} \Delta x_{[l]} = 0$  with  $\Delta t_{[l]} = \lambda \Delta x_{[l]}$ ,  $\lambda \equiv \text{const}$ . For the first and the last term of the previous equation, the classical proof of Lax and Wendroff should be applied, as outlined in [30]. This means that as  $l \rightarrow \infty$ , the assumed  $L^1_{loc}$  convergence of  $U^{[l]}$  to  $u$ , together with the smoothness of  $\varphi$ , ensures the convergence of the first summand to  $\int_0^\infty \int_{-\infty}^\infty \varphi_t(x, t) u(x, t) dx$ . In addition, the choice of the initial data  $U_i^0 = \psi_i$  enables us to conclude the convergence of the RHS to  $-\int_{-\infty}^\infty \varphi(x, 0) u(x, 0) dx$ . The remaining task is to ascertain the limit for the middle term. In particular, we are interested in showing the convergence of the new numerical flux to the physical flux and therefore in estimating

$$\left| F_g^{\delta, [l]}(\mathbf{x}_i, t_n) - f(U^{[l]}(x_i, t_n)) \right| \quad (18)$$

for  $l \rightarrow \infty$  where  $\mathbf{x}_i = (x_i - p\Delta x, \dots, x_i + q\Delta x)$ . Here,  $F_g^{\delta, [l]}(\mathbf{x}_i, t_n)$  is a piece-wise constant reconstruction of the values  $F_{g_{i+\frac{1}{2}}}^\delta(U^n)$  given in (11) by substituting

$$g_{i+\frac{1}{2}}^{n, [l]} = g(U^{[l]}(\mathbf{x}_i, t_n)) = g(U^{[l]}(x_i - p\Delta x, t_n), \dots, U^{[l]}(x_i + q\Delta x, t_n))$$

for  $i = 1, \dots, N_l$  and  $n = 0, \dots, n_{T,l}$  for  $g_{i+\frac{1}{2}}^n$ . Adding and subtracting

$$g(U^{(l)}(\mathbf{x}_i, t_n)) = \max\{g(U^{(l)}(\mathbf{x}_i, t_n)), 0\} + \min\{g(U^{(l)}(\mathbf{x}_i, t_n)), 0\}$$

to (18) leads to

$$\begin{aligned} |F_{g_{i+\frac{1}{2}}}^\delta(U^n) - f(U^{(l)}(x_i, t_n))| &= \left| \max\{g(U^{(l)}(\mathbf{x}_i, t_n)), 0\} \delta_i^{(l)} + \min\{g(U^{(l)}(\mathbf{x}_i, t_n)), 0\} \delta_{i+1}^{(l)} - f(U^{(l)}(x_i, t_n)) \right| \\ &\leq \left| \max\{g(U^{(l)}(\mathbf{x}_i, t_n)), 0\} \delta_i^{(l)} - \max\{g(U^{(l)}(\mathbf{x}_i, t_n)), 0\} \right| + \left| \min\{g(U^{(l)}(\mathbf{x}_i, t_n)), 0\} \delta_{i+1}^{(l)} - \min\{g(U^{(l)}(\mathbf{x}_i, t_n)), 0\} \right| \\ &\quad + \left| \max\{g(U^{(l)}(\mathbf{x}_i, t_n)), 0\} + \min\{g(U^{(l)}(\mathbf{x}_i, t_n)), 0\} - f(U^{(l)}(x_i, t_n)) \right| \\ &= \left| \max\{g(U^{(l)}(\mathbf{x}_i, t_n)), 0\} \delta_i^{(l)} - \max\{g(U^{(l)}(\mathbf{x}_i, t_n)), 0\} \right| + \left| \min\{g(U^{(l)}(\mathbf{x}_i, t_n)), 0\} \delta_{i+1}^{(l)} - \min\{g(U^{(l)}(\mathbf{x}_i, t_n)), 0\} \right| \\ &\quad + \left| g(U^{(l)}(\mathbf{x}_i, t_n)) - f(U^{(l)}(x_i, t_n)) \right|, \end{aligned}$$

where the first two addends vanish a.e. as  $l \rightarrow \infty$  due to the convergence of the MPE weights based on the TTV boundedness property, as stated in Lemma 2.

Finally, as only the original numerical flux  $g$  appears in the remaining term  $|g(U^{(l)}(\mathbf{x}_i, t_n)) - f(U^{(l)}(x_i, t_n))|$ , we can follow the reasoning of the Lax–Wendroff theorem of [30]. Consistency of  $g$  is provided, i.e., we have the estimate

$$\begin{aligned} |g(U^{(l)}(\mathbf{x}_i, t_n)) - f(U^{(l)}(x_i, t_n))| &= \\ |g(U^{(l)}(x_i - p\Delta x, t_n), \dots, U^{(l)}(x_i + q\Delta x, t_n)) - f(U^{(l)}(x_i, t_n))| &\leq K \max_{-p \leq j \leq q} |U^{(l)}(x_i + j\Delta x, t_n) - U^{(l)}(x_i, t_n)| \end{aligned}$$

with a Lipschitz constant  $K$ , where the total variation boundedness (TVB) property of  $U^{(l)}(\cdot, t)$  leads to

$$\max_{-p \leq j \leq q} |U^{(l)}(x + j\Delta x, t) - U^{(l)}(x, t)| \rightarrow 0$$

as  $l \rightarrow \infty$  for almost every value of  $x$ . Altogether we proved the a.e. convergence of the numerical flux function  $F_g^{\delta, (l)}(\mathbf{x}_i, t_n)$  to the physical flux function  $f(U^{(l)}(x_i, t_n))$ . This gives, together with some standard estimates, the convergence of (17) to the weak form (16) and shows, since it is true for any test function  $\varphi \in C_0^1$ , that  $u$  is indeed a weak solution.

□

#### 4.1. Extension to Patankar-type RK

In this section, we draw the steps to extend the previous proof to the multi-stage Patankar-type schemes.

**Definition 7 (Total time variation for Runge–Kutta).** The total time variation of a numerical solution  $\{U_i^n\}_{i,n} = \{U(x_i, t_n)\}_{i,n}$  obtained by an  $s$ -stage RK method (TTVRK) over a domain  $[a, b] \times [0, T]$  for  $i = 1, \dots, N$  and  $n = 0, \dots, n_T$  with  $x_1 = a$ ,  $x_N = b$ ,  $t_0 = 0$  and  $t_{n_T} = T$  is

$$\text{TTVRK}(U(x, [0, T])) := \sum_{n=0}^{n_T} \sum_{k=1}^s |U^{(k)}(x, t_n) - U^{(k-1)}(x, t_n)| + \sum_{n=1}^{n_T} |U(x, t_n) - U^{(s)}(x, t_{n-1})|, \quad (19)$$

where  $U^{(0)}(x, t_n) := U(x, t_n)$  and  $U^{(k)}(x, t_n)$  denotes the piece-wise constant reconstruction in space of the  $k$ -th stage computed from the numerical solution  $\mathbf{U}^n$  at time level  $n$ .

297 **Lemma 4** (Extended Patankar weights convergence). Let  $U^{(l)}(x_i, t_n)$  be the approximation of problem (1)  
 298 at the space  $x_i^{(l)}$  and time  $t_n^{(l)}$  for a mesh discretization  $\Delta x_{(l)} \sim \Delta t_{(l)} \rightarrow 0$  for  $l \rightarrow \infty$ . For every  $l$ , consider  
 299  $U^{(l)}(x, t)$  a piece-wise constant reconstruction of the numerical solution  $U^{(l)}(x_i, t_n)$ . Let  $\text{TTVRK}(U^{(l)}(x, [0, T]))$   
 300 be uniformly bounded for all  $x \in [a, b]$  and for all  $l \in \mathbb{N}$ . Suppose that  $U^{(l)} \rightarrow u$  in  $L^1_{\text{loc}}([a, b] \times [0, T])$ . Defin-  
 301 ing  $\delta^{(l)}(x, t)$  as the piece-wise constant reconstruction of  $\delta^{(l)}(x_i, t_n) = U^{(l)}(x_i, t_{n+1})/\sigma^{(l)}(x_i, t_n)$  and  $\gamma^{(k), (l)}(x, t)$   
 302 the one of  $\gamma^{(k), (l)}(x_i, t_n^k) = U^{(k), (l)}(x_i, t_n^k)/\pi^{(k), (l)}(x_i, t_n^k)$  for all  $i$  and  $n$ . Then,  $\delta^{(l)}(x, t)$  and  $\gamma^{(k), (l)}(x, t) \rightarrow 1$  almost  
 303 everywhere in  $[a, b] \times [0, T]$  if  $\sigma^{(l)}(x, t)$ , respectively,  $\pi^{(k), (l)}(x, t) \rightarrow u(x, t)$  as  $l \rightarrow \infty$ .

*Proof.* First we investigate the weights of the update

$$\delta^{(l)}(x, t) := \frac{U^{(l)}(x, t + \Delta t_{(l)})}{\sigma^{(l)}(x, t)} = 1 + \frac{U^{(l)}(x, t + \Delta t_{(l)}) - \sigma^{(l)}(x, t)}{\sigma^{(l)}(x, t)}.$$

Now, we can follow the same reasoning as in the proof of Lemma 2 under usage of the uniformly bounded  
 TTVRK and the convergence of the numerical solutions to a weak solution  $u$  in  $L^1_{\text{loc}}$ , to show  $\lim_{l \rightarrow \infty} U^{(l)}(x, t + \Delta t_{(l)}) = u(x, t)$  a.e.. Furthermore we assumed that  $\sigma^{(l)}(x, t) \rightarrow u(x, t)$  which finally leads to  $\lim_{l \rightarrow \infty} U^{(l)}(x, t + \Delta t_{(l)}) - \sigma^{(l)}(x, t) = 0$  and thus to  $\delta^{(l)}(x, t) \rightarrow 1$ . We proceed considering the weights of the stages

$$\gamma^{(k), (l)}(x, t) = \frac{U^{(k), (l)}(x, t)}{\pi^{(k), (l)}(x, t)} = 1 + \frac{U^{(k), (l)}(x, t) - \pi^{(k), (l)}(x, t)}{\pi^{(k), (l)}(x, t)}.$$

304 Once again, we use the fact that TTVRK is uniformly bounded, together with the convergence to a weak  
 305 solution  $u$  for the limit  $\lim_{l \rightarrow \infty} U^{(k), (l)}(x, t) = u(x, t)$ . Additionally, we assumed  $\pi^{(k), (l)}(x, t) \rightarrow u(x, t)$  as  
 306  $l \rightarrow \infty$ , so this leads to  $\lim_{l \rightarrow \infty} U^{(k), (l)}(x, t) - \pi^{(k), (l)}(x, t) = 0$  and thus to the desired result.  $\square$

307 In the previous Lemma 4, some convergence for the PWDs was assumed, that is  $\sigma^{(l)}(x, t)$  respectively  
 308  $\pi^{(k), (l)}(x, t) \rightarrow u(x, t)$  almost everywhere for  $l \rightarrow \infty$  and  $(x, t) \in [a, b] \times [0, T]$ .

**Theorem 5** (Lax–Wendroff for MPRK schemes). Assume the PWDs of an MPRK method continuously  
 depend on the  $i$ -th component of the stages, i.e.

$$\pi_i^{(k)} = \pi_i^{(k)}(U_i^n, U_i^{(1)}, \dots, U_i^{(k-1)}) \quad \text{and} \quad \sigma_i = \sigma_i(U_i^n, U_i^{(1)}, \dots, U_i^{(s)}).$$

Furthermore we assume that the PWDs are consistent in the sense that

$$\pi_i^{(k)}(U_i^n, U_i^n, \dots, U_i^n) = U_i^n \quad \text{and} \quad \sigma_i(U_i^n, U_i^n, \dots, U_i^n) = U_i^n.$$

309 Then, the PWDs fulfill the extended Patankar weights convergence of Lemma 4.

310 *Proof.* Recall that  $U^{(k), (l)} \rightarrow u$  for  $l \rightarrow \infty$  by means of the uniform boundedness of the TTVRK together  
 311 with  $U^{(l)} \rightarrow u$  in  $L^1_{\text{loc}}([a, b] \times [0, T])$ , as mentioned in the proof of Lemma 4. Hence, the arguments of the  
 312 PDWs tend to  $u$ , and thus, the continuity and consistency of the PWDs yield  $\sigma^{(l)}(x, t)$ ,  $\pi^{(k), (l)}(x, t) \rightarrow u(x, t)$   
 313 almost everywhere for  $l \rightarrow \infty$  and  $(x, t) \in [a, b] \times [0, T]$ .  $\square$

314 **Remark 8.** The assumptions made in the previous theorem are exactly those of [20] and [4] which cover  
 315 common methods such as the ones presented in [23, 24, 35]. Altogether Theorem 5 shows that the PWDs  
 316 from the literature already satisfy the assumption from Lemma 4.

317 In the following example, we illustrate that the assumption of Theorem 5 are fulfilled for a second-order  
 318 family of MP methods.

**Example 2** (Verification of PWD Hypothesis for MPRK22( $\alpha$ )). Consider the two stage second-order family of MPRK methods introduced in [23] and denoted by MPRK22( $\alpha$ ), where  $\alpha \geq \frac{1}{2}$  is a free parameter and the PWDs are given by  $\pi_i^{(2)} = U_i^n$  and  $\sigma_i = (U_i^n)^{1-\frac{1}{\alpha}} (U_i^{(2)})^{\frac{1}{\alpha}}$ .

For  $l \rightarrow \infty$  we find  $\gamma^{(2),\{l\}}(x, t) = \frac{U^{(2),\{l\}}(x, t)}{U^{n,\{l\}}(x, t)} \rightarrow 1$  and thus  $\sigma_i^{\{l\}} \rightarrow (U_i^n)^{1-\frac{1}{\alpha}} (U_i^n)^{\frac{1}{\alpha}} = U_i^n$  i.e.  $\delta^{\{l\}} \rightarrow 1$  almost everywhere.

**Remark 9 (Lax–Wendroff proof: extension to Patankar-type RK sketch).** The extension to RK methods with more stages differs only in the investigation of equation (18), where we have, instead of the MPE weighted numerical flux (11), the generalized version (13), i.e.,

$$\left| \bar{F}_{g_{i+\frac{1}{2}}}^{n,\delta,\{l\}} - f(U^{\{l\}}(x_i, t_n)) \right|$$

which, using the definition of  $\bar{F}_{g_{i+\frac{1}{2}}}^{n,\delta,\{l\}}$  and  $\sum_{r=1}^s b_r = 1$ , reads

$$\left| \sum_{r=1}^s b_r F_g^{\delta,\{l\}}(U^{(r),\{l\}}(\mathbf{x}_i, t_n^{(r)})) - f(U^{\{l\}}(x_i, t_n)) \right| \leq \sum_{r=1}^s |b_r| \left| F_g^{\delta,\{l\}}(U^{(r),\{l\}}(\mathbf{x}_i, t_n^{(r)})) - f(U^{\{l\}}(x_i, t_n)) \right|.$$

Thus, defining again  $\mathbf{x}_i = (x_i - p\Delta x, \dots, x_i + q\Delta x)$ , we take an arbitrary summand  $r$  and decompose it analogously to the proof of Theorem 3, which leads to

$$\begin{aligned} \left| F_g^{\delta,\{l\}}(U^{(r),\{l\}}(\mathbf{x}_i, t_n^{(r)})) - f(U^{\{l\}}(x_i, t_n)) \right| &\leq \left| \max\{g(U^{(r),\{l\}}(\mathbf{x}_i, t_n^{(r)})), 0\} \delta_i^{\{l\}} - \max\{g(U^{(r),\{l\}}(\mathbf{x}_i, t_n^{(r)})), 0\} \right| \\ &\quad + \left| \min\{g(U^{(r),\{l\}}(\mathbf{x}_i, t_n^{(r)})), 0\} \delta_{i+1}^{\{l\}} - \min\{g(U^{(r),\{l\}}(\mathbf{x}_i, t_n^{(r)})), 0\} \right| \\ &\quad + \left| g(U^{(r),\{l\}}(\mathbf{x}_i, t_n^{(r)})) - f(U^{\{l\}}(x_i, t_n)) \right|. \end{aligned}$$

The first two lines vanish reasoned by  $\delta_i^{\{l\}} \rightarrow 1$  for  $l \rightarrow \infty$ , i.e., we need to prove that in the limit  $l \rightarrow \infty$  the last summand

$$\begin{aligned} &\left| g(U^{(r),\{l\}}(\mathbf{x}_i, t_n^{(r)})) - f(U^{\{l\}}(x_i, t_n)) \right| \\ &= \left| g(U^{(r),\{l\}}(x_i - q\Delta x, t_n^{(r)}), \dots, U^{(r),\{l\}}(x_i + p\Delta x, t_n^{(r)})) \right| \leq L \max_{-p \leq j \leq q} \left| U^{(r),\{l\}}(x_i + j\Delta x, t_n^{(r)}) - U^{\{l\}}(x_i, t_n) \right| \rightarrow 0, \end{aligned}$$

for a.e.  $x_i$ , where, independently of the argument, we can apply the Lipschitz continuity of  $g$  with Lipschitz constant  $L$  to get the estimation.

Note in this case the RK stages are arguments of the numerical flux. Therefore, we first need the TTVRK boundedness property to apply Lemma 4 and ensure that the stages converge to the same solution  $\lim_{l \rightarrow \infty} U^{(r),\{l\}}(x_i + j\Delta x, t_n^{(r)}) - U^{\{l\}}(x_i + j\Delta x, t_n) = 0$  and proceed with the proof of the original Lax–Wendroff theorem to obtain the desired result.

**Remark 10 (Lax–Wendroff for NSARK).** In the case of MP methods we assumed in Lemma 4 and Theorem 5 that the PWDs fulfill certain properties to conclude that the NSW grid functions  $\delta^{\{l\}}, \gamma^{(k),\{l\}}$  satisfy

$$\lim_{l \rightarrow \infty} \delta^{\{l\}}(x, t) = 1 \quad \text{and} \quad \lim_{l \rightarrow \infty} \gamma^{(k),\{l\}}(x, t) = 1 \quad (20)$$

almost everywhere. This property represents the main ingredient in the proof outlined in Remark 9. Indeed, to extend the Lax–Wendroff theorem also to general NSARK schemes with general solution-dependent coefficients of the form (8), it is sufficient to add the condition (20) to the classical assumptions of the Lax–Wendroff theorem. In the case of MPRK schemes, we used the TTVRK boundedness together with special properties of the PWDs (see Theorem 5) to deduce (20) in Lemma 4.

**Remark 11 (On the TTV boundedness assumption).** The difference with respect to the classical Lax–Wendroff theorem consists of the assumption that the numerical solutions are TTV bounded uniformly in space and in the limit process. The authors could not easily remove this assumption to prove the Lax–Wendroff theorem. Different strategies have been pursued but all of them ended up requiring some extra hypotheses. In particular, the technique proposed by Birken and Linders [8] shows that, using iterative methods for implicit schemes, the Lax–Wendroff theorem can be applied also in the context of Patankar-type methods. Nevertheless, in their work the case where both the limit of iterations of the linear solver and the mesh refinement level go to infinity is not treated. In particular, it is hard to prove the  $\lim_{l \rightarrow \infty} \lim_{z \rightarrow \infty}$  case, where  $l$  is the mesh refinement level and  $z$  is the linear solver iteration. Indeed, this leads to a dependency of the composed numerical flux in one step of the RK on the whole domain. This destroys the argument used in the classical Lax–Wendroff theorem to prove the flux convergence almost everywhere.

The hypotheses of being TVB in space and time for a given spatial and temporal discretization is not trivial to study, as one needs to find a bound uniformly with the mesh refinement, which in a discrete setting is virtually impossible to check. What one can do in some cases is to prove a stronger result, i.e., the total variation diminishing (TVD) (in space) property. In Appendix A, we study the TVD property of the MPE scheme applied to an upwind first order FV spatial discretization for the nonviscous Burgers’ equation. We also propose some numerical tests to validate the findings.

## 5. Numerical results

In this section, we will run some simple Riemann problem tests to validate the convergence towards exact solutions. What we are looking for is to test the regimes where the MP weights are large and change significantly the scheme with respect to explicit ones. Clearly, this happens close to discontinuity and close to 0 values of the solution, while everywhere else the MP weights quickly converges to 1. Hence, we choose to test Riemann problems with very low value of one of the two sides.

We will start from first order methods with the MPE time discretization and then we will proceed with higher order discretizations.

### 5.1. First order tests

#### 5.1.1. Burgers’ equation

We start the discussion with Burgers’ equation

$$\partial_t u + \partial_x \left( \frac{u^2}{2} \right) = 0, \quad x \in [-1, 1], t \in [0, T], \quad (21)$$

with a double Riemann problem for initial conditions, i.e.,

$$u(x, 0) = \begin{cases} u_1, & \text{if } -0.5 < x < 0.5, \\ u_2, & \text{else,} \end{cases} \quad (22)$$

with both  $u_1, u_2 > 0$  and periodic boundary conditions. The exact solution of this problem if  $u_1 > u_2$  is a discontinuity traveling at speed  $c = \frac{u_1^2 - u_2^2}{2(u_1 - u_2)}$  on the right and a rarefaction wave on the left. For the spatial discretization, we use a simple upwind numerical flux that, given the positivity of the solution, will sum up to  $g_{i+\frac{1}{2}}(U^n) = U_i^n$ . In Appendix A, we show that MPE with the upwind spatial discretization is TVD for an implicit CFL bound:  $\text{CFL} < 2$ .

To check the convergence of the scheme, we look at a very strong Riemann problem where  $u_1 = 10^4$  and  $u_2 = 10^{-30}$ , so that we expect the MP weights to be particularly large. We run some simulations with

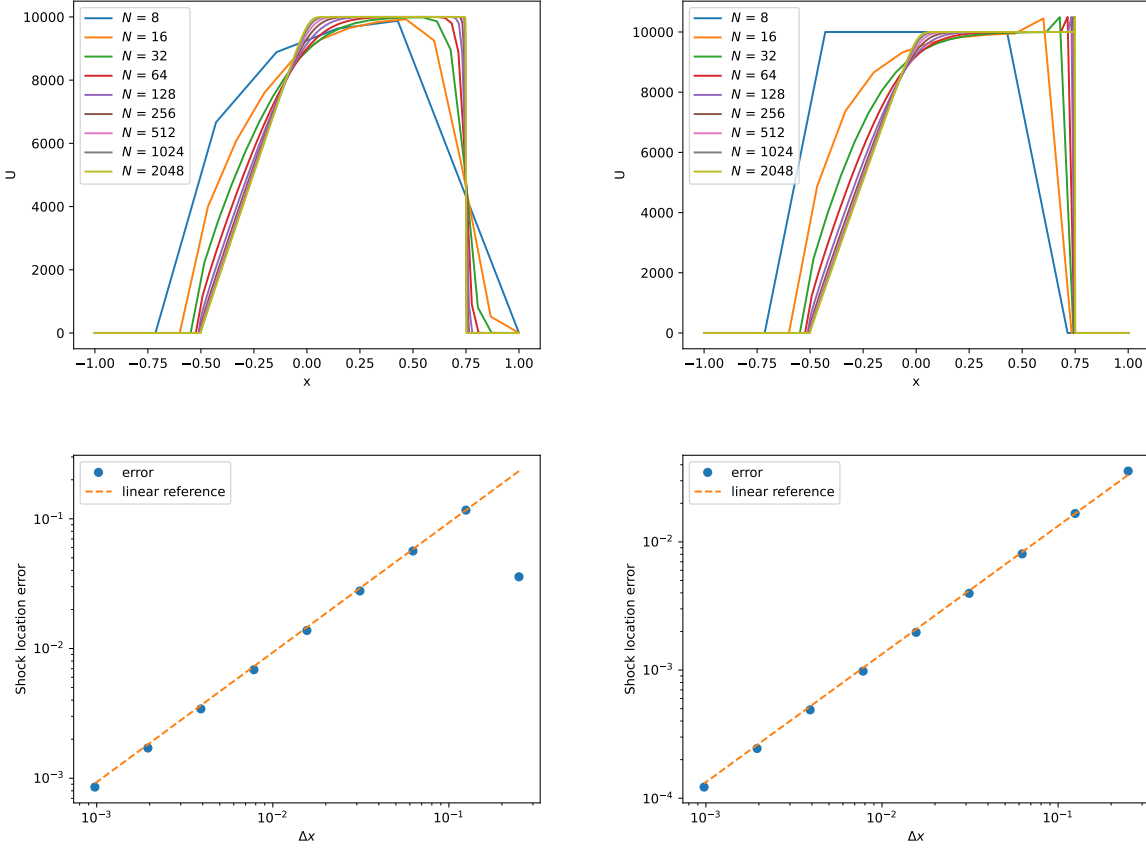


Figure 1: Simulations (top) and numerical shock location errors (bottom) for Burgers' equation (21) with upwind numerical flux and MPE. CFL=1 on the left and CFL=2.1 on the right

different number of cells  $N$  at CFL = 1 and CFL = 2.1. For each of these simulations we look for the interface with largest derivative and we set this point to be the “numerical shock location”  $s_N$ . We measure then  $|s_N - s|$ , where  $s$  is the exact shock location. For both CFLs (so, also in a non TVD regime), we obtain that the numerical shock locations converge linearly to the exact shock location.

In Figure 1, we show the simulations and the error of the shock locations, proving the convergence of the numerical solutions to the exact solution of the problem.

### 5.1.2. Buckley–Leverett equation

Then, we test the Buckley–Leverett equation that approximates a phase concentration in multiphase flows, i.e.,

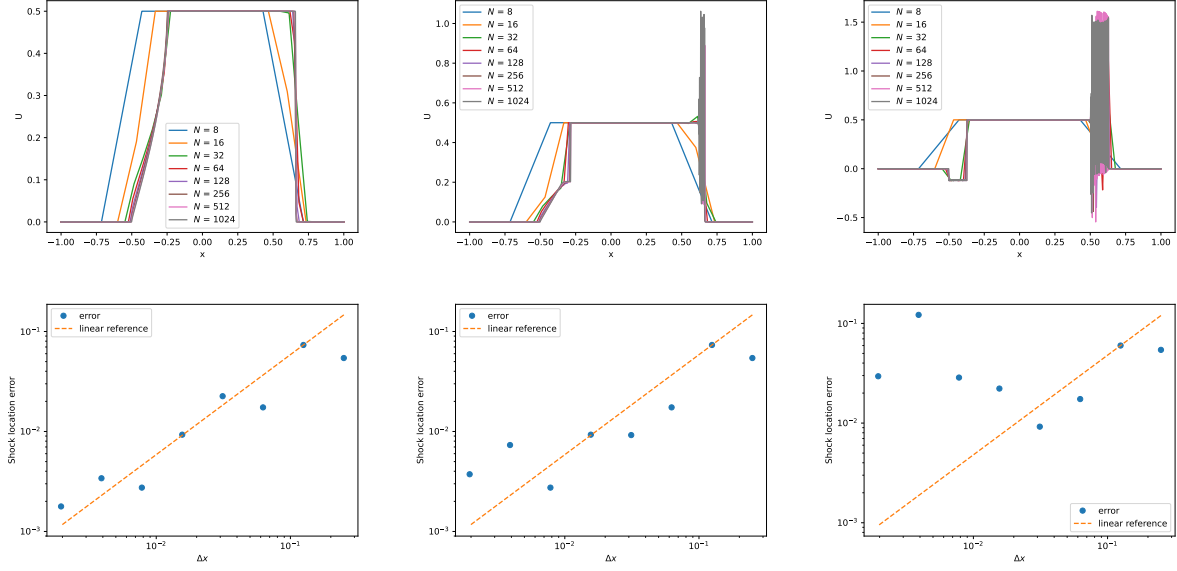
$$\partial_t u + \partial_x f(u) = 0, \quad \text{with } f(u) = \left( \frac{u^2}{u^2 + a(1 - u^2)} \right) \quad (23)$$

with the IC given by the Riemann problems (22), with  $u_1 = 0.5$  and  $u_2 = 10^{-30}$  and  $a = 0.5$ . Again, the right discontinuity travels with speed  $c = \frac{f(u_1) - f(u_2)}{u_1 - u_2}$ , while a rarefaction fan spreads from the left discontinuity.

In this case, we could prove that the scheme is TVD and, from the numerical simulations, we guess that the MPE is TVD for CFL  $\leq 1$ , as well as the explicit Euler method.

In Figure 2, we show the simulations of explicit Euler and MPE at various  $N$  for CFL = 0.99 (left), CFL = 1.2 (center) and CFL = 1.99 (right). Below each figure, we also plot the error of the numerical shock

## Explicit Euler



## MPE

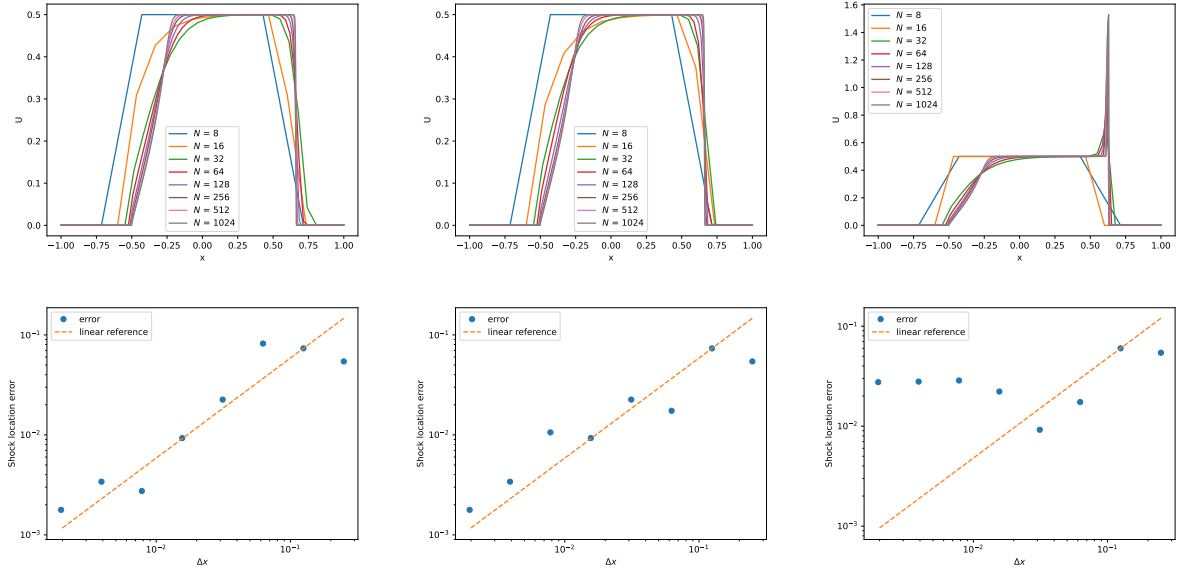


Figure 2: Simulations for explicit Euler (first line), MPE (third line) and numerical shock location errors for explicit Euler (second line) and MPE (last line) for Buckley–Leverett equation (23) with upwind numerical flux and MPE. CFL = 0.99 (left), CFL = 1.20 (center) and CFL = 1.99 (right)



location. MPE clearly has some qualitative improvements with respect to explicit Euler, in particular for CFL = 1.2 where we are outside the stability region for the explicit method, but we can still obtain reliable simulations for MPE. In the explicit method, we observe not only oscillations close to the shock, but also a wrong weak solution at the rarefaction wave (not the entropy solution), while the MPE gives correct results at both waves. Numerically, we have seen that MPE seems to be TVD up to CFL = 1, while explicit Euler is not TVD for CFL larger than 1. In any case, in the regime where explicit Euler converges to the exact solution, we also have that the MPE converges to the exact solution.

## 5.2. High order tests

In this section, we test with higher order schemes the findings of this work. We will first have a high order in time and upwind first order space reconstruction and then we will move to high order in space and time using a WENO spatial reconstruction.

### 5.2.1. High order in time

To test the high order in time Patankar-type schemes, we will use the MPDeC with different order of accuracy, while still using an upwind first order discretization. In Figure 3, we show the simulations for the same double Riemann problem studied in the previous section. As for the MPE, we notice a good convergence towards the exact solution, in particular in the shock location at an  $O(\Delta x)$  rate. For CFL = 2.1 there is still convergence, even if the shock location does not converge monotonically to the exact one. It seems that there are two first-order error lines on which the error bounces.

Moreover, numerically, we can observe that also the second order MPDeC(2) for Burgers with upwind seems TVD up to CFL 2, while higher order MPDeC are never TVD even for very small CFLs.

In figure (4), we check the TTV boundedness for different time discretization methods at different CFL number (0.99, 1.99, 2.99, 10) for a double RP (22) with  $u_1 = 2$  and  $u_2 = 10^{-13}$ . The TTV boundedness is the hypothesis that we added to the Lax–Wendroff theorem in order to be proven for Patankar-type methods. Here, we check whether this hypothesis is reasonable for such methods. In particular, we observe that the MP methods seem to have bounded TTV also for very large CFLs, even if the bound is larger than the exact solution one (which is bounded by 2). This does not prove that the method is uniformly bounded for large CFLs, but it gives good hopes, while for low CFLs it seems that the TTV is bounded by the exact TTV. On the other side, the explicit Euler method is TTV only for CFL  $\leq 1$ , while for larger CFLs has an exploding behavior.

### 5.2.2. High order in space and time with WENO-FV

Finally, in this section we propose a space-time high order discretization which involves a FV WENO spatial reconstruction [42] in each cell, with a positivity limiter [38] to guarantee the positivity in each reconstruction point. Typically these limiters are set to work with the explicit Euler method and SSPRK methods, where the CFL has to be adjusted according to the weights of the underlying quadrature, e.g. for order 5 one needs a CFL of  $\frac{1}{12}$ . In the context of Patankar-type schemes, it has been shown that this restriction of the CFL is not required [12].

Here, we use the discretization proposed in [12] where the FV-WENO discretization order matches the time integration order of accuracy. As time discretization, we use various Patankar-type methods, i.e. MPDeC [35] of various order and the optimal third order MP strong-stability-preserving Runge–Kutta method [19] (MPSSPRK3).

Also here, we propose two tests for different conservation laws with discontinuities with a state close to 0, for Burgers' equations and shallow water (SW) equations. Differently from the previous test cases, in this code a minimum quantity for the positive variable is set at  $10^{-14}$  to avoid blow-ups in the velocity computation (in SW).

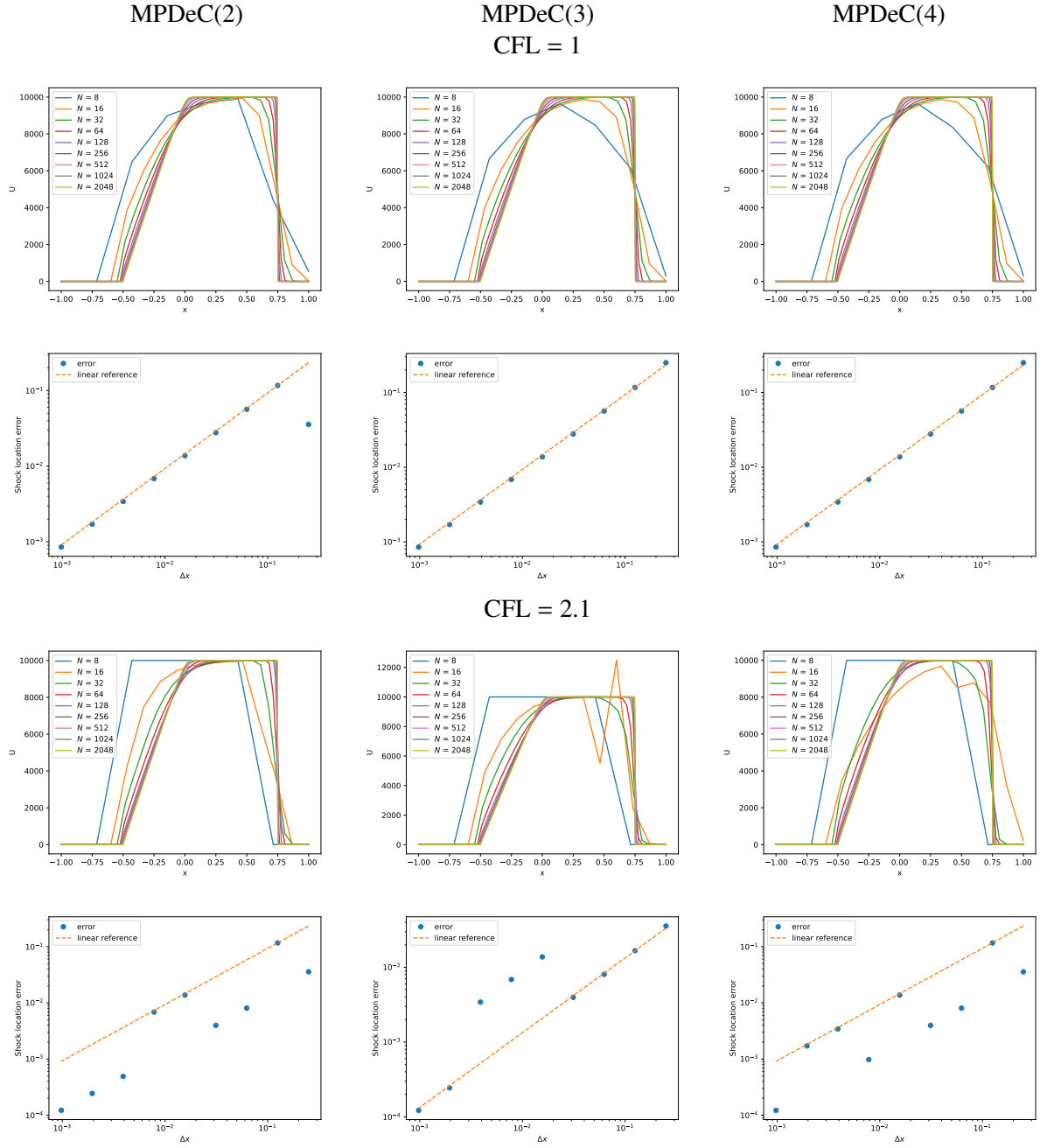


Figure 3: Burgers' equation (21) simulations and error of the shock location with MPDeC of order 2 (left), 3 (center) and 4 (right) and upwind numerical flux with different CFLs

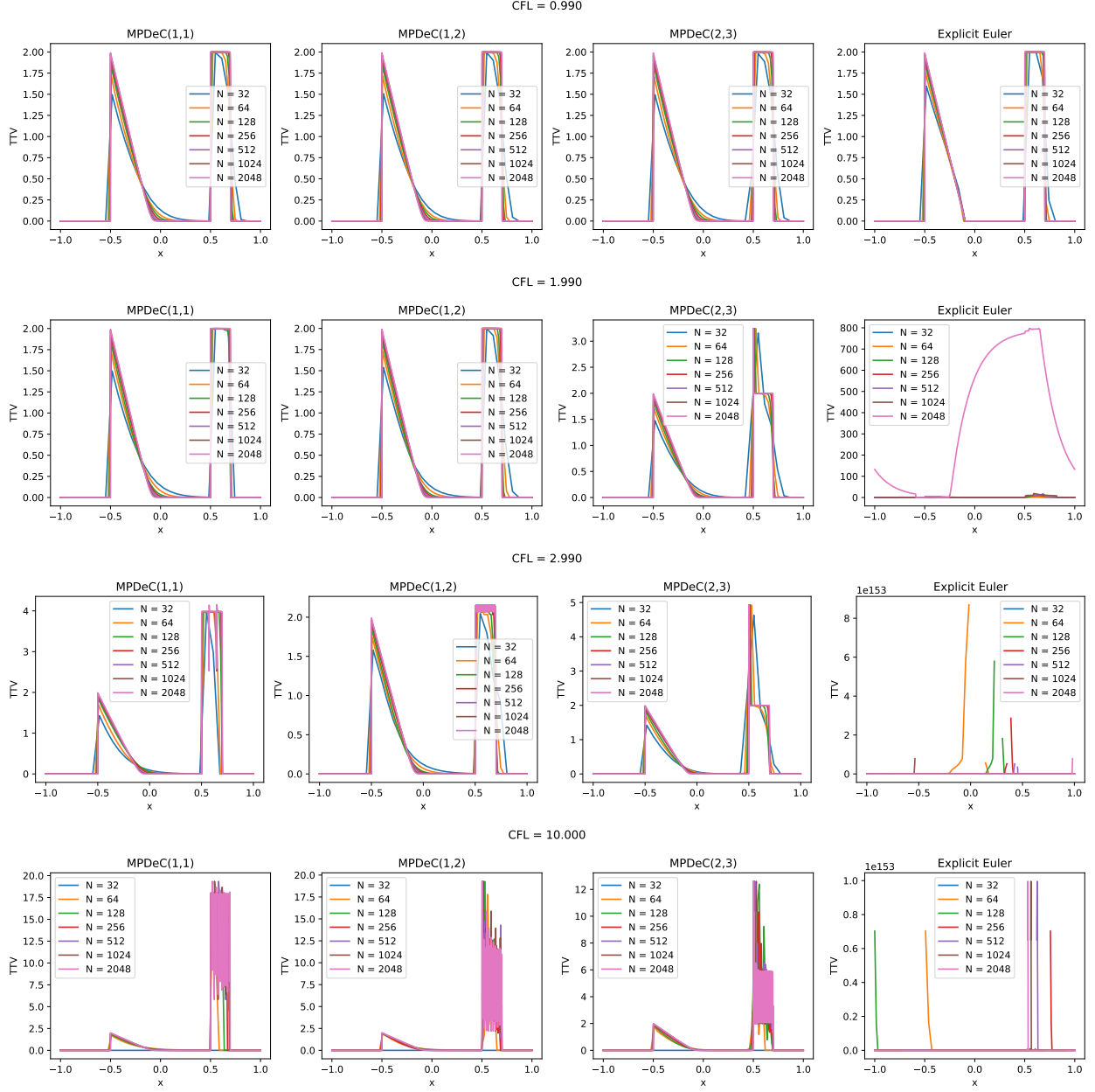
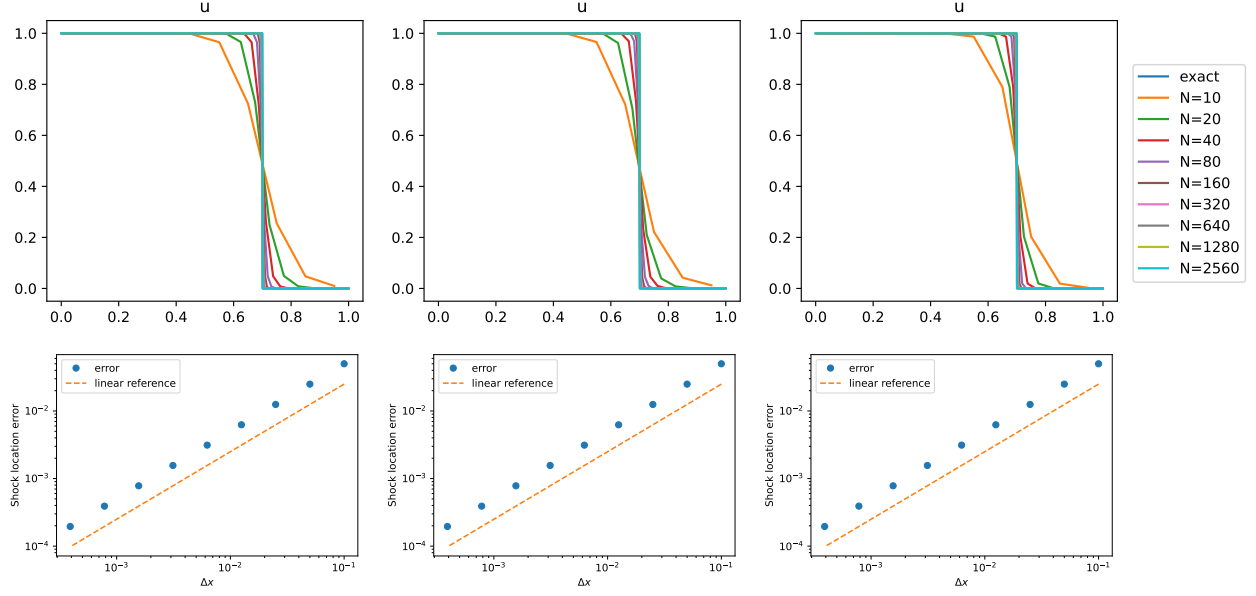


Figure 4: Total time variation for various methods and different CFL number for Burgers' equation (21) with IC (22) with  $u_1 = 2$  and  $u_2 = 10^{-13}$ . Observe that for different methods different scales are used.

## Burgers' equations



## Shallow water equations

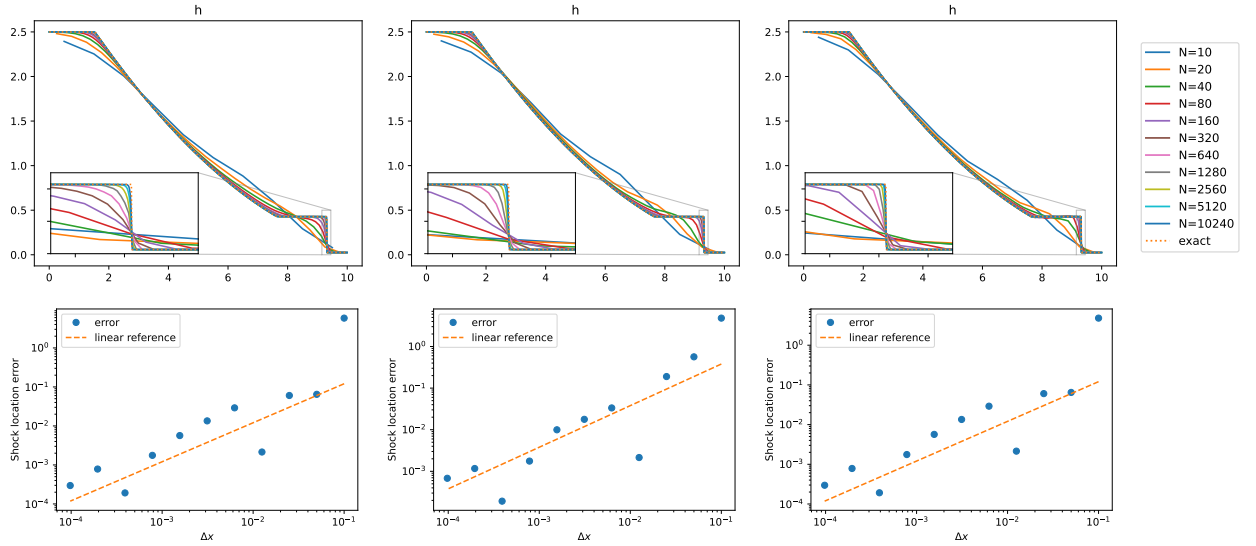


Figure 5: Burgers' (top) and SW dam break (bottom), solutions and shock location error refining the mesh: MPDeC2 (left), MPSSPRK3 (center) and MPDeC5 (right)

We start with the inviscid Burgers' equations  $\partial_t u(x, t) + \partial_x(u^2(x, t)/2) = 0$  for  $u : [0, 1] \times \mathbb{R}^+ \rightarrow \mathbb{R}$  with IC and exact solution

$$u(x, 0) = \begin{cases} 1 & x \leq 0.2, \\ 10^{-16} & \text{else,} \end{cases} \quad u(x, t) = \begin{cases} 1 & x \leq 0.2 + \frac{1}{2}t, \\ 10^{-16} & \text{else.} \end{cases}$$

The solution of this problem is clearly positive, hence it is reasonable to apply the MP procedure on the variable  $u$ . In Fig. 5 (top), we show the simulations for various methods at time  $T = 1$  obtained with  $\text{CFL} = 0.99$ . In all schemes, we see a convergence of the shock towards the exact solution, in accordance with Remark 9 and, hence, with the shock traveling at the right speed. This is a symptom of a conservative method and consistent numerical fluxes.

We then proceed with a shallow water equation test of a slightly wet dam break

$$\begin{cases} \partial_t h + \partial_x(hu) = 0, \\ \partial_t(hu) + \partial_x(hu^2 + \frac{9.8}{2}h^2) = 0, \end{cases} \quad h(x, 0) = \begin{cases} 2.5 & x \leq 5, \\ 0.025 & x > 5, \end{cases} \quad u(x, 0) = 0. \quad (24)$$

Clearly, for this test only  $h$  is part of a positive production-destruction system and is put in an MP solver, while  $hu$  is solved with the underlying explicit method. The choice of a wet dam break is dictated by the fact that dry dam breaks do not show a discontinuity close to the dry area, but rather a rarefaction wave, making it impossible to validate the speed of the shock close to the dry region. Moreover, we remark that, in particular in the initial times, the MP weights close to the shock are proportional to the ratio of the left and the right state of  $h$ , which in this case is 100, see (24). We believe that this jump would already influence the behavior of the solution. Larger jumps would make the shock smaller and the numerical diffusion would smear out their behavior, making the validation of the result impossible.

In Fig. 5 (bottom), we depict the simulation at time  $T = 0.7$  obtained with  $\text{CFL} = 0.99$  for different MP solvers next to the exact solutions as well as the numerical shock location error. Here, we also observe that the solutions converge to the exact one as the mesh is refined.

In [12, 11], other more challenging shallow water tests have been solved using the same approach with MPDeC. There, multiple shocks and their interactions in wet/dry regimes have been tested and in all those simulations the solutions showed reliable results in comparison with classical benchmarks.

## 6. Conclusions and perspectives

We have extended the Lax-Wendroff theorem to Patankar-type temporal discretization of difference methods for conservation laws. The proof of the theorem required an additional hypothesis, with respect to the classical one, that bounds the total time variation of the numerical solutions uniformly in space and refinements. The numerical results show that, if convergent, the Patankar-type methods converge to the same weak solution of the explicit methods, and sometimes they converge also when the explicit methods fail at converging, bringing in some extra stabilization.

It would be interesting to extend this result, for example, proving that the extra hypothesis is not needed. In this context, ideas similar to [8] could be explored, even if there is no guarantee that the two limits of linear solver iterations going to infinity and mesh refinements can be exchanged. Some of the authors are also working on defining an entropy-preserving version of Patankar-type methods. This will be of particular interest for the development of Patankar-type methods that guarantee the positivity of quantities that are not directly evolved by the conservation law but instead are derived by an additional equation of state. For instance, the pressure in the context of the compressible Euler equations would be positive if the entropy is dissipated along the simulation and the density is positive, see for example [7]. This still requires some additional work and it is not a trivial modification of Patankar-type schemes for PDS.

## Acknowledgements

P. Ö. is supported by the DFG within SPP 2410, project 525866748 and under the personal grant 520756621. D. T. was supported by the Ateneo Sapienza projects 2023 “Modeling, numerical treatment of hyperbolic equations and optimal control problems”.

## In memoriam

This paper is dedicated to the memory of Prof. Arturo Hidalgo López (\*July 03<sup>rd</sup> 1966 - †August 26<sup>th</sup> 2024) of the Universidad Politecnica de Madrid, organizer of HONOM 2019 and active participant in many other editions of HONOM. Our thoughts and wishes go to his wife Lourdes and his sister María Jesús, whom he left behind.

## Appendix A. Total Variation Diminishing property

In this section, we will prove an interesting result for a very specific combination of space-time discretization on Burgers’ equation. We use MPE as time discretization, an upwind numerical flux (being  $u$  positive, this is quite straightforward) with the definition of PDS of (9). We can prove that this scheme is total variation diminishing (TVD) for an *implicit* CFL  $< 2$ .

**Theorem 6.** *Consider the upwind method with the following PDS distribution*

$$\partial_i U_i = -\frac{U_i^2 - U_{i-1}^2}{2\Delta x}, \quad \text{with } p_{i,i-1} = \frac{U_{i-1}^2}{2\Delta x}, \quad d_{i,i+1} = \frac{U_i^2}{2\Delta x}. \quad (\text{A.1})$$

*The MPE scheme applied to this semidiscretization is TVD with the implicit CFL condition*

$$\frac{\Delta t^n}{\Delta x} U_i^{n+1} \leq 2 \quad \forall i, \quad (\text{A.2})$$

for all times  $n$  with  $\Delta t^n = t_{n+1} - t_n$ .

*Proof.* Let us first write down explicitly the MPE scheme applied to the previous semidiscretization:

$$U_i^{n+1} = U_i^n - \frac{\Delta t}{2\Delta x} \left( (U_i^n)^2 \frac{U_i^{n+1}}{U_i^n} - (U_{i-1}^n)^2 \frac{U_{i-1}^{n+1}}{U_{i-1}^n} \right) \quad (\text{A.3})$$

$$= U_i^n - \frac{\Delta t}{2\Delta x} (U_i^n U_i^{n+1} - U_{i-1}^n U_{i-1}^{n+1}) = U_i^n - \frac{\Delta t}{2\Delta x} (U_i^n (U_i^{n+1} - U_{i-1}^{n+1}) + (U_i^n - U_{i-1}^n) U_{i-1}^{n+1}). \quad (\text{A.4})$$

Now, applying the difference of two consecutive values, we obtain

$$U_i^{n+1} - U_{i-1}^{n+1} = U_i^n - U_{i-1}^n - \frac{\Delta t}{2\Delta x} (U_i^n (U_i^{n+1} - U_{i-1}^{n+1}) + (U_i^n - U_{i-1}^n) U_{i-1}^{n+1} - U_{i-1}^n (U_{i-1}^{n+1} - U_{i-2}^{n+1}) - (U_{i-1}^n - U_{i-2}^n) U_{i-2}^{n+1}) \quad (\text{A.5})$$

and bringing on the left hand side the implicit term that we can collect, we obtain

$$\begin{aligned} & (U_i^{n+1} - U_{i-1}^{n+1}) \left( 1 + \frac{\Delta t}{2\Delta x} U_i^n \right) \\ &= (U_i^n - U_{i-1}^n) \left( 1 - \frac{\Delta t}{2\Delta x} U_{i-1}^{n+1} \right) + \frac{\Delta t}{2\Delta x} (U_{i-1}^n (U_{i-1}^{n+1} - U_{i-2}^{n+1}) + (U_{i-1}^n - U_{i-2}^n) U_{i-2}^{n+1}). \end{aligned} \quad (\text{A.6})$$

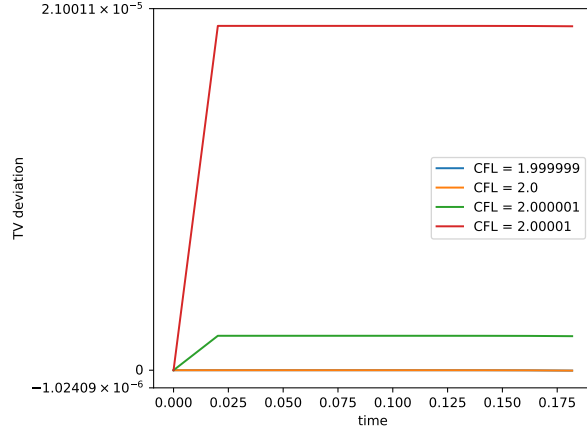


Figure A.6: Deviation of the total variation from the initial total variation (left) and maximum variation of TV (right) for different CFL numbers on a Burgers' double Riemann Problem test with MPE and upwind

Now, using (A.2) and the positivity of the solution, we obtain

$$\begin{aligned} \sum_i |U_i^{n+1} - U_{i-1}^{n+1}| \left(1 + \frac{\Delta t}{2\Delta x} U_i^n\right) &\leq \sum_i |U_i^n - U_{i-1}^n| \left(1 - \frac{\Delta t}{2\Delta x} U_{i-1}^{n+1}\right) \\ &\quad + \sum_i \frac{\Delta t}{2\Delta x} U_{i-1}^n |U_{i-1}^{n+1} - U_{i-2}^{n+1}| + \sum_i \frac{\Delta t}{2\Delta x} |U_{i-1}^n - U_{i-2}^n| U_{i-2}^{n+1} \end{aligned} \quad (\text{A.7})$$

$$\sum_i |U_i^{n+1} - U_{i-1}^{n+1}| \left(1 + \frac{\Delta t}{2\Delta x} U_i^n - \frac{\Delta t}{2\Delta x} U_i^n\right) \leq \sum_i |U_i^n - U_{i-1}^n| \left(1 - \frac{\Delta t}{2\Delta x} U_{i-1}^{n+1} + \frac{\Delta t}{2\Delta x} U_{i-1}^{n+1}\right) \quad (\text{A.8})$$

$$\Leftrightarrow \text{TV}(U^{n+1}) = \sum_i |U_i^{n+1} - U_{i-1}^{n+1}| \leq \sum_i |U_i^n - U_{i-1}^n| = \text{TV}(U^n). \quad (\text{A.9})$$

□

It is not so easy to use the same trick for other schemes and systems as the add-and-subtract trick we did was tailored for Burgers' equation and those Patankar weights. For more complex problems and schemes one would have further terms that are not so easy to handle.

**Remark 12 (On the implicit CFL condition).** The implicit CFL condition (A.2) is not the classical CFL condition imposed using the values of the solution at the previous time step, but it rather implies knowing the solution at the next time step, which is not possible. Nevertheless, from an experimental point of view, we have observed that also imposing the classical explicit CFL condition with CFL value equal to 2 is sharply guaranteeing the TVD property.

#### Appendix A.1. Numerical results

To test the TVD property, we run some simulations for the Burgers' equation with a double Riemann problem (22) as initial condition with different values for  $u_1$  and  $u_2$ . The numerical findings are independent on these values. We show the simulations for  $u_1 = 2$  and  $u_2 = 10^{-13}$  with  $N = 100$  discretization points and periodic boundary conditions. We test for different explicit CFL numbers close to the critical value 2 to check if the found condition is valid and sharp, i.e.,  $\frac{\Delta t^n}{\Delta x} \max_i U_i^n \leq \text{CFL}$ . In figure A.6, we plot the difference between the TV at a given time and the TV at the initial time. We clearly see that for  $\text{CFL} \leq 2$  the simulations are TVD while as soon as we cross the value 2, the simulations are not TVD. Moreover, the maximum of value of the difference between the TV at a time step and at the previous time step is reported in the right of

Figure A.6 and confirms once again that even imposing the CFL at an explicit level, we sharply obtain the TVD property for  $\text{CFL} \leq 2$ .

Numerically, we have observed that other high order schemes do not share the same property. For example, we could not find any CFL for which MPDeC of order 3 and 4 would be TVD, while for MPDeC of order 2 we obtained numerically the same CFL bound as MPE.

During the preparation of this work the authors used ChatGPT (OpenAI) in order to refer to the compactness of functions of bounded variation (Lemma 2). After using this tool/service, the author(s) reviewed and edited the content as needed and take full responsibility for the content of the published article.

## References

- [1] R. ABGRALL, *A personal discussion on conservation, and how to formulate it*, in Finite volumes for complex applications X – Volume 1. Elliptic and parabolic problems. FVCA 10, Strasbourg, France, October 30 – November 3, 2023. Invited contributions, Cham: Springer, 2023, pp. 3–19.
- [2] L. AMBROSIO, N. FUSCO, AND D. PALLARA, *Functions of bounded variation and free discontinuity problems*, Oxford university press, 2000.
- [3] E. AUDUSSE, F. BOUCHUT, M.-O. BRISTEAU, R. KLEIN, AND B. PERTHAME, *A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows*, SIAM J. Sci. Comput., 25 (2004), pp. 2050–2065.
- [4] A. I. ÁVILA, G. J. GONZÁLEZ, S. KOPECZ, AND A. MEISTER, *Extension of modified Patankar–Runge–Kutta schemes to nonautonomous production–destruction systems based on Oliver’s approach*, Journal of Computational and Applied Mathematics, 389 (2021), p. 113350.
- [5] A. I. ÁVILA, S. KOPECZ, AND A. MEISTER, *A comprehensive theory on generalized BBKS schemes*, Appl. Numer. Math., 157 (2020), pp. 19–37.
- [6] M. BEN-ARTZI, *GRP – a direct Godunov extension*, J. Comput. Phys., 519 (2024), p. 23. Id/No 113388.
- [7] C. BERTHON, V. MICHEL-DANSAC, AND A. THOMANN, *Towards a fully well-balanced and entropy-stable scheme for the euler equations with gravity: preserving isentropic steady solutions*, Mathematics of Computation, (2025).
- [8] P. BIRKEN AND V. LINDERS, *Conservation Properties of Iterative Methods for Implicit Discretizations of Conservation Laws*, J Sci Comput, 92 (2022), p. 60.
- [9] H. BURCHARD, E. DELEERSNIJDER, AND A. MEISTER, *A high-order conservative Patankar-type discretisation for stiff systems of production–destruction equations*, Appl. Numer. Math., 47 (2003), pp. 1–30.
- [10] M. H. CARPENTER AND T. C. FISHER, *High-Order Entropy Stable Formulations for Computational Fluid Dynamics*.
- [11] M. CIALLELLA, L. MICALIZZI, V. MICHEL-DANSAC, P. ÖFFNER, AND D. TORLO, *A high-order, fully well-balanced, unconditionally positivity-preserving finite volume framework for flood simulations*, GEM-International Journal on Geomathematics, 16 (2025), pp. 1–33.
- [12] M. CIALLELLA, L. MICALIZZI, P. ÖFFNER, AND D. TORLO, *An arbitrary high order and positivity preserving method for the shallow water equations*, Computers & Fluids, 247 (2022), p. 105630.



- [13] A. DUTT, L. GREENGARD, AND V. ROKHLIN, *Spectral Deferred Correction Methods for Ordinary Differential Equations*, BIT Numerical Mathematics, 40 (2000), pp. 241–266.
- [14] R. EYMARD, T. GALLOUËT, R. HERBIN, AND J.-C. LATCHÉ, *Finite volume schemes and lax–wendroff consistency*, Comptes Rendus. Mécanique, 350 (2022), pp. 1–13.
- [15] T. C. FISHER, M. H. CARPENTER, J. NORDSTRÖM, N. K. YAMALEEV, AND C. SWANSON, *Discretely conservative finite-difference formulations for nonlinear conservation laws in split form: theory and boundary conditions*, J. Comput. Phys., 234 (2013), pp. 353–375.
- [16] T. GALLOUËT, R. HERBIN, AND J.-C. LATCHÉ, *On the weak consistency of finite volumes schemes for conservation laws on general meshes*, SēMA J., 76 (2019), pp. 581–594.
- [17] E. GODLEWSKI AND P.-A. RAVIART, *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, vol. 118 of Applied Mathematical Sciences, Springer New York, New York, NY, 1996.
- [18] J. HUANG AND C.-W. SHU, *Positivity-Preserving Time Discretizations for Production–Destruction Equations with Applications to Non-equilibrium Flows*, Journal of Scientific Computing, 78 (2019), pp. 1811–1839.
- [19] J. HUANG, W. ZHAO, AND C.-W. SHU, *A third-order unconditionally positivity-preserving scheme for production–destruction equations with applications to non-equilibrium flows*, Journal of Scientific Computing, 79 (2019), pp. 1015–1056.
- [20] T. IZGIN, *A Unifying Theory for Runge-Kutta-like Time Integrators: Convergence and Stability*, PhD thesis, University of Kassel, 2024.
- [21] T. IZGIN, D. I. KETCHESON, AND A. MEISTER, *Order conditions for Runge–Kutta-like methods with solution-dependent coefficients*, Commun. Appl. Math. Comput. Sci., 20 (2025), pp. 29–66.
- [22] G. IZZO, E. MESSINA, M. PEZZELLA, AND A. VECCHIO, *Modified patankar linear multistep methods for production-destruction systems*, Journal of Scientific Computing, 102 (2025), p. 87.
- [23] S. KOPECZ AND A. MEISTER, *On order conditions for modified Patankar-Runge-Kutta schemes*, Appl. Numer. Math., 123 (2018), pp. 159–179.
- [24] ———, *Unconditionally positive and conservative third order modified Patankar-Runge-Kutta discretizations of production-destruction systems*, BIT, 58 (2018), pp. 691–728.
- [25] D. KRÖNER, *Numerical schemes for conservation laws*, Wiley-Teubner series, advances in numerical mathematics, Wiley; Teubner, Chichester; New York; Stuttgart, 1997.
- [26] D. KUZMIN AND H. HAJDUK, *Property-preserving numerical schemes for conservation laws*, Singapore: World Scientific, 2023.
- [27] D. KUZMIN, H. HAJDUK, AND A. RUPP, *Limiters-based entropy stabilization of semi-discrete and fully discrete schemes for nonlinear hyperbolic problems*, Comput. Methods Appl. Mech. Eng., 389 (2022), p. 28. Id/No 114428.
- [28] D. KUZMIN, M. LUKÁCOVA-MEDVID’OVÁ, AND P. ÖFFNER, *Consistency and convergence of flux-corrected finite element methods for nonlinear hyperbolic problems*. Preprint, arXiv:2308.14872 [math.NA] (2023), 2023.

- [29] P. LAX AND B. WENDROFF, *Systems of conservation laws*, Comm. Pure Appl. Math., 13 (1960), pp. 217–237.
- [30] R. J. LEVEQUE, *Numerical Methods for Conservation Laws*, Birkhäuser Basel, Basel, 1992.
- [31] V. LINDERS AND P. BIRKEN, *On the consistency of arnoldi-based krylov methods for conservation laws*, PAMM, 23 (2023), p. e202200157.
- [32] ———, *Locally conservative and flux consistent iterative methods*, SIAM Journal on Scientific Computing, 46 (2024), pp. S424–S444.
- [33] A. MARTIRADONNA, G. COLONNA, AND F. DIELE, *GeCo: Geometric Conservative nonstandard schemes for biochemical systems*, Appl. Numer. Math., 155 (2020), pp. 38–57.
- [34] A. MEISTER AND S. ORTLEB, *On unconditionally positive implicit time integration for the DG scheme applied to shallow water flows*, Internat. J. Numer. Methods Fluids, 76 (2014), pp. 69–94.
- [35] P. ÖFFNER AND D. TORLO, *Arbitrary high-order, conservative and positivity preserving Patankar-type deferred correction schemes*, Applied Numerical Mathematics, 153 (2020), pp. 15–34.
- [36] S. ORTLEB, *Positivity preserving implicit and partially implicit time integration methods in the context of the DG scheme applied to shallow water flows*, in Finite volumes for complex applications VII. Methods and theoretical aspects, vol. 77 of Springer Proc. Math. Stat., Springer, Cham, 2014, pp. 431–438.
- [37] S. ORTLEB AND W. HUNSDORFER, *Patankar-type Runge-Kutta schemes for linear PDEs*, vol. 1863, 07 2017, p. 320008.
- [38] B. PERTHAME AND C.-W. SHU, *On positivity preserving finite volume schemes for Euler equations*, Numerische Mathematik, 73 (1996), pp. 119–130.
- [39] M. RICCHIUTO AND A. BOLLERMANN, *Stabilized residual distribution for shallow water simulations*, J. Comput. Phys., 228 (2009), pp. 1071–1115.
- [40] C. SHI AND C.-W. SHU, *On local conservation of numerical methods for conservation laws*, Comput. Fluids, 169 (2018), pp. 3–9.
- [41] C. SHI AND C.-W. SHU, *On local conservation of numerical methods for conservation laws*, Computers & Fluids, 169 (2018), pp. 3–9. Recent progress in nonlinear numerical methods for time-dependent flow & transport problems.
- [42] C.-W. SHU AND S. OSHER, *Efficient implementation of essentially non-oscillatory shock-capturing schemes*, Journal of Computational Physics, 77 (1988), pp. 439–471.
- [43] E. TADMOR, *Entropy stability theory for difference approximations of nonlinear conservation laws and related time-dependent problems*, Acta Numer., 12 (2003), pp. 451–512.
- [44] D. TORLO, P. ÖFFNER, AND H. RANOCHA, *Issues with positivity-preserving Patankar-type schemes*, Appl. Numer. Math., 182 (2022), pp. 117–147.
- [45] G. TOTH, *Total of time variation diminishing principle for conservation laws*, J. Comput. Phys., 494 (2023), pp. Paper No. 112534, 14.
- [46] F. ZHU, J. HUANG, AND Y. YANG, *Bound-preserving discontinuous Galerkin methods with modified Patankar time integrations for chemical reacting flows*, Communications on Applied Mathematics and Computation, 6 (2024), pp. 190–217.