TVD IMEX Runge-Kutta schemes based on arbitrarily high order Butcher tableaux

Victor Michel-Dansac, Andrea Thomann

TVD IMEX Runge-Kutta schemes based on arbitrarily high order Butcher tableaux

Victor Michel-Dansac*, Andrea Thomann†

February 28, 2020

Abstract

The context of this work is the development of TVD Implicit-Explicit Runge-Kutta schemes to approximate the solution of hyperbolic multi-scale equations. A key feature of IMEX-RK schemes is that the resulting CFL condition does not depend on the stiff part of the considered equation, as long as this stiff part is treated implicitly. However, the negative result from Gottlieb et al. [17] states that it is not possible to construct an implicit Runge-Kutta scheme of order higher than one that is either L^∞ stable or TVD. We show that this result is also valid for IMEX-RK schemes. Therefore, rather than building a high-order TVD scheme, the goal of this work is to provide a way of improving the precision of a first-order IMEX-RK scheme, while retaining its L^∞ stability and TVD properties. To that end, we introduce a convex combination between an oscillatory high-order IMEX-RK scheme and the stable but diffusive first-order IMEX-RK scheme. We derive generic conditions to ensure the TVD property for a convex combination based on an arbitrarily high order Butcher tableau. To increase the precision, we combine our new TVD schemes with an optimal order detection strategy inspired by the MOOD framework. We compare them to L-stable methods from the literature, applied on a two-scale hyperbolic equation where we test the performance on continuous and discontinuous solutions.

1 Introduction

Multi-scale equations arise in a wide range of applications, such as shallow water equations studied e.g. in [5], magnetohydrodynamics [25], multi-material [1] or atmospheric flows [23]. When developing numerical methods for such applications, it is of prime importance to obtain physically admissible solutions under these multi-scale constraints.

In order to numerically treat these different scales, one must assess whether the fast scales are relevant to the physical solution. Indeed, accurately capturing these fast scales requires a very restrictive time step. This issue is discussed e.g. in [18] for the Euler equations. When the impact of the fast scales on the physical solution is less important, numerical methods which do not accurately capture these scales but follow only the slow scales are necessary. One option, which we will study in this paper, is to treat the slow dynamics explicitly and the fast dynamics implicitly in time, thus leading to an Implicit-Explicit (IMEX) scheme. Those schemes are well studied in the literature, see for instance [3] for efficient IMEX schemes applied on hyperbolic-parabolic problems, [30] for

*Institut de Mathématiques de Toulouse, Université Toulouse 3 Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex 9, France; INSA Toulouse, 135 avenue de Rangueil, 31077 Toulouse Cedex 4, France

†Dipartimento di Scienze e Alta Tecnologia, Università degli Studi dell'Insubria, Via Valleggio 11, 22100 Como, Italy; Marie Skłodowska-Curie fellow of the Istituto Nazionale di Alta Matematica Francesco Severi, Rome, Italy

IMEX schemes adapted to stiff relaxation source terms, or [28, 14, 8] for IMEX schemes designed for the low Mach regime of the Euler equations, as well as the references given therein. Most of those schemes possess the asymptotic preserving (AP) property. Such AP schemes are constructed in order to have the right limit behaviour when the fast wave speeds tend to infinity. A prominent example is the incompressible limit of the Euler equations as the acoustic wave speeds tend to infinity, which has been rigorously studied in the seminal work of [22]. Closely related to the AP property is the scale-independent diffusion of the numerical scheme, see for instance [34]. This guarantees an accurate description of the solution even in the limit regime. Even though we are only consider a linear two-scale scalar equation in this work, we can study the compatibility of our proposed IMEX discretisation with the AP property.

Another important property appearing when dealing with the approximation of physical phenomena is that the admissibility domain of the physical solution has to be preserved. Prominent examples for this are the positivity of the density and internal energy in the case of atmospheric flows, or the water height non-negativity for shallow water equations. Numerical techniques ensuring the positivity of these quantities are called positivity preserving. A lot of effort has been put into the design of such schemes, see e.g. [31, 33] for the Euler equations or [6, 4] for the shallow water equations. Widely used are explicit strong stability preserving (SSP) IMEX-Runge Kutta (IMEX-RK) schemes, whose development has been a major endeavour of the last two decades. They are usually built starting from the time integration of ordinary differential equations, see for instance [17] for SSP high-order time discretisation methods, [15] for SSPRK multi-step methods, [7] for the SSP property of a scheme combining a trapezoidal rule and a BDF2 discretisation, [10] for the development of some multi-step explicit SSPRK methods, and [21] for a study on the SSP property of RK time integrators for Godunov schemes.

Here, we are interested in even stronger stability properties, namely the total variation diminishing (TVD) property and the L^∞ stability. Both imply the aforementioned positivity preservation property. Explicit TVD RK schemes are widely available in the literature, for instance schemes up to order four are studied in [16]. However, in the seminal work by Gottlieb et al. [17], it has been proven that it is impossible to construct unconditionally stable higher order implicit TVD RK methods. Instead, one must either rely on a more restrictive time step, or forsake truly high-order implicit discretisations altogether. Unfortunately, this holds also for IMEX discretizations, as we will discuss in the beginning of this paper. In fact, this negative result is also observed in [13, 9] when attempting to construct second-order IMEX discretisations. Since our goal is to obtain a scale-independent time step restriction, this negative result consequently means that we cannot construct high-order TVD IMEX Runge-Kutta (IMEX-RK) schemes. Therefore, we seek to construct first-order TVD IMEX-RK schemes that are more precise than a TVD scheme based on first-order forward and backward Euler steps.

A first step in this direction was already taken in [13, 27], where the increase of precision is achieved by utilizing a convex combination of a first-order TVD scheme with an oscillatory second-order scheme. In [13], the ARS(2,2,2) scheme from [3] is used as a basis for the convex combination, and this result is extended to arbitrary second-order schemes in [27]. In this work, we extend it further to a convex combination with arbitrarily high order schemes. Let us emphasise once again that the scheme resulting from this convex combination is still first-order accurate, but is more precise than the usual first-order scheme. Note that convex combinations have already been used to recover first-order properties lost at higher orders, see for instance [19] to recover the positivity property or [26] for well-balanced problems.

In practical use, it turns out that it is not necessary to use a TVD scheme throughout the whole simulation but rather use it as a correction when the solution leaves the physical admissibility domain. Therefore, we increase the precision even further by adapting a MOOD-like method, introduced in [11], to the case of IMEX schemes.

The paper is organised as follows. In Section 2, we describe the problem of multi-scale equations, illustrated by a scalar linear hyperbolic equation. We shortly recall the IMEX formalism to numerically approximate those stiff equations, and we prove the negative result for the construction of higher order TVD IMEX-RK schemes. In addition, we shortly comment on the asymptotic behaviour of the equation and the associated numerical scheme as the fast wave speed goes to infinity. In Section 3, we derive a TVD IMEX scheme based on a convex combination between a second-order and a first-order IMEX update. The problem of higher order TVD space discretisation is also addressed in this section. The extension to TVD schemes based on arbitrarily high order Butcher tableaux is discussed in Section 4. There, we show that the convex combination on the time updates of the the first-order and higher order IMEX schemes is not enough to find a TVD scheme. Instead, we also apply a convex combination at each stage of the IMEX scheme. This method is illustrated by the construction of a TVD scheme based on third-order tableaux, combined with a third-order limiting procedure for the explicit space discretisation. Section 5 is devoted to numerical experiments using the schemes developed in the paper. First, to noticeably increase the precision of the developed scheme in these two sections, we introduce a MOOD procedure. Then, we suggest optimal values for the free parameters of the introduced TVD schemes, by compromising between precision and CPU time. To numerically validate that L^∞ stable and TVD schemes are really an improvement over widely used L-stable IMEX schemes, especially for small ε, we finally test the precision and performance of the schemes on continuous and discontinuous solutions of the scalar multi-scale equation. To complete the paper, a short conclusion, as well as plans for future work, are presented in Section 6.

2 Problem description

We consider the scalar linear two-scale initial value problem

$$\begin{cases} w_t + c_m w_x + \dfrac{c_a}{\varepsilon} w_x = 0, \\ w(0,x) = w^0(x), \end{cases} \tag{2.1}$$

where $w : (\mathbb{R}^+, \Omega) \to \mathbb{R}$, $\Omega \subset \mathbb{R}$. where the slow transport velocity is given by c_m and the fast transport velocity by c_a/ε, with c_m and c_a independent of the parameter $\varepsilon > 0$. Without loss of generality, we consider only the positive transport direction, where the transport velocities c_m and c_a are positive. Equation (2.1) mimics the wave structure of e.g. the non-dimensional Euler equations, see [13] where Equation (2.1) was used as a simple model for the Euler equations, with c_a/ε and c_m respectively representing the fast acoustic and slow material velocities. Thereby ε corresponds to the square of the Mach number M. Nevertheless, when developing numerical methods for the simplified scalar case (2.1) with small $\varepsilon > 0$, one faces similar challenges as for the Euler equations in the low Mach regime. Treating both derivatives in (2.1) explicitly would lead to a CFL condition depending on ε to ensure stability:

$$\Delta t \leqslant \varepsilon \frac{\Delta x}{\varepsilon c_m + c_a}.$$

Therefore, we adopt an IMEX approach and treat the derivative associated with the slow speed explicitly and the one associated with the fast speed implicitly. Since our main goal is to derive an L^∞ stable and TVD scheme, we have to use an upwind discretisation for both derivatives. This is motivated by the results in [14], where it is shown that for a non-linear system centred differences destroy the L^∞ stability. Although our setting is linear, we avoid centred differences to be able to apply the approach developed here on non-linear systems as e.g. the Euler equations.

The space and time discretisation follows the usual finite difference framework, although it can be easily translated into the finite volume setting. The space domain Ω is partitioned in N uniformly

spaced points $(x_j)_{j\in\{1,\dots,N\}}$ with the step size Δx. We discretise the time variable with $t^n = n\Delta t$, where Δt denotes the time step. Then the solution $w(t,x)$ at (t^n, x_j) is approximated by w_j^n. A semi-discrete first-order approximation of (2.1) in space with $\Delta_j(t) = w_j(t) - w_{j-1}(t)$ is given by

$$\partial_t w_j(t) + \frac{c_m}{\Delta x}\Delta_j(t) + \frac{c_a}{\varepsilon\Delta x}\Delta_j(t) = 0. \tag{2.2}$$

Later on, to extend the space discretisation in (2.2) to higher order, we will use a higher order reconstruction combined with a limiting procedure to ensure the TVD property.

For the time integration of (2.2), we use the IMEX-RK framework. The time update for an s-stage IMEX-RK scheme for equation (2.2) is given by

$$w_j^{n+1} = w_j^n - \lambda\sum_{k=1}^{s}\tilde{b}_k\Delta_j^{(k)} - \mu_\varepsilon\sum_{k=1}^{s} b_k\Delta_j^{(k)}, \tag{2.3}$$

where $\lambda = \frac{\Delta t}{\Delta x}c_m$, $\mu_\varepsilon = \frac{\Delta t}{\Delta x}\frac{c_a}{\varepsilon}$ and $\Delta_j^{(k)} = w_j^{(k)} - w_{j-1}^{(k)}$, and where the stages are defined as

$$w_j^{(k)} = w_j^n - \lambda\sum_{l=1}^{k-1}\tilde{a}_{kl}\Delta_j^{(l)} - \mu_\varepsilon\sum_{l=1}^{k} a_{kl}\Delta_j^{(l)}. \tag{2.4}$$

The CFL constraint is then only determined by λ, which is independent of ε. The weights $\tilde{a}_{kl}, a_{kl}$ appearing in the definition of the stages $w^{(k)}$ and $\tilde{b}_k, b_k$ appearing in the update w^{n+1}, as well as the intermediate time steps c_k, $\tilde{c}_k$, are summarized in two triplets $(\tilde{A}, \tilde{b}, \tilde{c})$ and (A, b, c), with $\tilde{A}, A \in \mathbb{R}^{s\times s}$, $\tilde{b}, b \in \mathbb{R}^s$ and $\tilde{c}, c \in \mathbb{R}^s$. Here, we consider the matrix associated with the explicit part $\tilde{A}$ to be lower triangular with zeros on the diagonal, and the matrix connected to the implicit part A to be lower triangular. Since we are considering multi-scale equations, we wish, for computational efficiency, for the CFL condition of the resulting scheme to only depend on the slow scale associated with λ. In addition, for the sake of simplicity, we consider an IMEX-RK method of type CK (Carpenter and Kennedy) [20], i.e. we take the first row of A to be zero. As shown in detail in [27] for a generic second-order CK method, we need the first column of A, as well as b_1, to be zero for the CFL condition not to depend on ε. This is also the recommended structure in [17]. We illustrate the structure in the following Butcher tableaux notation:

$$\text{explicit:}\quad \begin{array}{c|cccc} 0 & 0 & 0 & \cdots & 0 \\ \tilde{c}_2 & \tilde{a}_{21} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \tilde{c}_s & \tilde{a}_{s1} & \cdots & \tilde{a}_{s,s-1} & 0 \\ \hline & \tilde{b}_1 & \cdots & \tilde{b}_{s-1} & \tilde{b}_s \end{array} \qquad \text{implicit:}\quad \begin{array}{c|cccc} 0 & 0 & 0 & \cdots & 0 \\ c_2 & 0 & a_{22} & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 \\ c_s & 0 & a_{s2} & \cdots & a_{ss} \\ \hline & 0 & b_2 & \cdots & b_s \end{array}, \tag{2.5}$$

where the coefficients $\tilde{c}$ and c are respectively connected to $\tilde{A}$ and A via

$$\tilde{c}_i = \sum_{j=1}^{i-1}\tilde{a}_{ij} \quad \text{and} \quad c_i = \sum_{j=1}^{i} a_{ij}. \tag{2.6}$$

We are interested in higher order Butcher tableaux, which implies that the weights have to fulfil higher order compatibility conditions. The order conditions to obtain a scheme up to order three are given in Table 1 taken from [29]. For orders larger than three, we refer to the order conditions in [20].

First-order:	$\sum_{k=1}^{s} \tilde{b}_k = 1,$	$\sum_{k=1}^{s} b_k = 1$		
Second-order:	$\sum_{k=1}^{s} \tilde{b}_k \tilde{c}_k = \frac{1}{2},$	$\sum_{k=1}^{s} b_k c_k = \frac{1}{2},$	$\sum_{k=1}^{s} \tilde{b}_k c_k = \frac{1}{2},$	$\sum_{k=1}^{s} b_k \tilde{c}_k = \frac{1}{2}$
Third-order:	$\sum_{k=1}^{s} \tilde{b}_k \tilde{c}_k^2 = \frac{1}{3},$	$\sum_{k=1}^{s} b_k c_k^2 = \frac{1}{3},$	$\sum_{k=1}^{s} \tilde{b}_k \tilde{c}_k c_k = \frac{1}{3},$	$\sum_{k=1}^{s} b_k \tilde{c}_k c_k = \frac{1}{3},$
	$\sum_{k,l=1}^{s} \tilde{b}_k \tilde{a}_{kl} \tilde{c}_k = \frac{1}{6},$	$\sum_{k,l=1}^{s} b_k a_{kl} c_k = \frac{1}{6},$	$\sum_{k,l=1}^{s} \tilde{b}_k a_{kl} c_k = \frac{1}{6},$	$\sum_{k,l=1}^{s} b_k \tilde{a}_{kl} c_k = \frac{1}{6},$
	$\sum_{k,l=1}^{s} \tilde{b}_k a_{kl} \tilde{c}_k = \frac{1}{6},$	$\sum_{k,l=1}^{s} b_k a_{kl} \tilde{c}_k = \frac{1}{6},$	$\sum_{k,l=1}^{s} \tilde{b}_k \tilde{a}_{kl} \tilde{c}_k = \frac{1}{6},$	$\sum_{k,l=1}^{s} \tilde{b}_k \tilde{a}_{kl} \tilde{c}_k = \frac{1}{6}$

Table 1: Order conditions for IMEX-RK schemes up to third-order

2.1 Stability failure of IMEX-RK schemes

We are concerned with the construction of an IMEX-RK scheme based on p-th order Butcher tableaux (2.5) that is L^∞ stable and TVD. A scheme is said to be L^∞ stable if

$$\|w^{n+1}\|_\infty = \max_{j\in\{1,\dots,N\}} |w_j^{n+1}| \leqslant \|w^n\|_\infty. \tag{2.7}$$

and TVD if

$$\mathrm{TV}(w^{n+1}) = \sum_{j=1}^{N} \left| w_{j+1}^{n+1} - w_j^{n+1} \right| \leqslant \mathrm{TV}(w^n). \tag{2.8}$$

As proven in Gottlieb et al. [17], implicit RK schemes of higher order than one are neither L^∞ stable nor TVD, whereas explicit higher order RK schemes can be constructed with a CFL restriction on the time step [15]. We show, starting from the proof of Gottlieb et al. [17], that this immediately yields a negative result for IMEX-RK schemes as well, which have a CFL restriction only stemming from the slow waves. To that end, we write the IMEX update (2.3) as a convex combination of forward and backward Euler steps, with $h \in [0,1]$ and weights $\alpha_{ik} \geqslant 0$ fulfilling $\sum \alpha_{ik} = 1$ as

$$w^{n+1} = (1-h) \sum_{k=0}^{i-1} \left(\alpha_{ik} w^{(k)} + \Delta t \frac{\tilde{\beta}_{ik}}{1-h} c_m w_x^{(k)} \right) + h \left(\sum_{k=0}^{i-1} \alpha_{ik} w^{(k)} + \Delta t \frac{\beta_i}{h} \frac{c_a}{\varepsilon} w_x^{(i)} \right). \tag{2.9}$$

We assume $\beta_i > 0$ and $\tilde{\beta}_{ik} \geqslant 0$ without loss of generality. Indeed, negative β_i or $\tilde{\beta}_{ik}$ could still yield a TVD scheme, by changing the upwinding direction in the discretisation of the derivatives w_x. For simplicity, we also assume without loss of generality, in accordance with [17], that the non-diagonal entries of the implicit Butcher tableau are zero. We immediately see from (2.9) that the explicit part is a convex combination of TVD forward Euler (fE) steps, and is thus TVD under the CFL restriction $\Delta t \leqslant (1-h)\, \Delta t_{fE} \min(\alpha_{ik}/\tilde{\beta}_{ik})$. The proof of the TVD property relies on the fact that we can find non-negative weights α_{ik}, and still satisfy the order conditions. Unfortunately, Gottlieb et al. proved in their negative result [17] that this is impossible for unconditionally stable implicit schemes with order two or higher. Analogously, we show that this negative result also holds for the implicit part in the IMEX update (2.9).

Proposition 1. *For an at least second-order accurate IMEX-RK update (2.9), there is at least one negative* α_{ik}.

Proof. The order conditions for higher order RK schemes are included in the compatibility conditions for higher order IMEX-RK schemes. The second-order conditions read, for the implicit part of (2.9) as

$$\sum_{k=0}^{i-1} \alpha_{ik} = 1, \quad X_s = h, \quad Y_s = \frac{1}{2}h^2, \tag{2.10}$$

where $h \in (0,1)$, and where X_s, Y_s are defined recursively by

$$X_1 = \beta_1, \quad Y_1 = \beta_1^2, \quad X_s = \beta_s + \sum_{i=1}^{s-1} \alpha_{si} X_i, \quad Y_s = \beta_s X_s + \sum_{i=1}^{s-1} \alpha_{si} Y_i.$$

Following the proof in [17], we show now that, if $\alpha_{ik} \geqslant 0$ for all i,k, then we get

$$hX_s - Y_s < \frac{1}{2}h^2,$$

which contradicts (2.10). This contradiction is shown by using the formula

$$(1-a)hX_s - Y_s \leqslant \tau_s (1-a)^2$$

with arbitrary $a \in \mathbb{R}$ and

$$0 < \tau_1 = \frac{1}{4}h^2, \quad \tau_s = \frac{h^4}{4(h^2 - \tau_{s-1})}. \tag{2.11}$$

This estimate is shown by induction following the steps given in [17]. From (2.11), we find

$$0 < \tau_1 = \frac{1}{4}h^2 < \ldots < \tau_s < \frac{1}{2}h^2$$

which completes the proof. □

This negative result is illustrated in Figure 1, where we display the approximation of a discontinuous solution with the first-order scheme and the well-known second-order ARS(2,2,2) and third-order ARS(2,3,3) schemes, both proposed in [3]. For more detail on the numerical experiment, such as initial and boundary conditions, see Section 5. We clearly observe that the higher-order schemes present oscillations in the numerical solution, while the first-order solution is very diffusive. Let us underline that these oscillations are expected as soon as the scheme is more than first-order accurate, as per Proposition 1.

In order to construct a scheme that fulfils the L^∞ stability (2.7) and TVD property (2.8), we replace the update (2.3) by a convex combination of (2.3) and the first-order scheme, following [13]. The fully discrete first-order scheme reads

$$w_j^{n+1} = w_j^n - \lambda \Delta_j^n - \mu_\varepsilon \Delta_j^{n+1}. \tag{2.12}$$

Applying the convex combination with parameter with $\theta \in [0,1]$, the new update is then given by

$$w_j^{n+1} = w_j^n - \theta \left(\lambda \sum_{k=1}^{s} \tilde{b}_k \Delta_j^{(k)} + \mu_\varepsilon \sum_{k=1}^{s} b_k \Delta_j^{(k)} \right) - (1-\theta) \left(\lambda \Delta_j^n + \mu_\varepsilon \Delta_j^{n+1} \right). \tag{2.13}$$

We emphasise that the above convex combination (2.13) is only first-order accurate, due to the negative result proven in Proposition 1, but will have a higher precision than the usual first-order scheme (2.12).

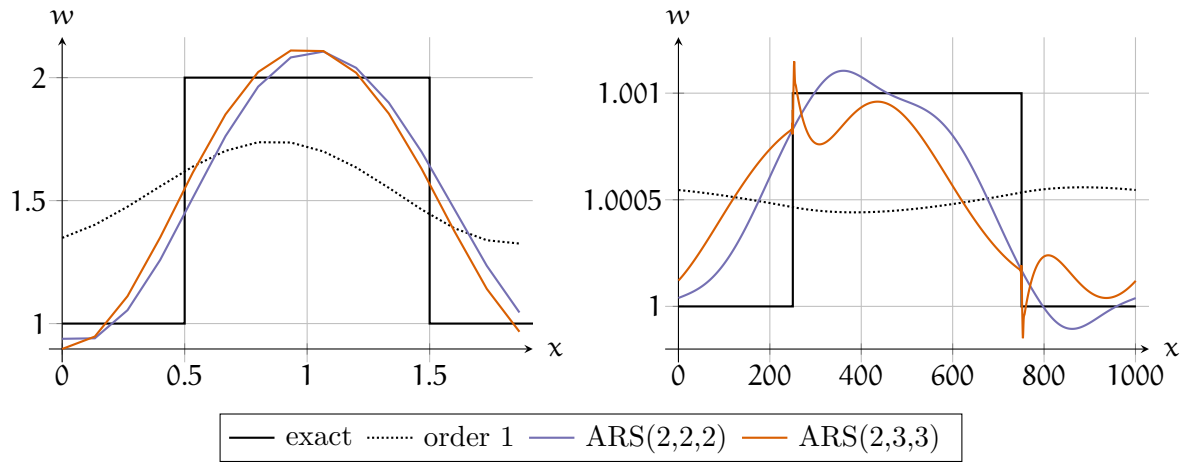

Figure 1: Approximation of a discontinuous solution using the first-order, second-order ARS(2,2,2) and third-order ARS(2,3,3) schemes. Left panel: $\varepsilon = 1$ and $N = 15$; right panel: $\varepsilon = 10^{-3}$ and $N = 2000$. In both cases, the higher-order approximations are oscillatory and the first-order one is diffusive. For more detail on the numerical experiment, see Section 5.

2.2 Asymptotic preservation properties

As our ultimate goal is to use the methods developed here on systems like the Euler equations in the low Mach regime, we now shortly address the issue of the asymptotic preserving (AP) property. For the analysis, following [12], we consider the Chapman-Enskog expansion of w in ε

$$w(t,x) = w_0(t,x) + \varepsilon\, w_1(t,x) + \mathcal{O}(\varepsilon^2). \tag{2.14}$$

Plugging the expansion (2.14) into equation (2.1), we find, by inspection of the $\mathcal{O}(\varepsilon^{-1})$ and $\mathcal{O}(\varepsilon^0)$ terms, that the limit equation of (2.1) as ε goes to zero is

$$\begin{cases} \partial_x w_0 = 0, & \text{(2.15a)} \\ \partial_t w_0 + c_a\, \partial_x w_1 = 0. & \text{(2.15b)} \end{cases}$$

Therefore, we formally find w_0 only depending on time, and we write $w_0(t,x) = w_0(t)$. Thus the well-prepared data without any further assumptions is given by $w = w_0(t) + \mathcal{O}(\varepsilon)$.

For the specific case of slipping and periodic boundary conditions, we have on a bounded domain $\int_\Omega \partial_x w_1\, dx = 0$. Then, integrating (2.15b) in space yields $\partial_t w_0 = 0$ and thus w_0 turns out to be constant. As a consequence, we find $\partial_x w_1 = 0$. Recursively applying this procedure, the limit equation in the case of periodic or slipping boundary conditions is given by a constant $w(x,t) = w^*$. As well-prepared data we define $w = w^* + \varepsilon w_1(t) + \mathcal{O}(\varepsilon^2)$, where w^* is constant and w_1 only depends on time.

To show that the numerical approximation is consistent with the limit equations in the low ε regime, we analyse, for simplicity, the one stage update (2.13) with $s = 1$. In the first case, we assume well-prepared data $w_j^n = (w_0)_j^n + \mathcal{O}(\varepsilon^1)$, which satisfies $(w_0)_j^n - (w_0)_{j-1}^n = 0$. Plugging the well-prepared data into the update (2.13), we find, from the $\mathcal{O}(\varepsilon^{-1})$ and $\mathcal{O}(\varepsilon^0)$ terms:

$$\begin{cases} (w_0)_j^{n+1} = (w_0)_{j-1}^{n+1}, & \text{(2.16a)} \\ (w_0)_j^{n+1} = (w_0)_j^n - \dfrac{\Delta t}{\Delta x} c_a \left((w_1)_j^{n+1} - (w_1)_{j-1}^{n+1} \right). & \text{(2.16b)} \end{cases}$$

Equation (2.16b) is a consistent implicit in time discretisation of the limit equation (2.15b), and from equation (2.16a) we note that the data at time t^{n+1} is still well-prepared.

Considering the case of slipping or periodic boundary conditions, we consider well-prepared data $w_j^n = (w_0)_j^n + \varepsilon (w_1)_j^n + \mathcal{O}(\varepsilon^2)$, with $(w_0)_j^n = w^*$ constant and $(w_1)_j^n - (w_1)_{j-1}^n = 0$. Plugging this data into the original equation (2.13), we arrive again at the discrete limit equations (2.16). Then, from (2.16a), we obtain that $(w_0)_j^{n+1} = (w_0)_{j-1}^{n+1}$ for all j. We denote this constant in space value by $(w_0)^{n+1}$. Then, summing (2.16b) over j, remarking that w_0 take the same value in every cell j, and arguing the boundary conditions prescribed on $(w_1)^{n+1}$, we obtain that $(w_0)^{n+1} = w^*$. Using this again in equation (2.16b), we get immediately that $(w_1)_j^{n+1} = (w_1)_{j-1}^{n+1}$. Therefore, we recover the well-prepared data $w_j^{n+1} = w^* + \varepsilon (w_1)_j^{n+1} + \mathcal{O}(\varepsilon^2)$ at time t^{n+1}.

In that sense, the scheme is considered to be asymptotically consistent. In addition, to check the numerical diffusion, we apply the upwind derivative on (2.16). We get, with well-prepared data and for any choice of boundary conditions

$$\frac{(w_0)_j^{n+1} - (w_0)_{j-1}^{n+1}}{\Delta x} = \frac{\Delta t}{\Delta x^2} c_a \left((w_1)_j^{n+1} - 2(w_1)_{j-1}^{n+1} + (w_1)_{j-2}^{n+1} \right).$$

We note that the diffusion does not depend on the inverse of ε. This property is important in order to also obtain an accurate solution for small ε. The case of more stages $s > 1$ is treated in a similar manner.

Even though we only study the one dimensional case in this paper, we would like to briefly mention the extension to multiple space dimensions. The multi-dimensional case is more interesting for practical applications, and the methodology used here to show the AP property can be easily extended to multiple space dimensions. Indeed, viewing w as a velocity, we can write a multi-dimensional version of (2.1) as

$$\partial_t \mathbf{w} + c_m \nabla \cdot \mathbf{w} + \frac{c_a}{\varepsilon} \nabla \cdot \mathbf{w} = 0$$

with $\mathbf{w} \in \mathbb{R}^d$, where d denotes the dimension. We remark that the asymptotic property $\partial_x w_0 = 0$ in one dimension translates to $\nabla \cdot \mathbf{w}_0 = 0$ for multiple dimensions. This divergence-free property is found in the incompressible Euler equations, viewed as the limit of the compressible Euler equations when the Mach number goes to 0, see for instance [22] for a rigorous analysis or [33] for an analysis of the AP property of an IMEX scheme for the Euler equations. This remark once again underlines the relevance of studying this simple toy problem.

3 L^∞ stable and TVD scheme based on second-order tableaux

The goal of this section is to provide a theoretical framework to construct L^∞ stable and TVD discretisations from a second-order Butcher tableau based on the general form (2.5). First, we discuss the stability properties of the convex combination scheme (2.13), with respect to the convex combination parameter θ. Then, we propose a strategy to increase the space accuracy of the resulting scheme.

3.1 TVD time integration

We apply the first- and second-order conditions from Table 1 and (2.6) on the Butcher tableaux given in (2.5) with $s = 3$. To reduce the number of computational steps, we assume in addition that the weights $\tilde{b}$ and b respectively coincide with the last rows of $\tilde{A}$ and A. This leads to the following

Butcher tableaux, where $\beta \neq \{0, 1\}$:

$$\text{explicit:} \quad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \beta & \beta & 0 & 0 \\ 1 & 1-\frac{1}{2\beta} & \frac{1}{2\beta} & 0 \\ \hline & 1-\frac{1}{2\beta} & \frac{1}{2\beta} & 0 \end{array}, \qquad \text{implicit:} \quad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \beta & 0 & \beta & 0 \\ 1 & 0 & \frac{1}{2(1-\beta)} & 1-\frac{1}{2(1-\beta)} \\ \hline & 0 & \frac{1}{2(1-\beta)} & 1-\frac{1}{2(1-\beta)} \end{array}. \tag{3.1}$$

Since the first rows of $\tilde{A}$ and A only contain zeros, the scheme consists only of two computational steps where $w^{(3)} = w^{n+1}$ and $w^{(1)} = w^n$. Using the stages given in (2.4), and the convex update (2.13), the scheme is given by

$$w_j^{(2)} + \mu_\varepsilon a_{22} \Delta_j^{(2)} = w_j^n - \lambda \tilde{a}_{21} \Delta_j^n, \tag{3.2a}$$

$$\begin{aligned} w_j^{n+1} + \mu_\varepsilon \left((1-\theta) + \theta a_{33}\right) \Delta_j^{n+1} &= w_j^n - \lambda \left((1-\theta) + \theta \tilde{a}_{31}\right) \Delta_j^n \\ &\quad - \theta \left(\lambda \tilde{a}_{32} + \mu_\varepsilon a_{32}\right) \Delta_j^{(2)}. \end{aligned} \tag{3.2b}$$

For the right-hand side to be independent from ε, we rewrite (3.2b) as

$$-\mu_\varepsilon \Delta_j^{(2)} = \frac{1}{a_{22}} \left(w_j^{(2)} - w_j^n\right) + \lambda \Delta_j^n.$$

Note that $\tilde{a}_{21} = a_{22}$. We have

$$w_j^{(2)} - \mu_\varepsilon a_{22} \Delta_j^{(2)} = (1 - \lambda a_{22}) w_j^n + \lambda a_{22} w_{j-1}^n, \tag{3.3a}$$

$$\begin{aligned} w_j^{n+1} - \mu_\varepsilon \left(1 + \theta(a_{33} - 1)\right) \Delta_j^{n+1} &= \left(1 - \lambda(1 + \theta(\tilde{a}_{31} - a_{32} - 1)) - \frac{\theta a_{32}}{a_{22}}\right) w_j^n \\ &\quad + \lambda \left(1 + \theta(\tilde{a}_{31} - a_{32} - 1)\right) w_{j-1}^n \\ &\quad + \theta \left(\frac{a_{32}}{a_{22}} - \lambda \tilde{a}_{32}\right) w_j^{(2)} \\ &\quad + \theta \lambda \tilde{a}_{32} w_{j-1}^{(2)}. \end{aligned} \tag{3.3b}$$

In total, we have three free parameters $\beta \neq \{0, 1\}$, $\lambda \geqslant 0$ and $\theta \in [0, 1]$. By setting $\beta \in (0, 1)$ with $\lambda < 1$ and $\theta \leqslant 2\beta(1-\beta)$, we find that all coefficients in front of $w_j^{(k)}, w_{j-1}^{(k)}$ on the right-hand sides of (3.3a) and (3.3b) are greater than or equal to zero. In (3.3a), we find:

$$\begin{aligned} w_j^n: &\quad 1 - \lambda a_{22} = 1 - \lambda\beta \geqslant 0, \\ w_{j-1}^n: &\quad \lambda a_{22} = \lambda\beta > 0, \end{aligned} \tag{3.4}$$

and, in (3.3b), we find

$$\begin{aligned} w_j^n: &\quad \left(1 - \lambda(1 + \theta(\tilde{a}_{31} - a_{32} - 1)) - \frac{\theta a_{32}}{a_{22}} = \left(1 - \frac{\theta}{2\beta(1-\beta)}\right) - \lambda \left(1 - \frac{\theta}{2\beta(1-\beta)}\right) \geqslant 0 \right. \\ w_{j-1}^n: &\quad \lambda \left(1 + \theta(\tilde{a}_{31} - a_{32} - 1)\right) = \lambda \left(1 - \frac{\theta}{2\beta(1-\beta)}\right) \geqslant 0 \\ w_j^{(2)}: &\quad \theta \left(\frac{a_{32}}{a_{22}} - \lambda \tilde{a}_{32}\right) = \theta \left(\frac{1}{2\beta(1-\beta)} - \lambda \frac{1}{2\beta}\right) > 0 \\ w_{j-1}^{(2)}: &\quad \theta \lambda \tilde{a}_{32} = \theta \lambda \frac{1}{2\beta} > 0. \end{aligned} \tag{3.5}$$

In addition, since μ_ε is non-negative, we have positive coefficients in front of $\Delta_j^{(2)}$ and Δ_j^{n+1} on the left hand side of equations (3.3a) and (3.3b). With the same notation as above we find

$$\begin{aligned} \Delta_j^{(2)}: &\quad \mu_\varepsilon a_{22} = \beta\mu_\varepsilon > 0, \\ \Delta_j^{n+1}: &\quad \mu_\varepsilon\,(1+\theta(a_{33}-1)) = \mu_\varepsilon\left(1-\frac{\theta}{2(1-\beta)}\right) > 0. \end{aligned} \tag{3.6}$$

The inequalities (3.4), (3.5) and (3.6) are the key element to show the L^∞ stability and TVD properties of scheme (3.2), because this ensures the proof only by using the triangle inequality $\|ax+by\| \leqslant a\|x\| + b\|y\|$ and reverse triangle inequality $a\|x\| - b\|y\| \leqslant \|ax-by\|$ for $x, y \in \mathbb{R}$ and $a, b \in \mathbb{R}$ with $a, b \geqslant 0$. We start with the L^∞ stability and show first that $\|w^{(2)}\|_\infty \leqslant \|w^n\|_\infty$. For periodic boundary conditions, we find with (3.4) and (3.6) that

$$\begin{aligned} \|w^n\|_\infty &= (1-\lambda a_{22})\,\|w^n\|_\infty + \lambda a_{22}\|w^n\|_\infty \\ &= (1-\lambda a_{22}) \max_j |w_j^n| + \lambda a_{22} \max_j |w_{j-1}^n| \\ &\geqslant \max_j \left|(1-\lambda a_{22})\,w_j^n + \lambda a_{22} w_{j-1}^n\right| \\ &= \max_j \left|w_j^n - \lambda a_{22}(w_j^n - w_{j-1}^n)\right| \\ &= \max_j \left|(1+\mu_\varepsilon a_{22})w_j^{(2)} - \mu_\varepsilon a_{22} w_{j-1}^{(2)}\right| \\ &\geqslant (1+\mu_\varepsilon a_{22})\|w^{(2)}\|_\infty - \mu_\varepsilon a_{22}\|w^{(2)}\|_\infty \\ &= \|w^{(2)}\|_\infty. \end{aligned}$$

Using (3.5) and (3.6), as well as the above estimate $\|w^{(2)}\|_\infty \leqslant \|w^n\|_\infty$, we can also prove analogously

$$\|w^{n+1}\|_\infty \leqslant \left(1-\frac{\theta a_{32}}{a_{22}}\right)\|w^n\|_\infty + \frac{\theta a_{32}}{a_{22}}\|w^{(2)}\|_\infty \leqslant \|w^n\|_\infty.$$

Thus, we have proven the L^∞ stability. We summarize the result in the following lemma.

Lemma 2. *For periodic boundary conditions under the CFL condition $\lambda < 1$, the scheme consisting of the Butcher tableaux (3.1) with the convex update (2.13) and the stages (2.4) with the parameters $\beta \in (0,1)$ and $\theta \leqslant 2\beta(1-\beta)$ is L^∞ stable.*

In addition, if the optimal value for θ is taken, that is if $\theta = \theta_{opt} = 2\beta(1-\beta)$, then the CFL condition relaxes to $\lambda \leqslant \min(\frac{1}{\beta}, \frac{1}{1-\beta})$.

Using the same arguments as for the proof of the L^∞ stability, we now show the TVD property. Assuming periodic boundary conditions, we write

$$\begin{aligned} \mathrm{TV}(w^n) &= (1-\lambda a_{22}) \sum_{j=1}^N \left|w_{j+1}^n - w_j^n\right| - \lambda a_{22} \sum_{j=1}^N \left|w_j^n - w_{j-1}^n\right| \\ &= \sum_{j=1}^N \left(\left|(1-\lambda a_{22})\,w_{j+1}^n - (1-\lambda a_{22})\,w_j^n\right| + \left|\lambda a_{22} w_j^n - \lambda a_{22} w_{j-1}^n\right|\right) \\ &\geqslant \sum_{j=1}^N \left|\left((1-\lambda a_{22})\,w_{j+1}^n - \lambda a_{22} w_j^n\right) - \left((1-\lambda a_{22})\,w_j^n - \lambda a_{22} w_{j-1}^n\right)\right| \\ &= \sum_{j=1}^N \left|\left((1+\mu_\varepsilon a_{22})\,w_{j+1}^{(2)} - \mu_\varepsilon a_{22} w_j^{(2)}\right) - \left((1+\mu_\varepsilon a_{22})\,w_j^{(2)} - \mu_\varepsilon a_{22} w_{j-1}^{(2)}\right)\right| \end{aligned}$$

$$
\begin{aligned}
&\geqslant \sum_{j=1}^{N} \left(\left| (1+\mu_\varepsilon a_{22}) \left(w_{j+1}^{(2)} - w_j^{(2)} \right) \right| - \left| \mu_\varepsilon a_{22} \left(w_j^{(2)} - w_{j-1}^{(2)} \right) \right| \right) \\
&= (1-\mu_\varepsilon a_{22}) \sum_{j=1}^{N} \left| w_{j+1}^n - w_j^n \right| - \mu_\varepsilon a_{22} \sum_{j=1}^{N} \left| w_j^n - w_{j-1}^n \right| \\
&= \mathrm{TV}(w^{(2)}).
\end{aligned}
$$

Using the above estimate, we now show the final TVD property. Since the proof is straightforward we only give the main steps

$$
\mathrm{TV}(w^{n+1}) \leqslant \left(1 - \frac{\theta a_{32}}{a_{22}} \right) \mathrm{TV}(w^n) + \frac{\theta a_{32}}{a_{22}} \mathrm{TV}(w^{(2)}) \leqslant \mathrm{TV}(w^n).
$$

This result is summarized as follows

Lemma 3. *For $\beta \in (0,1)$ and periodic boundary conditions, under the CFL condition $\lambda < 1$ for $\theta \leqslant 2\beta(1-\beta)$, and under the relaxed CFL condition $\lambda \leqslant \min(\frac{1}{\beta}, \frac{1}{1-\beta})$ for $\theta = \theta^{opt} = 2\beta(1-\beta)$, the scheme consisting of the Butcher tableaux (3.1) with the convex update (2.13) and the stages (2.4) is TVD.*

Note that for the proof of the Lemmata 2 and 3 we only used the positivity restrictions (3.4), (3.5) and (3.6), as well as the choice of the boundary conditions. This means that the TVD property will always hold under the exact same constraints as the L^∞ stability. Furthermore, the proof holds also for Neumann boundary conditions.

3.2 TVD reconstruction in space

To increase the accuracy of the spatial derivatives, we seek a second-order reconstruction of the point values w_j such that the resulting scheme is still L^∞ stable and TVD. We treat the explicit space derivatives before the implicit space derivatives.

Explicit space reconstruction. To obtain a second-order accurate approximation of the explicit spatial derivatives, we linearly reconstruct the values $w_j^{(k)}$ using the neighbouring point values, see for instance [24]. The reconstructed values $w_{j,-}^{(k)}$ and $w_{j,+}^{(k)}$ are then defined by

$$
\begin{cases}
w_{j,-}^{(k)} = w_j^{(k)} - \dfrac{\Delta x}{2} L\left(\sigma_{j+1/2}^{(k)}, \sigma_{j-1/2}^{(k)} \right), \\
w_{j,+}^{(k)} = w_j^{(k)} + \dfrac{\Delta x}{2} L\left(\sigma_{j-1/2}^{(k)}, \sigma_{j+1/2}^{(k)} \right),
\end{cases}
\tag{3.7}
$$

where $\sigma_{j+1/2}^{(k)}$ denotes the slope between the values of $w_j^{(k)}$ and $w_{j+1}^{(k)}$ given by

$$
\sigma_{j+1/2}^{(k)} = \frac{w_{j+1}^{(k)} - w_j^{(k)}}{\Delta x}.
$$

The function $L(\sigma_L, \sigma_R)$ is a slope limiter which should ensure that the reconstructed values still satisfy the maximum principle. For a three-point stencil the following estimate has to hold

$$
\min(|w_{j-1}^{(k)}|, |w_j^{(k)}|, |w_{j+1}^{(k)}|) \leqslant |w_{j,\pm}^{(k)}| \leqslant \max(|w_{j-1}^{(k)}|, |w_j^{(k)}|, |w_{j+1}^{(k)}|). \tag{3.8}
$$

A popular example of a second-order TVD slope limiter is the minmod limiter, defined for any two slopes σ_L and σ_R by

$$\text{minmod}(\sigma_L, \sigma_R) = \begin{cases} \min(\sigma_R, \sigma_L) & \text{if } \sigma_R > 0 \text{ and } \sigma_L > 0, \\ \max(\sigma_R, \sigma_L) & \text{if } \sigma_R < 0 \text{ and } \sigma_L < 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.9}$$

Using the reconstruction (3.7) and the notation $\Delta^{(k)}_{j,+} = w^{(k)}_{j,+} - w^{(k)}_{j-1,+}$, we write the stages and the update given in (3.2) as

$$\begin{aligned} w^{(2)}_j + \mu_\varepsilon a_{22} \Delta^{(2)}_j &= w^n_j - \lambda \tilde{a}_{21} \Delta^n_{j,+}, \\ w^{n+1}_j + \mu_\varepsilon \left((1-\theta) + \theta a_{33}\right) \Delta^{n+1}_j &= w^n_j - \lambda \left((1-\theta) + \theta \tilde{a}_{31}\right) \Delta^n_{j,+} \\ &\quad - \theta \left(\lambda \tilde{a}_{32} + \mu_\varepsilon a_{32}\right) \Delta^{(2)}_{j,+}. \end{aligned}$$

Due to the minmod limiting procedure, we immediately have from the estimate (3.8) that

$$\max_j \left|w^n_{j,+}\right| \leqslant \max_j \left|w^n_j\right| \quad \text{and} \quad \max_j \left|w^{(2)}_{j,+}\right| \leqslant \max_j \left|w^{(2)}_j\right|$$

for periodic boundary conditions. Using this estimates and following the analogue steps in the proofs of Lemma 2 and 3 it is easy to see, that under this reconstruction, the L^∞ stability and TVD property still hold.

Implicit space reconstruction. In the spirit of the reconstruction used to approximate the explicit derivatives, we could also increase the space accuracy of the implicit derivatives using TVD slope limiters. Note that the slopes are determined in general by a non-linear function, for example the minmod limiter (3.9). This would mean having to implicitly compute the reconstructed values (3.7). Such computations, if at all doable, would include an iterative process or a prediction correction method and therefore be extremely costly. We consider this increase in computational cost as too much in the sight of the actual gain in accuracy.

Another idea to approximate the implicit derivatives is the use of a Backward-Differencing-Formula (BDF), an implicit linear multi-step method [2]. For instance, the second-order BDF approximation reads

$$\frac{\partial w(x,t)}{\partial x} \approx \frac{1}{\Delta x}\left(3w_j - 4w_{j-1} + w_{j-2}\right), \tag{3.10}$$

while the third-order BDF approximation is given by

$$\frac{\partial w(x,t)}{\partial x} \approx \frac{1}{\Delta x}\left(\frac{11}{6}w_j - 3w_{j-1} + \frac{3}{2}w_{j-2} - \frac{1}{3}w_{j-3}\right). \tag{3.11}$$

Using the second-order BDF (3.10) in the first step of the scheme (3.2), we get

$$w^{(2)}_j + \mu_\varepsilon \frac{a_{22}}{2}\left(3w^{(2)}_j - 4w^{(2)}_{j-1} + w^{(2)}_{j-2}\right) = w^n_j - \lambda a_{22} \Delta^n_j.$$

Following the proof from Lemma 2, we have

$$\begin{aligned} \|w^n\|_\infty &\geqslant \max_j \left| \left(1 + \mu_\varepsilon \frac{3a_{22}}{2}\right) w^{(2)}_j - \mu_\varepsilon \frac{a_{22}}{2}\left(4w^{(2)}_{j-1} - w^{(2)}_{j-2}\right)\right| \\ &\geqslant \left(1 + \mu_\varepsilon \frac{3a_{22}}{2}\right) \|w^{(2)}\|_\infty - \mu_\varepsilon \frac{a_{22}}{2} \max_j \left|4w^{(2)}_{j-1} - w^{(2)}_{j-2}\right| \end{aligned}$$

To complete this step we need

$$\max_j \left| 4w_{j-1}^{(2)} - w_{j-2}^{(2)} \right| \leqslant 4\|w^{(2)}\|_\infty - \|w^{(2)}\|_\infty \tag{3.12}$$

which is a contradiction to the inverse triangular equation. Therefore using a second-order BDF will not lead to a TVD scheme. We can even extend this observation to a BDF of general order. As it is derived to match the Taylor series expansion up to an order p, its general form has alternating signs, and it can be written using $p+1$ coefficients $\kappa_i \geqslant 0$, $i = 0, \dots, p$, as in [2]

$$\frac{\partial w(x,t)}{\partial x} \approx \kappa_0 w_j - \kappa_1 w_{j-1} + \kappa_2 w_{j-2} - \dots + \kappa_p w_{j-p} \tag{3.13}$$

for an approximation of order p, where we have taken an even p for the moment. We use the BDF described by (3.13) for the approximation of the implicit space derivative, and we find in the estimate for the L^∞ stability:

$$\begin{aligned}
\|w^n\|_\infty &\geqslant \max_j \left| (1 + \mu_\varepsilon a_{22}\kappa_0)\, w_j^{(2)} - \mu_\varepsilon a_{22} \left(\kappa_1 w_{j-1}^{(2)} - \kappa_2 w_{j-2}^{(2)} + \kappa_3 w_{j-3}^{(2)} - \dots - \kappa_m w_{j-m}^{(2)} \right) \right| \\
&\geqslant (1 + \mu_\varepsilon a_{22}\kappa_0)\, \|w^{(2)}\|_\infty - \mu_\varepsilon a_{22} \max_j \left| \kappa_1 w_{j-1}^{(2)} - \kappa_2 w_{j-2}^{(2)} + \kappa_3 w_{j-3}^{(2)} - \dots - \kappa_m w_{j-m}^{(2)} \right| \\
&\geqslant (1 + \mu_\varepsilon a_{22}\kappa_1)\, \|w^{(2)}\|_\infty - \mu_\varepsilon a_{22} \max_j \left| \kappa_1 w_{j-1}^{(2)} - \kappa_2 w_{j-2}^{(2)} \right| - \dots \\
&\qquad - \mu_\varepsilon a_{22} \max_j \left| \kappa_{p-1} w_{j-p-1}^{(2)} - \kappa_p w_{j-p}^{(2)} \right|.
\end{aligned}$$

Analogously to (3.12), to achieve the right estimate, the inverse triangular inequality would be violated. The case of an odd p also fails.

This shows that treating the implicit spacial derivative with BDF is not an option here and, as will be seen in the numerical experiments, using the BDF approximation alone immediately leads to oscillatory solutions. Therefore, we keep the first-order upwind approximation of the implicit spatial derivatives. This is a loss of accuracy we are willing to take to obtain a TVD scheme, as due to the convex combination with the first-order scheme, the TVD scheme is only first-order accurate anyway.

We summarize the results of this section in the following result:

Theorem 4. *For $\beta \in (0,1)$ and periodic boundary conditions, the scheme consisting of the Butcher tableaux (3.1) with the convex update (2.13) and the stages (2.4), combined with the reconstruction procedure given by (3.7) and (3.9), is L^∞ stable and TVD under the CFL condition $\lambda < 1$ for $\theta \leqslant 2\beta(1-\beta)$, and the relaxed CFL condition $\lambda \leqslant \min(\frac{1}{\beta}, \frac{1}{1-\beta})$ for $\theta = \theta^{opt} = 2\beta(1-\beta)$.*

4 Extension to higher order tableaux

We start the construction of schemes using higher order tableaux by investigating the natural extension of the TVD scheme using third-order tableaux instead of second-order ones. To reduce the number of computational steps, we once again assume that the weights $\tilde{b}$ and b respectively coincide with the last rows of $\tilde{A}$ and A. Further, we assume $\tilde{c} = c$. The Butcher tableaux are given by

$$\text{explicit:}\quad \begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ c_2 & \tilde{a}_{21} & 0 & 0 & 0 \\ c_3 & \tilde{a}_{31} & \tilde{a}_{32} & 0 & 0 \\ c_4 & \tilde{a}_{41} & \tilde{a}_{42} & \tilde{a}_{43} & 0 \\ \hline & \tilde{a}_{41} & \tilde{a}_{42} & \tilde{a}_{43} & 0 \end{array}, \qquad \text{implicit:}\quad \begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ c_2 & 0 & a_{22} & 0 & 0 \\ c_3 & 0 & a_{32} & a_{33} & 0 \\ c_4 & 0 & a_{42} & a_{43} & a_{44} \\ \hline & 0 & a_{42} & a_{43} & a_{44} \end{array}. \tag{4.1}$$

Applying the third-order conditions given in Table 1 and (2.6) on the scheme given by (4.1) leads to the following tableaux, with $\gamma \neq \{\frac{1}{2}, \frac{2}{3}\}$:

$$\text{explicit:}\quad \begin{array}{c|ccccc} 0 & 0 & 0 & 0 & 0 \\ \frac{2-3\gamma}{3-6\gamma} & \frac{2-3\gamma}{3-6\gamma} & 0 & 0 & 0 \\ \gamma & \gamma - 2(3\gamma^2-3\gamma+1)\frac{2\gamma-1}{3\gamma-2} & 2(3\gamma^2-3\gamma+1)\frac{2\gamma-1}{3\gamma-2} & 0 & 0 \\ 1 & 0 & 1-\frac{1}{4(3\gamma^2-3\gamma+1)} & \frac{1}{4(3\gamma^2-3\gamma+1)} & 0 \\ \hline & 0 & 1-\frac{1}{4(3\gamma^2-3\gamma+1)} & \frac{1}{4(3\gamma^2-3\gamma+1)} & 0 \end{array} \tag{4.2}$$

$$\text{implicit:}\quad \begin{array}{c|ccccc} 0 & 0 & 0 & 0 & 0 \\ \frac{2-3\gamma}{3-6\gamma} & 0 & \frac{2-3\gamma}{3-6\gamma} & 0 & 0 \\ \gamma & 0 & 2\gamma-1 & 1-\gamma & 0 \\ 1 & 0 & 1-\frac{1}{4(3\gamma^2-3\gamma+1)} & \frac{1}{4(3\gamma^2-3\gamma+1)} & 0 \\ \hline & 0 & 1-\frac{1}{4(3\gamma^2-3\gamma+1)} & \frac{1}{4(3\gamma^2-3\gamma+1)} & 0 \end{array} \tag{4.3}$$

We now derive conditions on $\gamma \neq \{\frac{1}{2}, \frac{2}{3}\}$, $\lambda > 0$ and $\theta \in [0,1]$ such that the scheme given by (4.2), (4.3) and the convex combination with the first-order scheme (2.12) is L^∞ stable and TVD. From the first stage, we have $w^{(1)} = w^n$. The second stage with $c_2 = \frac{2-3\gamma}{3-6\gamma}$ is given by

$$w_j^{(2)} + \mu_\varepsilon c_2 \Delta_j^{(2)} = (1-\lambda c_2) w_j^n - \lambda c_2 w_{j-1}^n.$$

Following the proof of Lemmata 2 and 3, we require, in the fashion of (3.4)-(3.6):

$$\begin{aligned} c_2 > 0 &\iff \frac{2-3\gamma}{3-6\gamma} > 0 \iff \gamma < \frac{1}{2} \text{ or } \gamma > \frac{2}{3}, \\ 1-\lambda c_2 \geqslant 0 &\iff \lambda \leqslant \frac{1}{c_2} \iff \lambda \leqslant \frac{3-6\gamma}{2-3\gamma}. \end{aligned} \tag{4.4}$$

Note that the expressions in (4.4) are well-defined. The third stage, using

$$-\mu_\varepsilon \Delta_j^{(2)} = \frac{1}{c_2}(w_j^{(2)} - w_j^n) + \lambda \Delta_j^n,$$

is given by

$$\begin{aligned} w_j^{(3)} + \mu_\varepsilon a_{33} \Delta_j^{(3)} = &\left(1 - \frac{a_{32}}{c_2} - \lambda \tilde{a}_{31} + \lambda a_{32}\right) w_j^n + \lambda\left(\tilde{a}_{31} - a_{32}\right) w_{j-1}^n \\ &+ \left(\frac{a_{32}}{c_2} - \lambda \tilde{a}_{32}\right) w_j^{(2)} + \lambda \tilde{a}_{32} w_{j-1}^{(2)}. \end{aligned}$$

This leads to the following inequalities

$$\begin{aligned} a_{33} > 0 &\iff 1-\gamma > 0 \iff \gamma < 1, \\ \lambda \tilde{a}_{32} \geqslant 0 &\iff (3\gamma^2-3\gamma+1)\frac{2\gamma-1}{3\gamma-2} \geqslant 0 \iff \gamma \leqslant \frac{1}{2} \text{ or } \gamma \geqslant \frac{3}{2}, \\ \tilde{a}_{31} - a_{32} \geqslant 0 &\iff -\frac{\gamma(12\gamma^2-15\gamma+5)}{3\gamma-2} \geqslant 0 \iff \gamma \geqslant 0 \text{ and } \gamma < \frac{2}{3}. \end{aligned}$$

At this level, a temporary estimate for γ is $\gamma \in [0, \frac{1}{2})$. The next requirement is given by

$$\frac{a_{32}}{c_2} - \tilde{a}_{32}\lambda \geqslant 0 \iff -2(3\gamma^2-3\gamma+1)\lambda - 3(1-2\gamma) \geqslant 0. \tag{4.5}$$

This leads to a new condition on λ. We already found that $\gamma \in [0, \frac{1}{2})$, thus $3\gamma^2 - 3\gamma + 1 > 0$ and $1 - 2\gamma < 0$. Therefore the inequality in (4.5) cannot be fulfilled for a non-negative λ and consequently it is not possible to achieve the TVD property following the proof of Lemma 2 and Lemma 3. In the following, we propose a method to cure this defect and still keep the easy way of proving the TVD property.

4.1 Method of convex stages

As we have seen, the attempt to prove the L^∞ stability and TVD property already failed at the second step, while the convex combination with the first-order scheme is only applied on the final update. Therefore we propose a convex combination of each stage with a first-order update at time $t^n + c_k \Delta t$ for the k-th stage. To have only one time level, we set $\tilde{c} = c$. This framework allows for more free parameters $\theta_k \in [0, 1]$, where $k = 1, \dots, s$ denotes the stage in the IMEX scheme. To have the best precision possible, the goal is to choose as many θ_k as possible equal to one. Analogously to the convex update (2.13), the stages are given by

$$w_j^{(k)} + (1-\theta_k) c_k \mu_\varepsilon \Delta_j^{(k)} = w_j^n - \lambda \left((1-\theta_k) \tilde{c}_k \Delta_j^n + \theta_k \sum_{l=1}^{k-1} \tilde{a}_{kl} \Delta_j^{(l)} \right) - \mu_\varepsilon \theta_k \sum_{l=1}^{k} a_{kl} \Delta_j^{(l)}. \tag{4.6}$$

Note that, for the Butcher tableaux in the manner of (4.1), we immediately set $\theta_1 = 1$ to recover $w^{(1)} = w^n$. This means the convex stages appear earliest for $k = 2$. In the case where the weights $\tilde{b}$ and b respectively coincide with the last row of $\tilde{A}$ and A, the stage $w^{(s)}$ coincides with the final update w^{n+1}. In particular, we then have $\theta = \theta_s$.

In the spirit of the results from the second-order scheme, we seek a general framework on how to obtain TVD schemes with s stages using the IMEX formulation (2.13) – (4.6). Since the proof follows analogue steps as in Lemmata 2 and 3, we do not repeat the calculations and we directly give the final result.

Theorem 5. *Let $\tilde{A}, A \in \mathbb{R}^{s \times s}$, $\tilde{b}, b, \tilde{c}, c \in \mathbb{R}^s$ define two Butcher tableaux (2.5) fulfilling (2.6) and the p-th order compatibility conditions. Let $\tilde{b}$ and b coincide with the last rows of $\tilde{A}$ and A respectively, and let $\tilde{c} = c$. For $k = 1, \dots, s$ and $l = 1, \dots, k-1$, we define*

$$\mathcal{A}_k = \theta_k a_{kk} + (1-\theta_k) c_k, \quad \tilde{\mathcal{A}}_k = \theta_k a_{k1} + (1-\theta_k) c_k, \quad \mathcal{B}_{kl} = \frac{\theta_k a_{kl}}{\mathcal{A}_l}, \quad \tilde{\mathcal{B}}_{kl} = \theta_k \tilde{a}_{kl}.$$

In addition, we recursively define the following expressions:

$$\mathcal{C}_k = \tilde{\mathcal{A}}_k - \sum_{l=2}^{k-1} \mathcal{B}_{kl} \mathcal{C}_l, \qquad \mathcal{C}_{kl} = \tilde{\mathcal{B}}_{kl} - \sum_{r=l+1}^{k-1} \mathcal{B}_{kr} \mathcal{C}_{rl},$$

$$\mathcal{D}_k = 1 - \lambda \tilde{\mathcal{A}}_k - \sum_{l=2}^{k-1} \mathcal{B}_{kl} \mathcal{D}_l, \qquad \mathcal{D}_{kl} = \mathcal{B}_{kl} - \lambda \tilde{\mathcal{B}}_{kl} - \sum_{r=l+1}^{k-1} \mathcal{B}_{kr} \mathcal{D}_{rl}.$$

Then, with $\theta_1 = 1$ and under the following restrictions for $k = 1, \dots, s$ and $l = 1, \dots, k-1$,

$$\mathcal{A}_k > 0, \quad \mathcal{C}_k \geqslant 0, \quad \mathcal{D}_k \geqslant 0, \quad \mathcal{C}_{kl} \geqslant 0, \quad \mathcal{D}_{kl} \geqslant 0.$$

the scheme consisting of the stages (4.6) and the update (2.13), combined with a TVD limiter, is L^∞ stable and TVD under a CFL condition determined by $\lambda \geqslant 0$ where λ does not depend on ε.

We wish to remark that the obtained p-th order tableaux do not necessarily lead to stable schemes by themselves if they are not combined with the convex strategy. This is not a drawback since our goal is the L^∞ stability. For studies on A- or L-stability, we refer to [29].

The result from Theorem 5 can be extended to the case where the weights $\tilde{b}$ and b do not coincide with the respective last rows of $\tilde{A}$ and A. To be able to use the notation from Theorem 5, we view the update (2.13) as an additional explicit $(s+1)$-th stage of a scheme induced by Butcher tableaux (2.5) with $(s+1)\times(s+1)$ matrices with the diagonal entry $a_{s+1,s+1}=0$, where the weights $\tilde{b}$ and b respectively coincide with the last rows of the new $\tilde{A}$ and A. Then we define the convex parameter of the last stage as $\theta_{s+1}=\theta$. Theorem 5 is then applied to yield the L^∞ stability and the TVD property.

4.2 L^∞ stable and TVD scheme based on third-order tableaux

We now demonstrate that, with this method, a TVD scheme can be obtained based on the previous Butcher tableaux (4.2) – (4.3).

Determination of the third-order Butcher tableaux. Let us introduce one additional parameter $\theta_3\neq 1$, while keeping $\theta_1=\theta_2=1$. This means that we have the same stages for $w^{(1)}$ and $w^{(2)}$ as before. We recall that we obtained from the second stage $\gamma<\frac{1}{2}$ or $\gamma>\frac{2}{3}$ and $\lambda\leqslant\frac{3(1-2\gamma)}{2-3\gamma}$. Using the definition of the third stage given in (4.6), we now have with $w^{(1)}=w^n$:

$$\begin{aligned} w_j^{(3)}+\mu_\varepsilon\left((1-\theta_3)c_3+\theta_3 a_{33}\right)\Delta_j^{(3)} &= w_j^n-\lambda\left((1-\theta_3)c_3\Delta^n+\theta_3 c_3\left(\tilde{a}_{31}\Delta_j^n+\tilde{a}_{32}\Delta^{(2)}\right)\right)\\ &\quad+\theta_3 c_3 a_{32}\left(\frac{1}{c_2}(w_j^{(2)}-w_j^n)+\lambda\Delta_j^n\right)\\ &\Longleftrightarrow\\ w_j^{(3)}+\mu_\varepsilon c_3\left(c_3+\theta_3(a_{33}-c_3)\right)\Delta_j^{(3)} &= \left(1-\lambda(1-\theta_3)c_3-\theta_3\lambda(\tilde{a}_{31}-a_{32})-\frac{\theta_3 a_{32}}{c_2}\right)w_j^n\\ &\quad+\left(\lambda(1-\theta_3)c_3+\theta_3\lambda(\tilde{a}_{31}-a_{32})\right)w_{j-1}^n\\ &\quad+\theta_3\left(\frac{a_{32}}{c_2}-\lambda\tilde{a}_{32}\right)w_j^{(2)}+\theta_3\,\lambda\tilde{a}_{32}w_{j-1}^{(2)}. \end{aligned}$$

As in the previous case, we obtain

$$\lambda\tilde{a}_{32}\geqslant 0\iff\gamma\leqslant\frac{1}{2}\text{ or }\gamma\geqslant\frac{3}{2}.$$

For the requirement (4.5) that caused problems earlier, we choose from now on $\gamma>\frac{2}{3}$ and therewith $\lambda\leqslant\frac{3(2\gamma-1)}{2(3\gamma^2-3\gamma+1)}$. This choice is not in conflict with the coefficient in front of w_{j-1}^n as before and leads now to a restriction on θ_3 instead of on γ. We have

$$\begin{aligned} (1-\theta_3)c_3+\theta_3(\tilde{a}_{31}-a_{32})\geqslant 0 &\iff 3\theta_3(2\gamma-1)^2\leqslant 3\gamma-2\\ &\iff\theta_3\leqslant\frac{3\gamma-2}{3(2\gamma-1)^2}. \end{aligned}\tag{4.7}$$

The next restriction gives another estimate on θ_3, as follows:

$$\begin{aligned} c_3+\theta_3(a_{33}-c_3)\geqslant 0 &\iff\gamma-\theta_3(2\gamma-1)\geqslant 0\\ &\iff\theta_3\leqslant\frac{\gamma}{2\gamma-1}. \end{aligned}$$

It is easy to see that this condition on θ_3 is less restrictive than the one obtained from (4.7) for all $\gamma > \frac{2}{3}$. For a given γ, the largest value we can take for θ_3 is therefore given by

$$\theta_3^{\text{opt}} = \frac{3\gamma - 2}{3(2\gamma - 1)^2},$$

and θ_3 must satisfy $\theta_3 \leqslant \theta_3^{\text{opt}}$. The last restriction for the third stage is given by

$$1 - \lambda(1 - \theta_3)c_3 - \theta_3\lambda(\tilde{a}_{31} - a_{32}) - \frac{\theta_3 a_{32}}{c_2} \geqslant 0. \tag{4.8}$$

This condition is always fulfilled if we choose $\theta_3 = \theta_3^{\text{opt}}$. In doing so, we have the maximal allowed input from the original stages (2.4). Otherwise, (4.8) leads to another, more restrictive estimate for λ. We repeat this procedure for the last stage. We skip the lengthy but straightforward computations and give the final estimates on the free parameters $\gamma, \lambda, \theta_3$ and θ_4 directly in Corollary 6.

Explicit space reconstruction. To increase the space accuracy of the scheme, we use a TVD third-order space reconstruction satisfying (3.8). This merely amounts to setting the function L in the space reconstruction described in Section 3.2. We choose the third-order limiting procedure introduced in [32]. This procedure switches between the oscillatory non-limited third-order reconstruction and a third-order TVD limiter. Switching to the TVD limiter is triggered in the event where a non-physical oscillation represented by a non-smooth extremum is detected.

Here, we recall the expression of this slope limiter. For any two slopes σ_L and σ_R, define the third-order slope limiter with smoothness detection

$$L_3^{l,s}(\sigma_L, \sigma_R) = \begin{cases} L_3(\sigma_L, \sigma_R) & \text{if } \eta(\sigma_L, \sigma_R) < 1, \\ L_3^l(\sigma_L, \sigma_R) & \text{otherwise.} \end{cases} \tag{4.9}$$

In (4.9), we have introduced

- the unlimited third-order slope reconstruction L_3, defined by

$$L_3(\sigma_L, \sigma_R) = \frac{1}{3}(\sigma_L + 2\sigma_R); \tag{4.10}$$

- the third-order TVD limiter L_3^l, defined by

$$L_3^l(\sigma_L, \sigma_R) = \begin{cases} \max\left(0, \min\left(L_3(\sigma_L, \sigma_R), 2\sigma_L, \frac{3}{2}\sigma_R\right)\right) & \text{if } \sigma_R > 0 \text{ and } \sigma_L > 0, \\ \max\left(0, \min\left(L_3(\sigma_L, \sigma_R), -\sigma_L\right)\right) & \text{if } \sigma_R > 0 \text{ and } \sigma_L < 0, \\ \min\left(0, \max\left(L_3(\sigma_L, \sigma_R), -\sigma_L\right)\right) & \text{if } \sigma_R < 0 \text{ and } \sigma_L > 0, \\ \min\left(0, \max\left(L_3(\sigma_L, \sigma_R), 2\sigma_L, \frac{3}{2}\sigma_R\right)\right) & \text{if } \sigma_R < 0 \text{ and } \sigma_L < 0. \end{cases}$$

- a smoothness indicator η, defined by

$$\eta(\sigma_L, \sigma_R) = \frac{1}{\alpha}\sqrt{\frac{2}{5}}\sqrt{\sigma_L^2 + \sigma_R^2},$$

where $\alpha = \max_{x \in \Omega_s} |(w^0)''(x)|$ is the maximum value of second derivative of the initial condition w^0 on the domain $\Omega_s \subset \Omega$. Here, Ω_s denotes the set of points in Ω where the second derivative of the initial condition $w^0(x)$ is defined.

Since the limiter given in (4.9) is provably TVD according to [32], we apply Theorem 5 and immediately find the following result

Corollary 6. *The scheme consisting of the Butcher tableaux (4.2), (4.3), with the stages given in (4.6), the update in (2.13), and combined with the slope limiter (4.9), is* L^∞ *stable and TVD according to Theorem 5 with the following choice of parameters*

$$\gamma \geqslant \frac{3+\sqrt{3}}{6}, \quad \theta_1 = 1, \quad \theta_2 = 1, \quad \theta_3 = \frac{3\gamma-2}{3(2\gamma-1)^2}, \quad \theta_4 < \frac{4(3\gamma-2)(3\gamma^2-3\gamma+1)}{9(2\gamma-1)^3},$$

and under the CFL condition

$$\lambda \leqslant \frac{9(2\gamma-1)^3\theta_4 - 4(3\gamma-2)(3\gamma^2-3\gamma+1)}{(3\gamma-2)((24\gamma^2-24\gamma+7)\theta_4 - 4(3\gamma^2-3\gamma+1))}.$$

An analysis of the influence of the choice of the parameters $\gamma, \theta_3, \theta_4$ and λ will be conducted in the Section 5.2. Especially the balance between CPU time, i.e. the value of λ, and precision, expressed by the values of θ_3 and θ_4, will be discussed.

5 Numerical results

In this last section, we illustrate the capabilities of the schemes we have developed in Sections 3 and 4. To help referring to these methods, we introduce the following abbreviations.

- The IMEXp scheme denotes the scheme with an p-th order time discretisation and an p-th order space discretisation. Following this notation, the IMEX1 scheme is given by (2.12), the IMEX2 scheme corresponds to the Butcher tableaux (3.1), and the IMEX3 scheme corresponds to the Butcher tableaux (4.2) and (4.3). The second-order unlimited space discretisation (3.7) with $L(\sigma_L, \sigma_R) = \frac{1}{2}(\sigma_L + \sigma_R)$ is applied to the explicit part of the IMEX2 scheme, while the second-order BDF (3.10) is applied to its implicit part. The third-order unlimited space discretisation (3.7) with $L(\sigma_L, \sigma_R) = L_3(\sigma_L, \sigma_R)$ given by (4.10) is applied to the explicit part of the IMEX3 scheme, while the third-order BDF (3.11) is applied to its implicit part.
- The TVDp scheme is the TVD scheme constructed from the IMEXp tableau. The TVD2 scheme is obtained following Theorem 4, and the TVD3 scheme is given in Corollary 6.

For the remainder of this section, we consider several numerical experiments, with some common characteristics. In each experiment, we prescribe periodic boundary conditions, and we take $c_m = 1$ and $c_a = 1$. The value of the fast transport velocity therefore is $1/\varepsilon$. The values of ε will vary throughout the experiments to highlight how the results depend on ε. The space-time domain is taken such that the solution revolves exactly once with the periodic boundary conditions, i.e. we take the final time $t_{end} = 1$ and space domain $(0, c_m + \frac{c_a}{\varepsilon})$.

We introduce two exact solutions to Equation (2.1), which will help us demonstrate the properties of the schemes. First, we give a smooth solution $w^s(t,x)$ defined by

$$w^s(t,x) = 1 + \frac{\varepsilon}{2}\left(1 + \sin\left[2\pi\varepsilon\left(x - \left(c_m + \frac{c_a}{\varepsilon}\right)t\right)\right]\right), \tag{5.1}$$

which represents a sine function of amplitude ε, transported with the velocity $c_m + \frac{c_a}{\varepsilon}$. Second, a discontinuous solution $w^d(t,x)$ is given by

$$w^d(t,x) = \begin{cases} 1+\varepsilon & \text{if } \frac{1}{4} < \left(\frac{\left(x - \left(c_m + \frac{c_a}{\varepsilon}\right)t\right)}{c_m + \frac{c_a}{\varepsilon}} - \left\lfloor \frac{\left(x - \left(c_m + \frac{c_a}{\varepsilon}\right)t\right)}{c_m + \frac{c_a}{\varepsilon}} \right\rfloor\right) < \frac{3}{4} \\ 1 & \text{otherwise,} \end{cases} \tag{5.2}$$

which represents a rectangular bump of amplitude ε initially located in the space region $(\frac{1}{4}(c_m + \frac{c_a}{\varepsilon}), \frac{3}{4}(c_m + \frac{c_a}{\varepsilon}))$, transported with the velocity $c_m + \frac{c_a}{\varepsilon}$. These exact solutions will be taken as initial conditions by setting $t = 0$.

In the remainder of this paper, we first introduce in Section 5.1 a MOOD procedure to increase the precision of the TVDp scheme. Then, we study in Section 5.2 the influence of the free parameters in the schemes from Sections 3 and 4 on the precision and computational time. Finally, after having fixed the parameters, we study the behaviour of these schemes in Section 5.3 compared to IMEX schemes from the literature, when considering smooth and discontinuous solutions, for a wide range of ε. More specifically, we study the order of accuracy on smooth solutions, as well as the overshoot and undershoot magnitude on discontinuous ones.

5.1 Optimal order detection: MOOD-inspired procedure

The goal of this section is to introduce a MOOD-like procedure to increase the precision of the TVDp scheme without degrading its stability properties. The usual MOOD framework for explicit schemes, see e.g. [11], consists in locally and gradually lowering the order of the scheme when an oscillation is detected. In our IMEX case, the non-local nature of the implicit part prevents us from only recomputing the approximate solution on a few selected cells, and the solution has to be recomputed on the whole mesh. To avoid a prohibitive increase in the computation time, we instead suggest to directly use the TVDp scheme on the whole mesh as soon as an oscillation is detected in some cell. In addition, we state that an oscillation has been detected if the approximate solution does not satisfy the bounds of the initial condition.

This implicit MOOD framework is summarized in the following algorithm, which has also been stated in [13, 27].

Algorithm 7 (MOODp scheme)**.** *Equipped with the stable TVD*p *scheme, the MOOD*p *scheme consists in applying the following procedure at each time step:*

1. *Compute a candidate numerical solution* w_c^{n+1} *with the IMEX*p *scheme.*
2. *Detect whether an oscillation is present somewhere in the space domain, that is to say detect whether the discrete maximum principle is satisfied by the candidate solution:*

$$\|w_c^{n+1}\|_\infty \leqslant \|w^0\|_\infty. \tag{DMP}$$

3a. *If* (DMP) *holds, then set the numerical solution* w^{n+1} *equal to the candidate solution* w_c^{n+1}.

3b. *Otherwise, compute the numerical solution* w^{n+1} *with the* L^∞ *stable TVD*p *scheme.*

Applied at each time step, the procedure described in Algorithm 7 ensures that the numerical solution satisfies the maximum principle, i.e. $\|w^{n+1}\|_\infty \leqslant \|w^0\|_\infty$ for all $n \geqslant 0$.

5.2 Choice of the free parameters

We start these numerical experiments by suggesting optimal values of the free parameters in the schemes from Sections 3 and 4. To that end, we analyse the error produced by the schemes, as well as the CPU time taken, with respect to the free parameters. This analysis will help us give some insights on how to optimally choose these parameters, and on the trade-offs that must be made when making such choices.

Here, we study the effect of the time discretisation on the precision and computational time of our schemes. Therefore, we temporarily restrict ourselves to a first-order discretisation is space, in order to make sure only the effects of the time discretisation are studied. We compare the IMEX1 scheme

to the IMEX2, TVD2 and MOOD2 schemes in Section 5.2.1, and to the IMEX3, TVD3 and MOOD3 schemes in Section 5.2.2. In both cases, we set $\varepsilon = 0.1$ and we take $N = 400$ discretisation points, and the smooth exact solution (5.1) is considered. The conclusions of the forthcoming developments are unchanged if we consider another value of ε. Indeed, taking another ε would merely translate the curves without changing their relative positioning.

5.2.1 Choice of β in the TVD2 scheme

We consider the TVD2 scheme. According to Lemma 3, we can freely choose $\beta \in (0,1)$ and get a TVD scheme as long as $\theta = 2\beta(1-\beta)$ and $\lambda = \min(\frac{1}{\beta}, \frac{1}{1-\beta})$. These two quantities are displayed in Figure 2. We observe that, to maximize both θ and λ, one may be tempted to take $\beta = \frac{1}{2}$. In this case, the Butcher tableaux (3.1) degenerate to the Butcher tableaux of the ARS (1,2,2) midpoint scheme, see [3], and we get $\theta = \frac{1}{2}$, $\lambda = 2$. With these settings, the TVD2 scheme exactly reverts to two steps of the IMEX1 scheme, and we expect a loss of accuracy. Therefore, to base the TVD2 scheme on a truly second-order IMEX2 scheme, we have to take $\beta \neq \frac{1}{2}$.

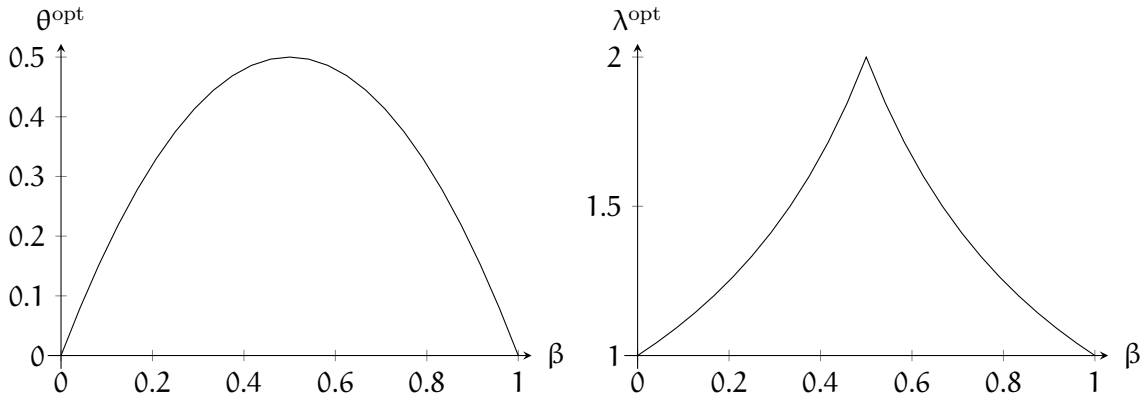

Figure 2: Values of the optimal convex combination parameter θ^{opt} (left panel) and the optimal CFL number λ^{opt} (right panel), with respect to the IMEX parameter β of the TVD2 scheme.

Let us now study the impact of the choice of β on the precision and speed of the numerical scheme. This study was partially performed, for $\beta < \frac{1}{2}$, in [27]. First, we check the CPU time taken by the four schemes with respect to β. We expect the CPU time taken by the IMEX1 and IMEX2 schemes not to be influenced by the value of β, unless $\beta = \frac{1}{2}$, since the midpoint scheme has only one implicit step. However, the CFL condition of the TVD2 and MOOD2 schemes is influenced by β, and we expect these two schemes to take more computational time when β is far from $\frac{1}{2}$. These observations are confirmed by Figure 3. We also observe that the MOOD2 scheme presents a sharp increase in CPU time around $\beta = 0.52$. This is due to the fact that the IMEX2 scheme is very unstable in this region of β, thus leading to more MOOD loops needed to correct its stability shortcomings.

These stability issues of the IMEX2 scheme are made apparent on the left panel of Figure 4, where we display the L^∞-error of the four schemes with respect to β. There, we observe that the L^∞-error of the IMEX2 scheme explodes around $\beta = 0.52$, even for this smooth solution. Decreasing the CFL condition improves this behaviour, without curing it completely. This highlights the need to use a stable scheme, such as the TVD2 scheme or the MOOD2 scheme. Furthermore, still in the left panel, we observe that the error of both the IMEX2 and the MOOD2 scheme increase sharply when $\beta > \frac{1}{2}$. Therefore, it seems sensible to restrict this study to $\beta < \frac{1}{2}$. In the right panels of Figure 4, we display zooms of the left panel error data for $\beta < \frac{1}{2}$. In the top right panel, we observe that the error of the TVD2 scheme reaches a minimum around $\beta = 0.3$; in the bottom right panel, we observe that the error of the MOOD2 scheme starts increasing around $\beta = 0.3$.

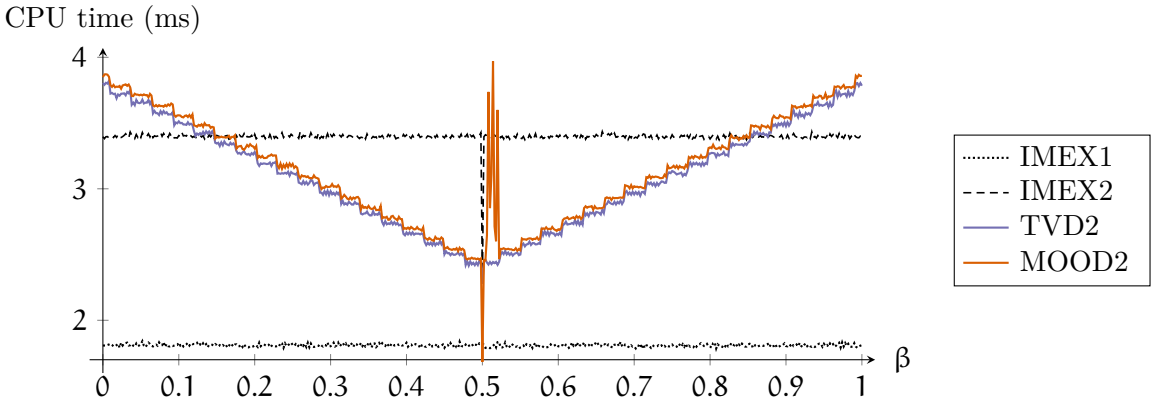

Figure 3: CPU time (in milliseconds) with respect to the IMEX parameter β, using the optimal values θ^{opt} and λ^{opt}, in the context of the test case presented in Section 5.2.1.

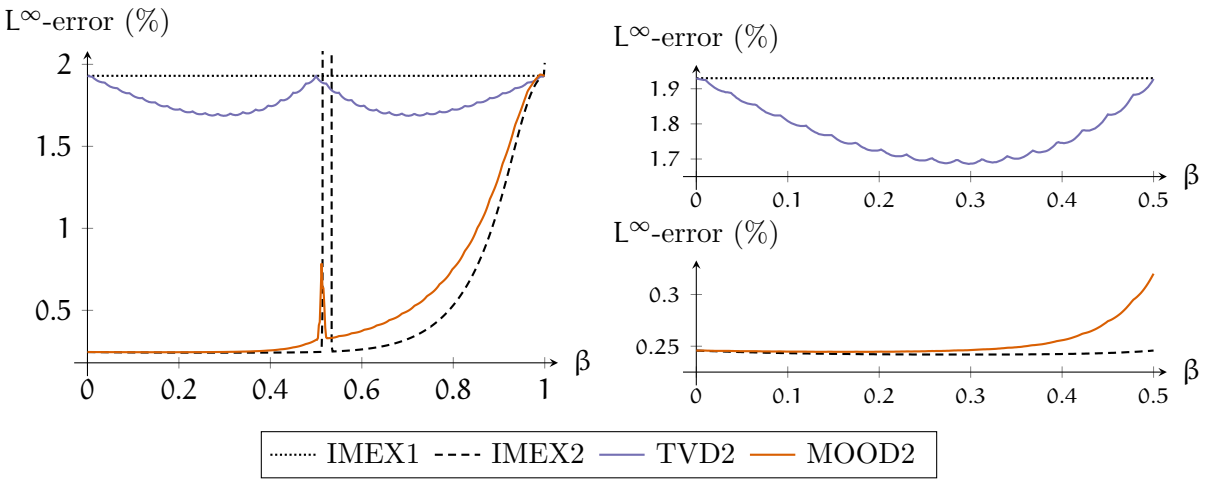

Figure 4: L^∞-error with respect to the IMEX parameter β, using the optimal values θ^{opt} and λ^{opt}, in the context of the test case presented in Section 5.2.1. The right panels contain a zoom on the left panel data, for $\beta \in [0, \frac{1}{2}]$.

Therefore, according to Figures 3 and 4, taking $\beta \simeq 0.3$ seems like a good compromise between error and CPU time taken. In fact, taking a finer β grid, we see that the error reaches a minimum for $\beta^{\text{opt}} = 1 - \frac{\sqrt{2}}{2}$, and we suggest taking this value as the optimal value of β for the IMEX2 scheme. This leads to the well-known ARS(2,2,2) scheme (see for instance [3, 29]), which incidentally was the base second-order scheme used to derive a TVD IMEX scheme in [13]. The Butcher tableaux (3.1) then become

$$\text{explicit:} \quad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1-\frac{\sqrt{2}}{2} & 1-\frac{\sqrt{2}}{2} & 0 & 0 \\ 1 & -\frac{\sqrt{2}}{2} & 1+\frac{\sqrt{2}}{2} & 0 \\ \hline & -\frac{\sqrt{2}}{2} & 1+\frac{\sqrt{2}}{2} & 0 \end{array}, \qquad \text{implicit:} \quad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1-\frac{\sqrt{2}}{2} & 0 & 1-\frac{\sqrt{2}}{2} & 0 \\ 1 & 0 & \frac{\sqrt{2}}{2} & 1-\frac{\sqrt{2}}{2} \\ \hline & 0 & \frac{\sqrt{2}}{2} & 1-\frac{\sqrt{2}}{2} \end{array}.$$

For the remainder of this article, we take

$$\beta = \beta^{\text{opt}} = 1 - \frac{\sqrt{2}}{2}.$$

5.2.2 Choice of γ and θ_4 in the TVD3 scheme

Now, regarding the TVD3 scheme, we have to set the values of θ_3, θ_4 and λ, constrained by Corollary 6. In this scheme, we have an optimal value of θ_3 given as a function of γ. However, we do not have one single optimal value of θ_4 and λ. Instead, we have an upper bound for these two quantities whose value depends on γ. Ideally, we would like θ_3, θ_4 and λ to be as large as possible. By inspection, we note that the maximum value of θ_3 is $\theta_3^{\text{opt}} = \frac{3}{8}$, obtained for $\gamma^{\text{opt}} = \frac{5}{6}$. The Butcher tableaux (4.2) and (4.3) then become

$$\text{explicit:}\quad \begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 1/4 & 1/4 & 0 & 0 & 0 \\ 5/6 & -13/18 & 14/9 & 0 & 0 \\ 1 & 0 & 4/7 & 3/7 & 0 \\ \hline & 0 & 4/7 & 3/7 & 0 \end{array}, \qquad \text{implicit:}\quad \begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 1/4 & 0 & 1/4 & 0 & 0 \\ 5/6 & 0 & 2/3 & 1/6 & 0 \\ 1 & 0 & 4/7 & 3/7 & 0 \\ \hline & 0 & 4/7 & 3/7 & 0 \end{array}. \tag{5.3}$$

Taking this value of γ in Corollary 6 yields the following bounds

$$0 < \theta_4 < \frac{7}{16} \quad \text{and} \quad 0 < \lambda < \frac{7 - 16\theta_4}{7 - 11\theta_4}. \tag{5.4}$$

These expressions lead us to formulating the following remark, which highlights the trade-off we need to operate between the value of θ_4 and that of λ.

Remark 8. *From (5.4), we note that λ is a decreasing function of θ_4, which implies that we are not able to use both a large θ_4 and a large λ. Therefore, there appears a trade-off between the CFL condition, which influences the CPU time taken by the scheme, and the convex combination parameter, which influences the precision of the scheme. Either we take a large λ and a small θ_4, to lower the CPU time but decrease the precision, or we take a large θ_4 and a small λ, to improve the precision but increase the CPU time. This remark also holds for any value of γ, by inspection of the formulas in Corollary 6.*

Now, let us quantify this balance between precision and CPU time. To address this issue, let us introduce $\alpha \in (0, 1)$, to rewrite (5.4) as follows

$$\theta_4 = \frac{7}{16}\alpha \quad \text{and} \quad \lambda = \frac{1 - \alpha}{1 - \frac{11}{16}\alpha}. \tag{5.5}$$

In Figure 5, we display the values of θ_4 and λ with respect to α. We indeed note that θ_4 increases and λ decreases when α increases.

We now repeat the experiments from Section 5.2.1, this time looking at the influence of α on the TVD3 scheme with $\gamma = \gamma^{\text{opt}} = \frac{5}{6}$. We first display in Figure 6 the CPU time with respect to α for the four schemes. Once again, as expected, the CPU time does not depend on α for the IMEX1 and IMEX3 schemes. In addition, since the CFL condition becomes more restrictive, the CPU time increases with α for the TVD3 and the MOOD3 schemes. Furthermore, note the presence of sharp increases in CPU time close to $\alpha = 0.12$. Like in the MOOD2 case, these increases show that more MOOD loops were necessary to compensate for an unstable IMEX3 scheme.

Now, in the left panel of Figure 7, we display the L^∞-error with respect to α for the four schemes under consideration. As expected, we observe that it decreases with α for the TVD3 scheme, since θ_4 increases. Comparing with Figure 3, we note that the error produced by the IMEX3 scheme is larger than the one produced by the IMEX2 scheme for small γ, on the same smooth test case. This is due to incurable instabilities appearing in the IMEX3 scheme, which plague the approximate solution with oscillations, thus increasing the error. Therefore, even on this smooth test case, it is crucial to

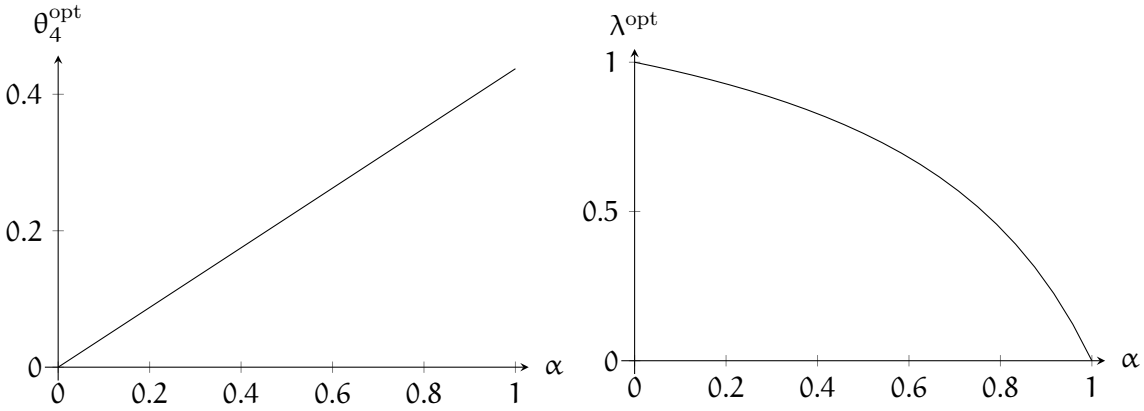

Figure 5: Values of the last convex combination parameter θ_4 (left panel) and the CFL number λ (right panel), with respect to the parameter α, for the TVD3 scheme with $\gamma = \gamma^{\text{opt}} = \frac{5}{6}$.

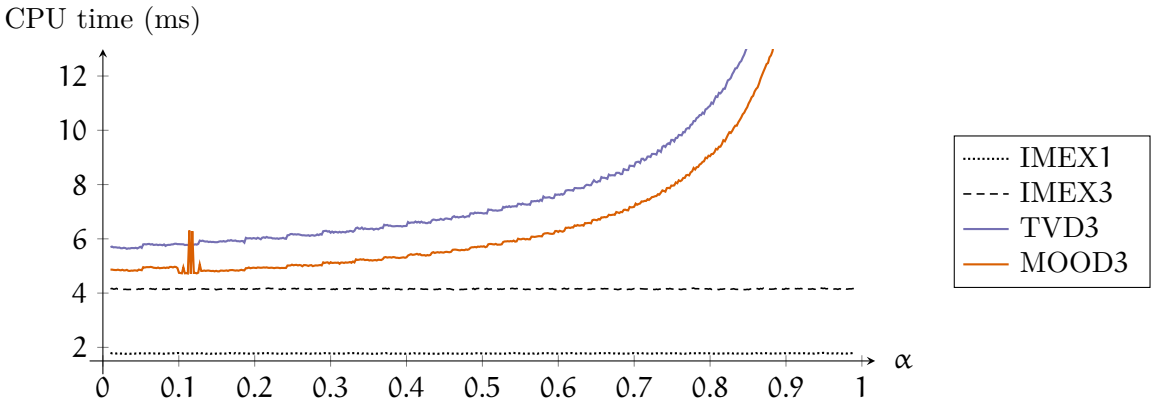

Figure 6: CPU time (in milliseconds) with respect to the parameter α, using $\gamma = \gamma^{\text{opt}} = \frac{5}{6}$, in the context of the test case presented in Section 5.2.2.

consider more stable schemes like the MOOD3 scheme, even though it also produces larger errors when α is close to 0.

In the right panel of Figure 7, we display a zoom on the CPU time and the L^∞-error produced by the IMEX3 and MOOD3 schemes, with respect to $0 < \alpha < 0.35$. We observe that the error stabilizes around $\alpha = 0.3$, and that the CPU time increases monotonically with α. Therefore, taking $\alpha = \frac{1}{3}$ seems to be a good compromise between precision and computational time. In the remainder of this article, we take

$$\gamma = \gamma^{\text{opt}} = \frac{5}{6} \quad \text{and} \quad \alpha = \alpha^{\text{opt}} = \frac{1}{3},$$

which leads to the following values for θ_3, θ_4 and λ:

$$\theta_3^{\text{opt}} = \frac{3}{8} = 0.375, \quad \theta_4^{\text{opt}} = \frac{7}{48} \simeq 0.146, \quad \text{and} \quad \lambda^{\text{opt}} = \frac{32}{37} \simeq 0.865.$$

5.3 Numerical tests

Now that the optimal values of the free parameters are established, let us test our scheme on a few numerical experiments. We first check in Section 5.3.1 the order of accuracy using the smooth solution (5.1), and we then study the behaviour of our schemes on the discontinuous solution (5.2) in

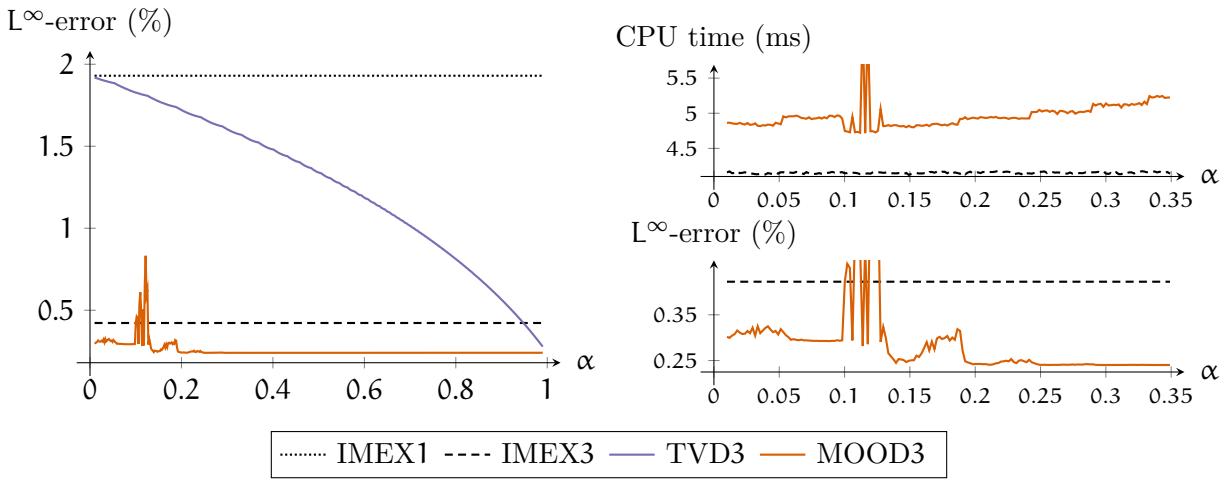

Figure 7: L^∞-error with respect to the parameter α, using $\gamma = \gamma^{opt} = \frac{5}{6}$, $\theta_3 = \frac{3}{8}$ and θ_4, λ given by (5.5). in the context of the test case presented in Section 5.2.2. For $\alpha \in (0, 0.35)$, the top right panel contains a zoom on the CPU time (data from Figure 6 and the bottom right panel contains a zoom on the L^∞-error (data from left panel).

Section 5.3.2. We expect the IMEXp schemes to behave well on smooth solutions, while their non-L^∞ stable nature should produce oscillations and destroy the numerical approximation of discontinuous solutions.

With the choice of β from Section 5.2.1, the IMEX2 scheme turns out to be the well-known ARS(2,2,2) scheme. However, the IMEX3 scheme, given by the tableaux (5.3), is not well-known in the literature. To provide a point of comparison, we introduce the ARS(2,3,3) scheme, reported in [3], Section 2.4, or [29], Table 5, given by the following tableaux

$$\text{expl.:}\quad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \delta & \delta & 0 & 0 \\ 1-\delta & \delta-1 & 2-2\delta & 0 \\ \hline & 0 & 1/2 & 1/2 \end{array}, \qquad \text{impl.:}\quad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \delta & 0 & \delta & 0 \\ 1-\delta & 0 & 1-2\delta & \delta \\ \hline & 0 & 1/2 & 1/2 \end{array}, \qquad \text{where } \delta = \frac{3+\sqrt{3}}{6}.$$

Note that this scheme falls within the framework of Section 4. Indeed, the above tableaux are nothing but the tableaux (4.2) and (4.3) with $\gamma = \frac{3-\sqrt{3}}{6}$. This value of γ does not satisfy the requirement of Corollary 6, and therefore we cannot prove the existence of convex combinations that make the ARS(2,3,3) scheme TVD and L^∞ stable, even though this ARS(2,3,3) scheme is actually L-stable. The following numerical experiments should therefore highlight that the property of L-stability is not enough to ensure non-oscillatory approximations.

Remark 9. *In the following numerical experiments, some values of the number of points* N *are large when* ε *is small. These large values of* N *have been chosen to ensure that more than* 10 *time iterations are needed to reach* t_{end}*. If fewer time iterations are considered, the time steps are too large to visually notice the differences between the schemes.*

5.3.1 Study of the order of accuracy

We now focus on the study of the order of accuracy of the schemes under consideration using the smooth solution (5.1). First, we consider the IMEX2 scheme, the TVD2 scheme and the MOOD2 schemes. Then, we study the ARS(2,3,3), IMEX3, TVD3 and MOOD3 schemes.

The IMEX2, TVD2 and MOOD2 schemes. In Figure 8, we display the convergence curves for the four schemes, for $\varepsilon = 1$ (left panel) and $\varepsilon = 10^{-3}$ (right panel). As expected, we observe that the IMEX1 and TVD2 schemes are both first-order accurate, with the TVD2 scheme being more precise than the IMEX1 scheme. In addition, both the MOOD2 and IMEX2 schemes are second-order accurate. This means that, in this context of a smooth solution, the MOOD correction allows us to get a second-order accurate scheme that also respects the maximum principle.

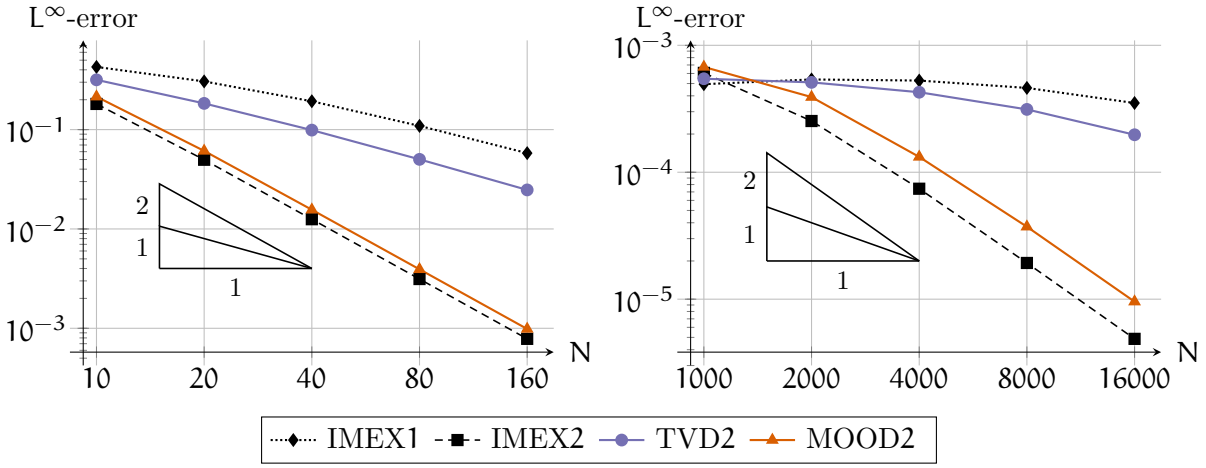

Figure 8: Error lines in L^∞ norm for the smooth solution (5.1) using the IMEX1, IMEX2, TVD2 and MOOD2 schemes. Left panel: $\varepsilon = 1$; right panel: $\varepsilon = 10^{-3}$.

The IMEX3, TVD3 and MOOD3 schemes. Now, let us consider the third-order IMEX3 scheme and the two schemes derived from this one. In Figure 9, we display the error to the exact solution, for $\varepsilon = 1$ in the left panel and $\varepsilon = 10^{-3}$ in the right panel. For $\varepsilon = 1$, we observe, as expected, that the TVD3 scheme is first-order accurate but more precise than the IMEX1 scheme, while the other three schemes are third-order accurate. For $\varepsilon = 10^{-3}$, we note that the error produced by the IMEX3 scheme starts to decrease slower than third-order when N becomes large. This is due to the instability of this IMEX3 scheme, and this problem is not experienced by the L-stable ARS(2,3,3) scheme. Due to these instabilities, the solution of the MOOD3 scheme is degraded since the MOOD algorithm switches more often to the TVD3 scheme than in the previous second-order case.

5.3.2 Approximation of a discontinuous solution

Now, we study the numerical approximation of the discontinuous solution (5.2). Like in the previous Section, we first study the IMEX2, TVD2 and MOOD2 schemes, before moving on to the IMEX3, TVD3 and MOOD3 schemes. Lastly, we perform an experiment to show that the BDF2 and BDF3 discretizations alone violate the maximum principle.

Here, to compute the order of accuracy of the scheme, we no longer focus on the L^∞ norm, which is not suited to the computation of an error between a discontinuous solution and its diffusive approximation. Instead, we turn to the L^1 norm, defined by

$$\|w^n\|_1 = \frac{1}{\Delta x} \sum_j |w_j^n|.$$

However, the above norm only measures the average deviation between the exact solution and the numerical approximation. Here, since we seek a measure of the maximum principle violation of the

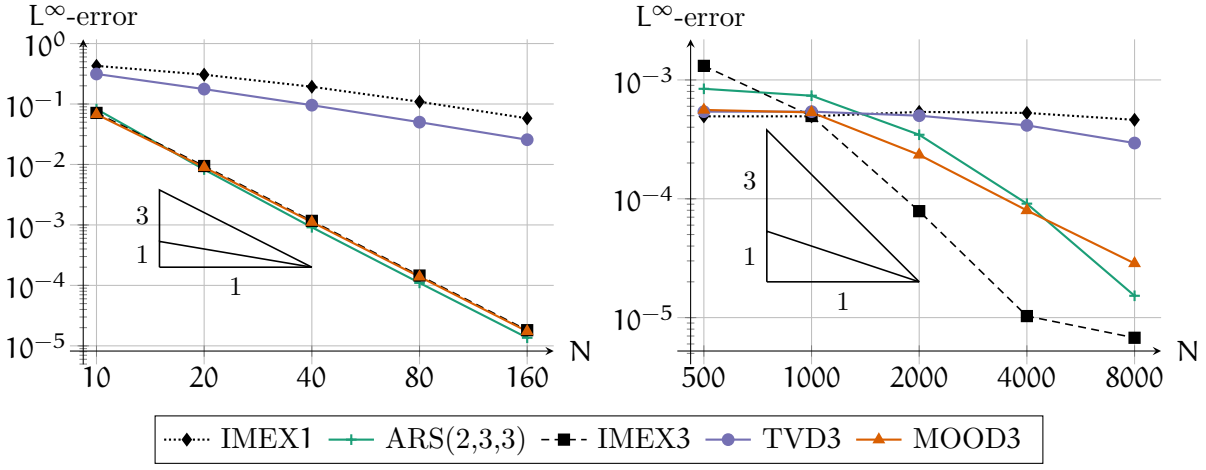

Figure 9: Error lines in L^∞ norm for the smooth solution (5.1) using the IMEX1, ARS(2,3,3), IMEX3, TVD3 and MOOD3 schemes. Left panel: $\varepsilon = 1$; right panel: $\varepsilon = 10^{-3}$.

IMEXp scheme, we instead consider the following modification of the L^1 norm:

$$\|w^n\|_{L^1_o} = \frac{1}{\Delta x}\sum_j \left(|w_j^n| + \max_{m\leqslant n}\left[\left(\max_j w_j^m - \min_j w_j^m\right) - \left(\max_j w_j^0 - \min_j w_j^0\right)\right]\right).$$

This quantity, although it does not satisfy the triangle inequality property of a norm, as it is in fact a quasinorm, allows us to add the impact of overshoots and undershoots to the usual measure of the average deviation between the solution and its approximation. Since the TVDp and MOODp methods are built to avoid such over- and undershoots, this additional term will vanish with these methods.

The IMEX2, TVD2 and MOOD2 schemes. In Figure 10, we display the results of the three schemes, and of the IMEX1 scheme for the sake of comparison, when approximating the discontinuous solution (5.2) (left panel: $\varepsilon = 1$, right panel: $\varepsilon = 10^{-3}$). In both cases, we observe that the IMEX2 scheme violates the maximum principle, while it is satisfied by the other three schemes. We observe that both phase and amplitude errors are present.

In Figure 11, we display the error lines in L^1 norm (left panels) and L^1_o quasinorm (right panels), for $\varepsilon = 1$ (top panels) and $\varepsilon = 10^{-3}$ (bottom panels). First, we observe that the theoretical order of convergence is not reached. At most, the schemes are order $\frac{1}{2}$. This is due to the fact that we approximate a discontinuous solution, where the numerical diffusion of the schemes considerably worsen the order of convergence, see for instance [24], Chapter 11. Second, as expected, the L^1-error of the IMEX2 scheme is lower than the one of the other schemes. Also, when taking the over- and undershoots into account thanks to the L^1_o quasinorm, we observe that the L^1_o quasinorm of the error produced by the IMEX1, TVD2 and MOOD2 schemes is the same as their L^1 norm. This was to be expected since no over- or undershoots are produced by these schemes. However, when looking at the L^1_o quasinorm of the error of the IMEX2 scheme, we observe that it stays roughly constant as N grows larger. This means that the improvement in L^1 norm, since the numerical solution is overall closer to the exact solution, is almost exactly compensated by an increase of the over- and undershoot magnitude. Therefore, even taking large N is not enough to ensure a good approximation of the exact discontinuous solution by the IMEX2 scheme.

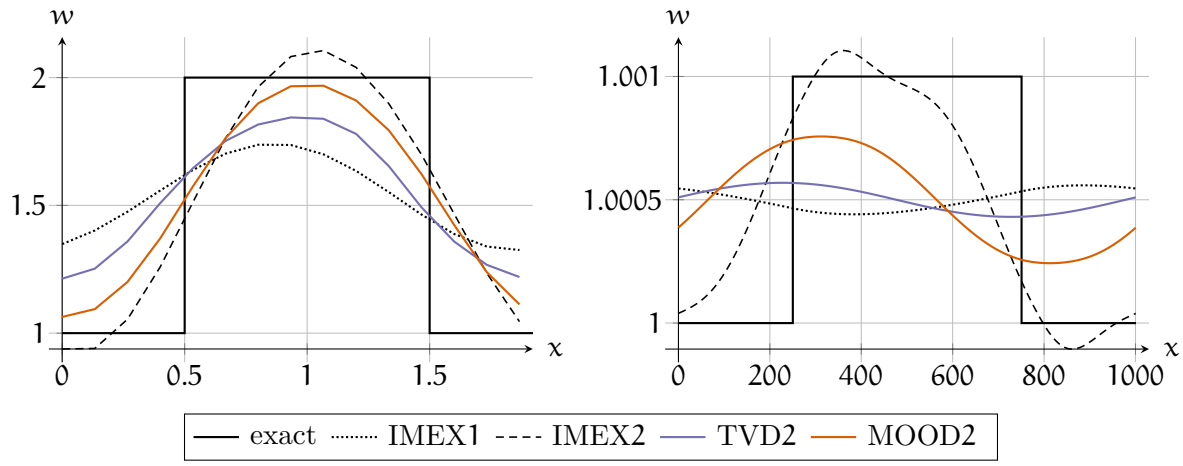

Figure 10: Approximation of the discontinuous solution (5.2) at time t_{end} using the IMEX1, IMEX2, TVD2 and MOOD2 schemes. Left panel: $\varepsilon = 1$ and $N = 15$; right panel: $\varepsilon = 10^{-3}$ and $N = 2000$.

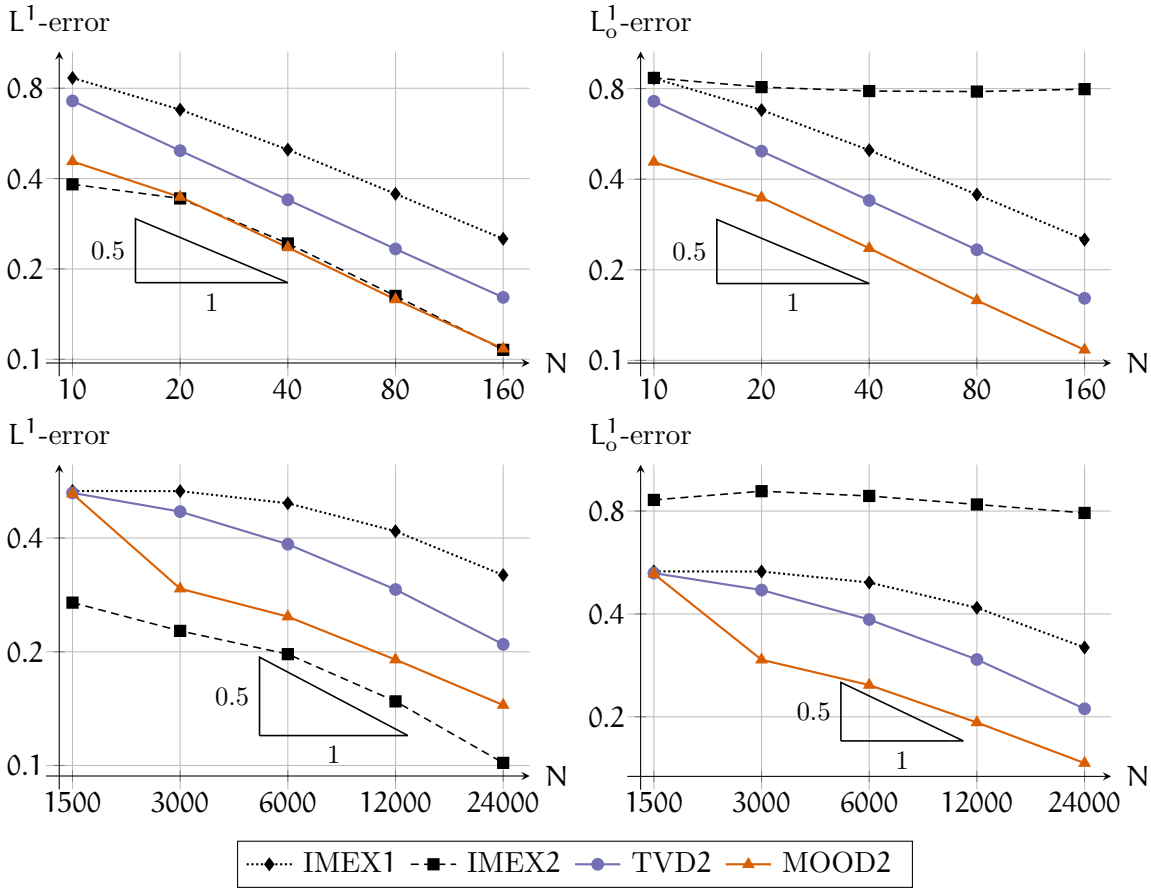

Figure 11: Error lines in L^1 norm (left panels) and L^1_o quasinorm (right panels) for the discontinuous solution (5.2) using the IMEX1, IMEX2, TVD2 and MOOD2 schemes. Top panels: $\varepsilon = 1$; bottom panels: $\varepsilon = 10^{-3}$.

The IMEX3, TVD3 and MOOD3 schemes. Now, we turn to Figure 12, where we have displayed the numerical approximation of the discontinuous solution by the IMEX1, ARS(2,3,3), IMEX3, TVD3 and MOOD3 schemes, for $\varepsilon = 1$ in the left panel and $\varepsilon = 10^{-3}$ in the right panel. Once again, we note that the pure IMEX high-order schemes are oscillatory and violate the maximum principle, while the other three schemes are in-bounds. A notable remark concerns the IMEX3 scheme when $\varepsilon = 10^{-3}$, in the right panel depicted by the dashed line. In this case, the scheme is so unstable that the numerical solution is unrecognisable. The MOOD3 scheme corrects this shortcoming. Here, we begin to see the limits of the MOOD detection criterion in Algorithm 7. Indeed, we force the MOOD3 solution to satisfy the maximum principle with respect to the initial condition, and therefore it is also TVD with respect to the initial condition, i.e. $\mathrm{TV}(w^n) \leqslant \mathrm{TV}(w^0)$, hence the rather small oscillations that has developed in the MOOD3 solution around $x = 300$ and $x = 800$.

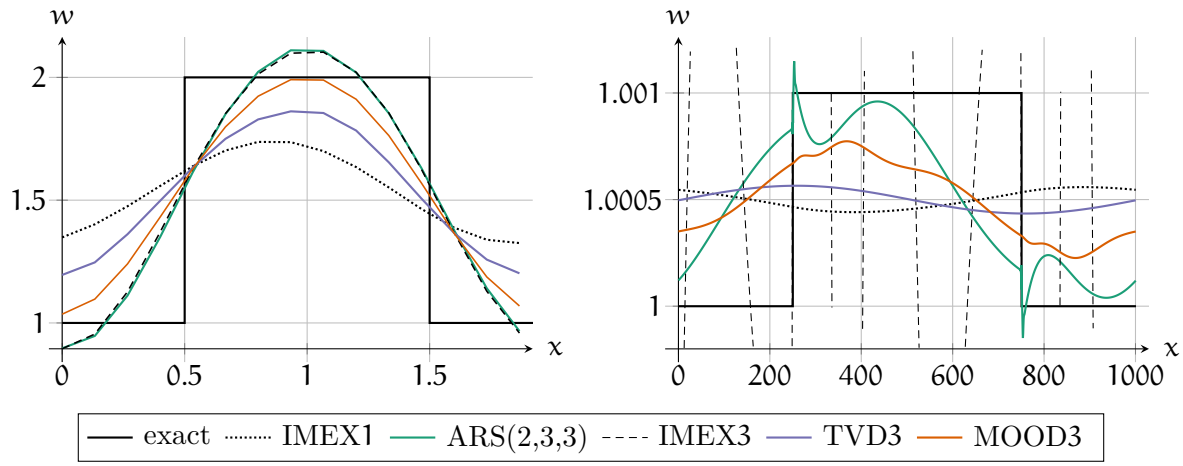

Figure 12: Approximation of the discontinuous solution (5.2) at time t_{end} using the IMEX1, ARS(2,3,3), IMEX3, TVD3 and MOOD3 schemes. Left panel: $\varepsilon = 1$ and $N = 15$; right panel: $\varepsilon = 10^{-3}$ and $N = 2000$. In the right panel, the errors produced by the IMEX3 scheme have destroyed the numerical approximation.

In Figure 13, we report the error produced by the five schemes, in the L^1 norm in the left panels and in the L^1_o quasinorm in the right panels, for $\varepsilon = 1$ in the top panels and $\varepsilon = 10^{-3}$, except for the IMEX3 scheme, whose error explodes in the bottom panels. Like in the case of the IMEX2, TVD2 and MOOD2 schemes, we observe that the theoretical order of convergence is not reached, and that the schemes are accurate up to order $\frac{1}{2}$ for the IMEX1, TVD3 and MOOD3 schemes, and up to order $\frac{3}{4}$ for the ARS(2,3,3) and IMEX3 schemes. In addition, once again, the L^1_o quasinorm for the ARS(2,3,3) and IMEX3 schemes stays roughly constant as N increases, which means that the L^1-error improvement is compensated by an increase in the over- and undershoot amplitude.

Failure of the TVD property for the BDF discretisations. This last experiment consists in computing the approximation of the discontinuous solution with the IMEX1 scheme, equipped only with the BDF2 or BDF3 space discretisations on the implicit part, respectively given by the BDF (3.10) and (3.11), instead of the backward Euler discretisation. In Figure 14, we display, with respect to the number of points N, the amplitude of the over- and undershoots produced by these two schemes, i.e. the following quantity

$$A^n = \max_{m \leqslant n} \left[\left(\max_j w_j^m - \min_j w_j^m \right) - \left(\max_j w_j^0 - \min_j w_j^0 \right) \right].$$

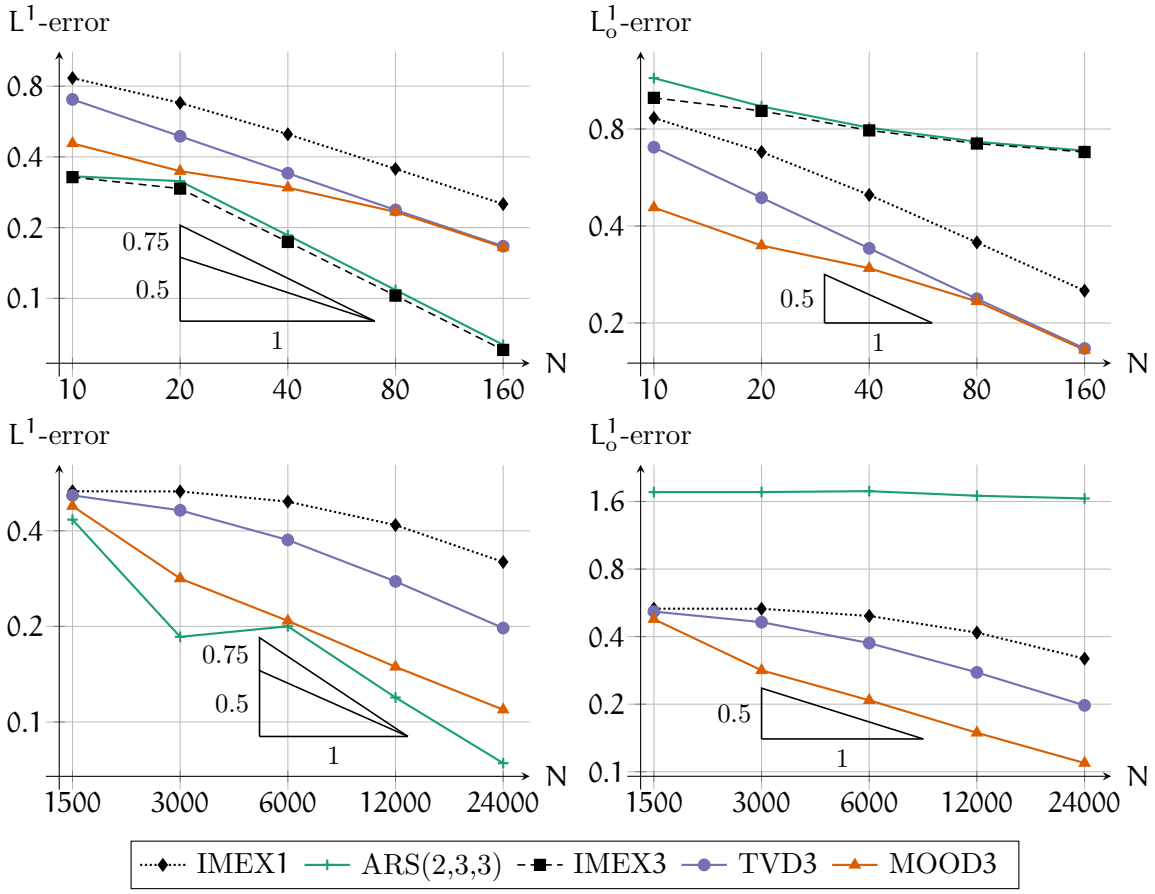

Figure 13: Error lines in L^1 norm (left panels) and L^1_o quasinorm (right panels) for the discontinuous solution (5.2) using the IMEX1, ARS(2,3,3), IMEX3, TVD3 and MOOD3 schemes. Top panels: $\varepsilon = 1$; bottom panels: $\varepsilon = 10^{-3}$. For $\varepsilon = 10^{-3}$, the IMEX3 error is so large that the error lines are not displayed (see Figure 12, right panel).

which was present in the definition of the L^1_o quasinorm. We observe that, although it decreases linearly when N increases, this amplitude is always non-zero, which means that the BDF2 and BDF3 discretisations alone are enough to violate the maximum principle, even when every other discretisation is first-order accurate. This numerical result ties in with the theoretical results expressed in Section 3.2.

6 Conclusions and future work

We have presented a way to construct highly precise TVD IMEX-RK schemes for multi-scale equations which are computationally efficient. To circumvent the negative result for the IMEX-RK schemes, we introduced a new class of TVD schemes consisting of convex combinations with a first-order TVD IMEX-RK scheme, for which the time step is only restricted by the slow wave speed.

The TVD property is crucial when approximating discontinuous solutions, as displayed in Figure 11 and Figure 13. Indeed, the bottom right panel of these figures show that even usual L-stable methods fail, producing large oscillations whose amplitude is not compensated by the decrease in L^1 error when the discretisation becomes finer.

To increase the precision of the schemes, we used the MOODp procedure, which gives a lot of

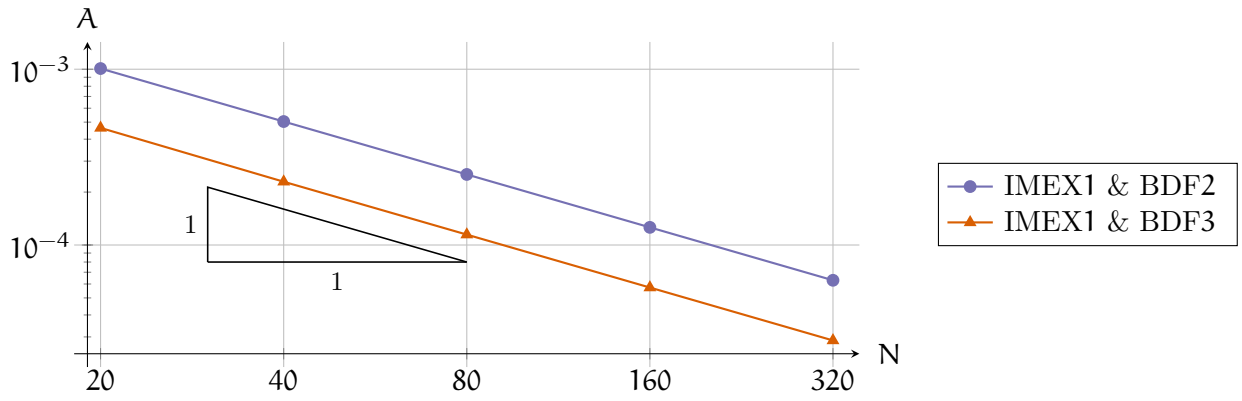

Figure 14: Evolution of the over- and undershoot amplitude for the discontinuous solution (5.2) using the IMEX1 scheme with BDF2 and BDF3 discretisations and $\varepsilon = 1$.

freedom to combine schemes to get highly precise numerical solutions. The most straightforward way of applying this MOOD technique, presented here, is to go directly from the IMEXp scheme to the TVDp scheme. However, as seen in the numerical experiments and especially in Figure 12, the IMEX3 scheme gives a very oscillatory approximation of the solution, especially when ε is small. Therefore, the MOOD procedure is activated quite often in this case, in order to dissipate this instability. It would be interesting to start from a more stable scheme, say the ARS(2,3,3) scheme, and to check how this translates to the numerical results. Similarly, a scheme that is more precise than the TVD3 scheme could also be used. Indeed, we could look into deriving a three-step TVD scheme based on a second-order tableau, where less restrictive order conditions could mean potentially higher values of θ. Promising preliminary investigations, using an optimisation algorithm, have already been undertaken in this direction.

Although we studied a one-dimensional scalar equation in this work, we find many parallels with non-linear multi-dimensional systems, like the non-dimensional Euler equations. Future work will concern the transfer of the one-dimensional linear scalar methods constructed here to multi-dimensional multi-scale non-linear systems.

Acknowledgements V. Michel-Dansac extends his thanks to the Service Hydrographique et Océanographique de la Marine (SHOM) for financial support. A. Thomann acknowledges the support of the INDAM-DP-COFUND-2015, grant number 713485. This work was started during the SHARK-FV conference (Sharing Higher-order Advanced Research Know-how on Finite Volume http://www.SHARK-FV.eu/) held in 2019. The authors would like to thank Gabriella Puppo for fruitful discussions and comments.

References

[1] E. Abbate, A. Iollo, and G. Puppo. An Implicit Scheme for Moving Walls and Multi-Material Interfaces in Weakly Compressible Materials. *Commun. Comput. Phys.*, 27(1):116–144, 2019.

[2] U. M. Ascher and L. R. Petzold. *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. SIAM: Society for Industrial and Applied Mathematics, 1998.

[3] U. M. Ascher, S. J. Ruuth, and R. J. Spiteri. Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations. *Appl. Numer. Math.*, 25(2-3):151–167, 1997. Special issue on time integration (Amsterdam, 1996).

[4] E. Audusse, C. Chalons, and P. Ung. A simple well-balanced and positive numerical scheme for the shallow-water system. *Commun. Math. Sci.*, 13(5):1317–1332, 2015.

[5] G. Bispen, K. R. Arun, M. Lukáčová-Medviďová, and S. Noelle. IMEX large time step finite volume methods for low Froude number shallow water flows. *Commun. Comput. Phys.*, 16(2):307–347, 2014.

[6] A. Bollermann, S. Noelle, and M. Lukáčová-Medviďová. Finite volume evolution Galerkin methods for the shallow water equations with dry beds. *Commun. Comput. Phys.*, 10(2):371–404, 2011.

[7] L. Bonaventura and A. Della Rocca. Unconditionally Strong Stability Preserving Extensions of the TR-BDF2 Method. *J. Sci. Comput.*, 70(2):859–895, aug 2016.

[8] S. Boscarino, G. Russo, and L. Scandurra. All Mach Number Second Order Semi-implicit Scheme for the Euler Equations of Gas Dynamics. *J. Sci. Comput.*, 77(2):850–884, 2018.

[9] F. Bouchut, E. Franck, and L. Navoret. A low cost semi-implicit low-Mach relaxation scheme for the full Euler equations. working paper or preprint, 2019.

[10] C. Bresten, S. Gottlieb, Z. Grant, D. Higgs, D. I. Ketcheson, and A. Németh. Explicit strong stability preserving multistep Runge–Kutta methods. *Math. Comput.*, 86(304):747–769, 2017.

[11] S. Clain, S. Diot, and R. Loubère. A high-order finite volume method for systems of conservation laws—Multi-dimensional Optimal Order Detection (MOOD). *J. Comput. Phys.*, 230(10):4028–4050, 2011.

[12] O. Delestre. *Simulation du ruisselement d'eau de pluie sur des surfaces agricoles*. PhD thesis, Université d'Orléans, 2010.

[13] G. Dimarco, R. Loubère, V. Michel-Dansac, and M.-H. Vignal. Second-order implicit-explicit total variation diminishing schemes for the Euler system in the low Mach regime. *J. Comput. Phys.*, 372:178–201, 2018.

[14] G. Dimarco, R. Loubère, and M.-H. Vignal. Study of a New Asymptotic Preserving Scheme for the Euler System in the Low Mach Number Limit. *SIAM J. Sci. Comput*, 39(5):A2099–A2128, 2017.

[15] S. Gottlieb. On high order strong stability preserving Runge-Kutta and multi step time discretizations. *J. Sci. Comput.*, 25(1-2):105–128, 2005.

[16] S. Gottlieb and C.-W. Shu. Total variation diminishing Runge-Kutta schemes. *Math. Comp.*, 67(221):73–85, 1998.

[17] S. Gottlieb, C.-W. Shu, and E. Tadmor. Strong stability-preserving high-order time discretization methods. *SIAM Rev.*, 43(1):89–112, 2001.

[18] H. Guillard and C. Viozat. On the behaviour of upwind schemes in the low Mach number limit. *Comput. & Fluids*, 28(1):63–86, 1999.

[19] X. Y. Hu, N. A. Adams, and C.-W. Shu. Positivity-preserving method for high-order conservative schemes solving compressible Euler equations. *J. Comput. Phys.*, 242:169–180, 2013.

[20] C. A. Kennedy and M. H. Carpenter. Additive Runge–Kutta schemes for convection–diffusion–reaction equations. *Appl. Numer. Math.*, 44(1-2):139–181, 2003.

[21] D. I. Ketcheson and A. C. Robinson. On the practical importance of the SSP property for Runge-Kutta time integrators for some common Godunov-type schemes. *Int. J. Numer. Methods Fluids*, 48(3):271–303, 2005.

[22] S. Klainerman and A. Majda. Singular limits of quasilinear hyperbolic systems with large parameters and the incompressible limit of compressible fluids. *Comm. Pure Appl. Math.*, 34(4):481–524, 1981.

[23] R. Klein. Scale-Dependent Models for Atmospheric Flows. *Annu. Rev. Fluid. Mech.*, 42(1):249–274, jan 2010.

[24] R. J. LeVeque. *Numerical methods for conservation laws*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 1992.

[25] W. H. Matthaeus and M. R. Brown. Nearly incompressible magnetohydrodynamics at low Mach number. *Phys. Fluids*, 31(12):3634, 1988.

[26] V. Michel-Dansac, C. Berthon, S. Clain, and F. Foucher. A well-balanced scheme for the shallow-water equations with topography or Manning friction. *J. Comput. Phys.*, 335:115–154, 2017.

[27] V. Michel-Dansac and A. Thomann. On high-accuracy L^∞-stable IMEX schemes for scalar hyperbolic multi-scale equations. submitted, 2019.

[28] S. Noelle, G. Bispen, K. R. Arun, M. Lukáčová-Medviďová, and C.-D. Munz. A weakly asymptotic preserving low Mach number scheme for the Euler equations of gas dynamics. *SIAM J. Sci. Comput.*, 36(6):B989–B1024, 2014.

[29] L. Pareschi and G. Russo. Implicit-explicit Runge-Kutta schemes for stiff systems of differential equations. In *Recent trends in numerical analysis*, volume 3 of *Adv. Theory Comput. Math.*, pages 269–288. Nova Sci. Publ., Huntington, NY, 2001.

[30] L. Pareschi and G. Russo. Implicit-Explicit Runge-Kutta schemes and applications to hyperbolic systems with relaxation. *J. Sci. Comput.*, 25(1-2):129–155, 2005.

[31] B. Perthame and C.-W. Shu. On positivity preserving finite volume schemes for Euler equations. *Numer. Math.*, 73(1):119–130, 1996.

[32] B. Schmidtmann, B. Seibold, and M. Torrilhon. Relations between WENO3 and third-order limiting in finite volume methods. *J. Sci. Comput.*, 68(2):624–652, 2015.

[33] A. Thomann, M. Zenk, G. Puppo, and C. Klingenberg. An all speed second order IMEX relaxation scheme for the Euler equations. *Commun. Comput. Phys.*, 2020. Accepted.

[34] E. Turkel. Preconditioned methods for solving the incompressible and low speed compressible equations. *J. Comput. Phys.*, 72(2):277–298, 1987.