On high-precision L∞-stable IMEX schemes for scalar hyperbolic multi-scale equations

Victor Michel-Dansac, Andrea Thomann

▶ **To cite this version:**

Victor Michel-Dansac, Andrea Thomann. On high-precision L∞-stable IMEX schemes for scalar hyperbolic multi-scale equations. Recent Advances in Numerical methods for Hyperbolic PDE Systems. Selected talks of Numhyp 2019, Jun 2019, Málaga, Spain. 10.1007/978-3-030-72850-2_4 . hal-02303491v3

HAL Id: hal-02303491
https://hal.science/hal-02303491v3

Submitted on 13 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

On high-precision L^∞-stable IMEX schemes for scalar hyperbolic multi-scale equations

Victor Michel-Dansac*, Andrea Thomann†

December 13, 2020

Abstract

We present a framework to build high-precision IMEX schemes that fulfill the maximum principle, applied to a scalar hyperbolic multi-scale equation. Motivated by the findings in [5] that implicit R-K schemes are not L^∞-stable, our scheme, for which we can prove the L^∞ stability, is based on a convex combination between a first- and a class of second-order IMEX schemes. We numerically demonstrate the advantages of our scheme, especially for discontinuous problems, and give a MOOD procedure to increase the precision.

1 Introduction

We consider the scalar multi-scale equation

$$w_t + c_e w_x + \frac{c_i}{\varepsilon} w_x = 0, \tag{1.1}$$

where we set the constants $c_e, c_i > 0$ and the parameter $\varepsilon > 0$. The model (1.1) mimics the behavior of the isentropic Euler equations with a slow speed c_e and a fast speed c_i/ε, where ε corresponds to the square of the Mach number. We treat the derivative w_x associated with the slow scale c_e explicitly, whereas w_x associated with the fast scale c_i/ε is treated implicitly in time due to the stiffness introduced by $\varepsilon < 1$. For computational efficiency, the resulting CFL condition, and therefore the time step, has to be independent of ε. In space, we apply an upwind discretization because, already having in mind the non-linear nature of e.g. the Euler equations, using a central scheme for the implicit part will not lead to a L^∞-stable scheme, as shown in [4] for a non-linear system.

The discretization of time and space follows the usual finite difference framework. The space domain $[x_1, x_N]$ is partitioned in N uniformly spaced points $(x_j)_{j\in[1,N]}$, with the step size Δx. We discretize the time variable with $t^n = n\Delta t$, where Δt denotes the time step. Then, the solution $w(t,x)$ of (1.1) at (t^n, x_j) is approximated by w_j^n. The first-order implicit-explicit (IMEX) discretization of (1.1) is given by

$$w_j^{n+1} = w_j - \lambda(w_j^n - w_{j-1}^n) - \mu_\varepsilon(w_j^{n+1} - w_{j-1}^{n+1}), \tag{1.2}$$

*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;
[Present affiliation] Université de Strasbourg, CNRS, Inria, IRMA, F-67000 Strasbourg, France;
victor.michel-dansac@inria.fr

†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;
[Present affiliation] Institute of Mathematics, Johannes Gutenberg University Mainz, Staudingerweg 9, 55128 Mainz, Germany;
athomann@uni-mainz.de

where we define $\lambda = c_e \frac{\Delta t}{\Delta x}$ and $\mu_\varepsilon = \frac{c_i}{\varepsilon} \frac{\Delta t}{\Delta x}$ for abbreviation. Note that $\lambda, \mu_\varepsilon > 0$.

We are interested in IMEX schemes that meet the maximum principle. Here, we focus on L^∞-stable schemes, where a scheme is said to be L^∞-stable if

$$\|w^{n+1}\|_\infty = \max_{j \in [\![1,N]\!]} |w_j^{n+1}| \leqslant \|w^n\|_\infty. \tag{1.3}$$

As proven in [3], the first-order scheme (1.2) is L^∞-stable and TVD. Furthermore, as proven in [5], implicit Runge-Kutta schemes, and consequently second-order IMEX schemes, are not L^∞-stable. Therefore, we would like to propose a convex combination of (1.2) with a second-order IMEX scheme and give conditions for the L^∞ stability for the resulting scheme. We define the convex combination between the first-order scheme $w_j^{n+1,1st}$ and a second-order update $w_j^{n+1,2nd}$ for a parameter $\theta \in (0,1)$ as:

$$w_j^{n+1} = (1-\theta)\, w_j^{n+1,1st} + \theta\, w_j^{n+1,2nd}. \tag{1.4}$$

2 IMEX formulation

Generic formulations of an IMEX scheme introduce two $s \times s$ matrices $A = (a_{ij})$ and $\tilde{A} = (\tilde{a}_{ij})$, as well as two vectors $b, \tilde{b} \in \mathbb{R}^s$. They are regrouped in two linked Butcher tableaux

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array}, \qquad \begin{array}{c|c} \tilde{c} & \tilde{A} \\ \hline & \tilde{b}^T \end{array}.$$

The coefficients $c, \tilde{c}$ are only necessary if the right hand side depends explicitly on time. In the following we will use the pairs (A, b) for the implicit and $(\tilde{A}, \tilde{b})$ for the explicit part. To reduce computational costs, we take A to be lower triangular and $\tilde{A}$ to be strictly lower triangular. Applying the IMEX formulation on (1.1), we obtain the following scheme:

$$w^{n+1} = w^n - c_e \Delta t \sum_{k=1}^{s} \tilde{b}_k w_x^{(k)} - \frac{c_i}{\varepsilon} \Delta t \sum_{k=1}^{s} b_k w_x^{(k)}, \tag{2.1}$$

with the stages

$$w^{(k)} = w^n - c_e \Delta t \sum_{l=1}^{k-1} \tilde{a}_{kl} w_x^{(l)} - \frac{c_i}{\varepsilon} \Delta t \sum_{l=1}^{k} a_{kl} w_x^{(l)}. \tag{2.2}$$

IMEX Runge-Kutta (R-K) schemes can be classified depending on the structure of the matrix A.

Definition 1 *An IMEX R-K method is said to be of type CK (Carpenter and Kennedy [6]) if the matrix $A \in \mathbb{R}^{s \times s}$ can be written as*

$$A = \begin{pmatrix} 0 & 0 \\ a & \hat{A} \end{pmatrix},$$

where $a \in \mathbb{R}^{s-1}$ and $\hat{A} \in \mathbb{R}^{(s-1)\times(s-1)}$ is invertible. In the case where $a = 0$, the scheme is said to be of ARS type (Asher, Ruuth and Spiteri [1]).

In the following we will consider a second-order 2-stage and a second-order 3-stage IMEX R-K method of type CK. To obtain a second-order scheme, there are the following compatibility conditions [9]:

$$\begin{aligned} &\sum_{k=1}^{s} \tilde{b}_k = 1\,; \quad \sum_{k=1}^{s} b_k = 1\,; \quad \forall k,\ \tilde{c}_k = \sum_{l=1}^{k-1} \tilde{a}_{kl}\,; \quad \forall k,\ c_k = \sum_{l=1}^{k-1} a_{kl}\,; \\ &\sum_{k=1}^{s} \tilde{b}_k \tilde{c}_k = \frac{1}{2}\,; \quad \sum_{k=1}^{s} \tilde{b}_k c_k = \frac{1}{2}\,; \quad \sum_{k=1}^{s} b_k \tilde{c}_k = \frac{1}{2}\,; \quad \sum_{k=1}^{s} b_k c_k = \frac{1}{2}. \end{aligned} \tag{2.3}$$

2.1 A 2-stage CK type IMEX R-K method

For a 2-stage CK type method, we have the following Butcher tableaux, with $a_{22} \neq 0$:

$$\text{explicit:}\quad \begin{array}{c|cc} 0 & 0 & 0 \\ \tilde{c}_2 & \tilde{a}_{21} & 0 \\ \hline & \tilde{b}_1 & \tilde{b}_2 \end{array}, \qquad \text{implicit:}\quad \begin{array}{c|cc} 0 & 0 & 0 \\ c_2 & a_{21} & a_{22} \\ \hline & b_1 & b_2 \end{array}. \tag{2.4}$$

With the compatibility conditions (2.3), we can simplify (2.4) to

$$\text{explicit:}\quad \begin{array}{c|cc} 0 & 0 & 0 \\ \alpha & \alpha & 0 \\ \hline & 1-\frac{1}{2\alpha} & \frac{1}{2\alpha} \end{array}, \qquad \text{implicit:}\quad \begin{array}{c|cc} 0 & 0 & 0 \\ \alpha & \gamma & \alpha-\gamma \\ \hline & 1-\frac{1}{2\alpha} & \frac{1}{2\alpha} \end{array}, \tag{2.5}$$

where $\alpha-\gamma \neq 0$ and $\alpha \neq 0$.

Using (2.1), (2.2) and (2.5), we can define the second-order discretization of (1.1) as

$$\begin{aligned} w_j^{(1)} &= w_j^n - \lambda\alpha(w_j^n - w_{j-1}^n) - \gamma\mu_\varepsilon(w_j^n - w_{j-1}^n) - \mu_\varepsilon(\alpha-\gamma)(w_j^{(1)} - w_{j-1}^{(1)}), \\ w_j^{n+1} &= w_j^n - \left(1-\frac{1}{2\alpha}\right)(\lambda+\mu_\varepsilon)(w_j^n - w_{j-1}^n) - \frac{1}{2\alpha}(\lambda+\mu_\varepsilon)(w_j^{(1)} - w_{j-1}^{(1)}). \end{aligned} \tag{2.6}$$

Due to the matrix structure of the CK type R-K scheme, we have only two computational steps. The first one computes $w^{(1)}$, and the second one w^{n+1}. The convex combination (1.4) between the schemes (1.2) and (2.6), with the shorter notation $\Delta = w_j - w_{j-1}$, is given by:

$$\begin{aligned} w_j^{(1)} &= w_j^n - \lambda\alpha\Delta^n - \gamma\mu_\varepsilon\Delta^n - \mu_\varepsilon(\alpha-\gamma)\Delta^{(1)}, \\ w_j^{n+1} &= w_j^n - \left(\lambda - \theta\frac{1}{2\alpha}(\lambda+\mu_\varepsilon) + \theta\mu_\varepsilon\right)\Delta^n \\ &\quad - \theta\frac{1}{2\alpha}(\lambda+\mu_\varepsilon)\Delta^{(1)} - (1-\theta)\mu_\varepsilon\Delta^{n+1}. \end{aligned} \tag{2.7}$$

We can sort (2.7) by grouping the w^{n+1} and $w^{(1)}$ terms:

$$\begin{aligned} (1+\mu_\varepsilon(\alpha-\gamma))w_j^{(1)} - \mu_\varepsilon(\alpha-\gamma)w_{j-1}^{(1)} &= (1-(\lambda\alpha+\gamma\mu_\varepsilon))w_j^n \\ &\quad + (\lambda\alpha+\gamma\mu_\varepsilon)w_{j-1}^n, \end{aligned} \tag{2.8}$$

$$\begin{aligned} (1+(1-\theta)\mu_\varepsilon)w_j^{n+1} - (1-\theta)\mu_\varepsilon w_{j-1}^{n+1} &= w_j^n - \left(\lambda - \theta\frac{1}{2\alpha}(\lambda+\mu_\varepsilon) + \theta\mu_\varepsilon\right)\Delta^n \\ &\quad - \theta\frac{1}{2\alpha}(\lambda+\mu_\varepsilon)\Delta^{(1)}. \end{aligned} \tag{2.9}$$

In the following, we will assume periodic boundary conditions. We will prove the L^∞ stability (1.3) by starting with the proof of $\|w^{(1)}\|_\infty \leqslant \|w^n\|_\infty$. For each time step, we will use the triangle inequality $|x+y| \leqslant |x|+|y|$ and the reverse triangle inequality $|x|-|y| \leqslant |x-y|$ for $x, y \in \mathbb{R}$. To apply use those inequalities, we require in (2.8)

$$\lambda\alpha + \gamma\mu_\varepsilon \geqslant 0 \tag{2.10}$$

$$1 - (\lambda\alpha + \gamma\mu_\varepsilon) \geqslant 0 \tag{2.11}$$

$$1 + \mu_\varepsilon(\alpha-\gamma) \geqslant 0 \tag{2.12}$$

$$\mu_\varepsilon(\alpha-\gamma) \geqslant 0. \tag{2.13}$$

Using equation (2.8), we can now write the following estimate for $\|w^n\|_\infty$:

$$\begin{aligned}\|w^n\|_\infty &= (1-(\lambda\alpha+\gamma\mu_\varepsilon))\|w^n\|_\infty + (\lambda\alpha+\gamma\mu_\varepsilon)\|w^n\|_\infty \\ &\geqslant \|(1-(\lambda\alpha+\gamma\mu_\varepsilon))w_j^n + (\lambda\alpha+\gamma\mu_\varepsilon)w_{j-1}^n\|_\infty \\ &= \|(1+\mu_\varepsilon(\alpha-\gamma))w_j^{(1)} - \mu_\varepsilon(\alpha-\gamma)w_{j-1}^{(1)}\|_\infty \\ &\geqslant (1+\mu_\varepsilon(\alpha-\gamma))\|w^{(1)}\|_\infty - \mu_\varepsilon(\alpha-\gamma)\|w^{(1)}\|_\infty \\ &= \|w^{(1)}\|_\infty.\end{aligned}$$

From requirement (2.10), we get that $\alpha c_e + \gamma\frac{c_i}{\varepsilon} \geqslant 0$. In order to get a Butcher tableau independent of ε, we require $\alpha > 0$ and $\gamma \geqslant 0$. Relation (2.11) leads to a CFL condition $\frac{\Delta t}{\Delta x}(\alpha c_e + \gamma\frac{c_i}{\varepsilon}) \leqslant 1$. Note that, due to computational efficiency, we seek a time step restriction independent of ε. Therefore, we must take $\gamma = 0$, which is compatible with the restriction $\gamma \geqslant 0$. With those settings, (2.12) and (2.13) are always fulfilled.

Let us prove now that $\|w^{n+1}\|_\infty \leqslant \|w^n\|_\infty$. First, we rewrite (2.8) as follows:

$$-\mu_\varepsilon\Delta^{(1)} = \frac{1}{\alpha}w_j^{(1)} - \frac{1}{\alpha}w_j^n + \lambda(w_j^n - w_{j-1}^n). \tag{2.14}$$

After inserting (2.14) into (2.9), we obtain further conditions given by

$$r_1 = 1 - \frac{\theta}{2\alpha^2} + \lambda\left(-1+\frac{\theta}{\alpha}\right) + \mu_\varepsilon\theta\left(-1+\frac{1}{2\alpha}\right) \geqslant 0, \tag{2.15}$$

$$r_2 = \lambda\left(1-\frac{\theta}{\alpha}\right) + \mu_\varepsilon\theta\left(1-\frac{1}{2\alpha}\right) \geqslant 0, \tag{2.16}$$

$$\frac{\theta}{2\alpha^2} - \frac{\theta\lambda}{2\alpha} \geqslant 0. \tag{2.17}$$

Using (2.9), as well as the above conditions, we obtain the following estimate

$$\begin{aligned}\|w^{n+1}\|_\infty &= (1+(1-\theta)\mu_\varepsilon)\|w^{n+1}\|_\infty - (1-\theta)\mu_\varepsilon\|w^{n+1}\|_\infty \\ &\leqslant \|(1+(1-\theta)\mu_\varepsilon)w_j^{n+1} - (1-\theta)\mu_\varepsilon w_{j-1}^{n+1}\|_\infty \\ &= \|r_1 w_j^n + r_2 w_{j-1}^n + (\mu_\varepsilon - \frac{\theta\lambda}{2\alpha})w_j^{(1)} + \frac{\theta\lambda}{2\alpha}w_{j-1}^{(1)}\|_\infty \\ &\leqslant \left(1-\frac{\theta}{2\alpha^2}\right)\|w^n\|_\infty + \frac{\theta}{2\alpha^2}\|w^{(1)}\|_\infty \\ &\leqslant \|w^n\|_\infty,\end{aligned}$$

which shows the L^∞ stability. From the constraints (2.15)-(2.17), we can compute the final estimates for the free parameters α, θ, λ. The condition (2.17) gives a CFL restriction of $\lambda \leqslant \frac{1}{\alpha}$. Since we want to avoid a dependence of c_e, c_i or ε on α and θ, we need in (2.16) $1-\frac{\theta}{\alpha} \geqslant 0$, that is $\alpha \geqslant \theta$ and $1-\frac{1}{2\alpha} \geqslant 0$, which leads to $\alpha \geqslant \frac{1}{2}$. With the same motivation, we need $-1+\frac{1}{2\alpha} \geqslant 0$ in (2.15), that is $\alpha \leqslant \frac{1}{2}$. Together it follows that $\alpha = \frac{1}{2}$ and we get from (2.15) the final CFL condition $\lambda \leqslant 1$. With $\alpha = \frac{1}{2}$ and $\gamma = 0$ fixed, we have recovered a 2-stage ARS type method with the midpoint rule as the implicit part, given by

$$\text{explicit: } \begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array}, \qquad \text{implicit: } \begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \hline & 0 & 1 \end{array}. \tag{2.18}$$

The above results are summed up in the following theorem:

Theorem 2 *For periodic boundary conditions and under the CFL condition*

$$\Delta t \leqslant \frac{\Delta x}{c_e},$$

the scheme consisting of the convex combination of the first-order scheme (1.2) and the second-order scheme constructed from (2.18) is L^∞-stable as long as $\theta \leqslant \frac{1}{2}$.

Remark 3 *In order to have the maximal input of the second-order scheme, we would want to set $\theta = \theta_{opt} = \frac{1}{2}$. With this choice of θ, the restriction (2.15) for $\alpha = \frac{1}{2}$ is satisfied immediately and we get the less restrictive CFL condition*

$$\Delta t \leqslant 2\frac{\Delta x}{c_e}.$$

Unfortunately, the midpoint rule with the above CFL condition and $\theta = \theta_{opt}$ exactly reduces to two steps of a first-order scheme. We therefore advise $\theta < \frac{1}{2}$ to get a second-order scheme.

Since $\gamma = 0$, the initial CK type method (2.5) reduces to an ARS type method (2.18). This observation is summarized in the following corollary

Corollary 4 *If there is a second-order CK type IMEX R-K scheme of the form (2.5) that is L^∞-stable in the convex combination with (1.2) under a CFL condition independent of ε, then it has to be of ARS type.*

2.2 A 3-stage CK type IMEX R-K method

In this section, we adapt the derivation of the 2-stage case to a 3-stage CK type method. It is described by the following Butcher tableaux, with $a_{22} \neq 0$ and $a_{33} \neq 0$:

$$\text{explicit:}\quad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \tilde{c}_2 & \tilde{a}_{21} & 0 & 0 \\ \tilde{c}_3 & \tilde{a}_{31} & \tilde{a}_{32} & 0 \\ \hline & \tilde{a}_{31} & \tilde{a}_{32} & 0 \end{array}, \qquad \text{implicit:}\quad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ c_2 & a_{21} & a_{22} & 0 \\ c_3 & a_{31} & a_{32} & a_{33} \\ \hline & a_{31} & a_{32} & a_{33} \end{array}, \tag{2.19}$$

To have the same number of computational steps as in the 2-stage scheme (2.1), we have set $b = (a_{3j})$ and $\tilde{b} = (\tilde{a}_{3j})$.

With the second-order compatibility conditions (2.3) and $a_{22} = \beta$ and $a_{33} = \alpha$, we introduce

$$\kappa = \frac{2(\gamma+\beta)(1-\alpha)+2\alpha-1}{2(\gamma+\beta)}$$

and simplify (2.19) to:

$$\text{explicit:}\quad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \gamma+\beta & \gamma+\beta & 0 & 0 \\ 1 & 1-\frac{1}{2(\gamma+\beta)} & \frac{1}{2(\gamma+\beta)} & 0 \\ \hline & 1-\frac{1}{2(\gamma+\beta)} & \frac{1}{2(\gamma+\beta)} & 0 \end{array}, \qquad \text{implicit:}\quad \begin{array}{c|ccc} 0 & 0 & 0 & \\ \gamma+\beta & \gamma & \beta & 0 \\ 1 & \kappa & \frac{1-2\alpha}{2(\gamma+\beta)} & \alpha \\ \hline & \kappa & \frac{1-2\alpha}{2(\gamma+\beta)} & \alpha \end{array}. \tag{2.20}$$

Analogously to (2.6), we can write the second-order scheme using (2.20) as

$$\begin{aligned} w_j^{(1)} + \mu_\varepsilon \beta \Delta^{(1)} &= w_j^n - (\lambda(\gamma+\beta) + \mu_\varepsilon \gamma)\Delta^n \\ w_j^{n+1} + \mu_\varepsilon \alpha \Delta^{n+1} &= w_j^n - \left(\lambda \frac{2(\gamma+\beta)-1}{2(\gamma+\beta)} + \kappa\mu_\varepsilon\right)\Delta^n \\ &\quad - \left(\lambda \frac{1}{2(\gamma+\beta)} + \mu_\varepsilon \frac{1-2\alpha}{2(\gamma+\beta)}\right)\Delta^{(1)}. \end{aligned}$$

We conduct an analogous analysis as in the 2-stage case, which results in the following ARS-type IMEX R-K method for $\beta \in (0, \frac{1}{2})$:

$$\text{explicit:} \quad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \beta & \beta & 0 & 0 \\ 1 & 1-\frac{1}{2\beta} & \frac{1}{2\beta} & 0 \\ \hline & 1-\frac{1}{2\beta} & \frac{1}{2\beta} & 0 \end{array}, \qquad \text{implicit:} \quad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \beta & 0 & \beta & 0 \\ 1 & 0 & \frac{1}{2(1-\beta)} & 1-\frac{1}{2(1-\beta)} \\ \hline & 0 & \frac{1}{2(1-\beta)} & 1-\frac{1}{2(1-\beta)} \end{array}. \tag{2.21}$$

One example for (2.21) is the widely used ARS(2,2,2) method with $\beta = 1 - \frac{\sqrt{2}}{2}$, see [1].

Theorem 5 *For periodic boundary conditions and under the CFL condition*

$$\Delta t \leqslant \frac{\Delta x}{c_e},$$

the scheme consisting of the convex combination of the first-order scheme (1.2) and the second-order scheme constructed from (2.21) with $\beta \in (0, \frac{1}{2})$ is L^∞-stable as long as $\theta \leqslant 2\beta(1-\beta)$.

Remark 6 *In order to have the maximal input of the second-order scheme, we set*

$$\theta_{opt} = 2\beta(1-\beta). \tag{2.22}$$

With the choice $\theta = \theta_{opt}$, we get the less restrictive CFL condition

$$\Delta t \leqslant C_{opt} \frac{\Delta x}{c_e}, \text{ where } C_{opt} = \frac{1}{1-\beta}. \tag{2.23}$$

The values of θ_{opt} and C_{opt} are displayed with respect to β in Figure 1.

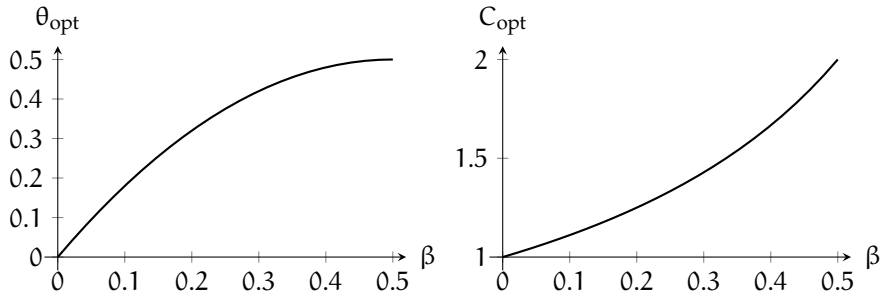

Figure 1: Values of the optimal convex combination parameter θ_{opt} (left panel) and the optimal CFL number C_{opt} (right panel), with respect to the IMEX parameter β.

Remark 7 *Allowing $\beta = \frac{1}{2}$, the 3-stage ARS type method (2.21) reduces to the 2-stage ARS type method using the midpoint rule (2.18). In addition, the choice $\beta = \frac{1}{2}$ maximizes both θ_{opt} and λ.*

Corollary 8 *If there is a second-order CK type IMEX R-K scheme of the form (2.21) that is L^∞-stable in the convex combination with (1.2) under a CFL condition independent of ε, then it has to be of ARS type.*

3 Numerical results

This section is dedicated to providing numerical experiments to test the schemes introduced above:

- The first-order scheme given by (1.2),
- The second-order scheme given by (2.21),
- The L^∞-stable scheme obtained via the convex combination with the parameter $\theta = \theta_{opt}$ given by (2.22), between the first-order scheme (1.2) and the second-order scheme (2.21),
- The MOOD scheme resulting from an optimal order detection procedure explained in Section 3.1 and applied to the L^∞-stable scheme.

Throughout this section, the space domain is given by $[0,1]$ and periodic boundary conditions are prescribed. The time domain is given by $[0, t_{end}]$, where t_{end} chosen such that the exact solution completes exactly one revolution of the space domain, as follows:

$$t_{end} = \frac{1}{c_e + \frac{c_i}{\varepsilon}}.$$

Unless otherwise mentioned, the space and time discretizations are linked with the optimal CFL condition defined by (2.23). The constants c_e and c_i are both taken equal to 1.

We start this section with an introduction to an order detection procedure in Section 3.1. Then, we provide a way to choose the parameter β in Section 3.2. Finally, in Section 3.3, we provide several numerical tests with smooth and especially non-smooth exact solutions. The smooth exact solution is given by

$$w_{ex}^{smooth}(t,x) = 1 + \frac{\varepsilon}{2}\left(1 + \sin\left(2\pi\left(x - \left(c_e + \frac{c_i}{\varepsilon}\right)t\right)\right)\right), \tag{3.1}$$

and describes the transport of a sine wave of amplitude ε. The discontinuous exact solution is given by

$$w_{ex} = \begin{cases} 1+\varepsilon & \text{if } x - \left(c_e + \frac{c_i}{\varepsilon}\right)t \in \left(\frac{1}{4}, \frac{3}{4}\right), \\ 1 & \text{otherwise.} \end{cases} \tag{3.2}$$

which corresponds to the transport of a square wave of amplitude ε.

3.1 Optimal order detection: a MOOD-like technique

The L^∞-stable scheme is a convex combination between the diffusive first-order scheme and the oscillatory second-order scheme. Since those oscillations may violate the maximum principle, we do not wish to use the second-order scheme everywhere in the computational domain. Using the L^∞-stable scheme introduces enough diffusion to get rid of the oscillations and to ensure the maximum principle. However, once the diffusion has been introduced, there is no need to add even more diffusion and the second-order scheme could be used until its result once again violates the maximum principle, at which point the L^∞-stable scheme is necessary once again.

The procedure outlined above is akin to the Multidimensional Optimal-Order Detection techniques developed in the MOOD framework (see for instance [2]). It results in the MOOD scheme, given by the algorithm below:

Algorithm 9 *If the exact solution is bounded between w_{min} and w_{max}, using the optimal CFL number (2.23), the MOOD scheme is given as a result of applying the following algorithm at each time step.*

1. *Compute the second-order solution.*
2. *Detect if this second-order solution breaks the maximum principle, i.e. if it oscillates below w_{min} or above w_{max}.*

3. *If the maximum principle is violated, compute and output the solution given by the L^∞-stable scheme; otherwise, output the second-order solution.*

This algorithm ensures a drastic improvement in the numerical results when this procedure is used instead of using the L^∞-stable scheme at each time step.

3.2 Choice of β in the 3-stage method

This first set of numerical experiments is dedicated to providing a way to choose an optimal value for β. At the moment, we know that $\beta \in (0, 1/2)$ and we are able to find a non-zero value of θ for all values of β. According to Figure 1, the optimal CFL number as well as the optimal θ increase as β goes to $1/2$. Therefore, it would be tempting to take β as close to $1/2$ as possible. To check whether this preliminary analysis is accurate, we study the CPU time and the L^∞ error of the scheme with respect to β, in order to suggest an optimal value of β.

Throughout this set of numerical experiments, we consider the smooth exact solution (3.1) with $\varepsilon = 10^{-1}$.

Study of the CPU time. The CPU time taken by our program is influenced by β because the CFL number number C_{opt}, given by (2.23), itself depends on β. Indeed, as β varies from 0 to $1/2$, C_{opt} ranges between 1 and 2, as evidenced in Figure 1.

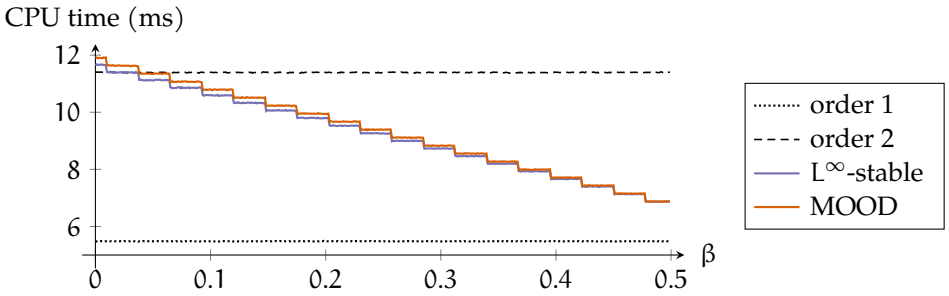

Figure 2: CPU time (in milliseconds) with respect to the IMEX parameter β, using the optimal values θ_{opt} and C_{opt}, in the context of the test case presented in Section 3.2.

In Figure 2, we note that the CPU time for the L^∞-stable and MOOD schemes decreases when β tends to $1/2$. This was expected as the CFL number C_{opt} is increasing with β, thus allowing for larger time steps. Let us also note that the MOOD procedure is not very costly for this smooth test case. Moreover, we remark that the second-order scheme takes twice as much CPU time as the first-order scheme, which is also expected due to the additional intermediate step.

Study of the L^∞ error. Now, we turn to the study of the L^∞ error with respect to β. For $\beta \in (0, 1/2)$, the L^∞-stable and MOOD schemes are L^∞-stable, but this property alone does not indicate their precision. From now on, we take the optimal CFL number C_{opt}.

In Figure 3, we observe that the second-order scheme is, as expected, much more precise than the first-order one. In addition, we note that the L^∞-stable scheme is more precise than the first-order one, but not by a large margin. Finally, we remark that the MOOD procedure is essential to improve the precision of the L^∞-stable scheme.

Regarding the choice of β, we note on the top right panel that the L^∞-stable scheme reduces to the first-order one in two cases. When $\beta = 0$, we get $\theta_{opt} = 0$, and the convex combination consists only in the first-order scheme. When $\beta = 1/2$, we get $\theta_{opt} = 0$ and $C_{opt} = 2$, and the convex combination actually coincides with

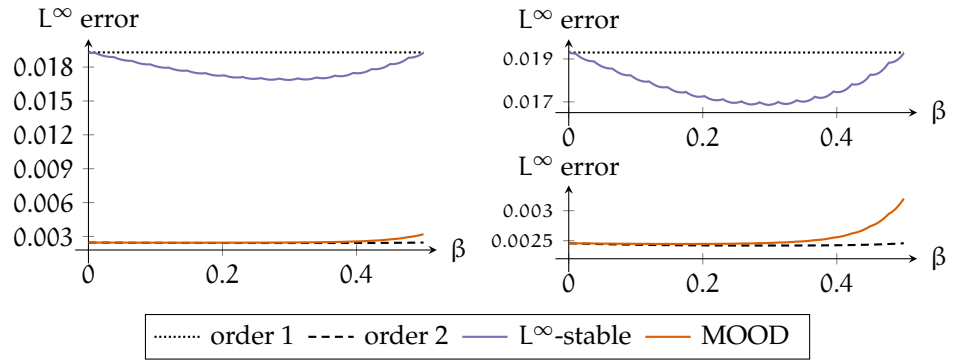

Figure 3: L^∞ error with respect to the IMEX parameter β, using the optimal values θ_{opt} and C_{opt}, in the context of the test case presented in Section 3.2. The right panels contain a zoom on the left panel data.

the first-order scheme. Between these two values, the L^∞ error produced by the L^∞-stable scheme reaches a minimum. Interestingly, this minimum is close to the point where the MOOD error starts increasing (see the bottom right panel). We note that this minimum is located around $\beta \simeq 1 - \sqrt{2}/2$, which is widely used e.g. in [1, 9, 3].

Conclusion: choice of β. In this first study, we have observed that:

- the CPU time gets smaller as β gets larger;
- the L^∞ error reaches a minimum at $\beta = 1 - \sqrt{2}/2$.

Based on this observations, we define β_{opt}, which will be used in the remainder of this article, as

$$\beta_{opt} = 1 - \frac{\sqrt{2}}{2}.$$

3.3 Numerical tests

Before we start with the numerical results, we want to remark that we do not consider an increase in space order. Such an increase, and its effect on smooth solutions, has been documented at length in [3]. Therein was concluded that, if ε is close to 1, then using a second-order scheme in time and a first-order scheme in space does not provide a significant and observable gain compared to a first-order scheme in time and space. Conversely, if ε is close to 0, then using a first-order scheme in time and a second-order scheme in space does not provide a significant and observable gain compared to a first-order scheme in time and space.

Therefore, we focus here only on second-order time accuracy whereas accuracy in space will be studied in forthcoming work.

3.3.1 Smooth solution: order of accuracy

To demonstrate that our schemes reach the desired order of accuracy, we compute L^∞ error curves with the smooth initial condition (3.1). In Figure 4, we display the L^∞ error with respect to the number of discretization points for the four schemes under consideration.

We note, as expected, that the first- and second-order schemes are respectively first- and second-order accurate. Moreover, the L^∞-stable scheme is first-order accurate, and the MOOD procedure greatly increases the

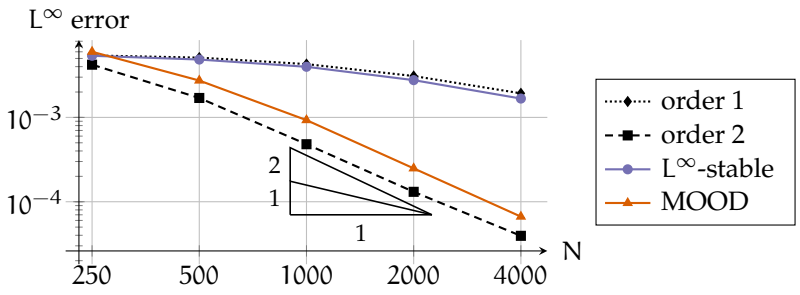

Figure 4: L^∞ error curves for the smooth solution (3.1), with $\varepsilon = 10^{-2}$.

precision of the L^∞-stable scheme, almost bringing it to the level of the second-order scheme. The loss of precision of the MOOD scheme compared to the second-order scheme is due to the fact that the MOOD scheme is L^∞-stable, contrary to the second-order scheme, and therefore it does not allow any violation of the maximum principle, even if such a violation would result in a precision increase.

As a consequence, the MOOD procedure is especially well-suited for smooth problems where the maximum principle is important. Let us now compare these approaches on a discontinuous solution, where we expect the L^∞-stable scheme to be of greater interest.

3.3.2 Discontinuous solution

We now consider the following discontinuous exact solution w_{ex}. In Figure 5, we display the results of the four schemes for different values of ε.

We first notice in the top left panel that the approximation of the exact solution is similar for all four schemes in the case of $\varepsilon = 1$.

In the other three panels, for $\varepsilon \in \{10^{-1}, 10^{-2}, 10^{-3}\}$, we note that the first-order scheme is always in-bounds, while the second-order scheme always violates the maximum principle. Here, we observe a clear improvement when using the L^∞-stable scheme, but the result is still somewhat diffusive. The MOOD procedure allows another gain in precision compared to the first-order scheme, while still staying in-bounds.

This underlines the necessity of L^∞-stable schemes when approximating discontinuous solutions. In addition, the MOOD procedure is useful when approximating continuous and discontinuous solutions with good precision, while respecting the maximum principle.

The final numerical experiment consists in quantifying how much better the result of the L^∞-stable scheme is, compared to both first- and second-order approximations, when considering a discontinuous solution. To address such an issue, we cannot simply compute the error in the L^∞ norm. Indeed, this norm is not well-suited for measuring the errors produced when approximating a discontinuous exact solution with a diffusive approximation. Instead, we turn to the L^1 norm, as well as a modification, the L^1_o quasinorm, which does not satisfy the triangle inequality property of a norm but enables us to measure relevant errors, defined as follows:

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

This quasinorm is the L^1 norm added to a quantity which has been designed to measure only overshoots and undershoots. This quantity encodes how much the numerical solution violates the maximum principle. Therefore, we expect this added term to vanish as soon as the L^∞-stable scheme, with or without MOOD, is employed.

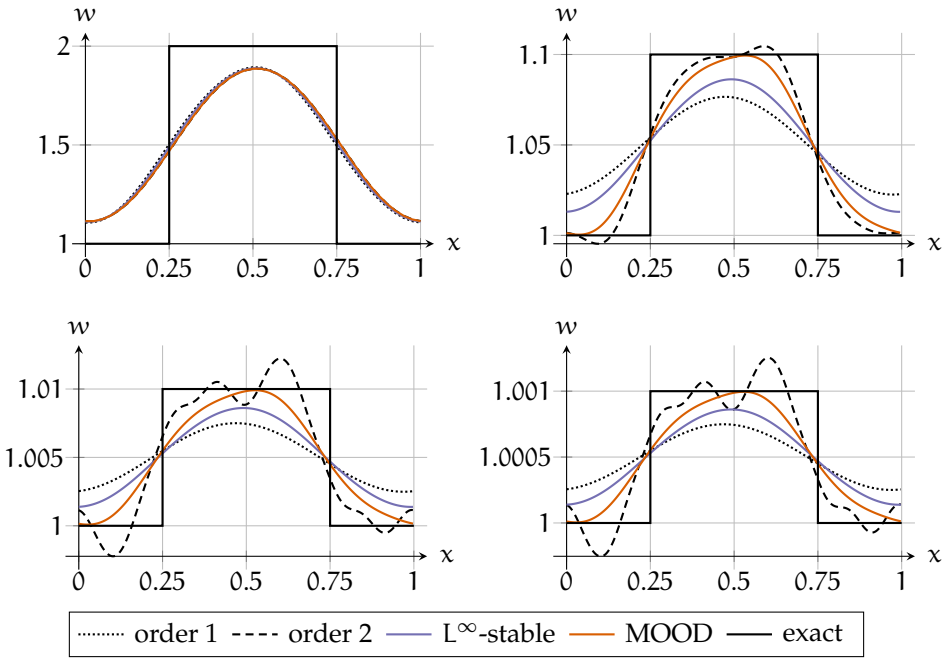

Figure 5: Approximation of the discontinuous solution (3.2). From left to right and top to bottom, we have taken: $\varepsilon = 1$ and $N = 40$, $\varepsilon = 10^{-1}$ and $N = 220$, $\varepsilon = 10^{-2}$ and $N = 2000$, $\varepsilon = 10^{-3}$ and $N = 20000$. These large values of N have been chosen to ensure that 20 time iterations are systematically needed to reach t_{end}. If smaller values are taken, the time steps are too large to visualize noticeable differences between the schemes.

In the top panels of Figure 6, we note that, for $\varepsilon = 1$, both errors take similar values for the four schemes under consideration. This is due to the fact that there are few spurious oscillations in this case (see Figure 5, top left panel). In addition, we observe that the scheme is accurate up to order 1/2 which is expected when approximating discontinuous solutions, see for instance [7].

Now, looking at the bottom left panel, we note that the L^1 error is lower for the second-order scheme than for the other ones and that the orders of accuracy of all schemes tend to 1/2 for large enough N. However, the bottom right panel, which takes into account the over- and undershoots when computing the error, paints another picture: the second-order scheme is actually the worst of all four. In addition, the error actually stays roughly constant when the number of discretization points increases. This means that, as N increases, the gains in L^1 error seem to be compensated by an increase of the overshoot and undershoot amplitude.

4 Conclusions and future work

We have presented a way of constructing L^∞-stable IMEX schemes that, combined with a MOOD procedure, yield high-precision approximate solutions for stiff and non-stiff systems. As we have demonstrated with simple numerical examples, for non-stiff systems higher order IMEX R-K schemes still give good results although violating the maximum principle, whereas for stiff systems they produce spurious oscillations and L^∞-stable schemes are needed to give accurate solutions. In this work, we have mainly focused on the time accuracy and have neglected higher order space discretizations. This, together with the extension to TVD and higher order IMEX schemes, is explored in [8]. In addition, for physical applications, asymptotic preservation properties, as

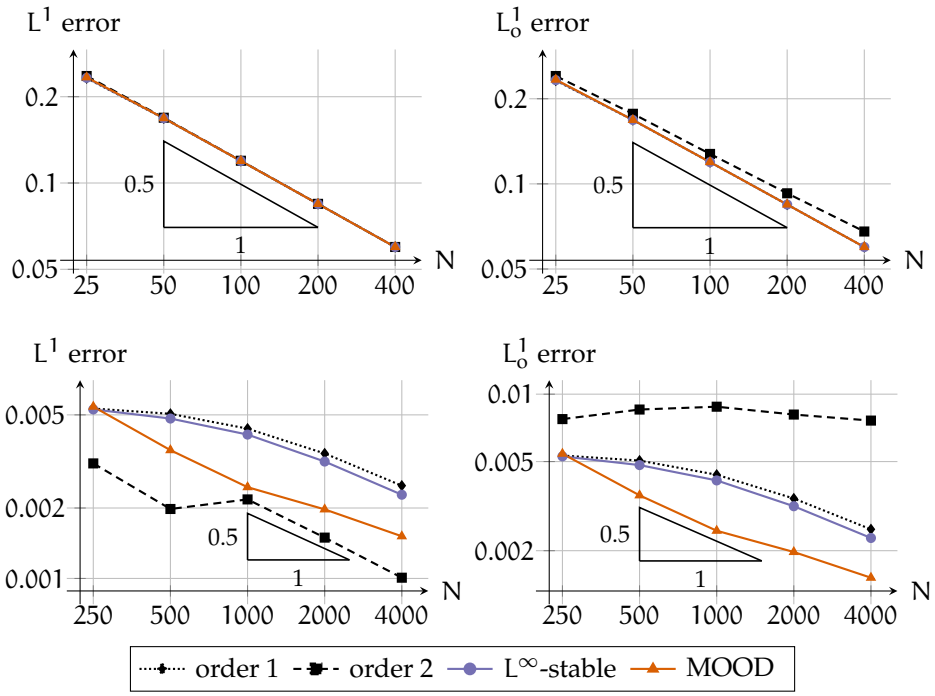

Figure 6: L^1 (left panels) and L^1_o (right panels) error curves for the discontinuous solution (3.2), for $\varepsilon = 1$ (top panels) and $\varepsilon = 10^{-2}$ (bottom panels).

well as scale-independent diffusion, will be studied.

Acknowledgments. V. Michel-Dansac extends his thanks to the Service d'Hydrographie et d'Océanographie de la Marine (SHOM) for financial support.

A. Thomann acknowledges the support of the INDAM-DP-COFUND-2015, grant number 713485.

References

[1] 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).

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

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

[8] V. Michel-Dansac and A. Thomann. TVD IMEX Runge-Kutta schemes based on arbitrarily high order Butcher tableaux. submitted, 2020.

[9] 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.