MATHEMATICS OF COMPUTATION
Volume 80, Number 276, October 2011, Pages 2097–2126
S 0025-5718(2011)02463-4
Article electronically published on February 4, 2011

ON STABILITY ISSUES FOR IMEX SCHEMES APPLIED TO 1D SCALAR HYPERBOLIC EQUATIONS WITH STIFF REACTION TERMS

R. DONAT, I. HIGUERAS, AND A. MARTÍNEZ-GAVARA

Abstract. The application of a Method of Lines to a hyperbolic PDE with source terms gives rise to a system of ODEs containing terms that may have very different stiffness properties. In this case, Implicit-Explicit Runge-Kutta (IMEX-RK) schemes are particularly useful as high order time integrators because they allow an explicit handling of the convective terms, which can be discretized using the highly developed shock capturing technology, together with an implicit treatment of the source terms, necessary for stability reasons.

Motivated by the structure of the source term in a model problem introduced by LeVeque and Yee in [J. Comput. Phys. 86 (1990)], in this paper we study the preservation of certain invariant regions as a weak stability constraint. For the class of source terms considered in this paper, the unit interval is an invariant region for the model balance law. In the first part of the paper, we consider first order time discretizations, which are the basic building blocks of higher order IMEX-RK schemes, and study the conditions that guarantee that $[0, 1]$ is also an invariant region for the numerical scheme. In the second part of the paper, we study the conditions that ensure the preservation of this property for higher order IMEX schemes.

1. Introduction

Many physical problems are governed by hyperbolic conservation laws with non-vanishing source terms. If the source terms model chemical reactions between different species and the reactions happen on time scales much faster than the fluid dynamic time scales, the solutions can develop thin reaction zones where the chemical-kinetics activity is concentrated. The time scales driving the reaction terms are typically several orders of magnitude faster than the scale on which the solution is evolving, and on which one would like to compute, so that such problems are said to have stiff source terms, in analogy with the classical case of stiff ordinary differential equations (ODEs).

In explicit methods, taking a time step appropriate for the slower scale of interest can result in violent numerical instabilities, caused by the fast scales. Stability, meaning the absence of violent oscillatory behavior, is often achieved by using implicit techniques in the numerical treatment of the source term. A variety of excellent implicit methods have been developed for solving stiff systems of ODEs, and in many cases these methods are used within a fractional splitting approach in

Received by the editor July 31, 2009 and, in revised form, April 26, 2010 and July 6, 2010.
2010 *Mathematics Subject Classification.* Primary 35L65, 65M20, 65L06, 65L20, 65M12.
The authors acknowledge support from projects MTM2008-00974, MTM2008-00785, MTM2006-01275 and MTM2009-07719.

order to obtain stable results for numerical simulations involving PDEs with stiff source terms.

Non-split methods, however, have some advantages and are sometimes preferred. Nevertheless, when a time-dependent PDE involves terms that need a differentiated numerical treatment, it is natural to employ different discrete strategies for each one of them. One example of this kind of situation is provided by PDEs of convection-diffusion type. Here, linear multistep Implicit Explicit (IMEX) methods were proposed and analyzed as far back as the late 1970s [31, 4]. Instances of these methods have been successfully applied to the incompressible Navier-Stokes equations [17] and in environmental modeling studies [32]. A comparative study for PDEs of convection-diffusion type was carried out in [1], and a corresponding study for reaction-diffusion problems arising in morphology is reported in [27].

In the context of hyperbolic conservation laws with stiff source terms, IMEX-Runge-Kutta (IMEX-RK) schemes appear as a natural extension of the Runge-Kutta time discretizations widely used as high order time integrators for homogeneous conservation laws. IMEX Runge-Kutta schemes have been successfully applied by Pareschi and Russo in [26] to relaxation schemes for hyperbolic conservation laws. The success of these techniques relies on the ability to treat the convective part in an explicit fashion, while still maintaining an implicit handling of the source terms, which gives a distinct advantage when designing a general purpose high order, high resolution, numerical scheme.

In general, when a PDE is solved numerically, it is natural to expect, or require that the numerical solution has as many qualitative properties of the analytical solution as possible. Stability requirements stem from the desire to have numerical schemes that preserve, at the discrete level, certain properties of the analytic solution of the problem to be solved.

For homogeneous conservation laws, Strong Stability Preserving (SSP) high order time discretizations provide an appropriate framework for nonlinear stability, and are widely used in practice. The basic assumption is the strong stability (in any norm, seminorm or convex functional) of the spatial discretization coupled with the Explicit Euler (EE) method for the time discretization, under a known step-size restriction $\triangle t \leq \tau_{EE}$. High order SSP-RK time discretizations preserve the strong stability properties of the EE scheme, while achieving higher order accuracy in time, perhaps under a modified step-size restriction of the form

$$\triangle t \leq C\, \tau_{EE}\,. \tag{1.1}$$

Nowadays, the theory of SSP-RK schemes is well developed (see e.g. [8, 6, 10, 11, 7]), and the connections between SSP theory and contractivity theory [5, 7, 19] has lead to important advances in the field. In particular, it is now known that for an explicit RK scheme, the largest step-size restriction for strong stability preservation is of the form (1.1), with C being the radius of absolute monotonicity of the RK method (see [6, 10, 11]).

In [29], the first paper where high order SSP-RK time discretizations for homogeneous conservation laws were proposed, the relevant seminorm was the Total Variation (TV) semi-norm. Since the true solution of a scalar, homogeneous, conservation law has the TVD (Total Variation Diminishing) property, imposing this requirement on the numerical scheme is appropriate. Indeed, Forward Euler time discretizations of a scalar, homogeneous, conservation law are TVD, under a

convenient CFL restriction, provided that the discrete divergence operator is appropriately chosen. In addition, the convergence of TVD schemes for homogeneous conservation laws follows from Helly's theorem (see e.g. [23]), hence, TVD requirements provide adequate stability constraints in the homogeneous case.

For a scalar balance law, the properties of the solution strongly depend on certain properties of the source term. For example, when the source term is a non-increasing function, the TV of the exact solution of the scalar balance law is also a non-increasing function, as in the homogeneous case. In this case, it is shown in [3] that some semi-implicit and fully implicit schemes do inherit this property, in a rather natural way, thanks to the non-increasing character of the source term.

In general, however, the source term might not be decreasing and these results are not applicable. The basic model problem in [24] provides an example of this situation. The source term has multiple zeros in the domain of interest, hence it is not a decreasing function, so that the arguments in [3] cannot be applied. The TV of the solution (or the L_∞ norm) of a hyperbolic PDE such as that of the LeVeque and Yee (LV&Y) model problem [24] is not guaranteed to decrease in time, hence no strong stability constraints can be expected to hold in first order time discretizations.

It is worth mentioning that clearly visible numerical oscillations can be observed in the numerical results obtained in [24], either with a semi-implicit generalization of MacCormack's scheme, including limiters, or with space-time splitting schemes, also including limiters. Motivated by these results, in this paper we consider instead other properties that can be imposed on the numerical scheme in order to guarantee the absence of spurious oscillatory behavior. In this paper we argue that the preservation of certain *invariant regions*, the interval $[0, 1]$ for the LV&Y model problem, may be turned into a *weak stability* concept that can easily be analyzed for explicit, implicit and semi-implicit first order schemes, as well as for Diagonally Implicit-Explicit Runge-Kutta (D-IMEX) schemes.

The analysis in this paper is strongly connected to the SSP theory for the homogeneous case. The work of Shu and Osher in the development of TVD Runge-Kutta schemes [29], and later on the development of SSP schemes (see [8, 7]), has put in evidence the importance of re-interpreting a Runge-Kutta scheme as a convex combination of Euler steps. In a similar way, we seek to obtain general conditions that guarantee the preservation of our *invariance* property, i.e., the preservation of $[0, 1]$ as an invariant region for the numerical scheme, provided that this interval is an invariant region for certain Euler-type time discretizations of each of the operators involved. For Runge-Kutta methods, this way of proceeding is closely related to the one followed in the context of numerical positivity [14, 15].

The paper is organized as follows: In section 2 we recall various theoretical results concerning the entropy solution of balance laws and review the model problem by LeVeque and Yee in [24] and its generalization in [23], which serve to establish a general class of source terms that have the unit interval as an invariant region. In section 3, we introduce the basic terminology concerning the Method of Lines (MOL) and we demonstrate that, for the class of source terms cosidered, the interval $[0, 1]$ is also an *invariant region* for the MOL system, when monotone fluxes are used in the discrete divergence operator. The necessary notation to deal with high order IMEX methods is introduced in section 4, as well as some considerations on stability issues. The invariance of the interval $[0, 1]$ for some first order schemes

is studied in section 5. The main result in the paper, which concerns the preservation of the numerical invariance of $[0,1]$ for certain high order IMEX schemes is developed in section 6. A key point is the ability to construct an appropriate splitting for the coefficient matrix of the implicit scheme. Section 7 is devoted to this important issue. Results in sections 6 and 7 are stated for general hyperbolic problems with source terms, and in section 8 we show how they can be applied to the LV&Y model problem when two concrete IMEX Runge-Kutta methods are used. Some numerical experiments done in section 8.2 confirm the agreement between the predicted and observed step-size restrictions. The paper ends with some conclusions and forthcoming work.

2. LeVeque and Yee model problem

The Cauchy problem for a scalar conservation law with a source term depending only on the solution can be stated as follows:

$$u_t + f(u)_x = s(u)\,, \quad u(x,0) = u_0(x), \qquad x \in \mathbb{R},\, t > 0. \tag{2.1}$$

The properties of the source term have a direct impact on certain important properties of the true solution. The following result (see e.g. [3, 20]) summarizes the properties of the solution most relevant to the analysis in this paper.

Theorem 2.1. *Consider the problem* (2.1) *with initial condition* $u(x,0) = u_0(x) \in L^\infty(\mathbb{R}) \cap L^1(\mathbb{R})$ *and let* γ *be such that*

$$\sup_u s'(u) \le \gamma\,. \tag{2.2}$$

Then (2.1) *possesses a unique entropy solution* $u(x,t)$ *satisfying*

i) $\|u(x,t)\|_{L^\infty(\mathbb{R})} \le e^{\gamma\, t}(\|u_0(x)\|_{L^\infty(\mathbb{R})} + |s(0)|t)$.

ii) *For any* $v_0 \in L^\infty(\mathbb{R})$ *and corresponding entropy solution* $v(x,t)$*, we have*

$$\|u(x,t) - v(x,t)\|_{L^1(\mathbb{R})} \le e^{\gamma\, t}\|u_0(x) - v_0(x)\|_{L^1(\mathbb{R})}\,.$$

iii) *If* $u_0(x) \le v_0(x)$*, then the corresponding entropy solutions satisfy*

$$u(x,t) \le v(x,t)\,. \tag{2.3}$$

In [3, 28], the authors study the behavior of certain first order schemes applied to the model balance law (2.1) when

$$s(0) = 0, \quad \sup_u s'(u) \le 0. \tag{2.4}$$

For this class of source terms, (i) above implies monotonicity in the L^∞ norm of the solution, and (ii) implies that the solution has the TVD property, just as in the homogeneous case. In [3], first order monotone (or second order TVD) discretizations for the convective derivative are combined with a centered, pointwise discretization of the source term. It is seen in [3] that the semi-implicit, or fully implicit, numerical schemes constructed in this way inherit the monotonicity (and TVD) properties of the true solution (through a CFL-type restriction in the semi-implicit case).

If the source term is not a decreasing function, these results can no longer be expected to hold, as the proofs strongly rely on the fact that $s'(u) < 0$ in the domain of interest. Examples of such source terms are provided by the scalar model proposed in [24] (see also [9]), and later generalized in [23] as follows:

$$u_t + f(u)_x = s(u), \quad s(u) = -\mu\, u\,(u-1)(u-\beta), \quad 0 < \beta < 1\,, \mu > 0\,. \tag{2.5}$$

The source term in (2.5) has multiple equilibrium points, hence it does not fit in the general framework established in [3]. In general, we cannot expect monotonicity for the L^∞ norm, or the TV semi-norm, of solutions to (2.5), and these properties cannot be expected to hold either in the numerical solutions obtained with the first or second order schemes considered in [3].

In fact, for a balance law such as (2.5), only the monotonicity property (iii) in Theorem 2.1 remains, with respect to the corresponding results for the homogeneous case, or for source terms satisfying (2.4). It is not difficult to analyze the restrictions on the time step that are necessary in order to ensure that the numerical schemes constructed in [3] are monotone, hence they satisfy that if $U^n \leq V^n$, then $U^{n+1} \leq V^{n+1}$ (see section 5). However, it is well known that monotone schemes are necessarily first order accurate (see e.g. [23, 30]), hence we cannot expect this property to hold for higher order IMEX schemes.

The model problem (2.5) with $f(u) = u$ and $\beta = 0.5$ was used in [24] to illustrate a current well-known deficiency of most numerical schemes for hyperbolic conservation laws with stiff source terms: the occurrence of numerical fronts that propagate at non-physical speeds. The parameter μ controls the stiffness of the model. For $\mu > 0$, the ODE $u' = -\mu\, u\,(u-1)(u-\beta)$ has stable equilibria at $u = 0$ and $u = 1$, and an unstable equilibrium at $u = \beta$. In [24, 23], Riemann IVPs with $u_L, u_R \in \{0, 1\}$ (the two stable points of the ODE $u' = s(u)$) are used to illustrate the pathologies that can occur in the numerical approximation of this model problem. When $u_L = 1$, $u_R = 0$, the solution is a discontinuous profile joining the left and right states, that moves with the speed determined by the homogeneous conservation law, i.e., 0.5 for $f(u) = u^2/2$, as in [23], and 1 for $f(u) = u$, as in [24]. The case $u_L = 0$ and $u_R = 1$ is examined in [23], section 17.15. In this case, Burgers equation with no source term gives a rarefaction wave that spreads out the initial discontinuity, while the source term opposes this smearing and tends to drive the intermediate values back towards 0 or 1. These competing effects balance out and result in a smooth solution that rapidly approaches a traveling wave of the form

$$u(x,t) = w(\mu(\beta\, t - x)) \tag{2.6}$$

that moves at speed β (see [23] for details). Increasing μ leads to traveling wave profiles that resemble discontinuous fronts, and whose numerical simulation suffers from the same pathological behavior as in the shock case.

It is worth noticing that property (iii) in Theorem 2.1 readily implies that the unit interval $[0, 1]$ is an invariant region for the model (2.5). Since $s(0) = 0 = s(1)$, it follows that $u(x,t) \equiv 0$ and $u(x,t) \equiv 1$ are solutions of the PDE, then application of part (iii) in Theorem 2.1 gives that

$$0 \leq u(x,0) \leq 1 \qquad \Rightarrow \qquad 0 \leq u(x,t) \leq 1\,, \quad \forall t > 0. \tag{2.7}$$

Thus, it seems reasonable to require similar inequalities to hold for the numerical solution, i.e.,

$$0 \leq U_i^n \leq 1\,, \qquad i = 1 \ldots N\,, \qquad n \geq 1\,, \tag{2.8}$$

since the numerical oscillations observed in the test problems considered in [24, 23], always correspond to a violation of (2.8).

Notice that (2.7) simply states that $[0, 1]$ is an *invariant domain* for the balance law (2.5) while (2.8) states that $[0, 1]$ is an *invariant domain* for the numerical

scheme (see e.g. [2]). We remark that this 'weak monotonicity' is not ensured by the use of classical flux-limiters in the discretization of the convective derivative, as demonstrated by the results in [24] (see Figures 1-2 in [24]).

In this paper, we concentrate on the conditions to be imposed on a numerical scheme in order to ensure that invariant regions, between two equilibrium points of the source term in the model problem (2.1), remain invariant also for the numerical solution obtained from the numerical scheme. In practice, this desirable property leads to non-oscillatory solutions, hence we consider it a weak form of stability.

The emphasis here is not on the specific form of the source term. Our assumptions on the source term will only be that $s(u) \in \mathcal{C}^2([0,1])$ and

$$s(0) = 0 = s(1), \qquad -m \le s'(u) \le M, \qquad u \in [0,1], \quad m \ge 0, M \ge 0, \tag{2.9}$$

so that $s'(u)$ might change sign in the domain of interest. The equilibrium points are kept as $u = 0$ and $u = 1$ for simplicity in the exposition, hence $\mathcal{I} = [0,1]$ is the *invariant region* to be preserved by the numerical scheme. The source terms considered in [24] and [23] belong to this class.

Lemma 2.2. *For $s(u) = -\mu\, u\,(u-1)\,(u-\beta)$, we have (2.9) with*

$$m = \mu \max(\beta, 1-\beta), \qquad M = \mu\,(1-\beta+\beta^2)/3. \tag{2.10}$$

The proof is trivial. Notice also that $m > 0$ and $M > 0$ for all values $\beta \in (0,1)$.

3. Method of lines (MOL) discretizations

Given a convection-reaction problem of the form (2.1), a standard application of the Method of Lines (MOL) reduces the PDE to an initial value problem for a system of ODEs of the form,

$$\frac{\partial U}{\partial t} = D(U(t)) + S(U(t))\,, \qquad U(0) = U_0\,. \tag{3.1}$$

In this paper, we shall consider that the initial vector $U_0 = (u(x_1,0), \ldots, u(x_N,0))^t$, with $\{x_i\}_{i=1}^N$ being the spatial computational mesh. Hence, the solution at time t is $U(t) = (U_1(t), U_2(t), \ldots, U_N(t))^t$, where $U_i(t) \approx u(x_i,t)$. The term $D(U)$ in (3.1) represents the discretization of the convective derivative term, $-f(u)_x$. It is well known that when shock computations are involved, conservative formulations,

$$D_i(U) = -\frac{F_{i+1/2} - F_{i-1/2}}{\triangle x} \tag{3.2}$$

must be considered. Here $F_{i+1/2} = F(U_{i-q+1}, \ldots, U_{i+q})$ is a numerical flux function consistent with the convective flux $f(u)$, i.e., $F(U, \ldots, U) = f(U)$, and F is Lipschitz continuous with respect to its arguments. The term $S(U)$ in (3.1) represents the discrete approximation of the source term, $s(u)$. In this paper, we always consider

$$S_i(U) = s(U_i)\,, \tag{3.3}$$

although some results obtained in the paper might be valid for some other discretizations, for example, those of the form

$$S_i(U) = g(U_{i-1}, U_i, U_{i+1})\,, \tag{3.4}$$

where the function $g(u,v,w)$ satisfies at least that $g(u,u,u) = s(u)$ (see e.g. [9]).

3.1. **Monotonicity in MOL Discretizations.** The aim of this section is to show that the choice of monotone discretization operators in the divergence operator, as in [3], ensures that $[0,1]$ is an *invariant domain* for the ODE system obtained from the MOL discretization of the model problem (2.1), when $s(0) = 0 = s(1)$, i.e.,

$$0 \le U_i(t) \le 1, \qquad i = 1, \dots, N, \quad t \ge t_0, \tag{3.5}$$

We include the basic definitions and, as in [3], restrict ourselves to the case of three-point monotone schemes, for the sake of simplicity.

A monotone three-point divergence operator has the form

$$D_i(U) = -\frac{F(U_i, U_{i+1}) - F(U_{i-1}, U_i)}{\triangle x}\,, \tag{3.6}$$

where the numerical flux $F(U,V)$ is a non-decreasing function of its first argument U and a non-increasing function of its second argument V. We will denote this property by $F(\uparrow,\downarrow)$ and refer to such numerical flux functions as *monotone*. Well known monotone fluxes (under an appropriate CFL condition) include the Lax-Friedrichs, the Godunov and the Engquist-Osher fluxes.

The proof of property (3.5) for the MOL system with monotone flux functions follows neatly from the theory of monotone dynamical systems. Here, we only give a brief description of the necessary concepts. For full details on the underlying theory we refer the reader to [13, 16].

Let us consider the initial value problem

$$U'(t) = G(U,t)\,, \qquad U(t_0) = U_0\,, \tag{3.7}$$

where $G : D \times J \to \mathbb{R}^N$ is a locally Lipschitz vector-valued function, $D \subset \mathbb{R}^N$ an open set, and $J \subset \mathbb{R}$ a non-trivial open interval. The theory of monotone dynamical systems studies when, given initial values U_0 and V_0 such that $U_0 \le V_0$, the solutions $U(t)$ and $V(t)$ satisfy $U(t) \le V(t)$, i.e.,

$$U_0 \le V_0 \qquad \Longrightarrow \qquad U(t) \le V(t)\,, \quad \forall t \in J\,. \tag{3.8}$$

In the above expressions, and in the rest of the paper, the vector inequalities should be understood componentwise. If the ODE system (3.7) satisfies property (3.8), then it is called *monotone*. A sufficient and necessary condition for monotonicity is the Kamke-Müller condition: for all $(U,t), (V,t) \in D \times J$,

$$U \le V \quad \text{ and } \quad U_i = V_i \text{ for some } i \qquad \Longrightarrow \qquad G_i(U,t) \le G_i(V,t)\,.$$

We shall use the Kamke-Müller characterization to prove the monotonicity of the MOL system when the numerical fluxes are monotone.

Theorem 3.1. *The MOL system* (3.1)-(3.6) *is monotone provided that the numerical divergence* (3.6) *is computed with a monotone numerical flux function, and the discrete approximation of the source term satisfies*

$$S_i(U) \le S_i(V) \quad \textit{whenever} \quad U \le V \ \textit{ with } \ U_i = V_i\,. \tag{3.9}$$

Proof. We prove that the function

$$G_i(U) = \frac{F(U_{i-1}, U_i) - F(U_i, U_{i+1})}{\triangle x} + S_i(U)$$

satisfies the Kamke-Müller condition. To this aim, we consider vectors $U = (U_j)_{j=1}^N$, $V = (V_j)_{j=1}^N$, such that $U \leq V$ and $U_i = V_i$. Then,

$$G_i(U) = \frac{F(U_{i-1}, U_i) - F(U_i, U_{i+1})}{\triangle x} + S_i(U) = \frac{F(U_{i-1}, V_i) - F(V_i, U_{i+1})}{\triangle x} + S_i(U)$$
$$\leq \frac{F(V_{i-1}, V_i) - F(V_i, V_{i+1})}{\triangle x} + S_i(V) = G_i(V)\,. \qquad \square$$

The desired result follows easily from the previous theorem.

Corollary 3.2. *Under the assumptions of Theorem* 3.1*, if* $S(0) = S(1) = 0$*, then the solution of the system of ODEs given by* (3.1)-(3.6) *satisfies the following property:*

$$0 \leq U_i(0) \leq 1\,, \quad i = 1, \ldots, N \qquad \Longrightarrow \qquad 0 \leq U_i(t) \leq 1\,, \quad i = 1, \ldots, N\,.$$

Proof. The consistency of the numerical flux function, together with the fact that $S(0) = S(1) = 0$ imply that for $U_i(0) = 0$, $\forall i$ we have $U_i(t) = 0$, $\forall i$, $\forall t$, and also $U_i(0) = 1$, $\forall i$ leads to $U_i(t) = 1$, $\forall i$, $\forall t$. The result follows immediately from the monotonicity of the system of ODEs, granted in Theorem 3.1. $\square$

Notice that, provided $D(U)$ is monotone, $[0, 1]$ is an invariant region for the MOL system whenever the discretization of the source term satisfies the consistency condition $S(0) = S(1) = 0$, which holds trivially for (3.3). In a similar way, discretizations of the source terms of the form (3.4) fulfill condition (3.9) if the function $g(u, v, w)$ is a nondecreasing function of its first, and third arguments.

4. Time stepping procedures

The semi-discrete MOL approach (3.1) decouples the issues of spatial and temporal accuracy and allows for different numerical treatments to be applied to the convective derivative and the source term. It is widely accepted that stiff source terms should be handled in an implicit fashion, in order to avoid stability problems related with the fast scales. On the other hand, there are a number of robust, and rather specialized, numerical flux functions that can be used if discontinuous, or nearly discontinuous, solutions need to be computed. These numerical flux functions are often non-linear and quite complex, hence their use is simpler in an explicit framework. These observations lead, in a rather natural way, to consider IMEX Runge-Kutta schemes for the time integration of the MOL system (3.1).

4.1. **Diagonally Implicit Explicit Runge-Kutta schemes.**

Implicit Explicit Runge-Kutta (IMEX-RK) schemes fall within the wider class of additive RK methods (ARK), which are often employed for solving ODEs like (3.1) when each one of the terms on the right-hand side requires a specific numerical treatment. When simplicity and efficiency in solving the algebraic equations corresponding to the implicit terms of the IMEX-RK process is of importance, Diagonally Implicit Explicit Runge-Kutta (D-IMEX) schemes are often considered, since, in this case, the non-linear system that results from the implicit handling of the source term can be solved stage to stage.

The general form of an s-stage D-IMEX scheme system (3.1) is as follows (see e.g., [26])

$$\begin{aligned} U^{(i)} &= U^n + \triangle t \sum_{j=1}^{i-1} a_{ij}\, D(U^{(j)}) + \triangle t \sum_{j=1}^{i} \tilde{a}_{ij}\, S(U^{(j)})\,, \qquad 1 \le i \le s\,, \\ U^{n+1} &= U^n + \triangle t \sum_{i=1}^{s} b_i\, D(U^{(i)}) + \triangle t \sum_{i=1}^{s} \tilde{b}_i\, S(U^{(i)})\,, \end{aligned} \tag{4.1}$$

where $U^{(i)}$ represent the internal stages of the method.

In the specialized literature concerning ARK schemes, it is customary to use a compact matrix notation to represent the method. Here, following [12], we denote $\mathcal{A} = (a_{ij})$, $\tilde{\mathcal{A}} = (\tilde{a}_{ij})$, $b = (b_i)$ and $\tilde{b} = (\tilde{b}_i)$, and define the matrices

$$\mathbb{A} = \begin{pmatrix} \mathcal{A} & 0 \\ b^t & 0 \end{pmatrix}, \qquad \tilde{\mathbb{A}} = \begin{pmatrix} \tilde{\mathcal{A}} & 0 \\ \tilde{b}^t & 0 \end{pmatrix}.$$

With this notation, the matrix $\mathbb{A}$ contains all the information associated to the explicit scheme, while the matrix $\tilde{\mathbb{A}}$, is directly related to the implicit one. Observe that for a D-IMEX scheme (4.1), $\mathbb{A}$ must be a strictly lower triangular matrix, while $\tilde{\mathbb{A}}$ is only required to be lower triangular.

With the notation above, (4.1) can be expressed in *compact form* as

$$\mathcal{U} = e \otimes U^n + \triangle t\, (\mathbb{A} \otimes I)\, \mathcal{D}(\mathcal{U}) + \triangle t\, (\tilde{\mathbb{A}} \otimes I)\, \mathcal{S}(\mathcal{U}), \tag{4.2}$$

where the symbol $\otimes$ denotes the Kronecker product, $e = (1, \dots, 1)^t \in \mathbb{R}^{s+1}$ and $\mathcal{U} = ({U^{(1)}}^t, \dots, {U^{(s)}}^t, (U^{n+1})^t)^t \in \mathbb{R}^{(s+1)N}$; see [5] and [12] for specific details. Here, $\mathcal{D}(\mathcal{U}) = (D(U^{(1)})^T, \dots, D(U^{(s)})^T, 0^T)^T \in \mathbb{R}^{(s+1)N}$, with analogous notation for $\mathcal{S}(\mathcal{U})$. In the rest of the paper, we only consider centered discretizations of the source term, $S_i(U) = s(U_i)$.

4.2. **Stability issues concerning high order Runge-Kutta methods.** As mentioned in the introduction, Strong Stability Preserving (SSP) high order time discretizations provide an appropriate framework for stability when dealing with homogeneous conservation laws. Given an explicit RK scheme with coefficient matrix $\mathbb{A}$, it is now well known that the largest step-size restriction for strong stability preservation is of the form

$$\triangle t \le \mathcal{R}(\mathbb{A})\, \tau_{EE}\,, \tag{4.3}$$

where $\mathcal{R}(\mathbb{A})$ is the radius of absolute monotonicity of the RK method (see [6, 10, 11]). This important quantity is defined in [19, Equation (4.1)] as the largest value r_1 such that matrix $I + r_1\, \mathbb{A}$ is regular and inequalities

$$(I + r_1\, \mathbb{A})^{-1} e \ge 0\,, \qquad (I + r_1\, \mathbb{A})^{-1} \mathbb{A} \ge 0 \tag{4.4}$$

hold. Furthermore, any $r_1 \le \mathcal{R}(\mathbb{A})$ would satisfy (4.4) [19, Lemma 4.4]. The connections between SSP theory and the theory of contractivity have lead to the development of optimal and efficient SSP schemes [7].

We note that (4.3) implies that for a given RK scheme, non-trivial step-size restrictions are obtained only if $\mathcal{R}(\mathbb{A}) > 0$. It is worth mentioning that, in particular, $\mathcal{R}(\mathbb{A}) > 0$ implies the sign condition $\mathbb{A} \ge 0$ [19, Theorem 4.2].

For ARK methods $(\mathbb{A}, \tilde{\mathbb{A}})$, the concept of region of absolute monotonicity $\mathcal{R}(\mathbb{A}, \tilde{\mathbb{A}})$ extends the concept of radius of absolute monotonicity, in the sense that if EE

discretizations of each additive term in the ODE system are strongly stable, strong stability is preserved for the ARK method, under step-size restrictions that depend on the ones for the EE discretizations and the region of absolute stability of the ARK scheme (see [12] for details).

It is shown in [12] that non-trivial regions of absolute monotonicity are required for an ARK method to be SSP, however, it is interesting to notice that the two preferred IMEX schemes proposed by Pareschi and Russo in [26] have $\mathcal{R}(\mathbb{A}, \tilde{\mathbb{A}}) = \{0\}$ (see [12] and section 8), so that no time step restrictions can be imposed for strong stability preservation. These IMEX schemes are not SSP, according to the results in [12], however, the numerical solutions shown in [26] are perfectly well behaved, with no spurious oscillatory behavior. Hence, we may conclude that using IMEX schemes with non-empty regions of absolute stability might result in stability constraints that are too stringent in practice.

In this paper, we adapt the tools used in the context of SSP-RK and SSP-ARK schemes, to establish sufficient conditions for the *preservation* of a convex property $\mathcal{P}$ that, in practice, guarantees the absence of numerical oscillations. In our case, property $\mathcal{P}$ is the numerical invariance of the unit interval, $[0, 1]$.

We point out that the preservation of this property for RK methods (non-additive), can be obtained from the results in [14, 15] on numerical positivity of RK schemes. In this paper, we propose a non-standard extension of this result to the class of D-IMEX schemes. As in the SSP theory, assuming that $\mathcal{P}$ holds, under appropriate step-size restrictions, for first order time integrators applied to each additive term in (3.1), we shall prove that $\mathcal{P}$ can be ensured for the IMEX scheme, perhaps under a different time-step restriction. Our extension is specifically adapted to the class of source terms that satisfies (2.9).

The preservation of convex properties, such as our property $\mathcal{P}$, in an IMEX scheme will rely on the ability to write the internal stages in (6.2) as a convex combination of Euler steps. We analyze next the properties of first order time stepping methods, which are the basic building blocks of higher order IMEX schemes.

5. First order time-stepping processes

The simplest possible IMEX scheme, is given by the first-order Semi-Implicit (SI) method

$$U^{n+1} = U^n + \triangle t\, D(U^n) + \triangle t\, S(U^{n+1})\,. \tag{5.1}$$

In [3], Chalabi studies the behavior of the numerical solutions obtained when applying (5.1) to a balance law (2.1) for which $s'(u) < 0$. For such source terms, Theorem 2.1 ensures that the L_∞ norm and the TVD semi-norm of the true solution do not increase with time. It is proven in [3] that if the convective derivative is discretized by a monotone, three-point, discrete divergence operator (3.6), the scheme (5.1) is well defined (U^{n+1} can always be computed), and the *Strong Stability* properties of the true solution also hold for the numerical solution, i.e., the L_∞ norm and the TVD semi-norm of the numerical solution do not increase with the number of steps, under an appropriate CFL restriction which depends only on the discrete divergence operator. We remark that these properties allow for a standard convergence proof to the entropy solution of the balance law based on Helly's Theorem, just as in the homogeneous case (see [3] for details).

As mentioned earlier, these results no longer hold for more general source terms, since the proofs rely heavily on the fact that $s'(u) < 0$ in the domain of interest. In this section we assume that the source term satisfies (2.9), and study what restrictions on the time-step are necessary to ensure that $[0, 1]$ is an *invariant region* for the first order IMEX scheme (5.1). The results that we obtain can be considered as extensions of the results in [3].

In addition, we also analyze the Explicit Euler (EE) scheme,

$$U^{n+1} = U^n + \triangle t\, D(U^n) + \triangle t\, S(U^n)\,, \tag{5.2}$$

which was also considered in [28]. Even though it is usual to resort to an implicit treatment of the source terms for stability reasons, explicit techniques might be of interest for mildly stiff problems. Moreover, they are also building blocks of higher order IMEX schemes. Hence, we shall also study the properties of the first order EE scheme (5.2) for the time discretization of (3.1).

In general, we do not assume monotonicity for the discrete divergence operator, as in [3]. However, if the convective derivative is treated in an explicit fashion, consistency with the homogeneous case $U^{n+1} = U^n + \triangle t\, D(U^n)$, demands that the first-order Forward Euler time stepping also preserves the interval $[0, 1]$. Hence, we shall always assume that there exists $\tau_{_D} > 0$ such that

$$0 \le U^n \le e \qquad \Longrightarrow \qquad 0 \le U^n + \tau\, D(U^n) \le e\,, \quad \forall \tau \le \tau_{_D}. \tag{5.3}$$

We remark that numerical fluxes leading to TVD schemes in the homogeneous case will also lead to (5.3), under an appropriate CFL restriction on the time-step.

5.1. **The Semi-Implicit (SI) scheme.** The general result concerning the numerical invariance of $[0, 1]$ for the SI numerical scheme is stated below.

Proposition 5.1. *Consider the semi-implicit scheme* (5.1) *applied to the balance law* (2.1) *with a source term satisfying* (2.9)*, and assume that there exists* τ_D *such that* (5.3) *is fulfilled. Then,* $[0, 1]$ *is an invariant region for* (5.1)*, provided that the time-step is restricted so that*

$$0 \le \triangle t \le \min\left\{\frac{1}{M}, \tau_D\right\}. \tag{5.4}$$

Proof. Let us assume that $0 \le U_i^n \le 1$, for $i = 1, \ldots, N$. Notice that each equation in system (5.1) can be written as $U_i^{n+1} = \tilde{U}_i + s(U_i)$, where $\tilde{U} = U^n + \triangle t D(U^n)$ satisfies $0 \le \tilde{U} \le e$ when $\triangle t \le \tau_D$, thanks to (5.3).

For a given number δ, $0 \le \delta \le 1$, consider the function $g(u) = u - \triangle t\, s(u) - \delta$. Notice that $g(0) = -\delta \le 0$ and $g(1) = 1 - \delta \ge 0$, hence $g(u)$ has at least one root in $[0, 1]$. Notice that $g'(u) = 1 - \triangle t\, s'(u) \ge 1 - M\, \triangle t \ \ \forall u \in [0, 1]$, by (2.9). Hence, if $0 \le \triangle t \le 1/M$, we have that $g'(u) > 0\ \forall u \in [0, 1]$, so that $g(u)$ has at most one root in $[0, 1]$. Hence U_i^{n+1} is uniquely determined for each i, and it belongs to $[0, 1]$, provided that $\triangle t$ satisfies (5.4). □

Remark 5.2. The condition $\triangle t \le 1/M$ guarantees solvability of the non-linear equations to be solved for the internal stages.

For a source term satisfying (2.9), we cannot expect to enforce the strong stability properties considered in [3], i.e., we cannot expect the L_∞ norm or the TVD seminorm of the solution to diminish with time. In fact, only the monotonicity property remains, with respect to the results in Theorem 2.1. We can also prove that

if the discrete divergence operator is monotone, the numerical solution obtained with the SI scheme is monotone, under an appropriate CFL restriction. This result generalizes Proposition 2.2 in [3] to the class of source terms in (2.5).

Proposition 5.3. *The semi-implicit scheme* (5.1) *applied to the balance law* (2.1)*, with a source term satisfying* (2.9) *is monotone if* $D(U)$ *is monotone and* $\triangle t$ *satisfies* (5.4)*.*

Proof. Let $U^n \leq V^n$. Since $D(U)$ is a monotone operator, $\tilde{V}^n - \tilde{U}^n = (V^n + \triangle t D(V^n)) - (U^n + \triangle t D(U^n)) \geq 0$. We recall that for $\triangle t$ satisfying (5.4), we can guarantee that U^{n+1} and V^{n+1} can be computed and satisfy $0 \leq U^{n+1}, V^{n+1} \leq e$. We can then write

$$V^{n+1} - U^{n+1} = \tilde{V}^n - \tilde{U}^n + \triangle t \left(S(V^{n+1}) - S(U^{n+1}) \right),$$

hence, for each internal stage we have

$$(1 - \triangle t\, s'(\xi_i))(V_i^{n+1} - U_i^{n+1}) = \tilde{V}_i^n - \tilde{U}_i^n,$$

for some ξ_i which also belongs to $[0,1]$ which readily implies $U^{n+1} \leq V^{n+1}$. □

5.2. **The explicit Euler (EE) method.** The stability properties of the EE scheme (5.2) will also depend on the ability to establish a stability property for the forward Euler time discretization of the ODE $u' = s(u)$. For the EE scheme, in addition to (5.3), we require that there exists $\tau_+^s > 0$ such that

$$0 \leq U^n \leq e \qquad \Longrightarrow \qquad 0 \leq U^n + \tau\, S(U^n) \leq e\,, \quad \forall \tau \leq \tau_+^s\,. \tag{5.5}$$

Then, we can establish the following result.

Proposition 5.4. *Let* U^n *denote the numerical approximation obtained for the ODE system* (3.1) *with the explicit Euler discretization*

$$U^{n+1} = U^n + \tau\, D(U^n) + \tau\, S(U^n)\,. \tag{5.6}$$

Assume that (5.3) *and* (5.5) *hold for the convective derivative and source term discretizations. Then, we can ensure that*

$$0 \leq U^0 \leq e \qquad \Longrightarrow \qquad 0 \leq U^n \leq e\,, \quad \forall n \geq 0\,, \tag{5.7}$$

under the step-size restriction

$$0 \leq \tau \leq \frac{\tau_+^s}{\tau_{_D} + \tau_+^s} \tau_{_D}\,. \tag{5.8}$$

Proof. Let $\alpha, \beta \in \mathbb{R}$ be such that $0 < \alpha, \beta < 1$ and $\alpha + \beta = 1$. We can write

$$\begin{aligned} U^{n+1} &= U^n + \tau\, D(U^n) + \tau\, s(U^n) \\ &= \alpha \left(U^n + \frac{\tau}{\alpha} D(U^n) \right) + \beta \left(U^n + \frac{\tau}{\beta} s(U^n) \right). \end{aligned}$$

Given any $0 < \alpha < 1$, (5.3) and (5.5) imply that the two terms in the convex combination above remain in $[0,1]$ provided that

$$0 \leq \frac{\tau}{\alpha} \leq \tau_{_D}\,, \qquad 0 < \frac{\tau}{1-\alpha} \leq \tau_+^s\,.$$

Hence, for any given α in $(0,1)$, the components of U^{n+1} remain in $[0,1]$ provided that $\tau \le \min\left\{\alpha\,\tau_{_D}\,,\,(1-\alpha)\,\tau^s_+\right\}$. For

$$\alpha = \frac{\tau^s_+}{\tau_{_D}+\tau^s_+} \tag{5.9}$$

we have $\alpha\,\tau_{_D} = (1-\alpha)\,\tau^s_+$, which proves the result. □

Remark 5.5. When no source term is present we can consider $\tau^s_+ \to \infty$ so that α in (5.9) can be taken as 1 and the step-size restriction (5.8) is that of the scheme for the homogeneous conservation law, i.e., $0 \le \tau \le \tau_{_D}$. Similarly, when the convective term is not present, the step-size restriction (5.8) becomes $0 \le \tau \le \tau^s_+$.

Property (5.5) depends on the specific form of the source term, and has to be checked in each case. For the class of source terms satisfying (2.9), the following result can easily be shown

Lemma 5.6. *If $s(u)$ satisfies* (2.9)*, then*

$$0 \le u + \tau\, s(u) \le 1\,, \qquad \textit{for} \quad 0 \le \tau \le \frac{1}{M}\,, \tag{5.10}$$

$$0 \le u - \tau\, s(u) \le 1\,, \qquad \textit{for} \quad 0 \le \tau \le \frac{1}{m}\,. \tag{5.11}$$

Observe that (5.10) is an explicit Euler step in forward time for $u' = s(u)$, so that for the class of source terms satisfying (2.9) we always have $\tau^s_+ \ge \frac{1}{M}$. In addition, (5.11) is an explicit Euler step in backward time, so that we also have a similar $\tau^s_- \ge \frac{1}{m}$ for backward Euler steps. This interesting property of the class of source terms considered will be relevant later on, in our study of IMEX schemes.

6. Higher order D-IMEX schemes

Following the ideas in the SSP theory, the preservation of a convex property, $\mathcal{P}$, by an IMEX scheme will rely on the ability to write the internal stages in (6.2) as a convex combination of Euler steps. The following technical lemma shows that, under rather mild assumptions, the numerical solution and each internal stage in an additive scheme can be written as a linear combination of Euler steps for the convective and source terms. The proof is straightforward and shall be omitted.

Lemma 6.1. *Consider an additive RK scheme* (4.2) *with coefficient matrices $(\mathbb{A},\tilde{\mathbb{A}})$ and a splitting of the matrix $\tilde{\mathbb{A}}$ as $\tilde{\mathbb{A}} = \tilde{\mathbb{A}}_+ - \tilde{\mathbb{A}}_-$. Let $r_1, r_2, r_3 \in \mathbb{R}$, be non-zero numbers such that the matrix*

$$\mathbb{B} := r_1\,\mathbb{A} + r_2\,\tilde{\mathbb{A}}_+ + r_3\,\tilde{\mathbb{A}}_- \tag{6.1}$$

satisfies that $(I+\mathbb{B})$ is invertible. Then scheme (4.2) *can be rewritten as*

$$\begin{aligned}\mathcal{U} = {} & (I+\mathbb{B})^{-1} e \otimes U^n + r_1\left((I+\mathbb{B})^{-1}\mathbb{A}\otimes I\right)\left(\mathcal{U} + \frac{\triangle t}{r_1}\mathcal{D}(\mathcal{U})\right)\\ & + r_2\left((I+\mathbb{B})^{-1}\tilde{\mathbb{A}}_+\otimes I\right)\left(\mathcal{U} + \frac{\triangle t}{r_2}\mathcal{S}(\mathcal{U})\right)\\ & + r_3\left((I+\mathbb{B})^{-1}\tilde{\mathbb{A}}_-\otimes I\right)\left(\mathcal{U} - \frac{\triangle t}{r_3}\mathcal{S}(\mathcal{U})\right).\end{aligned} \tag{6.2}$$

Lemma 6.1 is valid for general splittings $\tilde{\mathbb{A}} = \tilde{\mathbb{A}}_+ - \tilde{\mathbb{A}}_-$ and real triplets $(r_1, r_2, r_3) \in \mathbb{R}^3$, provided that the matrix $(I + \mathbb{B})$ is invertible. If this is the case, expanding the relation $(I + \mathbb{B})^{-1}(I + \mathbb{B})\, e = e$ we get the following relation:

$$(I + \mathbb{B})^{-1} e + r_1\, (I + \mathbb{B})^{-1} \mathbb{A}\, e + r_2\, (I + \mathbb{B})^{-1} \tilde{\mathbb{A}}_+ e + r_3\, (I + \mathbb{B})^{-1} \tilde{\mathbb{A}}_-\, e = e\,. \tag{6.3}$$

Hence, in order to have a convex combination in (6.2) we require also that the following assumption holds true:

ASSUMPTION SP. *For an additive RK scheme $(\mathbb{A}, \tilde{\mathbb{A}})$, and a given splitting $\tilde{\mathbb{A}} = \tilde{\mathbb{A}}_+ - \tilde{\mathbb{A}}_-$, we can find $r_1, r_2, r_3 \geq 0$ such that matrix $I + \mathbb{B}$ is regular and*

$$(I + \mathbb{B})^{-1} e \geq 0\,, \qquad (I + \mathbb{B})^{-1} \mathbb{A} \geq 0\,, \qquad (I + \mathbb{B})^{-1} \tilde{\mathbb{A}}_\pm \geq 0\,. \tag{6.4}$$

Remark 6.2. For $\tilde{\mathbb{A}} = 0$ (that is, if $s(u) = 0$ or we use a non-additive RK scheme), (6.4) reduces to (4.4). Hence, assumption SP is satisfied for all $r_1 \leq \mathcal{R}(\mathbb{A})$, provided that $\mathcal{R}(\mathbb{A}) > 0$. In other words, the explicit scheme must be SSP. This condition is satisfied by the IMEX schemes considered by Pareschi and Russo in [26].

It is worth noticing the similarities between inequalities (6.4) and those arising in the context of monotone additive RK methods. In particular, given an ARK method $(\mathbb{A}, \tilde{\mathbb{A}})$ and a partition $\tilde{\mathbb{A}} = \tilde{\mathbb{A}}_+ - \tilde{\mathbb{A}}_-$, with a proof similar to the one in [12, Proposition 3.4], we obtain that if there exist values $r_1, r_2, r_3 > 0$ such that inequalities (6.4) hold, we must have the following sign properties:

$$\mathbb{A} \geq 0\,, \qquad \tilde{\mathbb{A}}_+ \geq 0\,, \qquad \tilde{\mathbb{A}}_- \geq 0\,.$$

For ease of reference, we establish the following notation:

$$\begin{gathered} (I + \mathbb{B})^{-1} e = (\alpha_i)\,, \qquad (I + \mathbb{B})^{-1} \mathbb{A} = (\beta_{i,j})\,, \\ (I + \mathbb{B})^{-1} \tilde{\mathbb{A}}_+ = (\tilde{\beta}^+_{ij})\,, \qquad (I + \mathbb{B})^{-1} \tilde{\mathbb{A}}_- = (\tilde{\beta}^-_{ij})\,. \end{gathered} \tag{6.5}$$

The structure of these matrices for D-IMEX schemes and splittings with $\tilde{\mathbb{A}}_-$ strictly lower triangular is summarized below.

Lemma 6.3. *Let $(\mathbb{A}, \tilde{\mathbb{A}})$ be the coefficient matrix of a D-IMEX scheme, and assume that $\tilde{\mathbb{A}} = \tilde{\mathbb{A}}_+ - \tilde{\mathbb{A}}_-$ with $\tilde{\mathbb{A}}_-$ strictly lower triangular. Then*

1. *The matrix $(I + \mathbb{B})$ is lower triangular and $(I + \mathbb{B})_{ii} = 1 + r_2 \tilde{a}_{ii}$. Hence, $\tilde{a}_{ii} \geq 0\ \forall i$, and $r_2 \geq 0$ readily ensure that $(I + \mathbb{B})$ is invertible.*
2. *The matrices $(I + \mathbb{B})^{-1}\mathbb{A}$ and $(I + \mathbb{B})^{-1}\tilde{\mathbb{A}}_-$ are strictly lower triangular; $(I + \mathbb{B})^{-1}\tilde{\mathbb{A}}_+$ is a lower triangular matrix, whose diagonal elements are given by $\tilde{\beta}^+_{ii} = \tilde{a}_{ii}/(1 + r_2 \tilde{a}_{ii})$.*
3. *If assumption SP holds true, then $0 \leq (I + \mathbb{B})^{-1} e \leq 1$.*

Proof. Parts (1) and (2) are trivial. For part (3), note that, under these restrictions, and with the notation established in (6.5), the componentwise version of system (6.3) can be written as

$$\alpha_i + r_1 \sum_{j=1}^{i-1} \beta_{ij} + r_2 \sum_{j=1}^{i} \tilde{\beta}^+_{ij} + r_3 \sum_{j=1}^{i-1} \tilde{\beta}^-_{ij} = 1\,, \qquad i = 1, \ldots, s+1\,. \tag{6.6}$$

The positivity assumption (6.4) implies that all terms in the sum are positive, hence they are all in $[0, 1]$. In particular, we readily get that $0 \leq \alpha_i \leq 1$. □

Given an IMEX scheme $(\mathbb{A}, \tilde{\mathbb{A}})$, and a splitting $\tilde{\mathbb{A}} = \tilde{\mathbb{A}}_+ - \tilde{\mathbb{A}}_-$ with $\tilde{\mathbb{A}}_-$ a strictly lower triangular matrix, Lemma 6.3 ensures that (6.2) provides an expression of the IMEX scheme written as a convex combination of explicit (forward) Euler steps for $\mathcal{D}(U)$, and explicit (forward), explicit (backward) and implicit Euler steps for $\mathcal{S}(U)$, provided that assumption SP holds. The possibility of having explicit Euler steps in forward and backward time for the source term will allow us to obtain the preservation of the convex property $\mathcal{P}$ for the class of source terms satisfying (2.9), when $\mathcal{R}(\tilde{\mathbb{A}}) = 0$.

A second key ingredient in the proof of our main result is the ability to solve the implicit steps involved in the IMEX process. For D-IMEX schemes (4.1), there is only one implicit Euler step involving the source term for each internal stage. Hence, we shall assume that the source term $s(u)$ satisfies the following assumption:

ASSUMPTION IES. *If* $0 \le u^n \le 1$, *then* u^{n+1} *can be computed from*

$$u^{n+1} = u^n + \tau\, s(u^{n+1})$$

for all $0 < \tau \le \tau_{_{IE}}$, *and satisfies* $0 \le u^{n+1} \le 1$.

We notice that the class of source terms considered in [3] fulfills this assumption for $\tau_{_{IE}} = +\infty$. For a source term satisfying (2.9) we have shown that $\tau_{_{IE}} \ge \min\{1/M, \tau_D\}$ (see Proposition 5.1).

The main result in this section establishes a set of sufficient conditions for numerical invariance of the interval $[0, 1]$ in D-IMEX schemes. Since all numerical evidence points out that when numerical oscillations do occur, the values on the numerical wave profile do not lie in $[0, 1]$, non-oscillatory results are expected when these conditions are satisfied.

Theorem 6.4. *Consider a D-IMEX method of the form* (4.2), *with coefficient matrices* $(\mathbb{A}, \tilde{\mathbb{A}})$, *for the ODE* (3.1). *Assume that:*

i) *There exist constants* $\tau_{_D} > 0$, $\tau^s_+ > 0$ *and* $\tau^s_- > 0$ *so that*

$$0 \le U^n \le e\,, \quad \tau \le \tau_{_D} \qquad \Longrightarrow \qquad 0 \le U^n + \tau\, D(U^n) \le e\,, \tag{6.7}$$

$$0 \le U^n \le e\,, \quad \tau \le \tau^s_+ \qquad \Longrightarrow \qquad 0 \le U^n + \tau\, S(U^n) \le e\,, \tag{6.8}$$

$$0 \le U^n \le e\,, \quad \tau \le \tau^s_- \qquad \Longrightarrow \qquad 0 \le U^n - \tau\, S(U^n) \le e\,. \tag{6.9}$$

ii) *The source term* $s(u)$ *satisfies the assumption IES above.*

iii) $\mathbb{A} \ge 0$, *and we have constructed a partition* $\tilde{\mathbb{A}} = \tilde{\mathbb{A}}_+ - \tilde{\mathbb{A}}_-$, *with* $\tilde{\mathbb{A}}_-$ *strictly lower triangular, such that* $\tilde{\mathbb{A}}_+ \ge 0$, $\tilde{\mathbb{A}}_- \ge 0$ *for which the assumption SP is satisfied, i.e., there exists* $r_1, r_2, r_3 \ge 0$ *such that inequalities* (6.4) *hold.*

Then

$$0 \le U^n \le e \qquad \Longrightarrow \qquad 0 \le U^{n+1} \le e\,,$$

under the step-size restriction

$$\triangle t \le \min\{r_1\, \tau_{_D}\,,\, r_2\, \tau^s_+\,,\, r_3\, \tau^s_-, \gamma\, \tau_{_{IE}}\}, \tag{6.10}$$

where $\gamma = 1/\max\{\tilde{a}_{ii}, 1 \le i \le s\}$.

Proof. Notice that, under the hypothesis on the splitting $\tilde{\mathbb{A}} = \tilde{\mathbb{A}}_+ - \tilde{\mathbb{A}}_-$, the matrix $(I + \mathbb{B})$ is invertible for any $r_2 \ge 0$. Given $r_1, r_2, r_3 \in \mathbb{R}^+$, we use Lemma 6.1 to

rewrite (4.2) as (6.2). The hypothesis on $\mathbb{A}$, $\tilde{\mathbb{A}}$ and $\tilde{\mathbb{A}}_{\pm}$, imply that each component $U^{(i)}$, $i = 1, \ldots, s$, of (6.2) has the form

$$U^{(i)} = \alpha_i\, U^n + r_1 \sum_{j=1}^{i-1} \beta_{ij} \left(U^{(j)} + \frac{\triangle t}{r_1} D(U^{(j)}) \right) + r_2 \sum_{j=1}^{i-1} \tilde{\beta}^+_{ij} \left(U^{(j)} + \frac{\triangle t}{r_2} S(U^{(j)}) \right) + r_2\, \tilde{\beta}^+_{ii}\, U^{(i)} + \triangle t\, \tilde{\beta}^+_{ii}\, S(U^{(i)}) + r_3 \sum_{j=1}^{i-1} \tilde{\beta}^-_{ij} \left(U^{(j)} - \frac{\triangle t}{r_3} S(U^{(j)}) \right),$$

which we can rewrite as

$$(1 - r_2\, \tilde{\beta}^+_{ii})\, U^{(i)} = \widehat{U}^{(i)} + \triangle t\, \tilde{\beta}^+_{ii}\, S(U^{(i)}), \tag{6.11}$$

with $\widehat{U}^{(i)}$ defined as

$$\widehat{U}^{(i)} = \alpha_i\, U^n + r_1 \sum_{j=1}^{i-1} \beta_{ij} \left(U^{(j)} + \frac{\triangle t}{r_1} D(U^{(j)}) \right) + r_2 \sum_{j=1}^{i-1} \tilde{\beta}^+_{ij} \left(U^{(j)} + \frac{\triangle t}{r_2} S(U^{(j)}) \right) + r_3 \sum_{j=1}^{i-1} \tilde{\beta}^-_{ij} \left(U^{(j)} - \frac{\triangle t}{r_3} S(U^{(j)}) \right). \tag{6.12}$$

We shall prove the result by an induction process over the internal stages. The following observations will be used in this process.

Notice that we can rewrite (6.6) as follows:

$$\alpha_i + r_1 \sum_{j=1}^{i-1} \beta_{ij} + r_2 \sum_{j=1}^{i-1} \tilde{\beta}^+_{ij} + r_3 \sum_{j=1}^{i-1} \tilde{\beta}^-_{ij} = 1 - r_2\, \tilde{\beta}^+_{ii}, \qquad i = 1, \ldots, s. \tag{6.13}$$

Using part (2) in Lemma 6.3, it is simple to deduce that $1 - r_2\, \tilde{\beta}^+_{ii} = (1 + r_2\, \tilde{a}_{ii})^{-1}$, thus $0 < 1 - r_2\, \tilde{\beta}^+_{ii} < 1$. In addition, $\tilde{a}_{ii} = \tilde{\beta}^+_{ii}/(1 - r_2\, \tilde{\beta}^+_{ii})$. Hence we can we can write (6.11) as

$$U^{(i)} = \bar{U}^{(i)} + \triangle t\, \tilde{a}_{ii}\, S(U^{(i)}), \qquad i = 1, \ldots, s, \tag{6.14}$$

where

$$\bar{U}^{(i)} = \frac{1}{1 - r_2\, \tilde{\beta}^+_{ii}}\, \widehat{U}^{(i)}. \tag{6.15}$$

We are now ready to start the induction process over the internal stages. For the first stage we have

$$U^{(1)} = \bar{U}^{(1)} + \triangle t\, \tilde{a}_{11}\, S(U^{(1)}), \tag{6.16}$$

with

$$\bar{U}^{(1)} = \frac{\alpha_1}{1 - r_2\, \tilde{\beta}^+_{11}}\, U^n = U^n$$

thanks to (6.13) for $i = 1$. Notice that, as $\tilde{a}_{11} \geq 0$, (6.16) represents an implicit Euler step in forward time for the ODE $u' = s(u)$. Our assumption on U^n together with assumption IES on the source term imply that $0 \leq U^{(1)} \leq e$ for all $0 \leq \triangle t\, \tilde{a}_{11} \leq \tau_{IE}$.

Suppose now that $0 \le U^{(j)} \le e$, $j = 1, \ldots, i-1$. Using (6.7)–(6.9), for $j = 1, \ldots, i-1$ we have

$$0 \le U^{(j)} + \frac{\triangle t}{r_1} D(U^{(j)}) \le e, \qquad \text{for } \frac{\triangle t}{r_1} \le \tau_{_D}, \tag{6.17}$$

$$0 \le U^{(j)} + \frac{\triangle t}{r_2} s(U^{(j)}) \le e, \qquad \text{for } \frac{\triangle t}{r_2} \le \tau_{_+}^s, \tag{6.18}$$

$$0 \le U^{(j)} - \frac{\triangle t}{r_3} s(U^{(j)}) \le e, \qquad \text{for } \frac{\triangle t}{r_3} \le \tau_{_-}^s. \tag{6.19}$$

If the triplet (r_1, r_2, r_3) is such that the positivity assumption SP is satisfied, then from (6.12) and (6.13), we get that

$$0 \le \widehat{U}^{(i)} \le (1 - r_2\, \tilde{\beta}_{ii}^{+})\, e\,,$$

under the step-size restriction $\triangle t \le \min\{r_1\, \tau_{_D}\, ,\, r_2\, \tau_{_+}^s\, , r_3\, \tau_{_-}^s\}$. Therefore, we deduce from (6.15) that $0 \le \bar{U}^{(i)} \le e$. Again, (6.14) is an implicit Euler step for the ODE $u' = s(u)$. Hence, our assumption IES on the source term and the fact that $\tilde{a}_{ii} \ge 0$ imply that $0 \le U^{(i)} \le e$ when $0 < \triangle t\, \tilde{a}_{ii} \le \tau_{_{IE}}$.

Once we have obtained that $0 \le U^{(i)} \le e$, $i = 1, \ldots, s$, we simply have to observe that U^{n+1}, the last component of (6.2), has the form

$$\begin{aligned} U^{n+1} =& \alpha_{s+1}\, U^n + r_1 \sum_{j=1}^{s} \beta_{s+1,j} \left(U^{(j)} + \frac{\triangle t}{r_1} D(U^{(j)}) \right) \\ &+ r_2 \sum_{j=1}^{s} \tilde{\beta}_{s+1,j}^{+} \left(U^{(j)} + \frac{\triangle t}{r_2} S(U^{(j)}) \right) + r_3 \sum_{j=1}^{s} \tilde{\beta}_{s+1,j}^{-} \left(U^{(j)} - \frac{\triangle t}{r_3} S(U^{(j)}) \right), \end{aligned}$$

that, together with (6.3) for $i = s+1$,

$$\alpha_{s+1} + r_1 \sum_{j=s}^{s} \beta_{s+1,j} + r_2 \sum_{j=1}^{s} \tilde{\beta}_{s+1,j}^{+} + r_3 \sum_{j=1}^{s} \tilde{\beta}_{s+1,j}^{-} = 1\,,$$

gives that $0 \le U^{n+1} \le e$ under step-size restriction (6.10). □

Remark 6.5. This way of proceeding also serves to establish sufficient conditions for numerical invariance of the interval $[0, 1]$ when higher order Runge-Kutta methods are used for the the time discretization of the ODE system (3.1). For this, we follow, e.g., [12] and rewrite a RK scheme with matrix coefficient $\mathbb{A}$ as a convex combination of Euler steps, as follows:

$$\mathcal{U} = (I + r\,\mathbb{A})^{-1} e \otimes U^n + r \left((I + r\,\mathbb{A})^{-1} \mathbb{A} \otimes I \right) \left(\mathcal{U} + \frac{\triangle t}{r} \mathcal{D}(\mathcal{U}) + \frac{\triangle t}{r} \mathcal{S}(\mathcal{U}) \right).$$

Provided that $r < \mathcal{R}(\mathbb{A})$, with $\mathcal{R}(\mathbb{A})$ the radius of absolute monotonicity of the RK scheme (see (4.4)), the ideas used in the proof of Theorem 6.4 lead to a direct proof of the numerical invariance of the interval $[0, 1]$ under the step-size restriction

$$\triangle t \le \frac{\tau_{_+}^s}{\tau_{_+}^s + \tau_{_D}}\, \tau_{_D}\, \mathcal{R}(\mathbb{A})\,, \tag{6.20}$$

which generalizes the result obtained for the explicit Euler method (5.8). We notice, however, that this result can also be obtained with the material in [14, 15]. Observe

that restriction (6.20) is also valid for implicit RK methods, provided that the non-linear systems associated to the implicit scheme can be solved. However, we do not consider here fully implicit schemes because of the practical difficulties in solving non-linear systems when non-trivial discretizations of the convective derivative term are used.

Remark 6.6. We recall that the class of source terms satisfying (2.9) satisfies the Assumption IES, with $\tau_{IE} \geq \min\{1/M, \tau_D\}$ and also (6.8)-(6.9) with $\tau_+^s \geq 1/m$, $\tau_-^s \geq 1/M$.

Remark 6.7. It should be noticed that Theorem 6.4, our main result, relies on the fact that $[0,1]$ is an invariant region for the explicit Euler method, applied to the ODE $u' = s(u)$, in *forward and backward time*. This way of proceeding is closely related to what is done by Shu and Osher in [29] in the context of RK-TVD time-discretization methods, when negative coefficients are required to write an explicit RK scheme $\mathbb{A}$ as linear combination of explicit Euler steps; in this case an auxiliary discrete divergence operator, which is assumed to be strongly stable for explicit Euler steps in backward time, is considered.

The step-size restriction $\triangle t \leq \gamma\, \tau_{IE}$ in (6.10) simply ensures the solvability of the non-linear equations involved in the D-IMEX process, and it is otherwise unrelated to the set of restrictions,

$$\triangle t \leq \min\{r_1\, \tau_D\, ,\, r_2\, \tau_+^s\, ,\, r_3\, \tau_-^s\}\, . \tag{6.21}$$

For practical purposes, the number of variable parameters in the determination of the step-size restriction (6.21) can be reduced by imposing that $r_1\tau_D = r_2\tau_+^s = r_3\tau_-^s$. Then, (6.21) becomes

$$\frac{\triangle t}{\tau_D} \leq r_1\, , \tag{6.22}$$

where r_1 needs to be determined for each given splitting. Under these premises, given a D-IMEX scheme $(\mathbb{A}, \tilde{\mathbb{A}})$ and a matrix splitting $\tilde{\mathbb{A}} = \tilde{\mathbb{A}}_+ - \tilde{\mathbb{A}}_-$, we can ensure the preservation of the numerical invariance of the interval $[0,1]$ for $(\mathbb{A}, \tilde{\mathbb{A}})$ if we can find $r_1 > 0$ so that the inequalities in (6.4) hold for the triplet $r_1(1, \tau_D/\tau_+^s, \tau_D/\tau_-^s)$.

Now the question of the choice on an *appropriate* splitting remains to be addressed. The next section is devoted to this issue.

7. Coefficient matrix splittings

If we bypass, for the time being, the time-step restriction for solvability, it is clear that any practical application of Theorem 6.4 requires a splitting, $\tilde{\mathbb{A}} = \tilde{\mathbb{A}}_+ - \tilde{\mathbb{A}}_-$, for which appropriate positive numbers r_1, r_2, r_3 can be found so that the inequalities in (6.4) are satisfied.

Notice that if we consider the trivial splitting $\tilde{\mathbb{A}}_+ = \tilde{\mathbb{A}}$, $\tilde{\mathbb{A}}_- = 0$, then for any $r_1 > 0$, $r_2 > 0$ (6.4) is equivalent to requiring that $(r_1, r_2) \in \mathcal{R}(\mathbb{A}, \tilde{\mathbb{A}})$, where $\mathcal{R}(\mathbb{A}, \tilde{\mathbb{A}})$ is the region of absolute monotonicity of the method (see [12] for details). Theorem 6.4 would then lead to non-trivial step-size restrictions only when $\mathcal{R}(\mathbb{A}, \tilde{\mathbb{A}})$ is non-trivial. As shown in [12], if $\mathcal{R}(\mathbb{A}, \tilde{\mathbb{A}})$ is non-trivial, the ARK scheme is SSP and Theorem 6.4 does not represent any improvement over the results obtained in [12].

If $\tilde{\mathbb{A}}$ contains negative elements below the diagonal (this would immediately imply that $\mathcal{R}(\mathbb{A},\tilde{\mathbb{A}})$ is trivial, see [12, Proposition 3.4]), we could consider the next simplest splitting, $\tilde{\mathbb{A}}_+ = (\max(\tilde{a}_{ij},0))$, $\tilde{\mathbb{A}}_- = (\min(\tilde{a}_{ij},0))$. However, there is no guarantee that this simple splitting will allow us to find $r_1 > 0$ such that $(r_1, r_2, r_3) = r_1(1, \tau_{_D}/\tau_+^s, \tau_{_D}/\tau_-^s)$ satisfies the assumption SP.

A splitting $\tilde{\mathbb{A}} = \tilde{\mathbb{A}}_+ - \tilde{\mathbb{A}}_-$ which might be appropriate for our purposes can be constructed by using a technique developed in [11], which allows us to obtain an optimal Shu-Osher representation of a RK scheme given by a coefficient matrix $\mathbb{A}$, from the point of view of the stability restrictions for Strong Stability (1.1), (4.3).

7.1. **Splittings based on Shu-Osher representations.** In [11, Propostion 2.7], it is seen that if $r = \mathcal{R}(\mathbb{A}) > 0$, the matrices $\Lambda = r\,\mathbb{A}\,(I + r\,\mathbb{A})^{-1}$ and $\Gamma = (I - \Lambda)\,\mathbb{A}$ provide the optimal Shu-Osher representation of a RK scheme with matrix coefficient $\mathbb{A}$. We refer the reader, e.g., to [11] for specific details on standard (Butcher) RK representations and Shu-Osher representations, and simply note here that these matrices satisfy the following positivity conditions: $\Lambda \geq 0$, $\Gamma \geq 0$, and $\Lambda - r\,\Gamma \geq 0$, $(I - \Lambda)\,e \geq 0$. We also point out that the link between the standard Butcher representation of a RK scheme, given by the matrix $\mathbb{A}$, and a Shu-Osher representation, given by matrices (Λ, Γ) relies on the fact that $\mathbb{A} = (I - \Lambda)^{-1}\,\Gamma$.

We show next how to use a Shu-Osher representation of $\mathbb{A}$, the explicit part of the IMEX scheme, in order to obtain a matrix splitting for $\tilde{\mathbb{A}}$. First we multiply (4.2) by $(I - \Lambda)$ to get

$$\mathcal{U} = \alpha \otimes U^n + (\Lambda \otimes I)\,\mathcal{U} + \triangle t\,(\Gamma \otimes I)\,\mathcal{D}(\mathcal{U}) + \triangle t\,\left((I - \Lambda)\,\tilde{\mathbb{A}} \otimes I\right)\mathcal{S}(\mathcal{U})\,, \tag{7.1}$$

with $\alpha = (I - \Lambda)\,e$. Defining $\tilde{\Gamma} := (I - \Lambda)\,\tilde{\mathbb{A}}$, this matrix has the same non-negative diagonal entries as $\tilde{\mathbb{A}}$. If there are negative off-diagonal entries, we split $\tilde{\Gamma}$ as $\tilde{\Gamma} = \tilde{\Gamma}_+ - \tilde{\Gamma}_-$, with $\tilde{\Gamma}_+, \tilde{\Gamma}_- \geq 0$. Accordingly, we split matrix $\tilde{\mathbb{A}}$ as $\tilde{\mathbb{A}} = \tilde{\mathbb{A}}_+ - \tilde{\mathbb{A}}_-$, where the matrices $\tilde{\mathbb{A}}_+ = (I - \text{Ł})^{-1}\tilde{\Gamma}_+$, $\tilde{\mathbb{A}}_- = (I - \text{Ł})^{-1}\tilde{\Gamma}_-$ satisfy the desired properties. In particular, since $(I - \Lambda)^{-1} \geq 0$ (see [11, Lemma 3.8]), we also have $\tilde{\mathbb{A}}_+ \geq 0$, $\tilde{\mathbb{A}}_- \geq 0$.

In principle, any Shu-Osher representation of $\mathbb{A}$ can be used. For obvious reasons, one is led to utilize the optimal one, and we shall do so in section 8. However, there is no guarantee that any of these choices will provide the largest $r_1 > 0$ such that $r_1(1, \tau_{_D}/\tau_+^s, \tau_{_D}/\tau_-^s)$ satisfies the assumption SP. This issue is considered in the next subsection.

7.2. **Optimal splittings.** Taking into account that τ_D is related to the spatial mesh size $\triangle x$ (through the usual CFL restriction for the operator $D(U)$ in the homogeneous case), the step-size restriction (6.22) for numerical invariance of the interval $[0, 1]$ takes the form of a CFL-like restriction. Since $r_1 > 0$ determines the condition on the mesh parameters to ensure numerical invariance of the interval $[0, 1]$, we can try to enlarge the step-size restriction (6.22) by optimizing the choice of the matrix splitting.

Several strategies can be attempted in order to design matrix splittings that lead to optimal values of r_1. In this section we show how to obtain optimal splittings by numerical search. The input data are the matrices $\mathbb{A}$ and $\tilde{\mathbb{A}}$ and given values of $\tau_{_D}$, τ_+^s and τ_-^s. In order to obtain splittings $\tilde{\mathbb{A}}_+$, $\tilde{\mathbb{A}}_-$ such that r_1 is as large as possible,

we set $y = \tau_D/\tau_+^s$, $z = \tau_D/\tau_-^s$ and solve the following optimization problem:

$$\max_{\tilde{\mathbb{A}}_+,\tilde{\mathbb{A}}_-} r_1$$
$$\text{subject to} \begin{cases} (I + r_1\,\mathbb{A} + r_1\,y\,\tilde{\mathbb{A}}_+ + r_1\,z\,\tilde{\mathbb{A}}_-)^{-1} e \geq 0\,, \\ (I + r_1\,\mathbb{A} + r_1\,y\,\tilde{\mathbb{A}}_+ + r_1\,z\,\tilde{\mathbb{A}}_-)^{-1} \mathbb{A} \geq 0\,, \\ (I + r_1\,\mathbb{A} + r_1\,y\,\tilde{\mathbb{A}}_+ + r_1\,z\,\tilde{\mathbb{A}}_-)^{-1} \tilde{\mathbb{A}}_\pm \geq 0\,, \\ \tilde{\mathbb{A}}_+, \tilde{\mathbb{A}}_- \geq 0\,,\ \tilde{\mathbb{A}} = \tilde{\mathbb{A}}_+ - \tilde{\mathbb{A}}_-\,. \end{cases}$$

In order to avoid inverses, we proceed as in [18], and reformulate the above problem as

$$\max_{K,\tilde{K}_\pm,\tilde{\mathbb{A}}_+,\tilde{\mathbb{A}}_-,\delta} r_1$$
$$\text{subject to} \begin{cases} (I + r_1\,\mathbb{A} + r_1\,y\,\tilde{\mathbb{A}}_+ + r_1\,z\,\tilde{\mathbb{A}}_-)\,\delta = e\,, \\ (I + r_1\,\mathbb{A} + r_1\,y\,\tilde{\mathbb{A}}_+ + r_1\,z\,\tilde{\mathbb{A}}_-)\,K = \mathbb{A}\,, \\ (I + r_1\,\mathbb{A} + r_1\,y\,\tilde{\mathbb{A}}_+ + r_1\,z\,\tilde{\mathbb{A}}_-)\,\tilde{K}_\pm = \tilde{\mathbb{A}}_\pm\,, \\ K, \tilde{K}_\pm, \tilde{\mathbb{A}}_+, \tilde{\mathbb{A}}_-, \delta \geq 0\,,\ \tilde{\mathbb{A}} = \tilde{\mathbb{A}}_+ - \tilde{\mathbb{A}}_-\,. \end{cases}$$

We can also include restrictions to obtain matrices $\tilde{\mathbb{A}}_+$ and $\tilde{\mathbb{A}}_-$ with the correct shape, namely, triangular and lower triangular, respectively. This problem can be solved, e.g., with the Matlab Optimization Toolbox.

8. Examples of Application: LV and Y model problem

For illustration purposes, in this section we shall consider two D-IMEX schemes considered by Pareschi and Russo in [26], namely, the SSP2(3,3,2), specified by the *Butcher tableaus*

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} & 0 \\ \hline \mathbb{A} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{array} \qquad \begin{array}{c|ccc} \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 1 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \hline \tilde{\mathbb{A}} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{array} \tag{8.1}$$

and the SSP2(3,2,2), with coefficient matrices

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \hline \mathbb{A} & 0 & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \begin{array}{c|ccc} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & -\frac{1}{2} & \frac{1}{2} & 0 \\ 1 & 0 & \frac{1}{2} & \frac{1}{2} \\ \hline \tilde{\mathbb{A}} & 0 & \frac{1}{2} & \frac{1}{2} \end{array} \tag{8.2}$$

In the notation SSP$k(s,\sigma,p)$, s is the number of stages of the implicit scheme, σ the number of stages in the explicit scheme, k is the order of the explicit scheme and p is the order of the IMEX scheme. As noted in [26], the explicit part of both methods is SSP.

Our aim here is to show how to obtain the different splittings considered in the previous section, and the associated time-step restrictions, obtained from applying

Theorem 6.4, that guarantee the preservation of $[0,1]$ as an invariant region for these IMEX schemes.

In order to fix the parameters to be used, we shall consider the model problem (2.5) with the source term considered in [24], i.e.,

$$u_t + f(u)_x = -\mu\, u\,(u-1)\,(u-0.5), \tag{8.3}$$

$$u(x,0) = \begin{cases} u_L & x < x_d, \\ u_R & x > x_d. \end{cases} \tag{8.4}$$

We shall assume that $f(u) = u$, as in [24], or $f(u) = u^2/2$ as in [23, 9].

For the source term in (8.3), a straight application of Lemma 2.2 ensures that (2.9) hold with $m = \mu/2$ and $M = \mu/4$, thus we have

$$\tau_{IE} \geq \frac{4}{\mu}, \quad \tau_+^s \geq \frac{2}{\mu}, \quad \tau_-^s \geq \frac{4}{\mu}\,.$$

In addition, we can consider $\tau_D = \triangle x / \max_u |f'(u)|$, which ensures (6.7) for any TVD discretization of the convective derivative. Hence, for Riemann IVPs so that $u_L, u_R \in \{0,1\}$, as in [24] and [23, section 17.15], $\tau_D = \triangle x$.

We recall that given a D-IMEX scheme $(\mathbb{A}, \tilde{\mathbb{A}})$ and a matrix splitting $\tilde{\mathbb{A}} = \tilde{\mathbb{A}}_+ - \tilde{\mathbb{A}}_-$, we can ensure the preservation of the numerical invariance of the interval $[0,1]$ provided that we can find $r_1 > 0$ so that the inequalities in (6.4) hold for the triplet $r_1(1, \tau_D/\tau_+^s, \tau_D/\tau_-^s)$.

Larger values of r_1, hence larger step-size restrictions, will result from considering the largest possible values for the parameters $\tau_D, \tau_+^s, \tau_-^s, \tau_{IE}$. For this particular source term, a larger value for τ_-^s can actually be used. The following result is proven in [25].

Lemma 8.1. *Let $s(u) = -\mu u(u-1)(u-0.5)$ with $\mu > 0$. If $0 \leq u \leq 1$, then we have that*

$$0 \leq u + \tau\, s(u) \leq 1\,, \qquad \text{for} \quad 0 \leq \tau \leq \frac{2}{\mu}\,, \tag{8.5}$$

$$0 \leq u - \tau\, s(u) \leq 1\,, \qquad \text{for} \quad 0 \leq \tau \leq \frac{16}{\mu}\,. \tag{8.6}$$

We can, hence, consider $\tau_D = \triangle x$, $\tau_+^s = 2/\mu$, $\tau_-^s = 16/\mu$. Thus, given a splitting $\tilde{\mathbb{A}} = \tilde{\mathbb{A}}_+ - \tilde{\mathbb{A}}_-$, if we can find $r_1 = r_1(\mu\triangle x) > 0$ so that the triplet $r_1(1\,, \mu\triangle x/2\,, \mu\triangle x/16)$ satisfies inequalities (6.4), Theorem 6.4 guarantees numerical invariance of the interval $[0,1]$ under the step-size restriction

$$\frac{\triangle t}{\triangle x} \leq \min\left\{ r_1, \frac{4\,\gamma}{\mu\triangle x}, \gamma \right\}, \tag{8.7}$$

where $\gamma = 3$ for the SSP2(3,3,2) scheme and $\gamma = 2$ for the SSP2(3,2,2) scheme.

The step-size restriction above can be compared to (6.20), the corresponding restriction for an explicit non-additive RK scheme with coefficient matrix $\mathbb{A}$, which becomes

$$\frac{\triangle t}{\triangle x} \leq \frac{2}{2 + \mu\triangle x}\, \mathcal{R}(\mathbb{A}). \tag{8.8}$$

SSP2(3,3,2) scheme. For this scheme, both the explicit and the implicit methods are SSP with $\mathcal{R}(\mathbb{A}) = 2$ and $\mathcal{R}(\tilde{\mathbb{A}}) = 12/5$; however, the region of absolute monotonicity of the additive RK scheme, $\mathcal{R}(\mathbb{A}, \tilde{\mathbb{A}})$, is trivial; [12, Example 1] and hence the splitting $\tilde{\mathbb{A}}_+ = \tilde{\mathbb{A}}$, $\tilde{\mathbb{A}}_- = 0$ does not allow us to find $r_1 > 0$ with the desired properties.

An appropriate splitting can be constructed, based on the optimal Shu-Osher representation, as specified in section 7.1. For this, we take $r = \mathcal{R}(\mathbb{A})$ and construct the matrices $\Lambda = r\mathbb{A}(I + r\mathbb{A})^{-1}$, $\Gamma = (I - \Lambda)\mathbb{A}$, $\alpha = (I - \Lambda)\mathbb{A}$,

$$\Lambda = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{2}{3} & 0 \end{pmatrix}, \ \Gamma = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 \end{pmatrix}, \ \alpha = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \frac{1}{3} \end{pmatrix}. \tag{8.9}$$

$$\tilde{\Gamma} = (I - \Lambda)\,\tilde{\mathbb{A}} = \begin{pmatrix} \frac{1}{4} & 0 & 0 & 0 \\ -\frac{1}{4} & \frac{1}{4} & 0 & 0 \\ \frac{1}{3} & \frac{1}{12} & \frac{1}{3} & 0 \\ \frac{1}{9} & \frac{1}{9} & \frac{1}{9} & 0 \end{pmatrix}. \tag{8.10}$$

Since the element (2,1) is negative, we consider the sign splitting $\tilde{\Gamma} = \tilde{\Gamma}_+ - \tilde{\Gamma}_-$. Then, the matrices $\tilde{\mathbb{A}}_+ = (I - \Lambda)^{-1}\tilde{\Gamma}_+$, $\tilde{\mathbb{A}}_- = (I - \Lambda)^{-1}\tilde{\Gamma}_+$, have the desired structure,

$$\tilde{\mathbb{A}}_+ = \begin{pmatrix} \frac{1}{4} & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ \frac{7}{12} & \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{3} & 0 \end{pmatrix}, \quad \tilde{\mathbb{A}}_- = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \frac{1}{4} & 0 & 0 & 0 \\ \frac{1}{4} & 0 & 0 & 0 \\ \frac{1}{6} & 0 & 0 & 0 \end{pmatrix}. \tag{8.11}$$

The next step is to check if it is possible to find $r_1 > 0$ such that inequalities (6.4) are satisfied for the triplet $r_1(1, \mu\triangle x/2, \mu\triangle x/16)$; furthermore, for each $\mu\triangle x$ we are interested in the largest value of r_1 fulfilling (6.4), denoted on the following by $r_1(\mu\triangle x)$. To achieve this aim, we have used a formal computer language, such as Mathematica. The function $r_1(\mu\triangle x)$ for $\mu\triangle x \in [0, 2]$ corresponding to this splitting is shown in Figure 1-left, with a dashed line. For reference, we mark the point $(1, 1)$ with the symbol $*$. Some values of this function, that will be used in the numerical experiments, are shown in Table 1.

SSP2(3,2,2) scheme. For this IMEX method, the explicit scheme is SSP with $\mathcal{R}(\mathbb{A}) = 1$. The matrix $\tilde{\mathbb{A}}$ contains a negative element, hence the implicit method cannot be SSP.

We notice that the simple sign splitting in $\tilde{\mathbb{A}}$ cannot be considered, since it is not possible to find $r_1 > 0$ such that $(I + \mathbb{B})^{-1}\tilde{\mathbb{A}}_- \geq 0$ holds for the triplet $r_1(1, \mu\triangle x/2, \mu\triangle x/16)$.

As before, we can easily compute a splitting based on optimal Shu-Osher representation. In this case,

$$\tilde{\Gamma} = \begin{pmatrix} \frac{1}{2} & 0 & 0 & 0 \\ -\frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{4} & \frac{1}{4} & 0 \end{pmatrix}. \tag{8.12}$$

With the sign splitting of $\tilde{\Gamma}$ we obtain

$$\tilde{\mathbb{A}}_{+} = \begin{pmatrix} \frac{1}{2} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}, \qquad \tilde{\mathbb{A}}_{-} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{4} & 0 & 0 & 0 \end{pmatrix}.$$

However, this splitting is not valid either: it is not possible to find $r_1 > 0$ such that $(I + \mathbb{B})^{-1}\tilde{\mathbb{A}}_{-} \geq 0$ holds for the triplet $r_1(1, \mu\triangle x/2, \mu\triangle x/16)$.

A different splitting of the matrix $\tilde{\Gamma}$ in (8.12) does produce, however, the desired result. Considering

$$\tilde{\Gamma}_{+} = \begin{pmatrix} \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{4} & \frac{1}{4} & 0 \end{pmatrix}, \qquad \tilde{\Gamma}_{-} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

we obtain

$$\tilde{\mathbb{A}}_{+} = \begin{pmatrix} \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}, \qquad \tilde{\mathbb{A}}_{-} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \end{pmatrix}. \tag{8.13}$$

In this case it is possible to obtain $r_1 > 0$ such that inequalities (6.4) hold. In Figure 1-right, this function is displayed, with dashed line, for $\mu\triangle x \in [0, 2]$, obtained using Mathematica, as before. The symbol $*$ at point $(1, 0.5)$ serves as reference. In Table 2 we show some values of this function for splitting (8.13).

The splittings above are based on the Shu-Osher representation for $\mathbb{A}$, but clearly their optimality, from the point of view of the step-size restrictions, cannot be claimed. In the next section we address this issue by a numerical search for these two D-IMEX schemes applied to the model problem.

8.1. **Optimal splittings by numerical search.** In this section we consider splittings such that $r_1(\mu\triangle x)$ is as large as possible. For each $\mu\triangle x$ we solve the optimization problem proposed in section 7.2 with $y = \mu\triangle x/2$ and $z = \mu\triangle x/16$; we use the Optimization Toolbox in Matlab.

In Figure 1, with a solid line, we show the results for the SSP2(3,3,2) and SSP2(3,2,2) schemes. It is worth noting that Figure 1 shows that when $\mu = 0$ (no source term), the optimal CFL is $\mathcal{R}(\mathbb{A}) = 2$, as it should be the case, according to (8.8). Notice that, for the values of $\mu\triangle x$ examined, the value of r_1 provides the

effective CFL restriction for the preservation of $[0,1]$ as an invariant region for the IMEX scheme, since $\min\{r_1, 4\gamma/(\mu\triangle x), \gamma\} = r_1$ in both cases.

In Tables 1 and 2 we show some of the optimal values obtained, that allow for a direct comparison. Observe that for the SSP2(3,3,2) method, the values obtained from the 'optimal search' process are quite similar to the ones obtained from the Shu-Osher splittings in section 8, but for the SSP2(3,2,2) scheme, significantly better values can be obtained by numerical search. We also observe that for SSP2(3,3,2) the values $r_1(\mu\triangle x)$ are larger than the ones obtained for scheme SSP2(3,2,2).

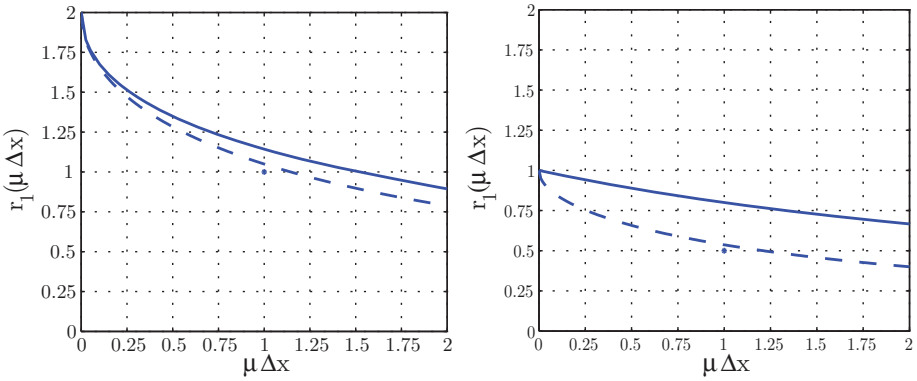

FIGURE 1. Function $r_1(\mu\triangle x)$ for SSP2(3,3,2) (left) and SSP2(3,2,2) (right) methods; the dashed line corresponds to splittings (8.11) and (8.13), and the solid line corresponds to numerical search.

TABLE 1. Largest values $r_1(\mu\triangle x)$ for SSP2(3,3,2) scheme.

Splitting	$\mu\triangle x = 1$	$\mu\triangle x = 2.5$	$\mu\triangle x = 10$
(8.11)	1.0497	0.7030	0.2743
Optimal	1.1429	0.8017	0.3188

TABLE 2. Largest values $r_1(\mu\triangle x)$ for SSP2(3,2,2) scheme.

Splitting	$\mu\triangle x = 1$	$\mu\triangle x = 2.5$	$\mu\triangle x = 10$
(8.13)	0.5369	0.3558	0.2003
Optimal	0.8000	0.6154	0.2857

Theorem 6.4 states only sufficient conditions for the preservation of the convex property $\mathcal{P}$. In what follows we perform a series of numerical experiments whose objective is to test the sharpness of the CFL restrictions just obtained.

8.2. **Numerical experiments for the LV&Y model problem.** We apply next the SSP(3,3,2) and SSP(3,2,2) IMEX schemes to the model problem (8.3)-(8.4), with different choices for the discrete divergence operator $D(U)$. Here, for the purpose of comparison, we shall consider the following:

(1) A first order upwind discretization, which leads to a monotone scheme in the homogeneous case.
(2) A standard ENO2 discretization (see [29]), which gives rise to a TVD scheme in the homogeneous case, under the usual CFL restriction.
(3) A standard ENO3 discretization (see [29]), which gives rise to an *Essentially Non-oscillatory* scheme in the homogeneous case. In this case, (6.7) cannot be theoretically ensured.

These are examples of non-linear reconstruction techniques, for which the use of a semi-implicit alternative, such as a D-IMEX scheme, is almost mandatory (it is mandatory for non-linear flux functions or non-linear operators in $D(U)$).

To test the sharpness of the CFL restrictions found in the previous section, we have performed a large number of numerical tests, covering typical situations of interest for the model problem (8.3). For $\mu = 10^{-3}$, we have compiled 'approximately' the observed CFL numbers for which the property $\mathcal{P}$ fails to be preserved, according to our numerical tests. Table 3 collects this information for the SSP2(3,3,2) scheme. The optimal step-size restriction obtained in the previous section is also displayed in the table for comparison purposes.

Each slot in the table displays the interval where the loss of preservation has been numerically observed: for the bottom value in the interval, the numerical solution lies in $[0, 1]$, while for the upper value this property has been lost. Some sample computations, corresponding to some of the entries on the table, are shown in Figure 2, for $f(u) = u$, and in Figure 3, for $f(u) = u^2/2$. The plots have been obtained with Matlab, so that when the numerical solution lies in the interval $[0, 1]$, the plot window is automatically set to $[0, 1]$; however, if some value is out of this interval, the plot window is automatically increased, which allows us to easily detect the existence of values out of the interval $[0, 1]$. In the bottom row of Figure 2 we can clearly observe the numerical delay, due to insufficient resolution for the value of $\mu\triangle x = 10$ considered.

The analogous results corresponding to the SSP2(3,2,2) scheme are compiled in Table 4.

The results in these tables confirm the theoretical behavior of the function $r_1(\mu\triangle x)$ displayed in Figure 1: the CFL restriction for numerical invariance of the interval $[0, 1]$ diminishes as $\mu\triangle x$ increases. In addition, the theoretical CFL restrictions obtained from the optimal splittings found by numerical search are not that far from the actual restrictions found in the test cases. It is also worth mentioning that even though we cannot ensure (6.7) for the ENO3 discrete divergence operator, we obtain non-oscillatory numerical profiles under the theoretically obtained CFL restriction in all cases. This situation is similar to what is numerically observed with high order spatial discretizations for the numerical divergence in SSP RK schemes for homogeneous conservation laws.

By looking both at the theoretical and numerical results obtained in this paper, it seems clear that the SSP2(3,3,2) scheme is more robust than the SSP2(3,2,2) scheme, which might explain why this is in fact the preferred scheme for numerical computations in [26].

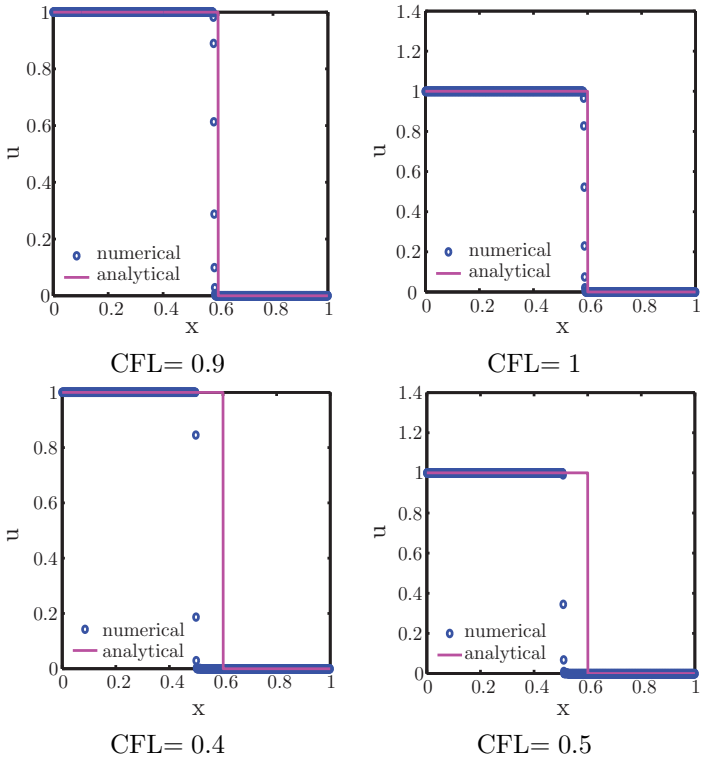

FIGURE 2. Numerical solution for the upwind numerical flux and the SSP2(3,3,2) scheme, $f(u) = u$ and $\triangle x = 10^{-3}$. Top: $\mu\triangle x = 2.5$, Bottom: $\mu\triangle x = 10$.

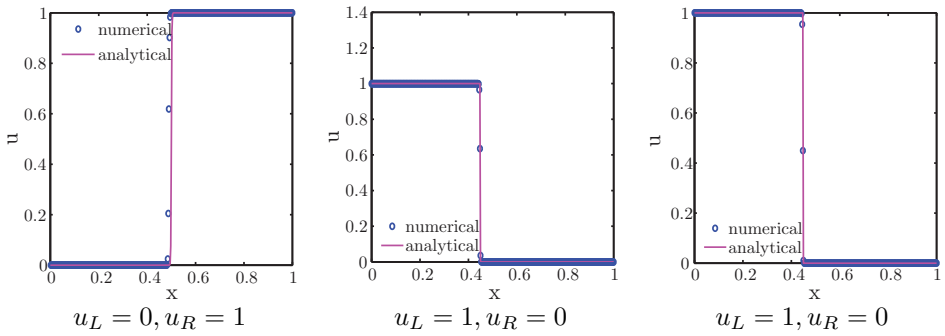

FIGURE 3. Numerical solution for the ENO2 numerical flux and the SSP2(3,3,2) scheme, $f(u) = u^2/2$. $\mu\triangle x = 2.5$ and $\triangle x = 10^{-3}$.

TABLE 3. SSP2(3,3,2) scheme for the model problem. Optimal and observed CFL for numerical preservation of $[0, 1]$.

	$\mu\triangle x = 1$	$\mu\triangle x = 2.5$	$\mu\triangle x = 10$
Optimal	1.1429	0.8017	0.3188
$f(u) = u, \quad u_L = 1, \quad u_R = 0$			
Obser. Upwind	$1.3 \le r_1 \le 1.4$	$0.9 \le r_1 \le 1.0$	$0.4 \le r_1 \le 0.5$
Obser. ENO2	$1.3 \le r_1 \le 1.4$	$0.9 \le r_1 \le 1.0$	$0.5 \le r_1 \le 0.6$
Obser. ENO3	$1.3 \le r_1 \le 1.4$	$0.9 \le r_1 \le 1.0$	$0.6 \le r_1 \le 0.7$
$f(u) = u^2/2, \quad u_L = 1, \quad u_R = 0$			
Obser. Upwind	$1.7 \le r_1 \le 1.8$	$1.4 \le r_1 \le 1.5$	$0.5 \le r_1 \le 0.6$
Obser. ENO2	$1.6 \le r_1 \le 1.7$	$1.1 \le r_1 \le 1.2$	$0.5 \le r_1 \le 0.6$
Obser. ENO3	$1.2 \le r_1 \le 1.3$	$0.9 \le r_1 \le 1.0$	$0.3 \le r_1 \le 0.4$
$f(u) = u^2/2, \quad u_L = 0, \quad u_R = 1$			
Obser. Upwind	$2.0 \le r_1 \le 2.1$	$1.9 \le r_1 \le 2.0$	$1.0 \le r_1 \le 1.1$
Obser. ENO2	$2.0 \le r_1 \le 2.1$	$1.9 \le r_1 \le 2.0$	$0.9 \le r_1 \le 1.0$
Obser. ENO3	$1.9 \le r_1 \le 2.0$	$1.9 \le r_1 \le 2.0$	$0.8 \le r_1 \le 0.9$

TABLE 4. SSP2(3,2,2) scheme for the model problem. Optimal and observed CFL for numerical preservation of $[0, 1]$.

	$\mu\triangle x = 1$	$\mu\triangle x = 2.5$	$\mu\triangle x = 10$
Optimal	0.8000	0.6154	0.2857
$f(u) = u, \quad u_L = 1, \quad u_R = 0$			
Obser. Upwind	$0.9 \le r_1 \le 1.0$	$0.8 \le r_1 \le 0.9$	$0.4 \le r_1 \le 0.5$
Obser. ENO2	$0.8 \le r_1 \le 0.9$	$0.7 \le r_1 \le 0.8$	$0.5 \le r_1 \le 0.6$
Obser. ENO3	$0.8 \le r_1 \le 0.9$	$0.7 \le r_1 \le 0.8$	$0.5 \le r_1 \le 0.6$
$f(u) = u^2/2, \quad u_L = 1, \quad u_R = 0$			
Obser. Upwind	$0.8 \le r_1 \le 0.9$	$0.7 \le r_1 \le 0.8$	$0.4 \le r_1 \le 0.5$
Obser. ENO2	$0.8 \le r_1 \le 0.9$	$0.7 \le r_1 \le 0.8$	$0.2 \le r_1 \le 0.3$
Obser. ENO3	$0.8 \le r_1 \le 0.9$	$0.7 \le r_1 \le 0.8$	$0.2 \le r_1 \le 0.3$
$f(u) = u^2/2, \quad u_L = 0, \quad u_R = 1$			
Obser. Upwind	$1.4 \le r_1 \le 1.5$	$1.3 \le r_1 \le 1.4$	$0.6 \le r_1 \le 0.7$
Obser. ENO2	$1.3 \le r_1 \le 1.4$	$1.3 \le r_1 \le 1.4$	$0.8 \le r_1 \le 0.9$
Obser. ENO3	$1.1 \le r_1 \le 1.2$	$1.3 \le r_1 \le 1.4$	$0.8 \le r_1 \le 0.9$

9. Conclusions and perspectives

When solving hyperbolic PDEs, it is nowadays common practice to discretize first the spatial variables to obtain a semi-discrete method of lines system of ODEs in the time variable. For hyperbolic PDEs with stiff source terms, stiff ODEs are obtained. In this context, IMEX-RK schemes provide a rather general framework for the time discretization of these ODEs due to their ability to treat the convective part in an explicit fashion, while still maintaining an implicit handling of the source terms. These schemes are good building blocks in general purpose numerical codes, in particular, when they are embedded in a larger framework, for example, as part of an AMR (for Adaptive Mesh Refinement) code.

In the context of numerical resolution of hyperbolic PDEs with stiff source terms, different issues can be addressed. In this paper, we have focused on the ability to maintain non-oscillatory reaction fronts, and we have obtained a set of sufficient conditions that ensure the preservation of certain relevant invariant regions for a class of source terms that include the model problem proposed by LeVeque and Yee in [24]. Our study extends previous ones in two aspects. The first one is the admissible source terms: we drop the non-increasing assumption on the source term and we relax it to condition (2.9); the second one is the class of time stepping methods considered: we consider not only first order stepping methods but higher order D-IMEX Runge-Kutta methods.

The results obtained have been used to compute CFL-like restrictions for the numerical preservation of the invariance of the interval $[0, 1]$ in two particular cases of D-IMEX schemes, previously considered by Pareschi and Russo in [26]. The results of various numerical tests on a model balance law indicate that the theoretical CFL-like conditions obtained are of practical use. We, hence, consider that our analysis provides new tools to test the robustness of a given scheme, from the point of view of certain qualitative behavior. Since experience says that it is safer to integrate complex problems with well-behaved schemes for test problems, the results obtained in this paper gain relevance in this context.

We have centered our analysis on the scalar balance law, however, we expect that the results obtained, and the techniques used in this paper, can be applied to a wider range of problems, as well as in the preservation of other 'weak' properties. Roughly speaking, the success in preserving a given property will depend on the ability to ensure that constants likewise in (6.7)-(6.9) exist. In particular, we are currently working on extending these techniques to 1D systems.

Finally, we remark that there are other numerical issues to be considered when dealing with the numerical solution of balance laws with stiff source terms. We mention here two of them that are part also of our current research. First, the issue of the spurious steady-state solutions studied by Lafon and Yee in [21, 22], which might be related to the well-balancing of the underlying scheme. Second, the numerical delay observed for very stiff source terms, which needs sufficient spatial refinement to be correctly modeled. In this respect, the combination of these techniques with an AMR code could be of great help in obtaining robust numerical simulations of complex problems.

Acknowledgment

The authors would like to thank D. Ketcheson for his help in the optimization scripts.

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** (1997), no. 2-3, 151–167. MR1485812 (98i:65054)
2. F. Bouchut, *Nonlinear stability of finite volume methods for hyperbolic conservation laws and well-balanced schemes for sources*, Frontiers in Mathematics, Birkhäuser Verlag, Basel, 2004. MR2128209 (2005m:65002)
3. A. Chalabi, *On convergence of numerical schemes for hyperbolic conservation laws with stiff source terms*, Math. Comp. **66** (1997), no. 218, 527–546. MR1397441 (97g:65178)
4. M. Crouzeix, *Une méthode multipas implicite-explicite pour l'approximation des équations d'évolution paraboliques*, Numer. Math. **35** (1980), no. 3, 257–276. MR0592157 (82b:65084)
5. K. Dekker and J. G. Verwer, *Stability of Runge-Kutta methods for stiff nonlinear differential equations*, CWI Monographs, North-Holland (1984). MR774402 (86g:65003)
6. L Ferracina and M. N. Spijker, *Stepsize restrictions for the Total-Variation-Diminishing property in general Runge–Kutta methods*, SIAM J. Numer. Anal. **42** (2004), no. 3, 1073–1093. MR2113676 (2005k:65126)
7. S. Gottlieb, D.I. Ketcheson, and C.W. Shu, *High order strong stability preserving time discretizations*, Journal of Scientific Computing **38** (2009), no. 3, 251–289. MR2475652 (2010b:65161)
8. S. Gottlieb, C.W. Shu, and E. Tadmor, *Strong stability preserving high-order time discretization methods*, SIAM Rev. **43** (2001), no. 1, 89–112. MR1854647 (2002f:65132)
9. D.F. Griffiths, A.M. Stuart, and H.C. Yee, *Numerical wave propagation in an advection equation with a nonlinear source term*, SIAM Journal on Numerical Analysis **29** (1992), no. 5, 1244–1260. MR1182730 (93h:65111)
10. I. Higueras, *On strong stability preserving time discretization methods*, J. Sci. Comput. **21** (2004), no. 2, 193–223. MR2069949 (2005d:65112)
11. ———, *Representations of Runge–Kutta methods and strong stability preserving methods*, SIAM J. Numer. Anal. **43** (2005), no. 3, 924–948. MR2177549 (2006j:65184)
12. ———, *Strong stability for additive Runge-Kutta methods*, SIAM J. Numer. Anal. **44** (2006), no. 4, 1735–1758. MR2257125 (2008c:65164)
13. M. W. Hirsch and H. Smith, *Monotone Dynamical Systems*, Handbook of differential equations: ordinary differential equations. Vol. II, Elsevier B. V., Amsterdam, 2005, pp. 239–357. MR2182759 (2006j:37017)
14. Z. Horváth, *Positivity of Runge-Kutta and diagonally split Runge-Kutta methods*, Applied numerical mathematics **28** (1998), no. 2-4, 309–326. MR1655167 (99i:65073)
15. ———, *On the positivity step size threshold of Runge–Kutta methods*, Applied Numerical Mathematics **53** (2005), no. 2-4, 341–356. MR2128530 (2005m:65137)
16. W. Hundsdorfer and J. G. Verwer, *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*, Springer, 2003. MR2002152 (2004g:65001)
17. G. E. Karniadakis, M. Israeli, and S. A. Orszag, *High-order splitting methods for the incompressible Navier-Stokes equations*, J. Comput. Phys. **97** (1991), 414–443. MR1137607 (92h:76066)
18. D. Ketcheson, C.B. Macdonald, and S. Gottlieb, *Optimal implicit strong stability preserving Runge-Kutta methods*, Appl. Numer. Math. **59** (2009), no. 2, 373–392. MR2484928 (2010a:65113)
19. J. F. B. M. Kraaijevanger, *Contractivity of Runge-Kutta methods*, BIT **31** (1991), no. 3, 482–528. MR1127488 (92i:65120)
20. S. N. Kruzkov, *First order quasi-linear equations in several independent variables*, Math. USSR-Sb. **10** (1970), 217–243.
21. A. Lafon and H. C. Yee, *Dynamical approach study of spurious steady-state numerical solutions of nonlinear differential equations part iii. the effects of nonlinear source terms in reaction-convection equations*, Comp. Fluid. Dyn. **6** (1996), 1–36.
22. ———, *Dynamical approach study of spurious steady-state numerical solutions of nonlinear differential equations part iv. stability vs methods of discretizing nonlinear source terms in reaction-convection equations*, Comp. Fluid. Dyn. **6** (1996), 89–123.
23. R. J. LeVeque, *Finite Volume Methods for Hyperbolic Problems*, Cambridge University Press, 2002. MR1925043 (2003h:65001)

24. R. J. LeVeque and H. C. Yee, *A study of numerical methods for hyperbolic conservation laws with stiff source terms*, J. Comput. Phys. **86** (1990), no. 1, 187–210. MR1033905 (90k:76009)
25. A. Martínez Gavara, *High Resolution Schemes for Hyperbolic Conservation Laws with Source Terms*, Ph.D. thesis, Universitat de València, http://www.tesisenxarxa.net/TDX-1111109-094349/, 2008.
26. L. Pareschi and G. Russo, *Implicit-Explicit Runge-Kutta schemes and applications to hyperbolic systems with relaxation*, J. Sci. Comput. **25** (2005), no. 1-2, 129–155. MR2231946 (2007b:65063)
27. S. Ruuth, *Implicit-explicit methods for resaction-diffusion problems in pattern formation*, J. Math. Biol. **34** (1995), no. 2, 148–176. MR1366356 (96j:92008)
28. H.J. Schroll and R. Winther, *Finite-difference schemes for scalar conservation laws with source terms*, IMA J. Numer. Anal. **16** (1996), no. 2, 201. MR1382716 (97g:65177)
29. C.W. Shu, *Total-Variation-Diminishing time discretizations*, SIAM J. Sci. Statist. Comput. **9** (1988), no. 6, 1073–1084. MR0963855 (90a:65196)
30. E. F. Toro, *Riemann Solvers and Numerical Methods for Fluid Dynamics*, Springer, Berlin, Germany, 1997. MR1474503 (98h:76099)
31. J. M. Varah, *Stability restrictions on second order, three level finite difference schemes for parabolic equations*, SIAM J. Numer. Anal. **17(2)** (1980), no. 2, 300–309. MR0567275 (81g:65121)
32. J. G. Verwer, J. G. Blom, and W. Hundsdorfer, *An implicit-explicit approach for atmospheric transport-chemistry problems*, Appl. Numer. Math. **20** (1996), no. 1-2, 191–209, Workshop on the method of lines for time-dependent problems (Lexington, KY, 1995). MR1385244

Departament de Matemàtica Aplicada, Universitat de València, 46100 Burjassot, Spain
E-mail address: donat@uv.es

Departamento de Ingeniería Matemática e Informática, Universidad Pública de Navarra, 31006 Pamplona, Spain
E-mail address: higueras@unavarra.es

Departamento de Matemática Aplicada I, Universidad de Sevilla, 41012 Sevilla, Spain
E-mail address: gavara@us.es