SIAM J. NUMER. ANAL.
Vol. 51, No. 1, pp. 163–190

FLUX-EXPLICIT IMEX RUNGE–KUTTA SCHEMES FOR HYPERBOLIC TO PARABOLIC RELAXATION PROBLEMS*

SEBASTIANO BOSCARINO† AND GIOVANNI RUSSO†

Abstract. We consider the development of Implicit-Explicit (IMEX) Runge–Kutta (R-K) schemes for hyperbolic systems of conservation laws with stiff diffusive relaxation. The asymptotic behavior of such systems as the relaxation parameter vanishes is governed by a reduced parabolic system, and it is desirable to have schemes that are able to capture the correct diffusive limit. The hyperbolic part becomes stiff when the system relaxes towards the parabolic equation. For this reason, in previous works part of the hyperbolic terms are treated implicitly [S. Jin, L. Pareschi, and G. Toscani, *SIAM J. Numer. Anal.*, 35 (1998), pp. 2405–2439], [S. Jin and L. Pareschi, *J. Comp. Phys.*, 161 (2000), pp. 312–330], [G. Naldi and L. Pareschi, *SIAM J. Numer. Anal.*, 37 (2000), pp. 1246–1270]. In particular, in [S. Boscarino, L. Pareschi, and G. Russo, *Implicit-explicit Runge-Kutta schemes for hyperbolic systems and kinetic equations in the diffusion limit*, submitted] the scheme relaxes to an implicit method for the limit diffusive problem, thus avoiding the $\Delta t \propto \Delta x^2$ restriction. It would be very desirable to have IMEX schemes which treat the hyperbolic part explicitly, because this allows the use of well-tested space discretization with no modification of the original system. However, the development of such methods presents the difficulty that the characteristic speeds diverge in the diffusive limit making the hyperbolic part very stiff. In this paper we show how to overcome these difficulties with the introduction of particular conditions on the coefficients of the IMEX R-K schemes such that we can treat the stiff component on the hyperbolic part explicitly without any manipulation of the original system. Moreover, the schemes proposed in this paper guarantee a CFL condition independent of the diffusive parameter where a CFL hyperbolic condition in the stiff regime is chosen. Such schemes are shown to have the correct diffusion limit and several numerical results confirm the theoretical analysis.

Key words. IMEX Runge–Kutta methods, hyperbolic conservation laws with sources, diffusion equations, stiff systems

AMS subject classifications. 65C20, 65M06, 76D05, 82C40

DOI. 10.1137/110850803

1. Introduction.

The development of numerical methods to solve hyperbolic systems in diffusive regimes has been a very active area of research in the last fifteen years (see, for example, [12, 8, 9, 11]). A strictly related field of research concerns the construction of schemes for the compressible Navier–Stokes limit (see [14] and the references therein). In such physical problems, the scaling parameter (mean free path) may vary by several orders of magnitude from the rarefied regime to the diffusive regime, and it is desirable to develop a class of robust numerical schemes that can work uniformly with respect to this parameter.

To understand such a hyperbolic system with diffusive relaxation, we consider a simple prototype of hyperbolic system with relaxation term given by

$$\begin{aligned} u_\tau + V_\xi &= 0, \\ V_\tau + p(u)_\xi &= -\frac{1}{\varepsilon}(V - Q(u)), \end{aligned}$$

where $u = u(\xi,\tau), V = V(\xi,\tau) \in \mathbb{R}$, $\varepsilon > 0$ is called the relaxation time.

*Received by the editors October 10, 2011; accepted for publication (in revised form) August 10, 2012; published electronically January 22, 2013.
http://www.siam.org/journals/sinum/51-1/85080.html

†Department of Mathematics and Computer Science, University of Catania, 95125 Catania, Italy (boscarino@dmi.unict.it, russo@dmi.unict.it).

Under the rescaling (diffusive scaling), we have

$$\tau = t/\varepsilon, \quad V = \varepsilon v,$$
$$x = \xi, \quad Q(u) = \varepsilon q(u),$$

we obtain a general diffusive relaxation system given by

$$\begin{aligned} u_t + v_x &= 0, \\ v_t + \frac{1}{\varepsilon^2} p(u)_x &= -\frac{1}{\varepsilon^2}(v - q(u)), \end{aligned} \tag{1.1}$$

where $p'(u) > 0$. This system is hyperbolic with two distinct real characteristic speeds $\pm\sqrt{p'(u)}/\varepsilon$. In the small relaxation limit, $\varepsilon \to 0$, the behavior of the solution to (1.1) is, at least formally, governed by

$$\begin{aligned} u_t + q(u)_x &= p(u)_{xx}, \\ v &= q(u) - p(u)_x. \end{aligned} \tag{1.2}$$

The so-colled subcharactheristic condition [22, 21] for system (1.2) becomes

$$|q'(u)|^2 < p'(u)/\varepsilon^2$$

and it is generally satisfied in the limit case, $\varepsilon \to 0$.

In this paper attention is devoted to the construction of methods for the numerical solution of system (1.1) that are able to capture the asymptotic behavior as $\varepsilon \to 0$.

Solving (1.1) numerically is challenging due to the stiffness of the problem both in the convection and in the relaxation terms. In general, Implicit-Explicit (IMEX) Runge–Kutta (R-K) schemes [5, 1, 2, 10, 15] represent a powerful tool for the time discretization of stiff systems. Unfortunately, since the characteristic speed of the hyperbolic part is of order $1/\varepsilon$, standard IMEX R-K schemes developed for hyperbolic systems with stiff relaxation [15, 3] fail in such parabolic scaling, because the CFL condition would require $\Delta t = \mathcal{O}(\varepsilon \Delta x)$. Of course, in the diffusive regime where $\varepsilon < \Delta x$, this is very restrictive since for an explicit method a parabolic condition $\Delta t = \mathcal{O}(\Delta x^2)$ should suffice.

Most previous works [12, 18, 11, 23, 24, 25] on asymptotic preserving schemes for hyperbolic system with diffusive relaxation are based on the separation of the hyperbolic part into a stiff and nonstiff component and combine the stiff one into the relaxation term which is treated implicitly. Moreover, in the limit of infinite stiffness, such schemes become consistent explicit schemes for the diffusive limit equation [12, 8, 9, 20, 11, 24, 25], therefore suffering from the usual stability restriction $\Delta t = \mathcal{O}(\Delta x^2)$. Schemes that avoid such time step restriction and provide fully implicit solvers in the case of transport equations have been analyzed in [19], where a new formulation of the problem (1.1) was introduced, which was based on the addition of two opposite diffusive terms, in the limit of large stiffness. The first term, added to part of the hyperbolic component, makes it nonstiff, therefore allowing an explicit treatment, while the other term is treated implicitly. The remaining component of the hyperbolic term is formally treated implicitly, thus avoiding stability restrictions. The resulting scheme is consistent and becomes an implicit scheme for underlying diffusion limit, therefore avoiding the typical parabolic restriction of previous methods.

The drawback of this approach is that the method requires an implicit treatment of some hyperbolic component. Even if such an implicit term can be explicitly computed in most cases, this procedure requires a reformulation of the discretization of the whole system. It would be desirable to construct IMEX R-K schemes in which the whole hyperbolic part is treated explicitly. At first sight this seems a formidable task, because of the divergence of the characteristic speeds.

Here we show that if suitable additional conditions on the IMEX R-K coefficients are satisfied, then one can indeed construct schemes which are fully explicit in the hyperbolic term, and which converge to an implicit high order scheme in the diffusion limit.

In the construction of the schemes we consider only the case $\varepsilon \approx 1$ (where we impose the classical order conditions) and $\varepsilon \to 0$. The issue of accuracy degradation for intermediate values of ε, which is quite well understood in the context of hyperbolic relaxation (see, for example, [3]), still needs further investigation and will be considered in future work.

The plan of the paper is the following. In section 2 we introduce a simple hyperbolic system with diffusion limit and study the performance of several first order IMEX R-K methods. In section 3 we provide a description and classification of the IMEX R-K schemes presented in the literature. Next we produce a convergence analysis of such IMEX R-K schemes that provides additional order conditions in order to achieve the correct diffusive limit. Section 4 is devoted to the construction of new IMEX R-K schemes that satisfy these additional order conditions. In section 5 we describe space discretization obtained by conservative finite difference schemes and, finally, numerical results are presented. We conclude the paper with some remarks and an appendix containing some proofs of the conditions and schemes presented in sections 3 and 4.

2. The simple IMEX Euler schemes.

We start by considering first order schemes for the simplest hyperbolic system with diffusive relaxation [16, 11, 20, 12, 19]:

$$\begin{aligned} u_t + v_x &= 0, \\ \varepsilon^2 v_t + u_x &= -v, \end{aligned} \tag{2.1}$$

which is a particular case of (1.1), with $p(u) = u$ and $q(u) = 0$. When $\varepsilon \to 0$, system (2.1) relaxes towards the heat equation

$$u_t = u_{xx}, \tag{2.2}$$

with $v = -u_x$.

Let us start to consider the following system:

$$y' = f(t, y) + g(t, y). \tag{2.3}$$

Two versions of first order IMEX Euler schemes are possible to apply to system (2.3).

The first version is described by the Butcher tableaux

$$\text{Explicit} : \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & 1 & 0 \end{array} \qquad \text{Implicit} : \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 0 & 1 \\ \hline & 0 & 1 \end{array}, \tag{2.4}$$

and we will call a scheme in this form IMEX-Euler(1) (see [5]).

The second one is described by the Butcher tableaux

$$\text{Explicit} : \begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array} \qquad \text{Implicit} : \begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}, \tag{2.5}$$

and we will call a scheme in this form IMEX-Euler(2) (see [15]).

Application of IMEX-Euler(1) to (2.3) leads to

$$y_{n+1} = y_n + \Delta t f(t_n, y_n) + \Delta t g(t_{n+1}, y_{n+1}),$$

where f is treated explicitly and g implicitly, and this means that we first compute the explicit term $y^\star = y_n + \Delta t f(t_n, y_n)$ and then solve the implicit equation for y_{n+1}.

Instead, application of IMEX-Euler(2) consists of computing the intermediate stage as $Y_1 = y_n + \Delta t g(t_{n+1}, Y_1)$, and then of adding the explicit term $y_{n+1} = Y_1 + \Delta t f(t_n, y_n)$.

Let us now consider system (2.1) in the following form:

$$\begin{aligned} u_t &= -v_x, \\ v_t &= -\frac{u_x}{\varepsilon^2} - \frac{v}{\varepsilon^2}; \end{aligned} \tag{2.6}$$

such a scheme is of the form (2.3), with

$$y = \begin{pmatrix} u \\ v \end{pmatrix}, \quad f(y) = -\begin{pmatrix} v_x \\ u_x/\varepsilon^2 \end{pmatrix}, \quad g(y) = \begin{pmatrix} 0 \\ -v/\varepsilon^2 \end{pmatrix}.$$

Both terms, f and g, are *stiff* for small values of the parameter ε. As $\varepsilon \to 0$, the first equation in (2.6) becomes (2.2), with $v = -u_x$. Now we apply both schemes to system (2.6), with simple second order central differencing for the discretization of the space derivative. We integrate the equation up to the time $T = 1$, with $N = 80$ grid points, initial conditions $u(x, 0) = \cos(x)$, $v(x, 0) = \sin(x)$, $x \in [-\pi, \pi]$, and $\Delta t = 0.5\Delta x^2$. In Figures 2.1 and 2.2 we report the results of the computation for $\varepsilon = 1, 10^{-1}, 10^{-2}, 10^{-4}$ for IMEX-Euler(2) and IMEX-Euler(1) schemes, respectively, compared with the exact solution.

Why does IMEX-Euler(1) work, and why does IMEX-Euler(2) not? In order to answer such question some analysis is required.

Let us look for the evolution of a Fourier mode of the form $u = \hat{u}(t)\exp(i\xi x)$, $v = \hat{v}(t)\exp(i\xi x)$, where we dropped the dependence of $\hat{u}$ and $\hat{v}$ from ξ. Inserting the *ansatz* into systems (2.6), and using the variable $\hat{w} = -i\hat{v}/\xi$ the system becomes

$$\begin{aligned} \hat{u}_t &= \xi^2 \hat{w}, \\ \epsilon^2 \hat{w}_t &= -\hat{u} - \hat{w}. \end{aligned} \tag{2.7}$$

Now applying IMEX-Euler(2) (2.5) to system (2.7), we obtain, after simple algebra, the following expression for the second component:

$$\zeta \hat{w}_{n+1} = \hat{w}_n \frac{\zeta^2}{1+\zeta} - \hat{u}_n,$$

where $\zeta = \varepsilon^2/\Delta t$. Then for small values of ε ($\varepsilon \to 0$, i.e., $\zeta \to 0$), the numerical solution $\hat{w}_{n+1}$ diverges as it is shown in Figure 2.1.

Similarly, applying IMEX-Euler(1) (2.4) to system (2.7), we obtain for the second component

$$\hat{w}_{n+1} = \frac{\zeta}{\zeta+1}\hat{w}_n - \frac{\hat{u}_n}{\zeta+1},$$

which becomes $\hat{w}_{n+1} = -\hat{u}_n$ in the limit $\varepsilon \to 0$.

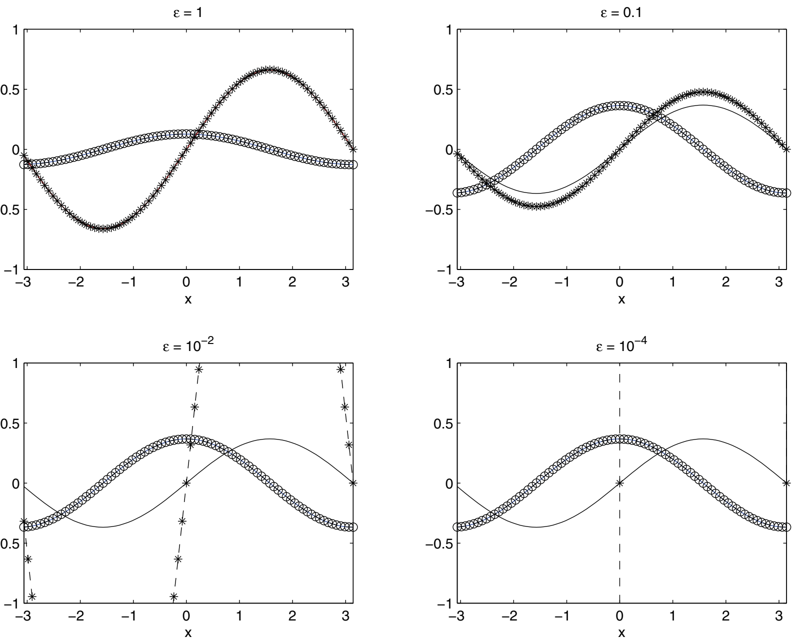

FIG. 2.1. *Numerical solution at time $T = 1$ for IMEX-Euler*(2), $\varepsilon = 1,\ 0.1,\ 10^{-2},\ 10^{-4}$ *with* $\Delta t = 0.5\Delta x^2$. *Solid lines are the exact solutions,* $\circ$ *numerical solutions for* u *component, and* $* - -$ *numerical ones for* v *component. The scheme is not asymptotic preserving for the variable* v.

This confirms that the scheme IMEX-Euler(1) captures well the behavior of the exact solution, as is shown in Figure 2.2. Since for the exact solution when $\varepsilon \to 0$ one has $w = -u$, we get consistency to the limit equation up to order one, i.e., $\hat{w}_{n+1} = -\hat{u}_{n+1} + \mathcal{O}(\theta)$ with $\theta = \xi^2 \Delta t$. Stability analysis on IMEX Euler(1) applied to system (2.7) gives condition $\Delta t \xi^2 \leqslant 1$ independent of ε, while similar analysis applied to the system (2.1) with derivatives computed by central differing gives $\Delta t/\Delta x^2 \leqslant 1$ (see Appendix A.1).

Motivated by this analysis, in the following section, we will introduce the main ingredients that guarantee good behavior of an IMEX R-K scheme in the limit case, i.e., $\varepsilon \to 0$.

The following questions naturally arise:

- What is the condition satisfied by scheme (2.4) but not by scheme (2.5) which guarantees consistency of an IMEX R-K scheme in the limit case $\varepsilon \to 0$?
- How can we use this condition to construct efficient higher order IMEX R-K schemes that capture the same limit?
- As $\varepsilon \to 0$, IMEX-Euler(1) becomes an explicit Euler for the diffusion equation with the classical stability restriction $\Delta t \propto \Delta x^2$. Is it possible to construct IMEX R-K schemes that, in the limit $\varepsilon \to 0$, converge to an implicit scheme for the limit diffusion equation, i.e., to schemes that are unconditionally stable in the limit?

In the following section we shall give answers to the above questions.

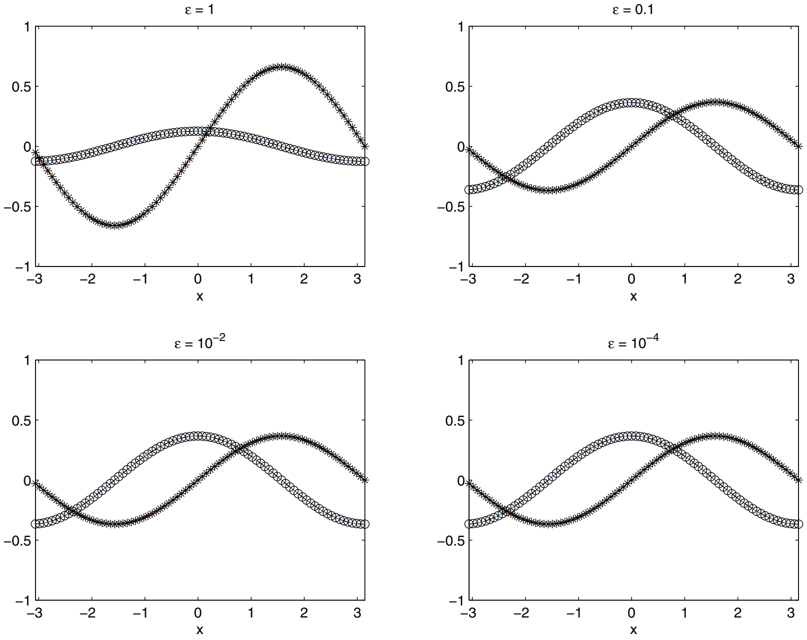

FIG. 2.2. *Numerical solution at time $T = 1$ for IMEX-Euler(1), for $\varepsilon = 1,\ 0.1,\ 10^{-2},\ 10^{-4}$ with $\Delta t = 0.5\Delta x^2$. Solid lines are the exact solutions, $\circ$ numerical solutions for u component, and $* - -$ numerical ones for v component. The scheme converges to the exact solution for all ε*

3. Analysis of IMEX R-K methods.

IMEX R-K schemes [5, 10, 15] have been widely used in the literature to treat problems that contain both stiff and nonstiff terms. Usually stiff terms are treated implicitly, while the nonstiff terms are treated explicitly, thus lowering the computational cost of the method. Standard application of IMEX R-K schemes to problem (2.6) would treat the first equation explicitly and the second implicitly [19]. Here we insist in treating the convection term of the second equation explicitly. We start from system (2.7) with $(\xi^2\hat{w}, -\hat{u}/\varepsilon^2)^T$ treated explicitly, while $(0, -\hat{w}/\varepsilon^2)^T$ is treated implicitly.

Then applying an IMEX R-K scheme to system (2.7) we obtain

$$\begin{aligned} \hat{u}_{n+1} &= \hat{u}_n + \Delta t \xi^2 \sum_{k=1}^{s} \tilde{b}_k \hat{W}_k, \\ \varepsilon^2 \hat{w}_{n+1} &= \varepsilon^2 \hat{w}_n - \Delta t \sum_{k=1}^{s} \tilde{b}_k \hat{U}_k - \Delta t \sum_{k=1}^{s} b_k \hat{W}_k \end{aligned} \tag{3.1}$$

for the numerical solution and

$$\begin{aligned} \hat{U}_k &= \hat{u}_n + \Delta t \xi^2 \sum_{j=1}^{k-1} \tilde{a}_{kj} \hat{W}_j, \\ \varepsilon^2 \hat{W}_k &= \varepsilon^2 \hat{w}_n - \Delta t \sum_{j=1}^{k-1} \tilde{a}_{kj} \hat{U}_j - \Delta t \sum_{j=1}^{k} a_{kj} \hat{W}_j \end{aligned} \tag{3.2}$$

for the internal stages.

Here the $s \times s$ matrices $\tilde{A} = (\tilde{a}_{ij})$, $A = (a_{ij})$ and the vectors $\tilde{b}$, $b \in \mathbb{R}^s$ characterize the scheme, and can be represented by a double *tableau* in the usual Butcher notation,

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

The coefficients $\tilde{c}$ and c are used if the right-hand side depends explicitly on time. We assume that they satisfy the usual relation: $\tilde{c}_i = \sum_{j=1}^{i-1} \tilde{a}_{ij}$, $c_i = \sum_{j=1}^{i} a_{ij}$. Matrix $\tilde{A}$ is lower triangular with zero diagonal, while matrix A is lower triangular, i.e., the implicit scheme is a diagonally implicit Runge–Kutta (DIRK); in this way we are sure that the term to which we apply the explicit scheme is actually evaluated explicitly [15, 3].

IMEX R-K schemes present in the literature can be classified in three different types characterized by the structure of the matrix $A = (a_{ij})_{i,j=1}^{s}$ of the implicit scheme. Following [1], we make use of the following definitions.

DEFINITION 3.1. *We call an IMEX R-K method of type A (see* [15]*) if the matrix $A \in \mathbb{R}^{s \times s}$ is invertible.*

DEFINITION 3.2. *We call an IMEX R-K method of type CK (see* [10]*) if the matrix $A \in \mathbb{R}^{s \times s}$ can be written as*

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

with the submatrix $\hat{A} \in \mathbb{R}^{(s-1) \times (s-1)}$ invertible.

IMEX R-K schemes, called of type ARS (see [5]), are a special case of the type CK with the vector $a = 0$.

3.1. Analysis of IMEX R-K schemes of type A. Hereafter we shall restrict our analysis to the type A IMEX R-K schemes for which the matrix A is invertible. This property greatly simplifies the analysis as compared to the case of IMEX-CK type (or, in particular, IMEX-ARS type), and this is the main motivation of our choice. Furthermore, type A methods are more robust against initial layer [15].

Now recalling also the conclusions of the analysis performed in section 2, we start this section by introducing a property which is important to guarantee asymptotic preservation.

We generalize the definition of *stiffly accurate* Runge–Kutta methods. We recall that an implicit scheme is called *stiffly accurate* if $b^T = e_s^T A$ with $e_s^T = (0, 0, \ldots, 1)$. This property is important, in particular, for the L-stability of the scheme [6], i.e., a method is called L-stable if it is A-stable and if, in addition, $\lim_{z\to\infty} R(z) = R(\infty) = 0$, where $R(z)$ is *the stability function* of the method, defined by $R(z) = 1 + zb^T(I - zA)^{-1}e$ (see [6, sect. IV.3]), and $b^T = (b_1, \ldots, b_s)$ and $e = (1, \ldots, 1)^T$.

Therefore, if an implicit Runge–Kutta method with matrix A invertible is stiffly accurate, then $R(\infty) = 0$. In fact, from the expression of $R(z)$, we have $R(\infty) = 1 - b^T A^{-1} e$. Now if the method is stiffly accurate, it follows $R(\infty) = 1 - b^T A^{-1} e = 1 - e_s^T e = 1 - 1 = 0$. Consequently, this makes A-stable methods L-stable.

We extend now the definition to IMEX schemes (see also [19]).

DEFINITION 3.3. *We say that an IMEX R-K scheme of type A (i.e., with nonsingular matrix A of the implicit part) is globally stiffly accurate if $b^T = e_s^T A$ and $\tilde{b}^T = e_s^T \tilde{A}$, with $e_s = (0, \ldots, 0, 1)^T$, and $c_s = \tilde{c}_s = 1$, i.e., the numerical solution is identical to the last internal stage value of the scheme.*

It is known in the literature that a special class of s-stage explicit Runge–Kutta schemes for which $\tilde{b}^T = e^T \tilde{A}$ is called *First Same As Last* (FSAL). Such schemes have the advantage that they require only $s-1$ function evaluations per time step, because the last stage of step n coincides with the first step of the step $n+1$ (see [7] for details). Then using this definition, we may say that an IMEX R-K scheme is *globally stiffly accurate* if the implicit scheme is stiffly accurate and the explicit scheme is FSAL.

Note that IMEX R-K schemes with the property $\tilde{b} = \tilde{A}^T e_s$, $b = A^T e_s$ were already introduced in [5].

3.2. Algebraic order conditions. In this section we consider schemes of type A and study their behavior as $\varepsilon \to 0$. In particular we show that a sufficient condition for the consistency in the limit case ($\varepsilon \to 0$) is that the method is globally stiffly accurate. We rewrite system (3.1) and (3.2) in vectorial form

$$\begin{aligned} \hat{u}_{n+1} &= \hat{u}_n + \theta \tilde{b}^T \hat{W}, \\ \zeta \hat{w}_{n+1} &= \zeta \hat{w}_n - \tilde{b}^T \hat{U} - b^T \hat{W}, \end{aligned} \tag{3.3}$$

where $\theta = \xi^2 \Delta t$, $\zeta = \varepsilon^2/\Delta t$, and

$$\begin{aligned} \hat{U} &= \hat{u}_n e + \theta \tilde{A} \hat{W}, \\ \zeta \hat{W} &= \zeta \hat{w}_n e - \tilde{A} \hat{U} - A \hat{W}, \end{aligned} \tag{3.4}$$

with $\hat{W} = (\hat{W}_1, \ldots, \hat{W}_s)^T$, $\hat{U} = (\hat{U}_1, \ldots, \hat{U}_s)^T$.

Solving for the stage values $\hat{W}$ we have $\hat{W} = (\zeta I + A)^{-1}(\zeta \hat{w}_n e - \tilde{A}\hat{U})$, which gives $\hat{W} = \zeta A^{-1} \hat{w}_n e - (I - \zeta A^{-1}) A^{-1} \tilde{A} \hat{U} + \mathcal{O}(\zeta^2)$.

Now substituting $\hat{W}$ in the numerical solution we have

$$\zeta \hat{w}_{n+1} = \zeta (1 - b^T A^{-1} e) \hat{w}_n + (b^T A^{-1} \tilde{A} - \tilde{b}^T) \hat{U} - \zeta b^T A^{-2} \tilde{A} \hat{U} + \mathcal{O}(\zeta^2).$$

Consistency as $\zeta \to 0$ implies

$$b^T A^{-1} \tilde{A} - \tilde{b}^T = 0. \tag{3.5}$$

Furthermore, if

$$1 - b^T A^{-1} e = 0, \tag{3.6}$$

one has

$$\hat{w}_{n+1} = -b^T A^{-2} \tilde{A} \hat{U}, \tag{3.7}$$

where the internal stages are given by

$$\hat{W} = -A^{-1} \tilde{A} \hat{U}. \tag{3.8}$$

Note that if the implicit scheme is stiffly accurate, then the condition (3.6) is satisfied. Then condition (3.5) is equivalent to $e_s^T \tilde{A} = \tilde{b}^T$, which means that the scheme is globally stiffly accurate.

Then a sufficient condition to guarantee that both (3.5) and (3.6) are satisfied is that the IMEX R-K of type A is globally stiffly accurate. Since for the exact solution one has $\hat{w} = -\hat{u}$ in the limit case $\varepsilon \to 0$, consistency up to order p requires

$$b^T A^{-2} \tilde{A} \hat{U} = -\hat{u}_{n+1} + \mathcal{O}(\theta^{p+1}).$$

From this requirement we derive the new additional order conditions on the IMEX R-K schemes of type A.

As usual, order conditions are obtained by matching the Taylor expansion of the exact solution and the numerical one, up to terms of the prescribed order. In the relaxed case, i.e., $\varepsilon \to 0$, system (2.7) reduces to

$$\hat{u}_t = -\xi^2 \hat{u}, \quad \hat{w} + \hat{u} = 0. \tag{3.9}$$

Equation (3.9) belongs to the class of differential algebraic systems for which a general theory has been studied in [1].

Now, inserting (3.8) into the first equation in (3.4) one can express $\hat{U}$ in terms of $\hat{u}_n$,

$$\hat{U} = \mathcal{B}^{-1} e \hat{u}_n, \tag{3.10}$$

where $\mathcal{B} = I - \theta\mathbb{A}, \mathbb{A} = \tilde{A}\mathcal{C}$, and $\mathcal{C} = -A^{-1}\tilde{A}$. This leads to the numerical solutions

$$\begin{aligned} \hat{u}_{n+1} &= (1 + \theta \tilde{b}^T \mathcal{C}\mathcal{B}^{-1} e)\hat{u}_n, \\ \hat{w}_{n+1} &= -b^T A^{-2} \tilde{A} \mathcal{B}^{-1} e \hat{u}_n. \end{aligned} \tag{3.11}$$

As usual, we write the order conditions starting from $n = 0$, and assuming that the initial conditions $\hat{u}_0$, $\hat{w}_0$ are consistent, i.e., $\hat{w}_0 + \hat{u}_0 = 0$.

The exact solution of the first equation in (3.9), after one time step, has the expression $\hat{u}(\theta) = \exp(-\theta)\hat{u}_0$, where $\hat{u}(t_0) = \hat{u}_0$ and by (3.9) we obtain $\hat{w}(\theta) = -\hat{u}(\theta)$. By Taylor's expansion we obtain $\hat{u}(\theta) = \hat{u}_0(1-\theta(1-\theta/2+\theta^2/6+\mathcal{O}(\theta^3)))$. In particular, it follows from (3.10) that $\mathcal{B}^{-1} = (I-\theta\mathbb{A})^{-1} = I+\theta\mathbb{A}+\theta^2\mathbb{A}^2+\mathcal{O}(\theta^3)$, and inserting this expansion into (3.11) we can compare the expressions for the numerical solution $\hat{u}_1$ and the exact one and equate equal powers of θ. This provides the following additional order conditions up to second order where, from now on, the vectorial relation $\tilde{c} = \tilde{A}e$ is used.

Consistency condition for w component:

$$b^T A^{-2} \tilde{c} = 1. \tag{3.12}$$

Order conditions for u-component up to second order:

$$-\tilde{b}^T \mathcal{C} e = 1, \qquad \tilde{b}^T \mathcal{C} \mathbb{A} e = \frac{1}{2}. \tag{3.13}$$

Order conditions for w-component up to second order:

$$b^T A^{-2} \tilde{A} \mathbb{A} e = -1, \quad b^T A^{-2} \tilde{A} \mathbb{A}^2 e = \frac{1}{2}. \tag{3.14}$$

We note that we have to add these additional order conditions to the classical order ones in order to achieve the expected order in the limit case, i.e., $\varepsilon \to 0$.

3.3. Removing parabolic stiffness. The procedure outlined above provides a scheme that converges to an explicit scheme for the parabolic equation in the limit case $\varepsilon \to 0$. Such schemes suffer from the standard CFL restriction $\Delta t = \mathcal{O}(\Delta x^2)$.

In order to remove such a restriction we adopt a technique similar to the one illustrated in [19], consisting in adding and subtracting the term $\mu(\varepsilon)u_{xx}$ to the first equation of system (2.1):

$$\begin{aligned} u_t &= -(v + \mu(\varepsilon)u_x)_x + \mu(\varepsilon)u_{xx}, \\ \varepsilon^2 v_t &= -u_x - v. \end{aligned} \tag{3.15}$$

Here $\mu(\varepsilon)$ is such that $\mu : \mathbb{R}^+ \to [0,\ 1]$ and $\mu(0) = 1$. When ε is not small there is no reason to add and subtract the term $\mu(\varepsilon)u_{xx}$, therefore $\mu(\varepsilon)$ will be small in such a regime, i.e., $\mu(\varepsilon) \approx 0$. The precise choice of $\mu(\varepsilon)$ will be specified later.

With this approach, the idea is that as $\varepsilon \to 0$, the quantity $v+\mu u_x \to 0$. Therefore, such a term can be treated explicitly, while the term $\mu(\varepsilon)u_{xx}$ will be treated implicitly in the first equation, i.e., we will treat the term $-((v + \mu u_x)_x, u_x/\varepsilon^2)^T$ explicitly, and the term $(\mu u_{xx}, -v/\varepsilon^2)^T$ implicitly, respectively.

The procedure can be justified by observing that in this way the scheme becomes an implicit scheme for the limit diffusion equation, while without adding and subtracting the term $\mu(\varepsilon)u_{xx}$, the scheme relaxes to an explicit one. The same approach has been used in [19], and similar techniques have been adopted in other contexts by several authors. See, for example, [26] where adding and subtracting elliptic terms is used to stabilize level set methods, or [27] for applications to kinetic problems.

Then performing a similar analysis as in the previous section for the system (3.15), we obtain additional order conditions that are exactly equal to the ones obtained for the w component, i.e., (3.12) and (3.14), while for the u component, the additional order conditions are different from (3.13). Below we list these new additional order conditions for the u component up to second order, where here we put $\mathbb{A} = \tilde{A}\mathcal{C} - A$ and $\mathcal{C} = I - A^{-1}\tilde{A}$.

Order conditions for u component up to second order:

$$(b^T - \tilde{b}^T\mathcal{C})e = 1, \quad (\tilde{b}^T\mathcal{C} - b^T)\mathbb{A}e = \frac{1}{2}. \tag{3.16}$$

3.4. Generalizations.

3.4.1. Nonzero flux, $q(u) \neq 0$. In this case the system relaxes to a convection-diffusion equation. We observe that, by adding and subtracting the term $\mu(\varepsilon)p(u)_{xx}$, (1.1) can be written as

$$\begin{aligned} u_t + (v + \mu(\varepsilon)p(u)_x)_x &= \mu(\varepsilon)p(u)_{xx}, \\ v_t + \frac{1}{\varepsilon^2}p(u)_x &= -\frac{1}{\varepsilon^2}(v - q(u)). \end{aligned} \tag{3.17}$$

The terms $(v + \mu(\varepsilon)p(u)_x)_x$ and $p(u)_x/\varepsilon^2$ on the left-hand side are treated explicitly, while the terms on the right-hand side, $(\mu(\varepsilon)p(u)_{xx},\ -(v - q(u))/\varepsilon^2$, are treated implicitly.

As $\varepsilon \to 0$, the scheme becomes an IMEX R-K scheme for the limit convection-diffusion equation (1.2), in which the convection term is treated explicitly and the diffusion one is treated implicitly.

By repeating the same analysis of sections 3.2 and 3.3 for the complete system (3.17) we obtain again the same algebraic order conditions (3.12), (3.14), and (3.16) and we have to include other additional algebraic order conditions. This analysis is performed in details in Appendix A.2, where nonlinear $p(u)$ and $q(u)$ are considered.

Here we explicitly give first and second additional order conditions

$$\begin{aligned} &\textit{order}\ \ 1\ \ \textit{w component}: && b^TA^{-2}\tilde{A}\tilde{c} = 1, \\ &\textit{order}\ \ 2\ \ \textit{u component}: && \tilde{b}^T\tilde{A}A^{-1}\tilde{c} = 1/2, \quad (b^T - \tilde{b}^T\mathcal{C})\tilde{c} = 1/2. \end{aligned} \tag{3.18}$$

We remark that the additional order conditions (3.12)–(3.14) and (3.18) are obtained in the limit case $\varepsilon = 0$, while classical order conditions are obtained for $\varepsilon = 1$. Convergence for intermediate values of ε (i.e., $0 < \varepsilon \ll 1$) are verified numerically without theoretical justification. An analysis of IMEX R-K schemes for hyperbolic relaxation aimed to study the behavior for intermediate regimes has been performed [3] leading to the development of uniformly accurate schemes. A similar analysis for the diffusive relaxation is presently under investigation.

We believe that for the general system (1.1), the construction of a second order IMEX R-K scheme that satisfies all the previous order conditions could be obtained. This question will be investigated in detail in a future work, since it is not trivial to construct a fourth stage second order scheme that satisfies all these conditions.

3.4.2. Stability considerations. Here we show that if the limit scheme for $\varepsilon \to 0$ is A-stable, then it is also L-stable. In order to show this we apply an IMEX R-K scheme of type A globally stiffly accurate to the evolution equations obtained inserting the Fourier solution in system (3.15), then we obtain for the numerical solution in vectorial form

$$\begin{aligned} \hat{u}_{n+1} &= \hat{u}_n + \theta \tilde{b}^T(\hat{W} + \hat{U}) - \theta b^T \hat{U}, \\ \zeta \hat{w}_{n+1} &= \zeta \hat{w}_n - \tilde{b}^T \hat{U} - b^T \hat{W}, \end{aligned} \tag{3.19}$$

and for the internal stages

$$\begin{aligned} \hat{U} &= \hat{u}_n e + \theta \tilde{A}(\hat{W} + \hat{U}) - \theta A \hat{U}, \\ \zeta \hat{W} &= \zeta \hat{w}_n e - \tilde{A} \hat{U} - A \hat{W}. \end{aligned} \tag{3.20}$$

We can rewrite (3.19), by using (3.20) as $\hat{\mathcal{U}}_{n+1} = \mathcal{R}(\theta, \zeta; \tilde{A}, A)\hat{\mathcal{U}}_n$, where $\hat{\mathcal{U}}_n = (\hat{u}_n, \hat{w}_n)^T$. In the limit case, i.e., $\varepsilon \to 0$ (or $\zeta \to 0$), the matrix $\mathcal{R}(\theta, \zeta; \tilde{A}, A)$ is reduced to

$$\hat{\mathcal{R}}(\theta; \tilde{A}, A) = \begin{pmatrix} R(\theta; \tilde{A}, A) & 0 \\ \tilde{r}_{21} & 0 \end{pmatrix}$$

with $R(\theta; \tilde{A}, A) = 1 + \theta \mathfrak{b}^T (I - \theta \mathbb{A})^{-1} e$, $\tilde{r}_{21} = -b^T A^{-2} \tilde{A} (I - \theta \mathbb{A})^{-1} e$, and $\mathfrak{b}^T = (\tilde{b}^T \mathcal{C} - b^T)$. The first component $\hat{u}_n$ satisfies $\hat{u}_{n+1} = R(\theta; \tilde{A}, A)\hat{u}_n$. Clearly, $\hat{u}_{n+1}$ remains bounded iff $|R(\theta; \tilde{A}, A)| \leq 1$.

Furthermore, we observe that $\lim_{\theta \to \infty} R(\theta; \tilde{A}, A) = 1 + \lim_{\theta \to \infty} \mathfrak{b}^T (\theta^{-1} I - \mathbb{A})^{-1} e = 1 - \mathfrak{b}^T \mathbb{A}^{-1} e$ and by the properties of globally stiffly accurate, we then have $\lim_{\theta \to \infty} |R(\theta; \tilde{A}, A)| = 0$. Indeed, $\mathfrak{b}^T \mathbb{A}^{-1} e = (\tilde{b}^T \mathcal{C} - b^T)(\tilde{A}\mathcal{C} - A)^{-1} e = e_s^T (\tilde{A}\mathcal{C} - A)(\tilde{A}\mathcal{C} - A)^{-1} e = e_s^T e = 1$. Then if the limit scheme is A-stable, this guarantees L-stability. The scheme proposed in this paper and presented in Appendix A.5 is indeed a globally stiffly accurate IMEX R-K scheme, and is A-stable as $\varepsilon \to 0$, therefore it is also L-stable. Stability domain of this scheme is shown in Figure A.1.

4. Construction of the scheme of type A.

In this section we construct a second order IMEX R-K scheme globally stiffly accurate, according to Definition 3.3, of type A that satisfies the classical second order conditions, the additional algebraic conditions (3.12), (3.14) for the second component and (3.16) for the first component. In particular we will prove that the construction of such scheme of second order is impossible with $s = 3$ internal stages, we therefore look for a four stage scheme.

Our task is now to determine the coefficients of a second order IMEX R-K scheme of type A globally stiffly accurate (AGSA). We first give lower bounds on the number of stages needed to reach second order. The results are summarized by the following two theorems.

Theorem 4.1. *There are no second order globally stiffly accurate IMEX R-K schemes of type A with three stages.*

A proof of this theorem is given in Appendix A.3.

Theorem 4.2. *There are no second order globally stiffly accurate IMEX R-K schemes of type A with four stages where the implicit part is singly diagonally implicit (SDIRK).*

The proof is reported in Appendix A.4.

For this reason we look for a family of second order IMEX R-K schemes of type A with a DIRK scheme in the implicit part, i.e., $a_{ii} \neq 0$ for all i. Then the Butcher *tableau* of an IMEX R-K scheme globally stiffly accurate of type A reads as

$$
\begin{array}{c|cccc}
0 & 0 & 0 & 0 & 0 \\
\tilde{c}_2 & \tilde{c}_2 & 0 & 0 & 0 \\
\tilde{c}_3 & \tilde{a}_{31} & \tilde{a}_{32} & 0 & 0 \\
1 & \tilde{b}_1 & \tilde{b}_2 & \tilde{b}_3 & 0 \\
\hline
 & \tilde{b}_1 & \tilde{b}_2 & \tilde{b}_3 & 0
\end{array}
\qquad
\begin{array}{c|cccc}
c_1 & c_1 & 0 & 0 & 0 \\
c_2 & a_{21} & a_{22} & 0 & 0 \\
c_3 & a_{31} & a_{32} & a_{33} & 0 \\
1 & b_1 & b_2 & b_3 & \gamma \\
\hline
 & b_1 & b_2 & b_3 & \gamma
\end{array}
$$

In order to simplify the computation of the coefficients, we choose $a_{22} = \gamma$ and $b_3 = 0$. Then, there is a total of 12 coefficients to be determined: 5 coefficients $\tilde{c}_2$, $\tilde{c}_3$, $\tilde{a}_{32}$, $\tilde{b}_2$, $\tilde{b}_3$ for the explicit part and 7 coefficients for the implicit one, γ, c_1, c_2, c_3, a_{32}, a_{33}, b_2.

Now we determine coefficients that satisfy the classical second order conditions

$$
\begin{aligned}
\tilde{b}^T e &= 1, & b^T e &= 1, \\
\tilde{b}^T \tilde{c} &= 1/2, & b^T c &= 1/2, \\
\tilde{b}^T c &= 1/2, & b^T \tilde{c} &= 1/2,
\end{aligned}
\tag{4.1}
$$

with $c = Ae$ and $\tilde{c} = \tilde{A}e$, conditions (3.12), (3.14) for the second component, and (3.16) for the first one, respectively.

Then the resulting system contains 9 equations and 12 unknowns with 3 free parameters. In particular we introduce a new equation given by the stability condition imposed on the explicit part of the scheme: $\sum_{i,j} \tilde{b}_i \tilde{a}_{ij} \tilde{c}_j = 1/6$, i.e., $\tilde{b}_3 \tilde{a}_{32} \tilde{c}_2 = 1/6$. This condition guarantees for the explicit part of the scheme a stability region as all third order explicit Runge–Kutta schemes. Finally, all remaining coefficients are obtained by $\tilde{b}_1 = 1 - \tilde{b}_2, -\tilde{b}_3$, $b_1 = 1 - b_2 - b_3 - \gamma$, $\tilde{a}_{31} = \tilde{c}_3 - \tilde{a}_{32}$, $a_{21} = c_2 - \gamma$, $a_{31} = c_3 - a_{33} - a_{32}$.

This IMEX R-K scheme of type A, globally stiffly accurate, with three stages for the explicit part, four stages for the implicit one and second order accurate will be denoted AGSA(3,4,2). Its coefficients are reported in Appendix A.5.

5. IMEX finite difference scheme.

In this section we describe the space discretization that is adopted in the paper. In view of future applications to quasi-linear hyperbolic systems of balance laws, we use conservative shock capturing schemes even if nonconservative discretization would be sufficient to treat systems with zero flux ($q(u) = 0$). In particular, we consider conservative finite difference discretization [13], which allows a simpler treatment of the source term with respect to finite volume type discretization.

We consider system (1.1) in the following form:

$$
\begin{aligned}
u_t &= -(v + \mu(\varepsilon) p(u)_x)_x + \mu(\varepsilon) p(u)_{xx}, \\
v_t &= -\frac{1}{\varepsilon^2} p(u)_x - \frac{1}{\varepsilon^2} v,
\end{aligned}
\tag{5.1}
$$

where we have added and subtracted the term $\mu(\varepsilon)\partial_{xx}p(u)$ and set $q(u) = 0$.

Here we describe a finite difference scheme for a system of the form

$$U_t + F(U)_x = G(U) \tag{5.2}$$

and apply it to the system (5.1) with

$$F(U) = \left(v + \mu p(u)_x, \frac{1}{\varepsilon^2}p(u)\right)^T, \quad G(U) = \left(\mu p(u)_{xx}, -\frac{1}{\varepsilon^2}v\right)^T.$$

Let Δx be the mesh width. We divide the computational domain into J cells $C_j = [x_{j-1/2}, x_{j+1/2}]$, $j = j_{min}, \ldots, j_{max}$, and denote by x_j the grid point located at the center.

Conservative finite difference for system (5.2) are written as follows [15]:

$$\frac{dU_j}{dt} = -\frac{\hat{F}_{j+\frac{1}{2}} - \hat{F}_{j-\frac{1}{2}}}{\Delta x} + G(U_j),$$

where $U_j(t) \approx U(x_j, t)$ is an approximation of the pointwise value of U at grid nodes, and the numerical flux at cell edge $x_{j+\frac{1}{2}}$ is computed as follows:

$$\hat{F}_{j+\frac{1}{2}} = \hat{F}_j^+(x_{j+\frac{1}{2}}) + \hat{F}_{j+1}^-(x_{j+\frac{1}{2}}).$$

The function $\hat{F}_j^+(x)$ and $\hat{F}_{j+1}^-(x)$ are suitable reconstructions defined, respectively, in cell j and in cell $j+1$. They are obtained as follows. First, we assume that the flux can be split into a positive and negative component

$$F(U) = F^+(U) + F^-(U),$$

with $\lambda(\nabla_U F^+(U)) \geq 0$, $\lambda(\nabla_U F^-(U)) \leq 0$. Here $\nabla_U F(U)$ denotes the Jacobian matrix, and $\forall\ A \in \mathbb{R}^{m\times m}$, $\lambda(A)$ denotes the spectrum of A. The quantity $F_j^\pm = F^\pm(U_j)$ are computed at cell center. Then $\hat{F}_j^\pm(x)$ are reconstructed from $\left\{F_j^\pm\right\}$ using high order essentially nonoscillatory reconstruction, such as ENO or WENO, that allows pointwise reconstruction of a function from its cell averages (see, e.g., [13] for details).

In all our examples we used the simple local Lax–Friedrichs flux decomposition, i.e., $F^+(U) = \frac{1}{2}(F(U) + \alpha U)$, $F^-(U) = \frac{1}{2}(F(U) - \alpha U)$, $\alpha \geq \max_U |\rho(\nabla_U F)|$, where $\rho(A) = \max_{1\leq i\leq m} |\lambda_i(A)|$ denotes the spectral radius of matrix $A \in \mathbb{R}^{m\times m}$, and the max defining α is taken for U varying in a suitable range in a neighborhood of each cell. Note that for large values of ε (e.g., $\varepsilon \approx 1$), we do not want to add and subtract term, which cause loss of accuracy when it is not needed. For such a reason a possible choice for μ can be

$$\mu(\varepsilon) = \begin{cases} 1 & \text{if} \quad \varepsilon^2 < \Delta x, \\ 0 & \text{if} \quad \varepsilon^2 \geq \Delta x \end{cases} \tag{5.3}$$

or some smoothed version of it (e.g., $\mu = \exp(-\varepsilon^2/\Delta x)$) as already used in [19].

For the diffusion term $\partial_{xx}p(u)$ we use the standard second order finite difference discretization.

5.1. Numerical tests. In this section we perform several numerical tests of our new IMEX R-K scheme of type A, i.e., AGSA(3,4,2). Our test problems are computed with coarse grids ($\Delta x, \Delta t \gg \varepsilon$), that do not resolve the small scales. In particular, in order to obtain high order schemes, we couple WENO schemes, e.g., third-second order WENO(3,2) reconstruction [13], for space discretization with IMEX R-K scheme AGSA(3,4,2) for time advancement. More precisely, WENO is applied to reconstruct the term $(-(v+\mu(\varepsilon)p(u)_x)_x, p(u)_x/\varepsilon^2)^T$ at the cell edges, while the term $\mu(\varepsilon)p(u)_{xx}$ is discretized by central differencing.

First we check the order of convergence by comparing the numerical solution of the system (2.1) to the exact solution. Since system (2.1) is linear, the exact solution can be computed by taking Fourier transform in space (Fourier series for boundary condition), and solving exactly the evolution equation for each Fourier mode. We use periodic boundary conditions in $x \in [-\pi, \pi]$ with initial conditions $u(x, 0) = \cos x$, $v(x, 0) = \sin x$. Two cases are considered $\varepsilon = 10^{-6}$ and $\varepsilon = 1$.

Numerical convergence rate is calculated by the formula

$$p = \log_2(E_{\Delta x_1}/E_{\Delta x_1}),$$

where $E_{\Delta x_1}$ and $E_{\Delta x_2}$ are the global errors computed with step Δx_1 and $\Delta x_2 = \Delta x_1/2$ respectively.

We computed the numerical solution using a hyperbolic stability restriction $\Delta t \leq \text{CFL}\Delta x$ with final time $t = 1$, $N = 40, 80, 160, 320, 640, 1280$ grid points. The L^∞ norms for the relative error among the numerical and exact solutions are computed and results for $\varepsilon = 10^{-6}$ are in Table 5.1(a). One can see that the expected convergence rate is reached. We remark that the convergence rate is also reached for $\varepsilon = 1$ as we can see in Table 5.1(b).

Table 5.1
L^∞-norms of the relative error and convergence rates of u.

(a) $\varepsilon = 10^{-6}$, CFL = 0.5

N	Error	Order
40	1.2014e-02	–
80	3.4484e-03	1.8007
160	9.3359e-04	1.8851
320	2.418e-04	1.9490
640	6.1579e-05	1.9733
1280	1.5557e-05	1.9849

(b) $\varepsilon = 1$, CFL = 0.5

N	Error	Order
40	7.3551e-02	–
80	2.1418e-02	1.7799
160	5.8368e-03	1.8756
320	1.5163e-03	1.9446
640	38663e-04	1.9715
1280	9.7681e-005	1.9848

In Figure 5.1 we report relative error as a function of CFL number for $\varepsilon = 10^{-6}$ and $N = 320$. Note that for CFL≈1 the error is quadratic in CFL, because the method is third order accurate in space, and second order accurate in time, therefore time error dominates. Now we perform a test proposed in [16] and [17] concerning nonlinear diffusion problems. We introduce the following relaxation system:

$$\begin{aligned} &\rho_t + j_x = 0, \\ &\varepsilon^2 j_t + Dp(\rho)_x = -j, \end{aligned} \tag{5.4}$$

where $p'(\rho) > 0$ and $D > 0$ is a diffusivity coefficient. System (5.4) is hyperbolic with two distinct real characteristic speeds $\pm\sqrt{p'(\rho)}/\varepsilon$. In particular, for small values of ε in the small relaxation limit ($\varepsilon \to 0^+$) one may get the following nonlinear degenerate parabolic equation:

$$\rho_t = Dp(\rho)_{xx}. \tag{5.5}$$

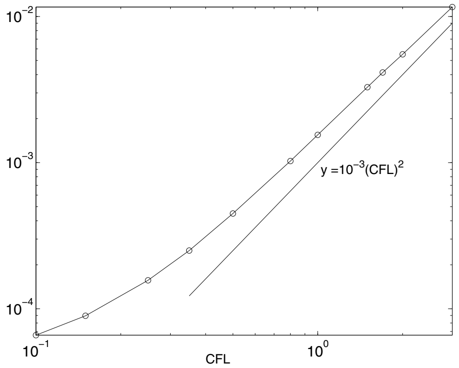

FIG. 5.1. *Relative error versus $CFL \equiv \Delta t/\Delta x$ in logarithmic scale.*

Equation (5.5) is degenerate if $p'(\rho^*) = 0$ for some ρ^*. In the case $p(\rho) = \rho^m$, with $m > 1$ the previous equation is the *porous media equation*. In this case the diffusion coefficient mu^{m-1} vanishes at the points where $u = 0$ and the governing parabolic equation degenerates there. At variance with classical diffusion equation, the support of the solution advances at constant speed. The interested reader may consult the book by Vázquez on the porous media equation [29]. For finite ε, the characteristic speeds vanish when $\rho = 0$, therefore linear waves are confined to the initial support. The support changes due to the formation of shocks. Of course the method works for $\rho > 0$. If there are regions with $\rho = 0$, because of truncation errors, negative numerical values may appear, and the code crashes. In order to avoid this at each time step the absolute value of ρ is used.

Another interesting case corresponds to $0 < m < 1$, and it is referred to as the *fast diffusion equation*. In our numerical tests we tested the problem with $m = 2, 3$.

Our aim now is to apply our scheme to the equivalent system

$$\begin{aligned} \rho_t + j_x &= \mu(\varepsilon) D(\rho^m)_{xx} - \mu(\varepsilon) D(\rho^m)_{xx}, \\ \varepsilon^2 j_t + D(\rho^m)_x &= -j, \end{aligned} \tag{5.6}$$

that for $\varepsilon \to 0^+$ relaxes towards (5.5) where we have put $\mu(0) = 1$.

We remark that if we apply an IMEX R-K scheme to the system (5.6) in the limit case ($\varepsilon \to 0^+$) we require solving the following equation:

$$\rho_t = D(\rho^m)_{xx} \tag{5.7}$$

implicitly.

Implicit Euler scheme applied to (5.7) reads

$$\rho_{i,n+1} = \rho_{i,n} + D\frac{\Delta t}{\Delta x^2}((\rho_{i-1,n+1})^m - 2(\rho_{i,n+1})^m + (\rho_{i+1,n+1})^m)$$

for $i = 0, 1, \ldots, N$. For an implicit method nonlinear equations must be solved at each time step, for example by the following iterative scheme:

$$\begin{aligned} \rho_{i,n+1}^{(k+1)} = \rho_{i,n}^{(k)} & \\ & + D\frac{\Delta t}{\Delta x^2}((\rho_{i-1,n+1}^{(k)})^{m-1}\rho_{i-1,n+1}^{(k+1)} - 2(\rho_{i,n+1}^{(k)})^{m-1}\rho_{i,n+1}^{(k+1)} \\ & + (\rho_{i+1,n+1}^{(k)})^{m-1}\rho_{i+1,n+1}^{(k+1)}) \end{aligned} \tag{5.8}$$

for $k = 0, 1, \ldots,$ where we can choose as initial guess $\rho^0_{i,n+1} = \rho_{i,n}$, for instance.

We can rewrite (5.8) as

$$\rho^{(k+1)}_{n+1} = \rho_n + D\frac{\Delta t}{\Delta x^2}\mathfrak{B}(\rho^{(k)}_{n+1})\rho^{(k+1)}_{n+1},$$

where the matrix $\mathfrak{B}(z)$ is a tridiagonal matrix with the ith row as

$$(0, \ldots, z^{m-1}_{i-1}, -2z^{m-1}_i, z^{m-1}_{i+1}, \ldots, 0).$$

With an abuse of notation, we denote by ρ_n the vector corresponding to the space discretization of the function $\rho(x, t_n)$. In the limit $k \to \infty$ the method converges to the solution obtained by implicit Euler, which is only first order accurate in time. In practice one iteration is enough to provide both stability and first order accuracy, and there is really no great advantage in using more than one iteration.

This statement is justified by the convergence of the iteration

$$\rho^{(k+1)}_{n+1} = \left(I - D\frac{\Delta t}{\Delta x^2}\mathfrak{B}(\rho^{(k)}_{n+1})\right)^{-1}\rho_n,$$

which is guaranteed provided that the spectral radius

$$\rho\left(\left(I - D\frac{\Delta t}{\Delta x^2}\mathfrak{B}(z)\right)^{-1}\right) < 1 \tag{5.9}$$

for all $z \in \mathbb{R}^{N+1}$ and $z > 0$, i.e., $\lambda(I - D\frac{\Delta t}{\Delta x^2}\mathfrak{B}(z)) > 1$. In order to guarantee inequality (5.9) we have to prove that the matrix $-\mathfrak{B}(z)$ is positive definite. We can write $\mathfrak{B}$ as $\mathfrak{B} = LZ$, where $Z = \mathrm{diag}(z_1, \ldots, z_N)$ and L is the tridiagonal matrix used for the discretization of the Laplacian in 1-D with Dirichlet boundary conditions. Then, by $\mathfrak{B}(z)u = \lambda u$, with any nonzero eigenvector u and $\lambda = (\lambda_0, \ldots, \lambda_N)$, with λ_i for $i = 0, \ldots, N$ eigenvalues, we get $LZu = \lambda u$.

Now, with the transformation $Z^{\frac{1}{2}}u = y$, we have $Z^{\frac{1}{2}}LZ^{\frac{1}{2}}y = \lambda y$; therefore λ are the eigenvalues of $Z^{\frac{1}{2}}LZ^{\frac{1}{2}}$, which are negative definite because $\lambda(L) < 0$ and $z > 0$. Since $\lambda(I - D\frac{\Delta t}{\Delta x^2}\mathfrak{B}(z)) = 1 - D\frac{\Delta t}{\Delta x^2}\lambda(\mathfrak{B}(z)) > 1$, this confirms the statement.

Furthermore, this approach is particularly effective in the case $m = 2$, since in this case we obtain a second order scheme by using just one iteration, i.e., by setting $\rho_{n+1} = \rho^{(1)}_{n+1}$, as can be checked analytically. In fact, from (5.7) by Taylor's expansion at $\rho(x, 0) = \rho_0(x)$ we have

$$\rho(x, \Delta t) = \rho_0(x) + D\Delta t((\rho_0(x))^2)_{xx} + D^2\Delta t^2(\rho_0(x)((\rho_0(x))^2)_{xx})_{xx} + \mathcal{O}(\Delta t^3),$$

and if we now use the standard centered approximations to the derivatives and drop the higher order terms, we obtain (we remark that we are interested here in what happens in just one time step) the following (using vector notation) discrete Taylor expansion:

$$\underline{\rho}_1 = \underline{\rho}_0 + \Delta t\mathcal{D}\underline{\rho}_0 + \Delta t^2\mathcal{D}^2\underline{\rho}^2_0 + \mathcal{O}(\Delta t^3, \Delta x^2)$$

where $\mathcal{D} = (D\mathfrak{B}(\underline{\rho}_0))/\Delta x^2$. Now putting $k = 0$ and $m = 2$ in (5.8), for one step we get

$$\underline{\rho}_1 = \underline{\rho}_0 + \Delta t\mathcal{D}\underline{\rho}_1$$

TABLE 5.2
L^1-norms of the relative error and convergence rates for ρ.

N	Error	Order
40	9.6371e-02	–
80	3.1304e-02	1.6222
160	9.6296e-03	1.7008
320	2.4073e-03	2.0000

So we have $\underline{\rho}_1 = (I + \Delta t\mathcal{D}\underline{\rho}_0 + \Delta t^2\mathcal{D}^2\underline{\rho}_0^2)\underline{\rho}_0 + \mathcal{O}(\Delta t^3)$ and dropping the higher order terms, we obtain

$$\underline{\rho}_1^{(1)} = \underline{\rho}_0 + \Delta t\mathcal{D}\underline{\rho}_0 + \Delta t^2\mathcal{D}^2\underline{\rho}_0$$

that is the statement. We remark that in the case $m \neq 2$, method (5.8) is only first order accurate, independently of the number of iterations. Better strategies, based on semi-implicit formulation [28], could be used, but they are beyond the scope of this paper.

In our numerical results we take $\Delta t = \mathcal{O}(\Delta x)$. The precise choices of Δt are reported in Figure 5.2 captions. Then we want to solve (5.6) with the initial data

$$\begin{aligned} \rho_0(x) &= (\cos(\pi x/2))^2 \quad \text{for } |x| < 1,\\ \rho_0(x) &= 0 \qquad\qquad\quad\; \text{for } |x| \geq 1. \end{aligned} \tag{5.10}$$

The computational domain is $|x| \leq 3$, the boundary conditions are periodic, and we choose $\varepsilon = 10^{-6}$.

First of all, we study the order of convergence of the scheme. As emphasized in [11] and [17], the front does not move for $t < 1/(3\pi^2) \approx 0.034$, then we choose a final time of the simulation $t_{fin} = 0.03$ to prevent the formation of the singularity of u_x from affecting the order of convergence. As a reference solution we use the one obtained numerically with $N = 2560$ grid points and we computed the L^1-norms of the errors of the numerical solutions with $N = 40, 80, 160, 320$ grid points. We tested the convergence rates by choosing $p(\rho) = \rho^2$. As confirmed by the data in Table 5.2, a second order is reached.

Furthermore, in Figure 5.2 we illustrate the propagation of the front for the numerical solution in the limit case ($\varepsilon = 10^{-6}$) with $p(\rho) = \rho^2$ and $p(\rho) = \rho^3$ with initial data (5.10) and $t \in [0,\ 2]$.

Now we test the scheme on the simple system (3.15) with $u = \rho$, $v = j$, by solving a Riemann problem with the initial data

$$\begin{aligned} \rho_L &= 2.0, \quad j_L = 0, \quad -1 < x < 0,\\ \rho_R &= 1.0, \quad j_R = 0, \quad 0 < x < 1. \end{aligned}$$

As ε goes to zero we have a pure diffusive linear problem $\rho_t = \rho_{xx}$, i.e., the problem becomes a classical Riemann problem for the heat equation.

In order to test our scheme we compute the numerical solution in the rarefied regime ($\varepsilon^2 > \Delta x$) and in the diffusive regime ($\varepsilon^2 \ll \Delta x$). Consequently, to solve this system in the rarefied and diffusive regime numerically we consider a smoothed version of μ, i.e., $\mu = \exp(-\varepsilon^2/\Delta x)$. This means that when ε is very large (i.e., rarefied regime) μ is very small, and on the other hand when ε is very small (i.e., diffusive regime) μ is equal to 1.

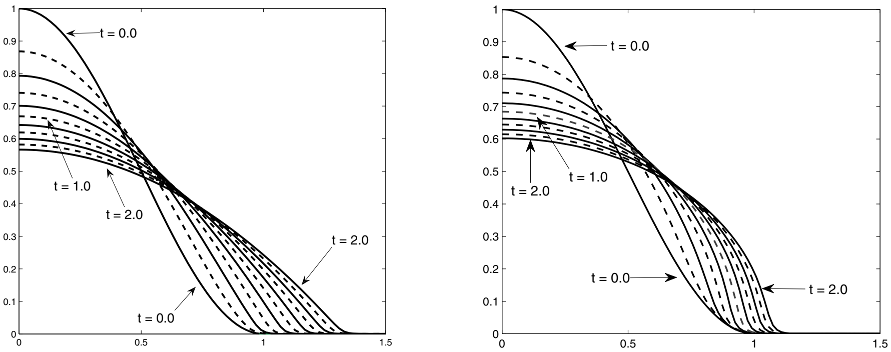

FIG. 5.2. *Numerical solution of the system* (5.6) *with* $\varepsilon = 10^{-6}$ *at time* $t = 2$. *On the left-hand side* $p(\rho) = \rho^2$, *on the right-hand side* $p(\rho) = \rho^3$. *The numerical solution are depicted at times* $t = 0, 0.2, 0.4, \ldots, 2.0$. *The solutions are obtained with* $N = 300$ *grid points for* $p(\rho) = \rho^2$, $\Delta x = 0.02$, *and* $CFL = 0.5$, *and with* $N = 360$ *grid points for* $p(\rho) = \rho^3$, $\Delta x = 0.0167$, *and* $CFL = 0.1$.

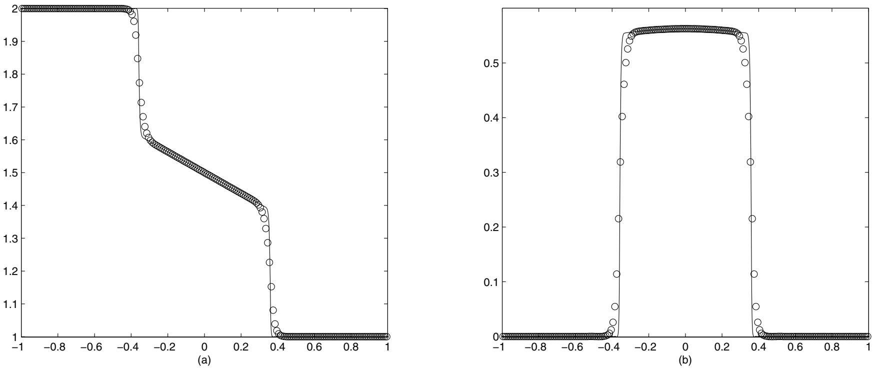

FIG. 5.3. *Numerical solution at time* $t = 0.25$ *in the rarefied regime* ($\varepsilon = 0.7$) *with* $\Delta t = 0.0025$, $CFL = 0.25$, *and* $\Delta x = 0.01$. *On the left-hand side the mass density* ρ (a) *and on the right-hand side the flow* j (b). *Solid line is the "exact" solution.*

Then we compute the scheme in the rarefied regime ($\varepsilon = 0.7$, see Figure 5.3) and in the diffusive regime (or stiff regime, see Figure 5.4) for $\varepsilon = 10^{-6}$. The numerical solutions for ρ and j in the rarefied and diffusive regime are depicted with an reference solution obtained using fine spacial grid of $N = 2000$ cells. The boundary conditions are of reflecting type. The solution is depicted at final time $t = 0.25$ in the rarefied regime and $t = 0.04$ in the diffusive regime.

In Figures 5.3 and 5.4 we can observe that the scheme developed here captures well the correct behavior of the solutions both in rarefied regime where this scheme provides an accurate description of the shock without oscillations near the discontinuities and in the diffusive regime where the numerical solutions match the reference solution very well.

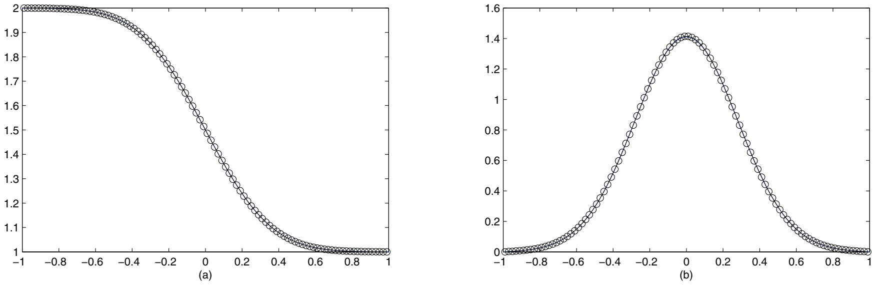

Fig. 5.4. *Numerical solution at time $t = 0.04$ in the parabolic regime ($\varepsilon = 10^{-6}$) with $\Delta t = 0.001$, $CFL = 0.1$, and $\Delta x = 0.02$. On the left-hand side the mass density* (a) *ρ and on the right-hand side the flow j* (b). *Solid line is the "exact" solution.*

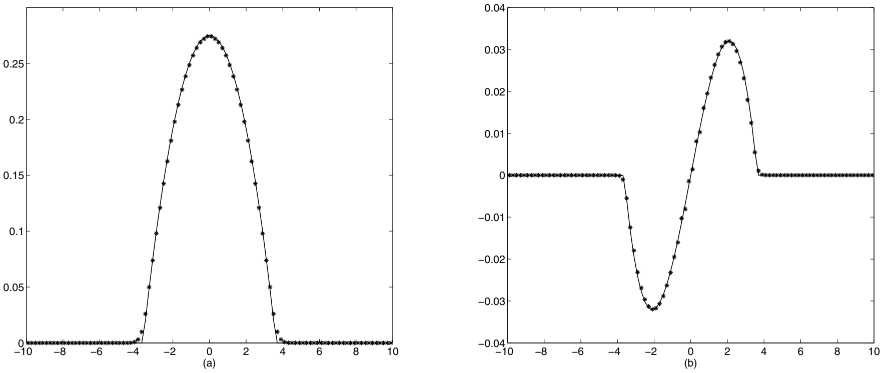

Fig. 5.5. *Numerical solution at time $t = 3.0$ for the Barenblatt problem* (5.12) *in the parabolic regime $\varepsilon = 10^{-6}$ with $\Delta x = 0.2$, $CFL = 0.5$, and $\Delta t = 0.1$. On the left-hand side the mass density ρ* (a) *and on the right-hand side the flow j* (b).

Finally we take the generalized Carlemann model

$$\rho_t + j_x = 0, \qquad (5.11)$$
$$\varepsilon^2 j_t + \rho_x = -2\rho^\kappa j$$

with $\kappa = -1$. This model relaxes to the porous media equation (5.7). In fact, when $\varepsilon \to 0$, the local equilibrium is given by

$$j = -\frac{\partial_x \rho}{2\rho^\kappa} = -\frac{\partial_x(\rho^{1-\kappa})}{2(1-\kappa)}$$

and the system relaxes to the nonlinear parabolic equation

$$\partial_t \rho = \frac{\partial_{xx}(\rho^{1-\kappa})}{2(1-\kappa)}.$$

Then we get (5.7) with $m = 2$ for $\kappa = -1$, where $D = 1/(2(1-\kappa))$.

Now we apply our scheme to the equivalent system

$$\begin{aligned} \rho_t + j_x &= -\mu(\varepsilon)\frac{\partial_{xx}(\rho^{1-\kappa})}{2(1-\kappa)} + \mu(\varepsilon)\frac{\partial_{xx}(\rho^{1-\kappa})}{2(1-\kappa)}, \\ \varepsilon^2 j_t + \rho_x &= -2\rho^{\kappa} j \end{aligned} \tag{5.12}$$

with $\kappa = -1$. Here we compare the numerical solution with the exact Barenblatt solution [4] for the porous media equation,

$$\begin{aligned} \rho(x,t) = \frac{1}{R(t)}\left[1 - \left(\frac{x}{R(t)}\right)^2\right], \quad j(x,t) = \rho\frac{2x}{R(t)^3}, \quad |x| < R(t), \\ \rho(x,t) = 0, \quad j(x,t) = 0, \quad |x| > R(t), \end{aligned} \tag{5.13}$$

where $R(t) = [12(t+1)]^{1/3}$, $t \geq 0$. We take $\Delta x = 0.2$ and $x \in\,]-10, 10[$. In Figure 5.5 the scheme AGSA(3,4,2) captures well the correct behavior of the exact solution.

6. Concluding remarks. In this paper we have introduced and analyzed a particular type of IMEX R-K schemes for hyperbolic systems with stiff diffusive relaxation that works well in the stiff regime, i.e., in the diffusive limit.

In particular our main goal is to construct IMEX R-K schemes that treat the convection term explicitly even though the characteristic speeds diverge. In this way we obtain a scheme that relaxes to an explicit scheme for the limit diffusion equation. Of course such an explicit scheme will suffer from the usual parabolic stability restriction $\Delta t \propto \Delta x^2$. Such a restriction can be overcome by reformulating the problem (2.1) in the equivalent form (3.15), by using a technique similar to the one adopted in [19]. We remark here that such a technique ensures stability only near relaxed regime, i.e., for $\varepsilon \ll 1$. For intermediate regimes hyperbolic CFL is not reached, and a more detailed analysis is needed in this case.

We derived additional order conditions up to second order that improve the accuracy in the limit case when $\varepsilon \to 0$. Construction of a second order IMEX R-K scheme globally stiffly accurate of type A has been proposed, and numerical tests on several hyperbolic systems with diffusive term reveal a good approximation of the equilibrium equations, i.e., it has the correct diffusive limit, and maintains second order accuracy in the limit.

Although we derived additional algebraic order conditions for the general system (3.17), we did not construct a scheme that satisfies them. This aspect will be investigated in future work.

Concerning higher order IMEX R-K schemes of type A, we note that it's not difficult to compute additional algebraic order conditions that have to be satisfied to construct a pth order IMEX R-K scheme of type A with $p \geq 3$. However, for larger orders $p \geq 3$, type A schemes are more complicated to construct than CK or ARS schemes [3].

Here we have concentrated on developing IMEX R-K scheme of type A because they are easier to analyze. It is of great interest to investigate other classes of IMEX R-K schemes, such as CK and ARS, in view of the construction of schemes of higher order. Furthermore many of the algebraic order conditions could be automatically satisfied or reduced by using the so-called simplifying assumptions (see [1, 2, 6] for details), and this is the reason why schemes of these types are more commonly used. The analysis of these types CK and ARS will be subject of future work.

In this paper we derived the scheme starting from a very specific model for diffusive relaxation. One of our goals is to extend a similar approach to a wider class of models, including the compressible Navier–Stokes equations and nonlinear diffusion model of LeFloch and Kawashima [28].

Appendix A.

A.1. Stability analysis of first order IMEX schemes. For the subsequent analysis we restrict to the linear case (2.6) and, in particular, using again a Fourier solution of the form $u = \hat{u}(t)\exp(i\xi x)$, $v = \hat{v}(t)\exp(i\xi x)$, and the new variable $\hat{w} = -i\hat{v}/\xi$ in place of $\hat{v}$, we obtain (2.7).

We apply IMEX-Euler(1) (2.4) to system (2.7) and we get

$$\begin{aligned} \hat{u}^{n+1} &= \hat{u}^n + \theta\hat{w}^n,\\ \zeta\hat{w}^{n+1} &= \zeta\hat{w}^n - \hat{u}^n - \hat{w}^{n+1}, \end{aligned} \tag{A.1}$$

which after manipulation can be written explicitly in the form

$$\begin{aligned} \hat{u}^{n+1} &= \hat{u}^n + \theta\hat{w}^n,\\ \hat{w}^{n+1} &= \frac{\zeta}{1+\zeta}\hat{w}^n - \frac{1}{1+\zeta}\hat{u}^n, \end{aligned} \tag{A.2}$$

where $\theta = \Delta t\xi^2$ and $\zeta = \varepsilon^2/\Delta t$.

In order to study the stability of the method we compute the eigenvalues of the stability matrix

$$R = \begin{pmatrix} 1 & \theta \\ -\dfrac{1}{1+\zeta} & \dfrac{\zeta}{1+\zeta} \end{pmatrix}. \tag{A.3}$$

We obtain for the characteristic polynomial the following expression:

$$(1+\zeta)\lambda^2 - \lambda(1+2\zeta) + (\zeta+\theta) = 0, \tag{A.4}$$

and it can be shown that the roots of the polynomial $|\lambda_\pm| < 1$ when $\theta \leqslant 1$, independently of ε.

A similar analysis applied to the system (2.6) with derivatives computed by central differing gives a stability matrix of the form

$$R = \begin{pmatrix} 1 & -ik\xi\,\mathrm{sinc}(\xi h) \\ -\dfrac{-ik\xi\,\mathrm{sinc}(\xi h)}{\varepsilon^2+k} & \dfrac{\varepsilon^2}{\varepsilon^2+k} \end{pmatrix}, \tag{A.5}$$

where $h \equiv \Delta x$, $k \equiv \Delta t$, and $\mathrm{sinc}(x) = \sin(x)/x$. We again obtain for the characteristic polynomial the expression

$$(1+\zeta)\lambda^2 - \lambda(1+2\zeta) + (\zeta + \theta\,\mathrm{sinc}^2(\xi h)) = 0, \tag{A.6}$$

and $|\lambda_\pm| < 1$ when $\theta\,\mathrm{sinc}^2(\xi h) \leqslant 1$, independently of ε, i.e., by

$$\xi^2 k \frac{\sin^2 \xi h}{h^2\xi^2} = \frac{k}{h^2}\sin^2 \xi h \leqslant 1,$$

we get $k/h^2 \leqslant 1$.

A.2. Additional order conditions. Here we derive the algebraic order conditions (3.12), (3.14), (3.16), and (3.18). We consider the equivalent system to (1.1) where we added and subtracted the term $\mu(\varepsilon)p(u)_{xx}$,

$$\begin{aligned} u_t &= -(v + \mu p(u)_x)_x + \mu p(u)_{xx}, \\ \varepsilon^2 v_t &= -p(u)_x - (v - q(u)). \end{aligned} \tag{A.7}$$

By using a method of lines approach (MOL), we discretize system (A.7) in space by a uniform mesh $\{x_i\}_{i=1}^N$ and $U_i(t) \approx u(x_i,t)$, $V_i(t) \approx v(x_i,t)$. We obtain a large system of ODEs,

$$\begin{aligned} U_t &= -D(V + \mu Dp(U)) + D(Dp(U)), \\ \varepsilon^2 V_t &= -Dp(U) - (V - Q(U)) \end{aligned} \tag{A.8}$$

with $U(t) = (U_1(t), U_2(t), \ldots, U_N(t))^T \in \mathbb{R}^N$ and $V(t) = (V_1(t), V_2(t), \ldots, V_N(t))^T \in \mathbb{R}^N$. Here DV and $Dp(U)$ (with a slight abuse of notation) denote the discretization of the terms v_x, $p(u)_x$, while $Q(U)$ represents the discretization of the term $q(u)$.

System (A.8) can be written in the following form:

$$\begin{aligned} U' &= f_1(U,V) + \hat{f}_2(U), \\ \varepsilon^2 V' &= g_1(U) - g_2(U,V), \end{aligned} \tag{A.9}$$

where the primes denote the time derivatives and $g_1(U) = -Dp(U)$, $g_2(U,V) = (V - Q(U))$, $f_1(U,V) = -D(V - g_1(U))$, and $\hat{f}_2(U) = -D(g_1(U))$. In the limit $\varepsilon \to 0$ we obtain $g_2(U,V) = g_1(U)$ or explicitly $V = g_1(U) + Q(U)$ that gives $f_1(U,V) = -DQ(U)$. Then, putting $\hat{f}_1(U) = -DQ(U)$, we read system (A.9),

$$U' = \hat{f}_1(U) + \hat{f}_2(U), \tag{A.10}$$

with $V = g_1(U) + Q(u)$.

Now applying one step of an IMEX R-K globally stiffly accurate of type A to system (A.9) w e get in vectorial form for the numerical solution

$$\begin{aligned} U_1 &= U_0 + \Delta t \tilde{b}^T f_1(\mathcal{U},\mathcal{V}) + \Delta t b^T \hat{f}_2(\mathcal{U}), \\ \zeta V_1 &= \zeta V_0 + \tilde{b}^T g_1(\mathcal{U}) - b^T g_2(\mathcal{U},\mathcal{V}), \end{aligned} \tag{A.11}$$

and for the internal stages

$$\begin{aligned} \mathcal{U} &= U_0 e + \Delta t \tilde{A} f_1(\mathcal{U},\mathcal{V}) + \Delta t A \hat{f}_2(\mathcal{U}), \\ \zeta \mathcal{V} &= \zeta V_0 e + \tilde{A} g_1(\mathcal{U}) - A g_2(\mathcal{U},\mathcal{V}), \end{aligned} \tag{A.12}$$

where $\zeta = \varepsilon^2/\Delta t$ with initial conditions $U_0 = U(t_0)$ and $V_0 = V(t_0)$. As usual we have denoted by $e = (1, \ldots, 1)^T$.

Starting from (A.12), by Definition 3.1, A is invertible and from the second equation in (A.12) we have $g_2(\mathcal{U},\mathcal{V}) = \zeta A^{-1}(V_0 e - \mathcal{V}) + A^{-1}\tilde{A} g_1(\mathcal{U})$. Now substituting this expression in the numerical solution V_1 we have

$$\zeta V_1 = \zeta(1 - b^T A^{-1} e)V_0 + \zeta b^T A^{-1}\mathcal{V} + (\tilde{b}^T - b^T A^{-1}\tilde{A}) g_1(\mathcal{U}). \tag{A.13}$$

By the global stiff accuracy property of the scheme of type A we get $\tilde{b}^T - b^T A^{-1}\tilde{A}e = 0$ and $R(\infty) = 1 - b^T A^{-1} e = 0$. Then we obtain $V_1 = e_s^T \mathcal{V}$ with

$e_s^T = (0, \dots, 1)$ that we make the definition of V_1 independent of ζ. Concerning the second equation in (A.12), as $\zeta \to 0$, i.e., $\varepsilon \to 0$, we get $g_2(\mathcal{U}, \mathcal{V}) = A^{-1}\tilde{A}g_1(\mathcal{U})$, or

$$\mathcal{V} = A^{-1}\tilde{A}g_1(\mathcal{U}) + Q(\mathcal{U}). \tag{A.14}$$

From the first equation in (A.12) we get explicitly $\mathcal{U} = U_0 e - \Delta t\tilde{A}D(\mathcal{V} - g_1(\mathcal{U})) - \Delta t AD(g_1(\mathcal{U}))$. By (A.14) we obtain for the internal stages

$$\mathcal{U} = U_0 e + \Delta t\tilde{A}\hat{f}_1(U) + \Delta t\mathbb{A}D(g_1(\mathcal{U})), \tag{A.15}$$

where $\mathbb{A} = \tilde{A}\mathcal{C} - A$, $\mathcal{C} = I - A^{-1}\tilde{A}$, and $\hat{f}_1(U) = -DQ(\mathcal{U})$, and for the numerical solution

$$U_1 = U_0 + \Delta t\tilde{b}^T\hat{f}_1(\mathcal{U}) + \Delta t(b^T - \tilde{b}^T\mathcal{C})\hat{f}_2(\mathcal{U}). \tag{A.16}$$

From now on we derive only order conditions up to the second order, an extension to the third order is trivial. In order to check the order of an IMEX R-K scheme, one has to compute the Taylor series expansion of the exact solutions $U(t_0 + \Delta t)$ and $V(t_0 + \Delta t)$ and the numerical ones U_1 and V_1 around to $\Delta t = 0$. First of all we compute the higher derivatives of the exact solutions U and V of (A.9) and (A.10) at the initial point t_0.

Let us start to evaluate the higher derivatives of the exact solution U. For this from (A.10), we get $U^{(q)}|_{\Delta t=0} = (\hat{f}_1(U))^{(q-1)} + (\hat{f}_2(U))^{(q-1)}$ and this gives for the first and second derivative

$$\begin{aligned} &U'(t_0) = \hat{f}_1(U_0) + \hat{f}_2(U_0), \\ &U''(t_0) = \hat{f}_1'(U_0)(\hat{f}_1(U_0) + \hat{f}_2(U_0)) + \hat{f}_2'(U_0)(\hat{f}_1(U_0) + \hat{f}_2(U_0)), \dots, \end{aligned} \tag{A.17}$$

and so on.

Now, by putting $\Delta t\hat{f}_i(\mathcal{U}) = K_i$ for $i = 1, 2$ we rewrite (A.15) and (A.16) as

$$\begin{aligned} &\mathcal{U} = U_0 e + \tilde{A}K_1 - \mathbb{A}K_2, \\ &U_1 = U_0 + \tilde{b}^T K_1 + (b^T - \tilde{b}^T\mathcal{C})K_2, \end{aligned} \tag{A.18}$$

where K_1, K_2, $\mathcal{U}$, and U_1 are functions of Δt.

Now we have to compute in (A.18) the derivatives of stage values and of the numerical solution at $\Delta t = 0$. By the Leibniz rule differentiating K_i, for $i = 1, 2$ with respect to Δt, this yields $K_i^{(q)} = \Delta t(\hat{f}_i(\mathcal{U}))^{(q)} + q(\hat{f}_i(\mathcal{U}))^{(q-1)}$ for $i = 1, 2$ and for $\Delta t = 0$ we get $K(0)_i^{(q)} = q(\hat{f}_i(\mathcal{U}))^{(q-1)}|_{\Delta t=0}$.

From (A.18) we get for the derivatives of the numerical solution

$$U_1^{(q)}(0) = \tilde{b}^T K_1^{(q)}(0) + (b^T - \tilde{b}^T\mathcal{C})K_2^{(q)}(0) \tag{A.19}$$

or explicitly up to second derivative

$$\begin{aligned} &U_1'(0) = \tilde{b}^T\hat{f}_1(U_0) + (b^T - \tilde{b}^T\mathcal{C})\hat{f}_2(U_0), \\ &U_1''(0) = 2\tilde{b}^T\hat{f}_1'(U_0)(\tilde{A}\hat{f}_1(U_0) - \mathbb{A}\hat{f}_2(U_0)) + 2(b^T - \tilde{b}^T\mathcal{C})\hat{f}_2'(U_0)(\tilde{A}\hat{f}_1(U_0) - \mathbb{A}\hat{f}_2(U_0)). \end{aligned} \tag{A.20}$$

Now in order to derive the order conditions we compare (A.17) and (A.20), and we obtain up to second order.

- First order conditions

$$\tilde{b}^T e = 1, \quad (b^T - \tilde{b}^T\mathcal{C})e = 1. \tag{A.21}$$

- Second order conditions

$$\begin{aligned} \tilde{b}^T \tilde{A} e &= 1/2, \quad (\tilde{b}^T\mathcal{C} - b^T)\mathbb{A}e = 1/2,\\ \tilde{b}^T \mathbb{A} e &= -1/2, \quad (b^T - \tilde{b}^T\mathcal{C})\tilde{A}e = 1/2. \end{aligned} \tag{A.22}$$

We note that $\tilde{b}^T e = 1$ and $\tilde{b}^T\tilde{A}e = \tilde{b}^T\tilde{c} = 1/2$ are some of the classical second order conditions (4.1) and $(b^T - \tilde{b}^T\mathcal{C})e = 1$ and $(\tilde{b}^T\mathcal{C} - b^T)\mathbb{A}e = 1/2$ are the additional order conditions (3.16) for the u component.

Now to make use of the definitions $\mathbb{A}$ and $\mathcal{C}$ and by the classical second order conditions (4.1), it is trivial to prove that from $\tilde{b}^T\mathbb{A}e = -1/2$, we can explicitly derive $\tilde{b}^T\tilde{A}A^{-1}\tilde{c} = 1/2$, and by $\tilde{c} = \tilde{A}e$ we get $(b^T - \tilde{b}^T\mathcal{C})\tilde{c} = 1/2$, i.e., the additional order conditions for the u component listed in (3.18).

Now we compute the derivatives of the exact solution V at the initial point t_0. Thus by (A.9), in the limit case $\varepsilon = 0$ we have $V = g_1(U) + Q(U)$ and computing the height derivatives at $\Delta t = 0$ this yields $V^{(q)}|_{\Delta t=0} = g_1(U)^{(q)} + Q(U)^{(q)} = G(U)^{(q)}$. This gives up to second derivative

$$\begin{aligned} &V'(t_0) = G'(U_0)(\hat{f}_1(U_0) + \hat{f}_2(U_0)),\\ &V'' = G''(U_0)(\hat{f}_1(U_0) + \hat{f}_2(U_0)) + G'(U_0)(\hat{f}_1'(U_0)(\hat{f}_1(U_0) + \hat{f}_2(U_0)) + \cdots\\ &\qquad + \hat{f}_2'(U_0)(\hat{f}_1(U_0) + \hat{f}_2(U_0))), \end{aligned} \tag{A.23}$$

with $G'(U_0) = g_1'(U_0) + Q'(U_0)$, $G^{''}(U_0) = g_1^{''}(U_0) + Q^{''}(U_0)$ and so on.

By the global stiff accuracy of the scheme of type A, we have the numerical solution $V_1 = e_s^T\mathcal{V}$ and by (A.14), we get $V_1 = b^T A^{-1}(A^{-1}\tilde{A}g_1(\mathcal{U}) + Q(\mathcal{U}))$. Now computing its high derivatives we obtain

$$V_1^{(q)}(0) = b^T A^{-1}(Q(\mathcal{U})^{(q)} + A^{-1}\tilde{A}g_1(\mathcal{U})^{(q)}).$$

Thus for the first derivative we get

$$V_1'(0) = b^T A^{-2}\tilde{A}g_1(U_0)\mathcal{U}' + b^T A^{-1}Q(U_0)\mathcal{U}', \tag{A.24}$$

with $\mathcal{U}' = \tilde{A}\hat{f}_1(0) - \mathbb{A}\hat{f}_2(0)$. Then comparing the exact and numerical solution we obtain the following order conditions up to first order:

- Consistency condition: $b^T A^{-2}\tilde{c} = 1$ (see (3.12));
- first order condition: $b^T A^{-2}\tilde{A}\mathbb{A}e = -1$ (see (3.14));
- additional algebraic order condition for the w-component $b^T A^{-2}\tilde{A}\tilde{c} = 1$ (see 3.18).

In particular, we obtain other additional algebraic order conditions:

$$b^T A^{-1}\tilde{A}e = 1, \quad b^T A^{-1}\mathbb{A}e = -1. \tag{A.25}$$

It is trivial to prove now that by the global stiff accuracy condition given in Definition 3.3, the two conditions (A.25) give again the order conditions in (A.21).

A.3. Proof of Theorem 4.1. We consider the classical second order conditions (4.1) and from Definition 3.3 the conditions $b^T = e_s^T A$ and $\tilde{b}^T = e_s^T \tilde{A}$.

For $s = 3$ the Butcher *tableau* of a stiffly accurate IMEX R-K of type A is

$$
\begin{array}{c|ccc}
0 & 0 & 0 & 0 \\
\tilde{c}_2 & \tilde{c}_2 & 0 & 0 \\
1 & \tilde{b}_1 & \tilde{b}_2 & 0 \\
\hline
 & \tilde{b}_1 & \tilde{b}_2 & 0
\end{array}
\qquad
\begin{array}{c|ccc}
c_1 & c_1 & 0 & 0 \\
c_2 & c_2 - a_{22} & a_{22} & 0 \\
1 & b_1 & b_2 & \gamma \\
\hline
 & b_1 & b_2 & \gamma
\end{array}
$$

(note that stiff accuracy implies $c_s = \tilde{c}_s = 1$). Now in order to solve the system (4.1) with $\tilde{b}^T = (\tilde{b}_1, \tilde{b}_2, 0)$, $b^T = (b_1, b_2, \gamma)$ and $\tilde{c} = (0, \tilde{c}_2, 1)^T$, $c = (c_1, c_2, 1)^T$, it is easy to compute the coefficients as follows:

$$\tilde{b}_2 = 1/(2\tilde{c}_2), \quad b_2 = (1 - 2\gamma)/(2\tilde{c}_2),$$

and

$$
\begin{aligned}
b_2(c_2 - c_1) &= 1/2 - \gamma - c_1 + \gamma c_1, \\
\tilde{b}_2(c_2 - c_1) &= 1/2 - c_1.
\end{aligned}
$$

Substituting $\tilde{b}_2$ and b_2 we get

$$\frac{(c_2 - c_1)}{2\tilde{c}_2} = \frac{1/2 - \gamma - c_1 - \gamma c_1}{1 - 2\gamma}, \qquad \frac{(c_2 - c_1)}{2\tilde{c}_2} = 1/2 - c_1.$$

Now, comparing and equating the two expressions we have $3\gamma c_1 = 0$, this yields that either γ or c_1 are zero and it is impossible because the matrix A is invertible. □

A.4. Proof of Theorem 4.2. By (3.12) and by the first equation in (3.16) we explicitly get

$$
\begin{aligned}
&\frac{1}{\gamma}(\tilde{b}_2\tilde{c}_2 + \tilde{b}_3\tilde{c}_3) - \frac{\tilde{b}_3 a_{32}\tilde{c}_2}{\gamma^2} = 1, \\
&-\frac{1}{\gamma^2}(b_2\tilde{c}_2 + b_3\tilde{c}_3) + \frac{b_3 a_{32}\tilde{c}_2}{\gamma^3} = 1 - \frac{1}{\gamma}.
\end{aligned}
\tag{A.26}
$$

The classical second order conditions, (4.1), become

$$
\begin{aligned}
&\tilde{b}_1 + \tilde{b}_2 + \tilde{b}_3 = 1, \quad b_1 + b_2 + b_3 + \gamma = 1, \\
&b_1\gamma + b_2 c_2 + b_3 c_3 + \gamma = \tfrac{1}{2}, \\
&\tilde{b}_2\tilde{c}_2 + \tilde{b}_3\tilde{c}_3 = \tfrac{1}{2}, \\
&b_2\tilde{c}_2 + b_3\tilde{c}_3 + \gamma = \tfrac{1}{2}, \\
&\tilde{b}_1\gamma + \tilde{b}_2 c_2 + \tilde{b}_3 c_3 = \tfrac{1}{2},
\end{aligned}
\tag{A.27}
$$

and we have from (A.26)

$$
\begin{aligned}
\tilde{b}_3 a_{32}\tilde{c}_2 &= \gamma(\tfrac{1}{2} - \gamma), \\
b_3 a_{32}\tilde{c}_2 &= \gamma(\gamma^2 - 2\gamma + \tfrac{1}{2}).
\end{aligned}
\tag{A.28}
$$

Now, dividing the two expressions in (A.28), we obtain

$$\tilde{b}_3 = \mathrm{C}(\gamma) b_3, \quad \text{with} \;\; \mathrm{C}(\gamma) := \frac{(\frac{1}{2} - \gamma)}{(\gamma^2 - 2\gamma + \frac{1}{2})}. \tag{A.29}$$

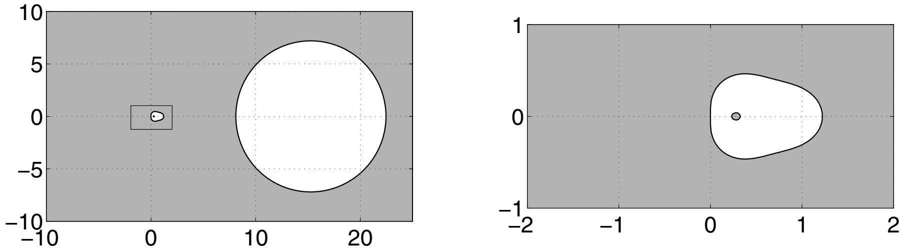

FIG. A.1. *Stability domain of AGSA$(3,4,2)$ scheme in the limit $\varepsilon \to 0$, in the complex θ plane. The gray region represents the stability region $|\hat{\mathcal{R}}(\theta)| \leq 1$. On the right-hand side: zoom of the region near the origin.*

If we replace (A.29) in (A.27) it follows that

$$\begin{aligned} &b_2(c_2-\gamma)+b_3(c_3-\gamma)=\tfrac{1}{2}-2\gamma+\gamma^2,\\ &\tilde{b}_2(c_2-\gamma)+\mathrm{C}(\gamma)b_3(c_3-\gamma)=\tfrac{1}{2}-\gamma,\\ &b_2\tilde{c}_2+b_3\tilde{c}_3=\tfrac{1}{2}-\gamma,\\ &\tilde{b}_2\tilde{c}_2+\mathrm{C}(\gamma)b_3\tilde{c}_3=\tfrac{1}{2}. \end{aligned} \tag{A.30}$$

Furthermore, by equations (3.14), we explicitly read

$$\begin{aligned} &\tilde{b}_3\tilde{a}_{32}+2\frac{\tilde{b}_3a_{32}\tilde{c}_2}{\gamma}-\frac{\tilde{b}_3\tilde{a}_{32}c_2}{\gamma}-\frac{b_3\tilde{a}_{32}\tilde{c}_2}{\gamma}+2\frac{\tilde{b}_3\tilde{a}_{32}\tilde{c}_2}{\gamma}=\frac{1}{2},\\ &b_3\tilde{a}_{32}+\frac{b_3a_{32}\tilde{c}_2}{\gamma}-\frac{b_3\tilde{a}_{32}c_2}{\gamma}+\frac{b_3\tilde{a}_{32}\tilde{c}_2}{\gamma}+\frac{\tilde{b}_3\tilde{a}_{32}\tilde{c}_2}{\gamma}=\frac{1}{2},\\ &\tilde{b}_3\tilde{a}_{32}+2b_3\tilde{a}_{32}-\tilde{b}_3a_{32}+3\frac{\tilde{b}_3\tilde{a}_{32}\tilde{c}_2}{\gamma}-\frac{\tilde{b}_3a_{32}\tilde{c}_2}{\gamma}+\frac{b_3a_{32}\tilde{c}_2}{\gamma}\\ &+2\frac{b_3\tilde{a}_{32}\tilde{c}_2}{\gamma}-\frac{\tilde{b}_3\tilde{a}_{32}c_2}{\gamma}-2\frac{b_3\tilde{a}_{32}c_2}{\gamma}+\frac{\tilde{b}_3a_{32}c_2}{\gamma}=1. \end{aligned} \tag{A.31}$$

Now, in order to determine the coefficients of the scheme, (A.28), (A.30), and (A.31) must be solved.

Thus, dividing the first and second expression in (A.31) follows $\tilde{b}_3=0$ that from (A.29) implies $b_3=0$ or $\mathrm{C}(\gamma)=0$. Then, if $b_3=0$, with $\tilde{b}_3=0$, system (A.31) is not satisfied. In a similar way if $\mathrm{C}(\gamma)=0$, i.e., $\gamma=1/2$, with $\tilde{b}_3=0$, we found no acceptable solution to the reduced system.

A.5. Coefficients of the AGSA(3,4,2) method and stability region. Below we list the coefficients of the scheme AGSA(3,4,2). The stability region of the scheme AGSA(3,4,2) is reported in Figure A.1.

$$\begin{aligned} &\tilde{c}_2=\tilde{a}_{21}=(-139833537)/38613965, && c_1=168999711/74248304,\\ &\tilde{a}_{31}=85870407/49798258, && \gamma=a_{22}=202439144/118586105,\\ &\tilde{a}_{32}=(-121251843)/1756367063, && a_{33}=12015439/183058594,\\ &\tilde{b}_2=1/6,\ \tilde{b}_3=2/3, && a_{31}=(-6418119)/169001713,\\ &a_{21}=44004295/24775207, && a_{32}=(-748951821)/1043823139,\\ &b_2=1/3,\ b_3=0,\ \tilde{b}_1=1-\tilde{b}_2-\tilde{b}_3, && b_1=1-\gamma-b_2-b_3. \end{aligned}$$

We observe that the number of stages of the explicit part is 3, not 4, because the scheme is FSAL, and therefore each time step requires three function evaluations.

REFERENCES

[1] S. Boscarino, *Error analysis of IMEX Runge–Kutta methods derided from differential algebraic systems*, SIAM J. Numer. Anal., 45 (2007), pp. 1600–1621.

[2] S. Boscarino, *On an accurate third order implicit-explicit Runge-Kutta method for stiff problems*, Appl. Numer. Math., 59 (2009), pp. 1515–1528.

[3] S. Boscarino and G. Russo, *On a class of uniformly accurate IMEX Runge–Kutta schemes and applications to hyperbolic systems with relaxation*, SIAM J. Sci. Comput., 31 (2009), pp. 1926–1945.

[4] G. I. Barenblatt, *On some unsteady motions of a liquid or a gas in a porous medium*, Akad. Nauk. SSSR Prikl. Math. Meh., 16 (1952), pp. 67–78 (in Russian).

[5] 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), pp. 151–167.

[6] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations* II: *Stiff and Differential Algebraic Problems*, Springer Ser. Comput. Math. 14, Springer-Verlag, Berlin, 1991, 2nd revised edition, 1996.

[7] E. Hairer, S. P. Norsett, and G. Wanner, *Solving Ordinary Differential Equation* I: *Nonstiff Problems*, Springer Ser. Comput. Math. 8, Springer-Verlag, Berlin, 1987, 2nd revised edition, 1993.

[8] A. Klar, *An asymptotic-induced scheme for nonstationary transport equations in the diffusive limit*, SIAM J. Numer. Anal., 35 (1998), pp. 1073–1094.

[9] P. Lafitte and G. Samaey, *Asymptotic-preserving projective integration schemes for kinetic equations in the diffusion limit*, SIAM J. Sci. Comput., 34 (2012), pp. A579–A602.

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

[11] G. Naldi and L. Pareschi, *Numerical schemes for hyperbolic systems of conservation laws with stiff diffusive relaxation*, SIAM J. Numer. Anal., 37 (2000), pp. 1246–1270.

[12] S. Jin, L. Pareschi, and G. Toscani, *Diffusive relaxation schemes for multiscale discrete-velocity kinetic equations*, SIAM J. Numer. Anal., 35 (1998), pp. 2405–2439.

[13] C. W. Shu, *Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws*, in Advanced Numerical Approximation of Nonlinear Hyperbolic Equations, Lecture Notes in Math. 1697, Springer, Berlin, 1998.

[14] M. Bennoune, M. Lemou, and L. Mieussens, *Uniformly stable numerical schemes for the Boltzmann equation preserving the compressible Navier-Stokes asymptotics*, J. Comput. Phys., 227 (2008), pp. 3781–3803.

[15] L. Pareschi and G. Russo, *Implicit-explicit Runge-Kutta schemes and applications to hyperbolic systems with relaxations*, J. Sci. Comput., 25 (2005), pp. 129–155.

[16] F. Cavalli, G. Naldi, G. Puppo, and M. Semplice, *High order relaxation schemes for nonlinear diffusion problems*, SIAM J. Numer. Anal., 45 (2007), pp. 2098–2119.

[17] J. L. Graveleau and P. Jamet, *A finite difference approach to some degenerate nonlinear parabolic equations*, SIAM J. Appl. Math., 20 (1971), pp. 199–223.

[18] S. Jin and L. Pareschi, *Discretization of the multiscale semiconductor Boltzmann equation by diffusive relaxation schemes*, J. Comput. Phys., 161 (2000), pp. 312–330.

[19] S. Boscarino, L. Pareschi, and G. Russo, *Implicit-Explicit Runge-Kutta schemes for hyperbolic systems and kinetic equations in the diffusion limit*, submitted.

[20] M. Lemou and L. Mieussens, *A new asymptotic preserving scheme based on micro-macro formulation for linear kinetic equations in the diffusion limit*, SIAM J. Sci. Comput., 31 (2008), pp. 334–368.

[21] T. P. Liu, *Hyperbolic conservation laws with relaxation*, Comm. Math. Phys., 108 (1987), pp. 153–175.

[22] G. Q. Chen, D. Levermore, and T. P. Liu, *Hyperbolic conservation laws with stiff relaxation terms and entropy*, Comm. Pure Appl. Math., 47 (1994), pp. 787–830.

[23] D. Aregba-Driollet, R. Natalini, and S. Tang, *Explicit diffusive kinetic schemes for nonlinear degenerate parabolic systems*, Math. Comp., 73 (2004), pp. 63–94.

[24] F. Cavalli, G. Naldi, G. Puppo, and M. Semplice, *Relaxed schemes based on diffusive relaxation for hyperbolic-parabolic problems: Some new developments*, in Numerical Methods for Balance Laws, Puppo Gabriella and Russo Giovanni, eds., 2009, pp. 175–195.

[25] F. Cavalli, G. Naldi, G. Puppo, and M. Semplice, *A family of relaxation schemes for*

nonlinear convection diffusion problems, Commun. Comput. Phys., 5 (2009), pp. 532–545.

[26] P. Smereka, *Semi-implicit level set methods for curvature and surface diffusion motion*, J. Sci. Comput., 19 (2003), pp. 439–456.

[27] F. Filbet and S. Jin, *A class of asymptotic-preserving schemes for kinetic equations and related problems with stiff sources*, J. Comput. Phys., 229 (2010), pp. 7625–7648.

[28] S. Boscarino, P. G. LeFloch, and G. Russo, *High-order asymptotic preserving methods for fully nonlinear relaxation problems*, SIAM J. Sci. Comput., submitted; preprint available online at http://arxiv.org/abs/1210.4761.

[29] J. L. Vázquez, *The Porus Media Equation. Mathematical Theory*, Oxford University Press, New York, 2006.