PREPRINT 417
FEBRUARY 2015

RWTH igpm

A new stable splitting for singularly perturbed ODEs

Jochen Schütz and Klaus Kaiser

Institut für
Geometrie und Praktische Mathematik
Templergraben 55, 52056 Aachen, Germany

updated version of July 7, 2015

A new stable splitting for singularly perturbed ODEs

Jochen Schütz, Klaus Kaiser

July 7, 2015

In this publication, we consider IMEX methods applied to singularly perturbed ordinary differential equations. We introduce a new splitting into stiff and non-stiff parts that has a direct extension to systems of conservation laws and investigate its performance analytically and numerically. We show that this splitting can in some cases improve the order of convergence, demonstrating that the phenomenon of order reduction is not only a consequence of the method but also of the splitting.

1. Introduction

The computation of extremely stiff ordinary differential equations has been the subject of extensive research over the last decades, see for example the standard textbook [25]. Our interest in stiff ordinary differential equations (ODE) stems from the approximation of compressible fluid flows [1, 47] using appropriate spatial discretization, such as for example the finite volume [20, 21, 34, 35] or the discontinuous Galerkin [9, 10, 11, 12] method. In particular, we are interested in the approximation of flows in the nearly incompressible regime, i.e., at very low Mach numbers ε [33]. This does not only lead to stiffness induced by a discretization parameter, but also to stiffness induced by the singular perturbation problem that constitutes the transition from compressible to incompressible flows [31, 43].

It has been recognized that handling such problems with explicit integration methods leads to a time step size depending on ε and thus to high computational costs if $\varepsilon \ll 1$. There are plenty of approaches in literature treating the stiffness. One example are split-explicit [18] methods that can increase computational efficiency for a fixed ε. However, time step restrictions are not independent of ε. Treating the equations fully implicit tends to remove the restriction on the time steps. In the context of computational fluid dynamics though, it usually leads to extreme damping [34]. Therefore in many cases, it might be beneficial to separate stiff (w.r.t. ε) from non-stiff terms and treat them implicitly and explicitly, respectively.

Examples of identifying stiff and non-stiff parts in the context of computational fluid dynamics can be found, e.g., in [5, 13, 16, 19, 22, 32, 36, 37, 41]; other examples, for instance from linear or elliptic equations can be found in [15, 44]. It has been recognized in [46] that decomposing the equations into stiff and non-stiff terms is not trivial. Even if both parts are stable individually, this does not necessarily mean that the overall algorithm is stable. For linear equations, the authors in [46] have found a uniformly stable scheme based on characteristic decomposition. This, however, can not easily be extended to nonlinear equations. To this end, we investigate a new, more general splitting based on the solution of the unperturbed ('incompressible') solution in this paper. This splitting has a direct extension to systems of conservation laws. Let us note that similar splittings have been presented in [5, 19, 36, 41, 48]. In particular, in [5], the authors linearize around the lake at rest solution and then treat the stiff, linear term implicitly. The systematic extension of this ansatz, which we pursue in this work, came out of discussions of the authors with Sebastian Noelle, Rupert Klein and Hamed Zakerzadeh, see [S. Noelle, R. Klein, J. Schütz, and H. Zakerzadeh. *RS-IMEX schemes: derivation and asymptotic stability.* IGPM Preprint, RWTH Aachen University, 2015 (in preparation)].

Splitting the equations into stiff and non-stiff parts leads to implicit / explicit (IMEX) time integration routines. Famous integrators include IMEX multistep methods, see, e.g., [4, 14, 26] and IMEX Runge-Kutta (RK) methods [3, 6, 7, 42]. We focus on IMEX BDF methods (to be explained in Sec. 4) and IMEX RK methods (to be explained in Sec. 5).

It is well-known that zero-stable IMEX BDF methods (up to six steps) are uniformly consistent, i.e., independent of the relation between the perturbation parameter ε and the time step Δt, the error converges with the correct order in Δt to zero. The same is not true for general IMEX RK methods [6] (with the exception of the specifically designed Runge-Kutta method by Boscarino [7]). This means that for $\Delta t \gg \varepsilon$, the error seems to exhibit some kind of degradation in convergence in general. In this work, we present some comparisons between a common splitting and the newly developed splitting, showing that order reduction is also a phenomenon of the splitting and not only the temporal integration.

The paper is organized as follows: In Sec. 2, we shortly comment on the singular perturbation problem used in this paper. Then in Sec. 3, we introduce the new splitting based on what we call the *reference solution* $w_{(0)}$; and we explain the important AP (asymptotic preserving) property [27, 28, 29]. In Sec. 4, we combine this splitting with the IMEX BDF method and show that it is AP. The analysis is equipped with numerical results. In Sec. 5, we apply the splitting to IMEX RK schemes, show again the AP property under suitable restrictions and present numerical results. Sec. 6 offers conclusions and outlook.

2. Singular perturbation problem and reference solution

In this work, we consider the ordinary differential equation

$$w'(t) = f(w(t)), \qquad w(0) = w_{in}, \tag{1}$$

where, for $\varepsilon > 0$,

$$w := \begin{pmatrix} y \\ z \end{pmatrix}, \qquad f(w) := \begin{pmatrix} z \\ \frac{g(y,z)}{\varepsilon} \end{pmatrix}. \tag{2}$$

(Note that extensions to somewhat more complex ODEs are often straightforward.) This equation constitutes a singularly perturbed problem (SPP), as formally, for $\varepsilon \to 0$ the equation changes its type to a differential algebraic equation. For a more detailed introduction to SPPs, we refer to [2, 24, 25, 38] and the references therein.

We assume that $g(\cdot,\cdot)$ is (at least) in class $C^2(\mathbb{R}^2)$, and that the logarithmic norm of $\partial_2 g(y,z)$ is bounded by a negative constant in the vicinity of the solution to guarantee the existence of an ε expansion, see [25]. (Note that in the scalar case, this amounts to $\partial_2 g(y,z) \leq -c$ for a positive constant c.)

One particular instance of (1) is *van der Pol* equation, defined by

$$g(y,z) := \left(1 - y^2\right) z - y. \tag{3}$$

Our interest is in the case of small (but finite) ε. Expanding w in terms of ε as

$$w(t) = w_{(0)}(t) + \varepsilon w_{(1)}(t) + \varepsilon^2 w_{(2)}(t) + \mathcal{O}(\varepsilon^3) \tag{4}$$

reveals that $w_{(0)} = (y_{(0)}, z_{(0)})^T$ fulfills the DAE

$$y'_{(0)}(t) = z_{(0)}(t), \qquad g(y_{(0)}, z_{(0)}) = 0. \tag{5}$$

Because $\partial_2 g(y,z)$ can be bound by a negative constant, we can guarantee that the latter equation is a differential algebraic equation (DAE) of index one [24]. Roughly speaking, this means that one can express $z_{(0)}$ as a function of $y_{(0)}$, and $y_{(0)}$ fulfills an ODE. We refer to $w_{(0)}$ as the *reference solution*:

Definition 1. *The solution* $w_{(0)} = (y_{(0)}, z_{(0)})^T$ *to* (5) *is called* reference solution *or RS, for short.*

Obviously, only carefully crafted initial conditions induce a well-posed DAE for $w_{(0)}$, and the same holds true for any $w_{(i)}$. Initial conditions w_{in} that 'survive' the limit as $\varepsilon \to 0$ are called *well-prepared*, they necessarily lie (to first order in ε) on the solution manifold of (5). One particular set of initial conditions for van der Pol equation from literature [25], that we are going to use in the sequel, is given by

$$w_{in} = \left(2, -\frac{2}{3} + \frac{10}{81}\varepsilon - \frac{292}{2187}\varepsilon^2 + \mathcal{O}(\varepsilon^3)\right)^T.$$

Remark 1. *Choosing arbitrary initial conditions that do not lie to first order on the solution manifold of* (5) *would lead to a boundary layer in the solution of* (3) *for finite but small* ε*. This, then again, is a solution that can* not *be represented via an expansion as in* (4)*, but it has to be represented by a more general expansion in both* ε *and* $\frac{t}{\varepsilon}$*. We refer to [25] for more details. Solutions that depend on* $\frac{t}{\varepsilon}$ *are not of interest in this work, as their dynamics is governed on a time scale of order* ε*, and a numerical scheme with uniform (w.r.t.* ε*) time step is arguably not useful in such a context.*

We note that, unlike for systems of conservation laws, the question of a suitable splitting for ordinary differential equations has usually not been discussed in literature. This is probably because for ordinary differential equations, there is a 'naive' splitting, given in Def. 2, that can be applied in a stable way (see [6]).

Definition 2. *The 'standard' (or 'naive') splitting usually employed in literature is given by*

$$f(w) = \begin{pmatrix} z \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \frac{g(y,z)}{\varepsilon} \end{pmatrix}, \tag{6}$$

where the second part is treated implicitly.

3. Splitting and AP property

3.1. Splitting

In this section, we define the newly-developed splitting. More formally, we introduce $\widehat{f}$ and $\widetilde{f}$, such that the right-hand side f of the ODE (1) can be split into

$$f(w) = \widehat{f}(w) + \widetilde{f}(w),$$

and we think about $\widetilde{f}(w)$ as a 'stiff' contribution to the flux. In (6), we already showed the standard way such a splitting is performed in literature. In this publication, we propose a splitting that has an extension to other types of ODEs and is related to the reference solution (RS) $w_{(0)}$, i.e., the (formal) limit solution for $\varepsilon \to 0$, see also Def. 1. The splitting is consequently called RS-IMEX splitting:

Definition 3. *RS-IMEX splitting: We define the following splitting for the right-hand side* f *of* (1)*:*

$$\widetilde{f}(w) = f(w_{(0)}) + f'(w_{(0)})(w - w_{(0)}), \qquad \widehat{f}(w) = f(w) - \widetilde{f}(w). \tag{7}$$

In the course of this work, $\widetilde{f}(w)$ will be treated implicitly, while $\widehat{f}(w)$ will be treated explicitly within an IMEX time integration method. There are a couple of remarks in order here:

Remark 2. *1. The motivation of this splitting is that for* ε *close to zero, the term* $w - w_{(0)}$ *is supposed to be* $\mathcal{O}(\varepsilon)$*, i.e., small. In particular,* $\widehat{f}$ *is supposed to be small. Therefore, we can have the hope that* $\widehat{f}$ *is 'non-stiff'.*

2. *Note that* $f(w_{(0)}) = (z_{(0)}, 0)^T$ *and*

$$f'(w_{(0)}) = \begin{pmatrix} 0 & 1 \\ \frac{1}{\varepsilon}\partial_1 g(w_{(0)}) & \frac{1}{\varepsilon}\partial_2 g(w_{(0)}) \end{pmatrix}.$$

From this, one can conclude that for $\Delta w := w - w_{(0)}$*, there holds*

$$\widetilde{f}(w) = \begin{pmatrix} z \\ \frac{1}{\varepsilon} g'(w_{(0)})\Delta w \end{pmatrix}, \qquad \widehat{f}(w) = \begin{pmatrix} 0 \\ \frac{1}{\varepsilon}\left(g(w) - g'(w_{(0)})\Delta w\right) \end{pmatrix}. \tag{8}$$

For future reference, we will define $\widetilde{g}(w)$ *as* $g'(w_{(0)})\Delta w$*, and* $\widehat{g}(w)$ *as* $g(w) - \widetilde{g}(w)$*.*

3. *In practice,* $w_{(0)}$ *is replaced by an approximation* $w_{(0)}^{app}$*.*

4. *Because* $w_{(0)}$ *depends on* t*, both* $\widehat{f}$ *and* $\widetilde{f}$ *depend on* t *as well. Most of the time, we will omit this dependence for the sake of simplicity.*

Remark 3. *The proposed splitting relies on a first-order Taylor expansion around the reference solution. The question of using a zeroth-order expansion directly springs to mind, i.e., taking the implicit part to be* $\widetilde{f}(w) = f(w_{(0)})$*. This, however, would render the scheme completely explicit, because the stiff part is independent of* w*, thus, there is no implicit stabilization mechanism present.*

We note once again that this splitting is universal in the sense that for any convergent singular perturbation problem, such a splitting is - at least formally - available.

3.2. Asymptotic preserving property

Consider any numerical method which computes an approximate solution $w_{\Delta t}^{n+1}$. This numerical solution should converge towards the exact solution w for $\Delta t \to 0$ and a finite value of ε: This is the standard consistency requirement. However, dealing with ε, another natural measure of consistency is whether the discrete solution, at a finite value of Δt, converges to some $w_{\Delta t,(0)}$ for $\varepsilon \to 0$, which constitutes an approximation to $w_{(0)}$. This consistency requirement is introduced in the sequel.

Definition 4. *An algorithm for the computation of a solution to* (1) *is called* asymptotic preserving *(AP) if the* discrete *limit (w.r.t.* $\varepsilon \to 0$*) algorithm is a consistent approximation to* (5)*.*

An illustration that is frequently shown in this context [28] can be seen in Fig. 1. If the diagram commutes (i.e., the order of limits $\Delta t \to 0$ and $\varepsilon \to 0$ can be changed) the algorithm is asymptotic preserving.

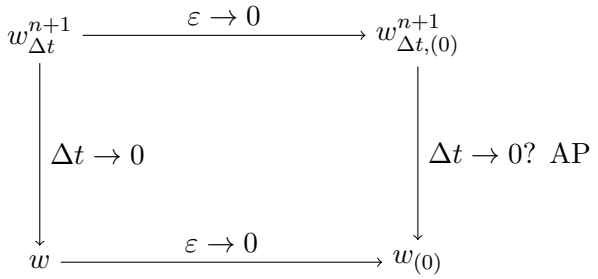

Figure 1: Illustration of the AP property. If $w_{\Delta t,(0)}^{n+1}$ converges toward $w_{(0)}$ for $\Delta t \to 0$, the algorithm is asymptotic preserving (AP).

4. IMEX BDF

In this section, we couple the splitting defined in Sec. 3 to an IMEX BDF scheme [26]. It is well-known that BDF schemes belong to the class of linear multistep schemes and are constructed in such a way that $w_{\Delta t}^{n+1}$ is given by the expression

$$\sum_{j=-1}^{s} \alpha_j w_{\Delta t}^{n-j} = \Delta t f(w_{\Delta t}^{n+1}).$$

Remark 4. *1. BDF schemes are zero-stable up to $s = 5$.*

2. Computing the coefficients is easily possible using the relation $A\vec{\alpha} = (0, 1, 0, \dots, 0)^T$ for $\vec{\alpha} = (\alpha_{-1}, \dots, \alpha_s)^T$ and matrix A with $A_{ij} = -\frac{(j-1)^{i-1}}{(i-1)!}$.

3. Obviously, these schemes are implicit.

The so-called extrapolated BDF scheme can be construced to be

$$\sum_{j=-1}^{s} \alpha_j w_{\Delta t}^{n-j} = \sum_{j=0}^{s} \Delta t \beta_j f(w_{\Delta t}^{n-j}).$$

These schemes are obviously explicit, and so they will constitute the explicit part of IMEX BDF.

Remark 5. *Again, the β_j fulfill a linear system of equations: $\vec{\beta} = (\beta_0, \dots, \beta_s)^T$ fulfills $B\vec{\beta} = (1, 0, 0, \dots, 0)^T$ for matrix B with $B_{ij} = (-1)^{i-1} \frac{j^{i-1}}{(i-1)!}$.*

Based on a splitting of f as in (7) it is obvious to construct the IMEX BDF scheme as

$$\sum_{j=-1}^{s} \alpha_j w_{\Delta t}^{n-j} = \Delta t \widetilde{f}(w_{\Delta t}^{n+1}) + \Delta t \sum_{j=0}^{s} \beta_j \widehat{f}(w_{\Delta t}^{n-j}). \tag{9}$$

Those schemes are usually indexed by their convergence order $s + 1$.

Remark 6. *The first-order IMEX scheme (*implicit-explicit Euler*)*

$$w_{\Delta t}^{n+1} = w_{\Delta t}^{n} + \Delta t \left(\widehat{f}(w_{\Delta t}^{n}) + \widetilde{f}(w_{\Delta t}^{n+1}) \right)$$

is an IMEX BDF scheme for $s = 0$.

We are now ready to show the main theorem of this section, namely, that the algorithm is *asymptotic preserving*.

4.1. Asymptotic Preserving Property

In this section we want to prove the AP property given in Def. 4.

Theorem 1. *The algorithm (9) with RS-IMEX splitting is asymptotic preserving with correct order $s + 1$.*

Proof. Let $w_{\Delta t}^{n}$ be expanded in terms of ε as

$$w_{\Delta t}^{n} = w_{\Delta t,(0)}^{n} + \varepsilon w_{\Delta t,(1)}^{n} + \mathcal{O}(\varepsilon^2)$$

for all n. We assume that start values $w^j_{\Delta t,(0)}$, $0 \le j \le s$ are consistent to the right order, i.e., $\Delta w^j_{\Delta t,(0)} := w^j_{\Delta t,(0)} - w_{(0)}(t^j) = \mathcal{O}(\Delta t^{s+1})$. Inserting the expansion of $w^{n-j}_{\Delta t}$, $0 \le j \le s$ into (9) leads to

$$\begin{aligned}
\sum_{j=-1}^{s} \alpha_j y^{n-j}_{\Delta t,(0)} &= \Delta t z^{n+1}_{\Delta t,(0)} + \mathcal{O}(\varepsilon),\\
\frac{\varepsilon}{\Delta t} \sum_{j=-1}^{s} \alpha_j z^{n-j}_{\Delta t,(0)} &= g'(w_{(0)}) \Delta w^{n+1}_{\Delta t,(0)} + \sum_{j=0}^{s} \beta_j \left(g(w^{n-j}_{\Delta t,(0)}) - g'(w_{(0)}) \Delta w^{n-j}_{\Delta t,(0)} \right) + \mathcal{O}(\varepsilon).
\end{aligned}$$

Then the (formal) limit algorithm for $\varepsilon \to 0$ is given by

$$\sum_{j=-1}^{s} \alpha_j y^{n-j}_{\Delta t,(0)} = \Delta t z^{n+1}_{\Delta t,(0)}, \tag{11a}$$

$$0 = g'(w_{(0)}) \Delta w^{n+1}_{\Delta t,(0)} + \sum_{j=0}^{s} \beta_j \left(g(w^{n-j}_{\Delta t,(0)}) - g'(w_{(0)}) \Delta w^{n-j}_{\Delta t,(0)} \right). \tag{11b}$$

The first equation can be rewritten as

$$\sum_{j=-1}^{s} \alpha_j y^{n-j}_{\Delta t,(0)} = \Delta t z_{(0)}(t^{n+1}) + \Delta t \Delta z^{n+1}_{\Delta t,(0)}.$$

This means that $y^{n+1}_{\Delta t,(0)} = y_{(0)}(t^{n+1}) + \mathcal{O}(\Delta t^{s+1}) + \mathcal{O}(\Delta t \Delta z^{n+1}_{\Delta t,(0)})$, which implies that $\Delta y^{n+1}_{\Delta t,(0)} = \mathcal{O}(\Delta t \Delta z^{n+1}_{\Delta t,(0)}) + \mathcal{O}(\Delta t^{s+1})$ for all n. Plugging this into (11b) and exploiting recursively $\Delta w^{n-j}_{\Delta t,(0)} = \mathcal{O}(\Delta t^{s+1})$ implies that $\Delta z^{n+1}_{\Delta t,(0)} = \mathcal{O}(\Delta t^{s+1})$. □

4.2. Numerical Results

In this section, we show numerical results based on van der Pol equation (see eqs. (1)–(3)) to show that the performed algorithm works as expected. In the numerical results, error has been computed as the two-norm of the difference to the solution at end time $T_{end} = 0.5$, i.e.,

$$e_{\Delta t} := \| w(T_{end}) - w^N_{\Delta t} \|_2,$$

where N is such that $N \Delta t = T_{end}$. We usually consider various ε, ranging from $\varepsilon = 10^{-1}$ to 10^{-7}, to show that the algorithms perform well for both relatively large and small values of ε.

We employ both IMEX BDF 2 and IMEX BDF 4, given by

$$\begin{aligned}
\frac{3}{2} w^{n+1}_{\Delta t} - 2 w^{n}_{\Delta t} + \frac{1}{2} w^{n-1}_{\Delta t} &= \Delta t \left(\widetilde{f}(w^{n+1}_{\Delta t}) + 2 \widehat{f}(w^{n}_{\Delta t}) - \widehat{f}(w^{n-1}_{\Delta t}) \right) \text{ and}\\
\frac{25}{12} w^{n+1}_{\Delta t} - 4 w^{n}_{\Delta t} + 3 w^{n-1}_{\Delta t} - \frac{4}{3} w^{n-2}_{\Delta t} + \frac{1}{4} w^{n-3}_{\Delta t} &= \Delta t \left(\widetilde{f}(w^{n+1}_{\Delta t}) + 4 \widehat{f}(w^{n}_{\Delta t}) - 6 \widehat{f}(w^{n-1}_{\Delta t}) + 4 \widehat{f}(w^{n-2}_{\Delta t}) - \widehat{f}(w^{n-3}_{\Delta t}) \right)
\end{aligned}$$

respectively.

Obviously, the most straightforward splitting (which, indeed, is usually employed in literature) is to split as in (6) and treat the ε-dependent part implicitly, i.e., take the stiff part to be $\left(0, \frac{g(y,z)}{\varepsilon}\right)^T$. We refer to this splitting as the *standard splitting*, see also Def. 2. For the RS-IMEX splitting as given in Def. 3, the reference solution $w_{(0)}$ is computed exactly. Note that this is possible analytically for van der Pol's example. In practical cases, however, $w_{(0)}$ (needed in (7)) is not readily available, so one has to compute an

approximation $w_{(0)}^{app}$ numerically. This approach, which we call RS-Approximate (or RSApp, for short), has also been implemented using a BDF discretization (always with corresponding order) of the limit differential algebraic equation. Initial steps needed for this multistep scheme are computed with a stiff integrator to extremely high precision.

In Fig. 2, numerical results are shown for IMEX BDF 2, where standard splitting (left), RS-IMEX splitting (middle) and RSApp are compared. Clearly, one can see that there is no order degradation in any cases. Furthermore, quantitatively and qualitatively, all plots behave nearly the same. An analogue observation can be made for IMEX BDF 4, shown in Fig. 3. All BDF schemes converge with their respective order of two and four uniformly in ε. Note that the irregularities one can observe for BDF 4 are cancellation effects often occurring for multistep schemes. They seem to be even more severe for the standard splitting.

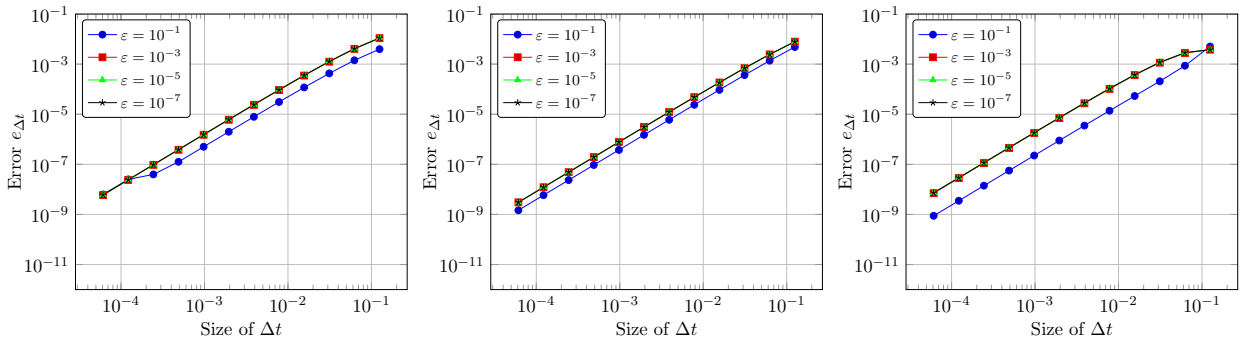

Figure 2: Convergence results for van der Pol equation for different values of ε, using the BDF 2 scheme coupled with the standard splitting (left), the RS-IMEX splitting with analytical $w_{(0)}$ (middle) and the RS-IMEX splitting with approximate $w_{(0)}^{app}$ (right). As an error measure, we choose $e_{\Delta t} := \|w(T_{end}) - w_{\Delta t}^N\|_2$ for $T_{end} = 0.5$ and N such that $T_{end} = N\Delta t$.

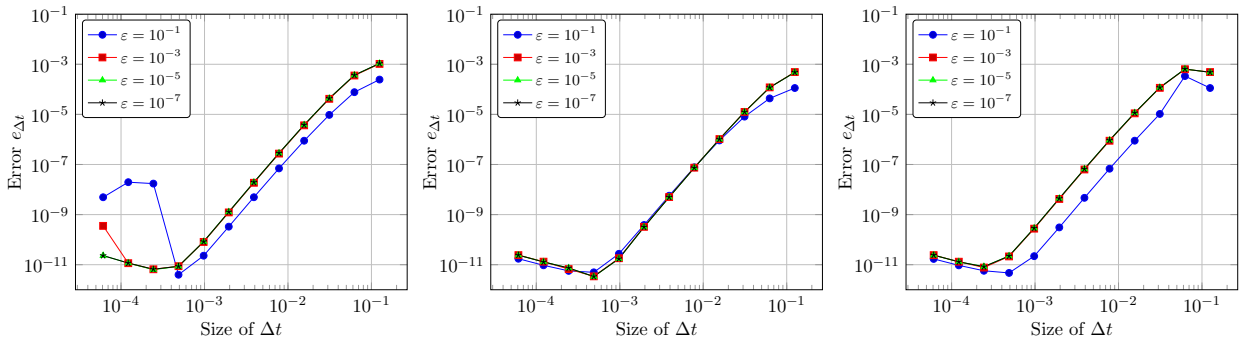

Figure 3: Convergence results for van der Pol equation for different values of ε, using the BDF 4 scheme coupled with the standard splitting (left), the RS-IMEX splitting with analytical $w_{(0)}$ (middle) and the RS-IMEX splitting with approximate $w_{(0)}^{app}$ (right). As an error measure, we choose $e_{\Delta t} := \|w(T_{end}) - w_{\Delta t}^N\|_2$ for $T_{end} = 0.5$ and N such that $T_{end} = N\Delta t$.

5. IMEX RK

As for BDF schemes, it is also possible to couple implicit and explicit Runge-Kutta methods with the splitting defined in Def. 3. In the following, we introduce the class of IMEX Runge-Kutta methods. (For a thorough discussion, we refer to [39] and the references therein.)

Definition 5 (IMEX Runge-Kutta scheme)**.** *For every $t^{n+1} = t^n + \Delta t$ do the following:*

1. *(Stages) For $i = 1, \dots, s$ solve*

$$w_{\Delta t,i} = w^n_{\Delta t} + \Delta t \sum_{j=1}^{i} \widetilde{A}_{i,j} k_j + \Delta t \sum_{j=1}^{i-1} \widehat{A}_{i,j} l_j,$$

   with $k_i = \widetilde{f}(w_{\Delta t,i}, t^n + \widetilde{c}_i \Delta t)$ and $l_i = \widehat{f}(w_{\Delta t,i}, t^n + \widehat{c}_i \Delta t)$. (Note that here, the dependence of $\widetilde{f}$ and $\widehat{f}$ on t is crucial, see also Rem. 2, which is why we make it explicit.)

2. *(Update) Finally evaluate*

$$w^{n+1}_{\Delta t} = w^n_{\Delta t} + \Delta t \sum_{j=1}^{s} \widetilde{b}_j k_j + \Delta t \sum_{j=1}^{s} \widehat{b}_j l_j.$$

The coefficients of the IMEX RK method are given by two Butcher tableaux, the one with overhats referring to the explicit, the other to the implicit method.

Remark 7. *We only consider IMEX Runge-Kutta schemes of* diagonally-implicit type*, i.e., our implicit Butcher matrix is a lower triangular matrix. Order conditions for such type of schemes can be found in [39].*

Remark 8. *Example tableaux can be found in A.*

In this work, we consider a variety of IMEX Runge-Kutta schemes. We consider both type A [40] and type CK [30] schemes:

Definition 6. *An IMEX Runge-Kutta scheme is*

- *of type A, if the matrix $\widetilde{A}$ is invertible.*
- *of type CK, if the matrix $\widetilde{A}$ is given by $\begin{pmatrix} 0 & 0 \\ \widetilde{\alpha} & \widetilde{\mathfrak{A}} \end{pmatrix}$ for $\widetilde{\alpha} \in \mathbb{R}^{s-1}$ and with $\widetilde{\mathfrak{A}} \in \mathbb{R}^{(s-1)\times(s-1)}$ invertible.*

See also Tbl. 1 for an overview.

$\widehat{c}$	$\widehat{A}$
	$\widehat{b}^T$

$\widetilde{c}$	$\widetilde{A}$
	$\widetilde{b}^T$

0	0	0
$\widehat{\mathfrak{c}}$	$\widehat{\alpha}$	$\widehat{\mathfrak{A}}$
	$\widehat{\beta}$	$\widehat{\mathfrak{b}}^T$

0	0	0
$\widetilde{\mathfrak{c}}$	$\widetilde{\alpha}$	$\widetilde{\mathfrak{A}}$
	$\widetilde{\beta}$	$\widetilde{\mathfrak{b}}^T$

Table 1: IMEX RK method of type A (left) and type CK (right). The left tableau of one pair refers to the explicit part.

The proof that our newly developed splitting is also AP in this context is valid for both type A and type CK schemes that are in addition both *globally stiffly accurate* and of diagonally-implicit type.

Definition 7 (Globally Stiffly Accurate)**.** *An IMEX RK method is called* globally stiffly accurate *(GSA) if $\widetilde{A}_{s,j} = \widetilde{b}_j$ and $\widehat{A}_{s,j} = \widehat{b}_j$ for $j = 1, \dots, s$.*

Corollary 1. *For an IMEX RK scheme that is GSA, the update $w^{n+1}_{\Delta t}$ is identical to the last stage, i.e., $w^{n+1}_{\Delta t} = w_{\Delta t,s}$.*

5.1. Toward the AP property: Example

To motivate the proof of the AP property, it is insightful to consider a specific IMEX RK method. More precisely, we consider a three-stage, second-order method taken from [3], named ARS-222. The corresponding Butcher tableaux are given in Tbl. 2. Coupling it to the RS IMEX splitting (see Def. 3 or eq. (8)), one

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \gamma & \gamma & 0 & 0 \\ 1 & \delta & 1\text{-}\delta & 0 \\ \hline & \delta & 1\text{-}\delta & 0 \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \gamma & 0 & \gamma & 0 \\ 1 & 0 & 1-\gamma & \gamma \\ \hline & 0 & 1-\gamma & \gamma \end{array}$$

Table 2: ARS-222 IMEX RK scheme with $\gamma = \frac{2-\sqrt{2}}{2} \approx 0.293$ and $\delta = 1 - \frac{1}{2\gamma} \approx -0.707$. Left: 'Explicit' tableau, right: 'implicit' tableau.

can explicitly write it as

$$\begin{aligned}
y_{\Delta t,1} &= y^n_{\Delta t} \\
z_{\Delta t,1} &= z^n_{\Delta t} \\
y_{\Delta t,2} &= y^n_{\Delta t} + \gamma \Delta t z_{\Delta t,2} \\
z_{\Delta t,2} &= z^n_{\Delta t} + \frac{\gamma \Delta t}{\varepsilon} \widetilde{g}(y_{\Delta t,2}, z_{\Delta t,2}) + \frac{\gamma \Delta t}{\varepsilon} \widehat{g}(y_{\Delta t,1}, z_{\Delta t,1}) \\
y^{n+1}_{\Delta t} \equiv y_{\Delta t,3} &= y^n_{\Delta t} + (1-\gamma)\Delta t z_{\Delta t,2} + \gamma \Delta t z_{\Delta t,3} \\
z^{n+1}_{\Delta t} \equiv z_{\Delta t,3} &= z^n_{\Delta t} + \frac{(1-\gamma)\Delta t}{\varepsilon} \widetilde{g}(y_{\Delta t,2}, z_{\Delta t,2}) + \frac{\gamma \Delta t}{\varepsilon} \widetilde{g}(y_{\Delta t,3}, z_{\Delta t,3}) + \frac{\delta \Delta t}{\varepsilon} \widehat{g}(y_{\Delta t,1}, z_{\Delta t,1}) + \frac{(1-\delta)\Delta t}{\varepsilon} \widehat{g}(y_{\Delta t,2}, z_{\Delta t,2})
\end{aligned}$$

Remark 9. *Obviously this scheme is of type CK and it is GSA. Furthermore, it fulfills the condition that* $\widehat{c} = \widetilde{c}$ *and may also be classified as a type ARS scheme [3], since* $\widetilde{\alpha} = 0$.

The first part of an AP proof (see Fig. 1) is to derive the limiting scheme, i.e., to express $w^n_{\Delta t}$ as $w^n_{\Delta t,(0)} + \mathcal{O}(\varepsilon)$ and find an expression for $w^{n+1}_{\Delta t,(0)}$.

Remark 10. *Subscript indices without brackets, such as* 1*, refer to stages of the Runge-Kutta method. Subscript indices with brackets, such as* (0)*, refer to terms of the asymptotic expansion.*

In a straightforward way, one can deduce the following lemma:

Lemma 1. *The limit ARS-222 scheme (coupled to the RS-IMEX splitting) is given by*

$$\begin{aligned}
y_{\Delta t,1,(0)} &= y^n_{\Delta t,(0)} \\
z_{\Delta t,1,(0)} &= z^n_{\Delta t,(0)} \\
y_{\Delta t,2,(0)} &= y^n_{\Delta t,(0)} + \gamma \Delta t z_{\Delta t,2,(0)} \\
0 &= \widetilde{g}(y_{\Delta t,2,(0)}, z_{\Delta t,2,(0)}) + \widehat{g}(y_{\Delta t,1,(0)}, z_{\Delta t,1,(0)}) \\
y_{\Delta t,3,(0)} &= y^n_{\Delta t,(0)} + (1-\gamma)\Delta t z_{\Delta t,2,(0)} + \gamma \Delta t z_{\Delta t,3,(0)} \\
0 &= (1-\gamma)\widetilde{g}(y_{\Delta t,2,(0)}, z_{\Delta t,2,(0)}) + \gamma \widetilde{g}(y_{\Delta t,3,(0)}, z_{\Delta t,3,(0)}) + \delta \widehat{g}(y_{\Delta t,1,(0)}, z_{\Delta t,1,(0)}) + (1-\delta)\widehat{g}(y_{\Delta t,2,(0)}, z_{\Delta t,2,(0)}) \\
y^{n+1}_{\Delta t,(0)} &= y_{\Delta t,3,(0)}, \qquad z^{n+1}_{\Delta t,(0)} = z_{\Delta t,3,(0)}.
\end{aligned}$$

The final thing one has to show is that the scheme given in La. 1 is a consistent discretization of (5). (Again, see Fig. 1.) We proceed by showing that the individual stages are consistent. Due to the GSA property, it then follows directly that the overall algorithm is a consistent discretization of the limit DAE (5). As a reminder, Δw denotes $w - w_{(0)}$ and similarly for other quantities. Since we are using exact initial values (after all this is a consistency analysis) there holds $\widehat{g}(y_{\Delta t,1}, z_{\Delta t,1}) = 0$.

Lemma 2. *The limit ARS-222 scheme (coupled to the RS-IMEX splitting) is a consistent approximation of the DAE (5).*

Corollary 2. *La. 2 implies that the ARS-222 scheme with splitting as given in Def. 3 is* asymptotic preserving *(AP).*

Proof. (Of La. 2)

Starting with the second stage (since first stage is equal to the initial conditions) of z, we insert the RS-IMEX splitting, see Def. 3. To simplify notation, we abbreviate $y_{(0)}(t^n + \Delta t\widetilde{c}_i) =: y_{i,(0)}$, note that $\widetilde{c} = \widehat{c}$ in this case. With this being said, one can equivalently rewrite the second stage as

$$0 = g'(w_{2,(0)})\Delta w_{\Delta t,2,(0)} \equiv \partial_1 g(y_{2,(0)}, z_{2,(0)})\Delta y_{\Delta t,2,(0)} + \partial_2 g(y_{2,(0)}, z_{2,(0)})\Delta z_{\Delta t,2,(0)}$$

and, because $\partial_2 g \neq 0$ (see Sec. 1), one can conclude that

$$\Delta z_{\Delta t,2,(0)} = \mathcal{O}(1)\Delta y_{\Delta t,2,(0)}. \tag{12}$$

The expression for $y_{\Delta t,2,(0)}$ can be augmented by $z_{2,(0)} - z_{2,(0)}$ to yield

$$y_{\Delta t,2,(0)} = y^n_{\Delta t,(0)} + \gamma\Delta t\left(z_{2,(0)} + \Delta z_{\Delta t,2,(0)}\right).$$

Subtracting $y_{2,(0)}$ on both sides and inserting the expression (12) for $z_{\Delta t,2,(0)}$ yields

$$\Delta y_{\Delta t,2,(0)} = y^n_{\Delta t,(0)} + \gamma\Delta t z_{2,(0)} - y_{2,(0)} - \mathcal{O}(\Delta t)\Delta y_{\Delta t,2,(0)},$$

which is equivalent to

$$\mathcal{O}(1)\Delta y_{\Delta t,2,(0)} = y^n_{\Delta t,(0)} + \gamma\Delta t z_{2,(0)} - y_{2,(0)}. \tag{13}$$

The right-hand side of (13) is the consistency error of the implicit Euler discretization of $y_{(0)} = z'_{(0)}$. Therefore, this term is $\mathcal{O}(\Delta t^2)$, and one can conclude that both $\Delta z_{\Delta t,2,(0)}$ and $\Delta y_{\Delta t,2,(0)}$, i.e., $\Delta w_{\Delta t,2,(0)}$, are also $\mathcal{O}(\Delta t^2)$. Before starting to show consistency of the third step, let us note that, because $\Delta w_{\Delta t,2,(0)} = \mathcal{O}(\Delta t^2)$, there holds

$$g(w_{\Delta t,2,(0)}) - g(w_{2,(0)}) - g'(w_{2,(0)})\Delta w_{\Delta t,2,(0)} = \mathcal{O}(\Delta t^4).$$

Starting from the second algebraic equation of the limit ARS-222 scheme, there holds

$$\begin{aligned} 0 &= (1-\gamma)g'(w_{2,(0)})\Delta w_{\Delta t,2,(0)} + \gamma g'(w_{3,(0)})\Delta w_{\Delta t,3,(0)} + (1-\delta)\left(g(w_{\Delta t,2,(0)}) - g(w_{2,(0)}) - g'(w_{2,(0)})\Delta w_{\Delta t,2,(0)}\right) \\ &= (1-\delta)g'(w_{3,(0)})\Delta w_{\Delta t,3,(0)} + \mathcal{O}(\Delta t^2). \end{aligned}$$

Similarly to before, one can conclude that

$$\Delta z_{\Delta t,3,(0)} = \mathcal{O}(1)\Delta y_{\Delta t,3,(0)} + \mathcal{O}(\Delta t^2). \tag{14}$$

One can rewrite the second stage for y as

$$y_{\Delta t,3,(0)} = y^n_{\Delta t,(0)} + (1-\gamma)\Delta t(z_{2,(0)} + \Delta z_{\Delta t,2,(0)}) + \gamma\Delta t(z_{3,(0)} + \Delta z_{\Delta t,3,(0)}).$$

Collecting $\Delta z_{\Delta t,i,(0)}$ terms and substituting them yields

$$y_{\Delta t,3,(0)} = y^n_{\Delta t,(0)} + \gamma \Delta t z_{3,(0)} + (1-\gamma)\Delta t z_{2,(0)} + \mathcal{O}(\Delta t)\Delta y_{\Delta t,3,(0)} + \mathcal{O}(\Delta t^2). \tag{15}$$

Again, the first three terms of the right-hand side of (15) form a discretization of $y'_{(0)} = z_{(0)}$, and therefore,

$$\Delta y_{\Delta t,3,(0)} = \mathcal{O}(\Delta t^2),$$

which, because of (14) implies that also $\Delta z_{\Delta t,3,(0)} = \mathcal{O}(\Delta t^2)$. This concludes the proof that the limit ARS-222 scheme is consistent with the DAE (5). □

The ideas presented in this short section are rather general, as they also apply for the more complex case of type A and type CK methods. In the sequel, we extend the proof to methods of type CK, provided they are GSA, fulfill $\widehat{c} = \widetilde{c}$ and are of diagonally-implicit type.

5.2. The AP property for type CK methods

Before actually starting to prove that type CK methods are AP, we have to introduce some notation first. In general, an IMEX RK method relies on s stages (see also Def. 5), namely $w_{\Delta t,1,(0)}, \dots, w_{\Delta t,s,(0)}$. A type CK method is designed in such a way that the first stage is equal to $w^n_{\Delta t,(0)}$. The following notation is rather standard in the analysis of IMEX RK methods:

Definition 8 (Notation)**.** *The vectors $\vec{y}_{\Delta t,(0)} \in \mathbb{R}^{s-1}$ and $\vec{z}_{\Delta t,(0)} \in \mathbb{R}^{s-1}$ denote*

$$\vec{y}_{\Delta t,(0)} := \begin{pmatrix} y_{\Delta t,2,(0)} \\ \vdots \\ y_{\Delta t,s,(0)} \end{pmatrix} \quad \text{and} \quad \vec{z}_{\Delta t,(0)} := \begin{pmatrix} z_{\Delta t,2,(0)} \\ \vdots \\ z_{\Delta t,s,(0)} \end{pmatrix}.$$

Please note that they start at stage two. Similar notation is applied for vectors $\vec{y}_{(0)}$ and $\vec{z}_{(0)}$, where, as seen in the last paragraph, $w_{i,(0)} := w_{(0)}(t^n + \Delta t \widetilde{c}_i)$. In addition, we define

$$\vec{y}^n_{\Delta t,(0)} := \begin{pmatrix} y^n_{\Delta t,(0)} \\ \vdots \\ y^n_{\Delta t,(0)} \end{pmatrix} \in \mathbb{R}^{s-1}.$$

It is natural and convenient to apply functions $g := \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ to such a vector, we define

$$g(\vec{y}_{\Delta t,(0)}, \vec{z}_{\Delta t,(0)}) := \begin{pmatrix} g(y_{\Delta t,2,(0)}, z_{\Delta t,2,(0)}) \\ \vdots \\ g(y_{\Delta t,s,(0)}, z_{\Delta t,s,(0)}) \end{pmatrix} \in \mathbb{R}^{s-1}.$$

As a last rather standard notation, we define $\operatorname{diag}(\vec{v})$ *of some vector $\vec{v}$ as the diagonal matrix with diagonal entries $\vec{v}_i$.*

With this being said, we can reformulate La. 1 for more general CK methods in the following way. Please note that we directly exploit the fact $w^{n+1}_{\Delta t,(0)} = w^{n+1}_{\Delta t,s,(0)}$, see also Cor. 1.

Lemma 3. *Let $\widetilde{c}_i = \widehat{c}_i$, and let the IMEX RK method (coupled to the RS-IMEX splitting) be of type CK , globally stiffly accurate and of diagonally-implicit type. Then, the limit scheme is given by*

$$y_{\Delta t,1,(0)} = y^n_{\Delta t,(0)}, \qquad z_{\Delta t,1,(0)} = z^n_{\Delta t,(0)} \tag{16}$$

$$\vec{y}_{\Delta t,(0)} = \vec{y}^n_{\Delta t,(0)} + \Delta t \widetilde{\mathfrak{A}} \vec{z}_{\Delta t,(0)} + \Delta t \widetilde{\alpha} z^n_{\Delta t,(0)} \tag{17}$$

$$0 = \widetilde{\mathfrak{A}} \widetilde{g}(\vec{y}_{\Delta t,(0)}, \vec{z}_{\Delta t,(0)}) + \widehat{\mathfrak{A}} \widehat{g}(\vec{y}_{\Delta t,(0)}, \vec{z}_{\Delta t,(0)}) + \widetilde{\alpha} \widetilde{g}(y^n_{\Delta t,(0)}, z^n_{\Delta t,(0)}) + \widehat{\alpha} \widehat{g}(y^n_{\Delta t,(0)}, z^n_{\Delta t,(0)}) \tag{18}$$

$$y^{n+1}_{\Delta t,(0)} = y_{\Delta t,s,(0)}, \quad z^{n+1}_{\Delta t,(0)} = z_{\Delta t,s,(0)}. \tag{19}$$

It is the goal of this section to prove that IMEX RK methods of type as described in La. 3 are AP. This is a formal consistency analysis, showing that the limit scheme is consistent with the DAE (5). As is classical in the theory of ODE integrators [23], one can set initial data to the exact solution of the ODE, and so one directly obtains that

$$\widetilde{g}(y^n_{\Delta t,(0)}, z^n_{\Delta t,(0)}) = 0 \qquad \text{and} \qquad \widehat{g}(y^n_{\Delta t,(0)}, z^n_{\Delta t,(0)}) = 0. \tag{20}$$

Furthermore, g was assumed to be smooth, and $\partial_2 g(\cdot) \neq 0$, so

$$\frac{\partial_1 g(w_{(0)})}{\partial_2 g(w_{(0)})} = \mathcal{O}(1) \tag{21}$$

holds for every reference solution $w_{(0)}$.

Definition 9 (Stage order)**.** *During the proof, we need the minimum stage order of both y and z, and so we define $m > 1$ to be such that*

$$\Delta\vec{y}_{\Delta t,(0)} = \mathcal{O}(\Delta t^m) \qquad \text{and} \qquad \Delta\vec{z}_{\Delta t,(0)} = \mathcal{O}(\Delta t^m). \tag{22}$$

Remark 11. *Def. 9 seems like circular reasoning on first sight, because it is actually that we want to show $m > 1$. However, the proceeding is as follows: As we are only using Runge-Kutta methods of diagonally-implicit type, we can consider the stages subsequently, starting by stage two, showing that it is consistent of order m, and then proceed to stage three and so on. This is similar to what has been seen in the example section 5.1, and is a consequence of the fact that with $\widetilde{\mathfrak{A}}$ being lower triangular, and $\widehat{\mathfrak{A}}$ being even strictly lower triangular, $\widetilde{\mathfrak{A}}^{-1}$ is also lower triangular, and $\widetilde{\mathfrak{A}}^{-1}\widehat{\mathfrak{A}}$ is even strictly lower triangular.*

Corollary 3. *Given that (22) holds, it is straightforward to show that*

$$g(\vec{y}_{\Delta t,(0)}, \vec{z}_{\Delta t,(0)}) - \operatorname{diag}\left(\partial_1 g(\vec{y}_{(0)}, \vec{z}_{(0)})\right) \Delta\vec{y}_{\Delta t,(0)} - \operatorname{diag}\left(\partial_2 g(\vec{y}_{(0)}, \vec{z}_{(0)})\right) \Delta\vec{z}_{\Delta t,(0)} = \mathcal{O}(\Delta t^{2\cdot m})$$

and

$$\operatorname{diag}\left(\partial_1 g(\vec{y}_{(0)}, \vec{z}_{(0)})\right) \Delta\vec{y}_{\Delta t,(0)} + \operatorname{diag}\left(\partial_2 g(\vec{y}_{(0)}, \vec{z}_{(0)})\right) \Delta\vec{z}_{\Delta t,(0)} = \mathcal{O}(\Delta t^m).$$

Theorem 2. *Let $\widetilde{c}_i = \widehat{c}_i$, and let the IMEX RK method (coupled to the RS-IMEX splitting) be of type CK, globally stiffly accurate and of diagonally-implicit type. Then, the limit scheme of a type CK GSA method is a consistent approximation of the DAE (5)*

Corollary 4. *With the conditions of Thm. 2, the CK Runge-Kutta scheme is* asymptotic preserving *(AP).*

Proof. (Of Thm. 2) We start with the algebraic equation (18), simplify it with the help of Cor. 3 and eqs. (20) and (8), and rearrange terms such that we get an expression for $\Delta\vec{z}_{\Delta t,(0)}$. So

$$\begin{aligned}
0 =& \widetilde{\mathfrak{A}} \left\{ \operatorname{diag}\left(\partial_1 g(\vec{y}_{(0)}, \vec{z}_{(0)})\right) \Delta\vec{y}_{\Delta t,(0)} + \operatorname{diag}\left(\partial_2 g(\vec{y}_{(0)}, \vec{z}_{(0)})\right) \Delta\vec{z}_{\Delta t,(0)} \right\} \\
&+ \widehat{\mathfrak{A}} \left\{ g(\vec{y}_{\Delta t,(0)}, \vec{z}_{\Delta t,(0)}) - \operatorname{diag}\left(\partial_1 g(\vec{y}_{(0)}, \vec{z}_{(0)})\right) \Delta\vec{y}_{\Delta t,(0)} - \operatorname{diag}\left(\partial_2 g(\vec{y}_{(0)}, \vec{z}_{(0)})\right) \Delta\vec{z}_{\Delta t,(0)} \right\} \\
\Rightarrow \Delta\vec{z}_{\Delta t,(0)} =& \mathcal{O}(1)\Delta\vec{y}_{\Delta t,(0)} + \mathcal{O}(\Delta t^m).
\end{aligned} \tag{23}$$

Please note that the error in a specific stage only relies on the error in previous stages, see also Rem. 11.

Continuing with the internal stages of $\vec{y}_{\Delta t,(0)}$, we can use Cor. 3, the previous result and rearrange the terms to obtain

$$\begin{aligned}
\vec{y}_{\Delta t,(0)} =& \vec{y}^n_{\Delta t,(0)} + \Delta t \widetilde{\mathfrak{A}} \left(\vec{z}_{\Delta t,(0)}\right) + \Delta t \widetilde{\alpha} \left(\vec{z}^n_{(0)}\right) \\
=& \vec{y}^n_{\Delta t,(0)} + \Delta t \left(\widetilde{\mathfrak{A}}\vec{z}_{(0)} + \widetilde{\alpha}\vec{z}^n_{(0)}\right) + \Delta t \widetilde{\mathfrak{A}} \left(\vec{z}_{\Delta t,(0)} - \vec{z}_{(0)}\right) \\
\Rightarrow \mathcal{O}(1)\Delta\vec{y}_{\Delta t,(0)} =& \vec{y}^n_{\Delta t,(0)} + \Delta t \left(\widetilde{\mathfrak{A}}\vec{z}_{(0)} + \widetilde{\alpha}\vec{z}^n_{(0)}\right) - \vec{y}_{(0)} + \mathcal{O}(\Delta t^{m+1})
\end{aligned}$$

As in Sec. 5.1, one can observe that the first terms on the right hand side correspond to a discretization error. More precisely, this is the discretization error of the implicit part of the Runge-Kutta method for $y' = z_{(0)}$. Therefore, and due to the definition of minimum stage order, see Def. 9, we observe that this term is $\mathcal{O}(\Delta t^m)$. This yields

$$\Delta \vec{y}_{\Delta t,(0)} = \mathcal{O}(\Delta t^m).$$

Using again eq. (23) we also get an error estimate for z,

$$\Delta \vec{z}_{\Delta t,(0)} = \mathcal{O}(\Delta t^m)$$

Every internal step is at least as good as an Euler step, and therefore, the minimal stage order is $m \geq 2$. Thus, the error per stage is at least $\mathcal{O}(\Delta t^2)$. Because the method is GSA, it follows that

$$\Delta w^{n+1}_{\Delta t,(0)} = \mathcal{O}(\Delta t^2).$$

This concludes the proof. □

There are a couple of remarks in order here:

Remark 12. *1. The proof is not quantitative in the sense that we only prove consistency. Order reduction as for example observed in [7] can - and in fact will - still appear.*

2. The condition $\widehat{c}_i = \widetilde{c}_i$ is not crucial for the consistency analysis, it just heavily simplifies the proof. In particular, for type A schemes, where $\widehat{c}_i \neq \widetilde{c}_i$ in general, Thm. 2 is still valid.

3. What is crucial, though, is the GSA condition, see also Def. 7. It is shown in the numerical results section that methods violating this property can give rise to unstable schemes in the context of an RS-IMEX splitting.

Remark 13. *The proof shows that, at least in this context, it is important to use the correct reference solution, i.e., the limit as $\varepsilon \to 0$. The proof does not apply to situations where $w_{(0)}$ is substituted by any other function.*

In the sequel, we will show numerical findings for a variety of Runge-Kutta methods.

5.3. Numerical results

In this section we show numerical results for van der Pol equation with different IMEX RK discretizations, and presenting interesting findings. Again, in all the results, we compare standard splitting versus the new RS-IMEX splitting versus a splitting that uses an approximate version of $w_{(0)}$ (RSApp, for short, see Rem. 15 below). We would like to point out already at this instant that there is hardly any difference between RS-IMEX and RSApp, but sometimes there is a tremendous difference between the standard splitting and the RS-IMEX.

Error has been computed as the two-norm of the difference to the solution at end time $T_{end} = 0.5$, i.e.,

$$e_{\Delta t} := \|w(T_{end}) - w^N_{\Delta t}\|_2,$$

where N is such that $N\Delta t = T_{end}$. (This is the same error measure as in Sec. 4 for BDF methods.) We usually consider various ε, ranging from $\varepsilon = 10^{-1}$ to 10^{-7}, to show that the algorithms perform well for both relatively large and small values of ε.

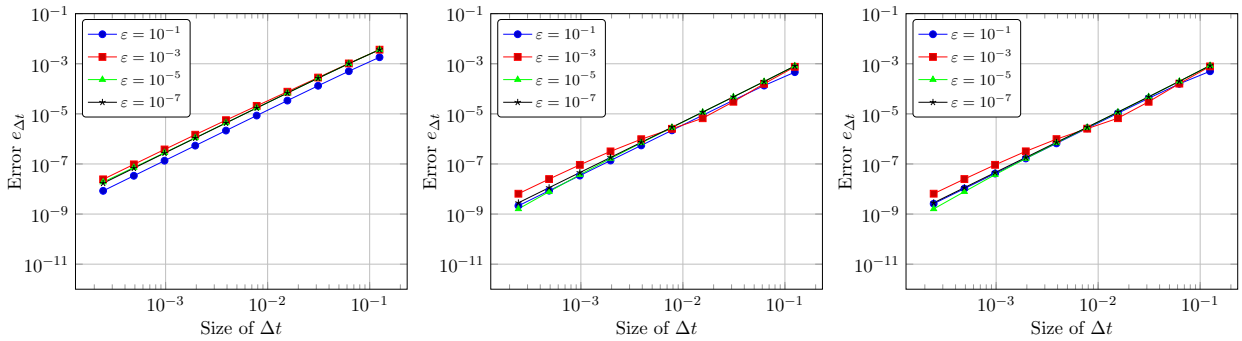

Figure 4: Convergence results for van der Pol equation for different values of ε, using the ARS-222 IMEX RK scheme coupled with the standard splitting (left), the RS-IMEX splitting with analytical $w_{(0)}$ (middle) and the RS-IMEX splitting with approximate $w_{(0)}^{app}$ (right). As an error measure, we choose $e_{\Delta t} := \|w(T_{end}) - w_{\Delta t}^N\|_2$ for $T_{end} = 0.5$ and N such that $T_{end} = N\Delta t$.

Remark 14. *Example tableaux used in the numerical results and an overview can be found in A. We only use* singly diagonally implicit schemes *(SDIRK) or* explicit singly diagonally implicit schemes *(ESDIRK) schemes for the implicit part. The first one means that $\widetilde{A}_{ii} = \widetilde{A}_{jj}$, while the latter one basically means the same, with the exception that the first stage is treated explicitly, i.e., $\widetilde{A}_{11} = 0$. (See also Def. 6).*

As already pointed out in the introduction, solving singularly perturbed equations can actually lead to the phenomenon of order reduction. This means that for common schemes and $\Delta t \gg \varepsilon$, the formal convergence order is not achieved. In the numerical results section, we compare different IMEX RK schemes for standard and RS-IMEX splitting. Again, we use van der Pol equation with both an 'exact' reference solution $w_{(0)}$ and an approximate reference solution $w_{(0)}^{app}$. The computation of this approximation is explained in the sequel.

Remark 15 (Approximation of $w_{(0)}^{app}$)**.** *Using the IMEX scheme as in Def. 5 with the standard splitting given in (6), allows to take the formal limit as $\varepsilon \to 0$. This yields a Runge-Kutta discretization for the differential-algebraic equation (5). In our numerical experiments, we compute the approximation $w_{(0)}^{app}$ using the same Runge-Kutta integrator as in the example considered, with exactly the same time steps. The stages of the computation of $w_{(0)}^{app}$ are saved and used in the computation of k_j and l_j needed in Def. 5. Note that it does not make sense to use the RS-IMEX splitting for the integration of this DAE (5), because its use necessitates the knowledge of $w_{(0)}$.*

ARS-222 and ARS-443 with RS-IMEX: Work as expected Our point of departure are the ARS-222 and ARS-443 schemes given in [3], see also the augmented Butcher tableaux in Tbls. 4 and 6. (443 refers to 4 stages implicit, 4 stages explicit, and third order convergence; similarly for 222) These schemes are GSA in the sense of Def. 7, and therefore, Thm. 2 applies and the schemes, with RS-IMEX (7) splitting, are AP. In Figs. 4 and 5, we plot results for the standard splitting (6), the new RS-IMEX splitting and its approximate version for various ε. It can be observed that the three plots do not show much difference, except that for ARS-222, RS-IMEX seems to have a slightly lower error. In particular, the order reduction (the non-uniform convergence as ε gets increasingly small) can be observed for ARS-443 for both the standard splitting and the RS-IMEX splitting. So far, this is what one would actually expect from our experiences with the IMEX BDF schemes, where the choice of splittings did hardly have an influence.

BHR-553 with RS-IMEX: Increased stability without update step The BHR-553 (5 stages, third order) scheme has been explicitly designed in [7] to circumvent order reduction. See Tbl. 8 for the corresponding

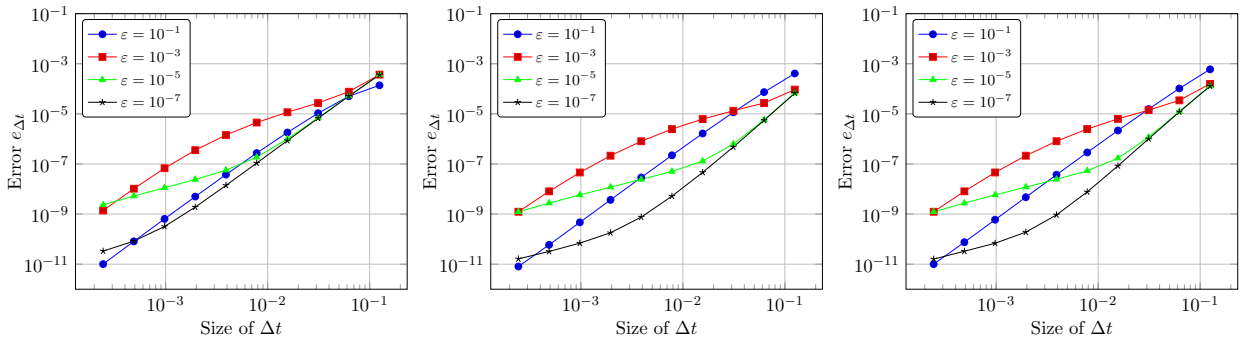

Figure 5: Convergence results for van der Pol equation for different values of ε, using the ARS-443 IMEX RK scheme coupled with the standard splitting (left), the RS-IMEX splitting with analytical $w_{(0)}$ (middle) and the RS-IMEX splitting with approximate $w_{(0)}^{app}$ (right). As an error measure, we choose $e_{\Delta t} := \|w(T_{end}) - w_{\Delta t}^N\|_2$ for $T_{end} = 0.5$ and N such that $T_{end} = N\Delta t$.

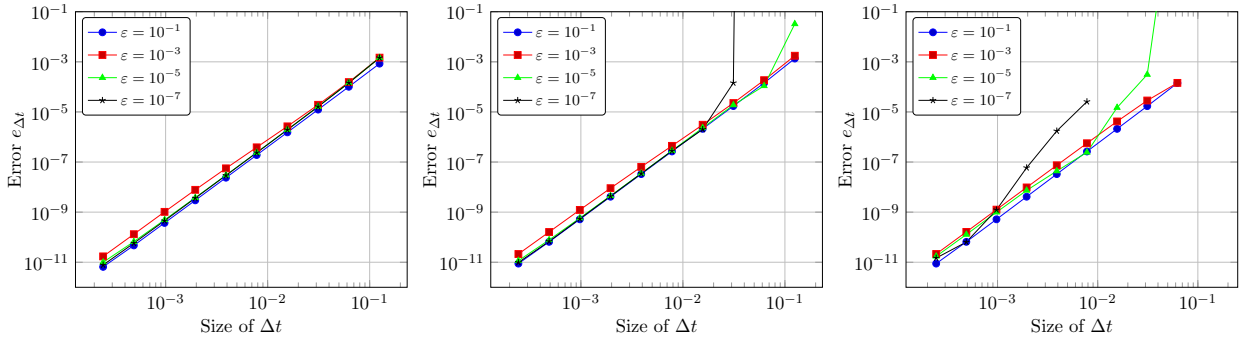

Figure 6: Convergence results for van der Pol equation for different values of ε, using the BHR-553 IMEX RK scheme coupled with the standard splitting (left), the RS-IMEX splitting with analytical $w_{(0)}$ (middle) and the RS-IMEX splitting with approximate $w_{(0)}^{app}$ (right). As an error measure, we choose $e_{\Delta t} := \|w(T_{end}) - w_{\Delta t}^N\|_2$ for $T_{end} = 0.5$ and N such that $T_{end} = N\Delta t$.

Butcher tableau. Note however that the scheme is not GSA (see Def. 7) as can be seen from the explicit tableau. Consequently, it is not covered by our AP analysis in Thm. 2.

In the numerical experiments, see Fig. 6, we can observe uniform third order convergence for both standard and RS-IMEX splitting. However, it seems that for extremely small values of ε, and 'large' values of Δt, the algorithm becomes unstable for the RS-IMEX, see Fig. 7. This might be due to either cancellation errors or due to the fact that the scheme is possibly not AP. Curiously, this effect does not show up if one chooses $w_{\Delta t}^{n+1}$ to be $w_{\Delta t,5}$, i.e., instead of computing an update, the last stage of the Runge-Kutta method is taken as new update, see Fig. 7. Additionally, third order is recovered, which is peculiar because the last stage only has a formal consistency of two.

For schemes that are GSA, where indeed $w_{\Delta t}^{n+1} = w_{\Delta t,s}$, there can occur some differences due to numerical roundoff errors. However, unlike in the case of BHR-553, they occur at small Δt.

BPR-353 and DPA-242 with RS-IMEX: Improved convergence The last paragraph is devoted to BPR-353 and DPA-242 schemes, presented in [8] and [17], respectively. Corresponding Butcher tableaux can be found in Tbl. 7 and Tbl. 5. Note that both schemes are GSA in the sense of Def. 7, and therefore Thm.

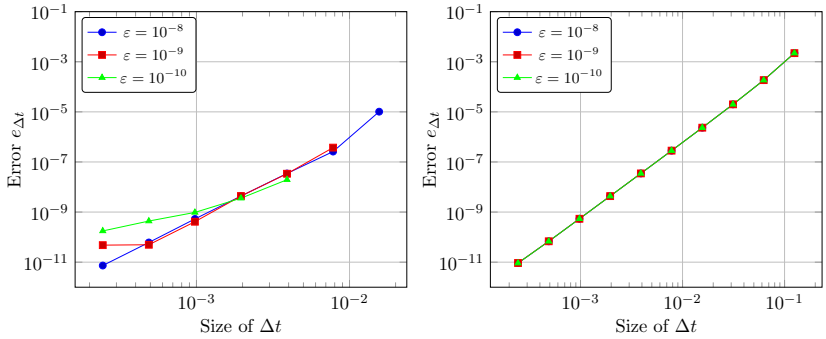

Figure 7: Convergence results for van der Pol equation for different values of ε, using the BHR-553 IMEX RK scheme coupled with the RS-IMEX splitting. Left: Full BHR-553 scheme which obviously exhibits instabilities (unplotted values are NaN). Right: Neglecting the update step, i.e., setting $w_{\Delta t}^{n+1} = w_{\Delta t,s}$. $w_{(0)}$ has been computed analytically. As an error measure, we choose $e_{\Delta t} := \|w(T_{end}) - w_{\Delta t}^N\|_2$ for $T_{end} = 0.5$ and N such that $T_{end} = N\Delta t$.

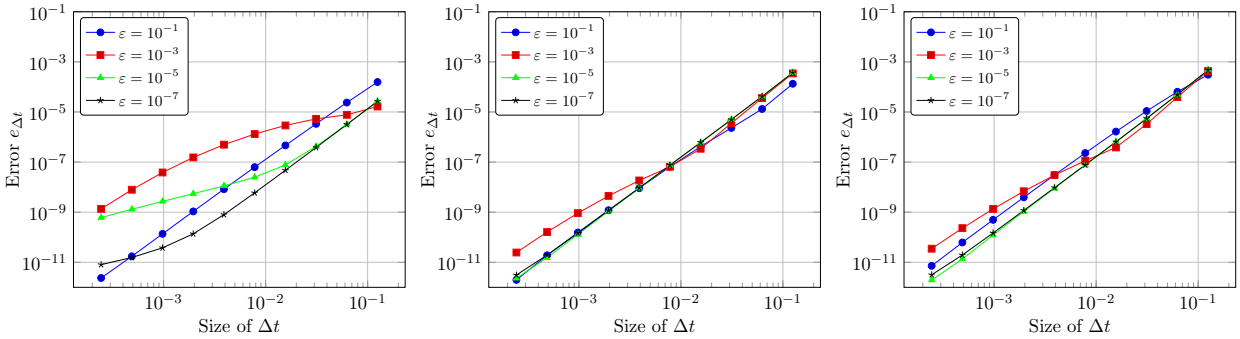

Figure 8: Convergence results for van der Pol equation for different values of ε, using the BPR-353 IMEX RK scheme coupled with the standard splitting (left), the RS-IMEX splitting with analytical $w_{(0)}$ (middle) and the RS-IMEX splitting with approximate $w_{(0)}^{app}$ (right). As an error measure, we choose $e_{\Delta t} := \|w(T_{end}) - w_{\Delta t}^N\|_2$ for $T_{end} = 0.5$ and N such that $T_{end} = N\Delta t$.

2 applies and the schemes are AP if coupled to the RS-IMEX splitting. Coupled to the standard splitting (6) on the other hand, it is well-known that these methods exhibit quite severe order loss as can be seen in Figs. 8 and 9 on the left sides. In particular, for the DPA scheme, this order loss is quite significant. Coupling both methods to the newly developed RS-IMEX splitting seems to yield uniform convergence in ε, see Figs. 8 and 9.

Remark 16. *We note that we did also test this on the Pareschi-Russo equation [39]*

$$y' = -z, \qquad z' = y + \frac{\sin(y) - z}{\varepsilon} \tag{24}$$

showing that the observed phenomenon is not a feature of van der Pol equation.

Up to now, we do not have a solid explanation for this effect, but we conjecture that it is because of the Taylor series approach employed in (7), that enforces the numerical solution to be close to $w_{(0)}$.

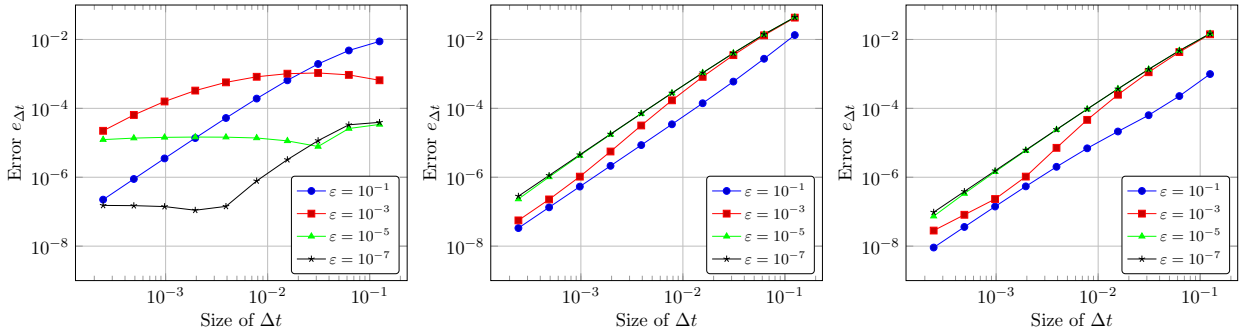

Figure 9: Convergence results for van der Pol equation for different values of ε, using the DPA-242 IMEX RK scheme coupled with the standard splitting (left), the RS-IMEX splitting with analytical $w_{(0)}$ (middle) and the RS-IMEX splitting with approximate $w_{(0)}^{app}$ (right). As an error measure, we choose $e_{\Delta t} := \|w(T_{end}) - w_{\Delta t}^N\|_2$ for $T_{end} = 0.5$ and N such that $T_{end} = N\Delta t$.

6. Conclusion and outlook

In this publication, we have developed a new splitting based on the reference solution and shown numerical comparison with the more established standard splitting. We have shown that for IMEX schemes of BDF type, the influence of the splitting is marginal. In contrast, for IMEX schemes of Runge-Kutta type, the splitting has indeed a broad influence. We have shown numerical results demonstrating that it is even possible to obtain faster-converging schemes. For both IMEX BDF and IMEX RK, we showed, under suitable conditions, that these schemes are AP.

Obviously, in this context, treating van der Pol equation is only an intermediate step. Our actual interest is in high-order approximation of singularly perturbed conservation laws. The next key step is therefore to apply the RS-IMEX splitting to the compressible Euler / Navier-Stokes equations. We anticipate that this will impose more severe problems - although preliminary results [45] already show a satisfactory behavior - as the discretization of the limit equation will most likely have a much higher influence on the quality and stability of an overall algorithm. Currently, this is work in progress.

Acknowledgements

The authors would like to thank Sebastian Noelle for fruitful discussions. Furthermore, we would like to thank the anonymous referee for suggestions that helped improving the paper substantially.

A. Butcher tableaux

In our numerical results section, we employ the IMEX RK methods shown in Tbl. 3. Note that except BHR-553, they are GSA (see Def. 7) and are therefore covered by Thm. 2. All schemes are of diagonally-implicit type.

For the sake of completeness, we show the augmented Butcher tableaux of the employed IMEX RK schemes. Left part of the tableaux (last example: top tableau) always refers to the explicit part, i.e., to $\widehat{A}$.

References

[1] J. D. Anderson. *Fundamentals of Aerodynamics*. McGraw-Hill New York, 3^{rd} edition, 2001.

Method	ARS-222	DPA-242	ARS-443	BPR-353	BHR-553
Type	CK	A	CK	CK	CK
GSA?	yes	yes	yes	yes	no
$\widetilde{c} = \widehat{c}$	yes	no	yes	yes	yes

Table 3: Employed IMEX RK methods.

0	0	0	0
γ	γ	0	0
1	δ	1-δ	0
	δ	1-δ	0

0	0	0	0
γ	0	γ	0
1	0	$1-\gamma$	γ
	0	$1-\gamma$	γ

Table 4: ARS-222 IMEX RK scheme with $\gamma = \frac{2-\sqrt{2}}{2} \approx 0.293$ and $\delta = 1 - \frac{1}{2\gamma} \approx -0.707$.

[2] U. M. Ascher and L. R. Petzold. *Computer methods for ordinary differential equations and differential-algebraic equations*, volume 61. Siam, 1998.

[3] U. M. Ascher, S. Ruuth, and R. Spiteri. Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations. *Applied Numerical Mathematics*, 25:151–167, 1997.

[4] U. M. Ascher, S. Ruuth, and B. Wetton. Implicit-Explicit methods for time-dependent partial differential equations. *SIAM Journal on Numerical Analysis*, 32:797–823, 1995.

[5] G. Bispen, K.R. Arun, M. Lukacova-Medvidova, and S. Noelle. IMEX large time step finite volume methods for low Froude number shallow water flows. *Communications in Computational Physics*, 2014. (in press).

[6] S. Boscarino. Error analysis of IMEX Runge-Kutta methods derived from differential-algebraic systems. *SIAM Journal on Numerical Analysis*, 45:1600–1621, 2007.

[7] S. Boscarino. On an accurate third order implicit-explicit Runge-Kutta method for stiff problems. *Applied Numerical Mathematics*, 59:1515–1528, 2009.

[8] S. Boscarino, L. Pareschi, and G. Russo. Implicit-explicit Runge–Kutta schemes for hyperbolic systems and kinetic equations in the diffusion limit. *SIAM Journal on Scientific Computing*, 35(1):A22–A51, 2013.

[9] B. Cockburn, S. Hou, and C.-W. Shu. The Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws IV: The multidimensional case. *Mathematics of Computation*, 54:545–581, 1990.

0	0	0	0	0
1/3	1/3	0	0	0
1	1	0	0	0
1	1/2	0	1/2	0
	1/2	0	1/2	0

1/2	1/2	0	0	0
2/3	1/6	1/2	0	0
1/2	-1/2	1/2	1/2	0
1	3/2	-3/2	1/2	1/2
	3/2	-3/2	1/2	1/2

Table 5: DPA-242 [17]

0	0	0	0	0	0
1/2	1/2	0	0	0	0
2/3	11/18	1/18	0	0	0
1/2	5/6	-5/6	1/2	0	0
1	1/4	7/4	3/4	-7/4	0
	1/4	7/4	3/4	-7/4	0

0	0	0	0	0	0
1/2	0	1/2	0	0	0
2/3	0	1/6	1/2	0	0
1/2	0	-1/2	1/2	1/2	0
1	0	3/2	-3/2	1/2	1/2
	0	3/2	-3/2	1/2	1/2

Table 6: ARS-443 [3]

0	0	0	0	0	0
1	1	0	0	0	0
2/3	4/9	2/9	0	0	0
1	1/4	0	3/4	0	0
1	1/4	0	3/4	0	0
	1/4	0	3/4	0	0

0	0	0	0	0	0
1	1/2	1/2	0	0	0
2/3	5/18	-1/9	1/2	0	0
1	1/2	0	0	1/2	0
1	1/4	0	3/4	-1/2	1/2
	1/4	0	3/4	-1/2	1/2

Table 7: BPR-353[8]

[10] B. Cockburn and S. Y. Lin. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws III: One dimensional systems. *Journal of Computational Physics*, 84:90–113, 1989.

[11] B. Cockburn and C.-W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws II: General framework. *Mathematics of Computation*, 52:411–435, 1988.

[12] B. Cockburn and C.-W. Shu. The Runge-Kutta local projection p^1-discontinuous Galerkin finite element method for scalar conservation laws. *RAIRO Mathematical modelling and numerical analysis*, 25:337–361, 1991.

[13] F. Cordier, P. Degond, and A. Kumbaro. An asymptotic-preserving all-speed scheme for the Euler and Navier-Stokes equations. *Journal of Computational Physics*, 231:5685–5704, 2012.

[14] M. Crouzeix. Une méthode multipas implicite-explicite pour l'approximation des équations d'évolution paraboliques. *Numerische Mathematik*, 35(3):257–276, 1980.

0	0	0	0	0	0
0.871733	0.871733	0	0	0	0
0.871733	0.435867	0.435867	0	0	0
2.34021	-0.800998	0	3.14121	0	0
1	0.356753	-0.19734	0.881949	-0.0413622	0
	0.412898	0	0.19734	-0.0461045	0.435867

0	0	0	0	0	0
0.871733	0.435867	0.435867	0	0	0
0.871733	0.435867	0	0.435867	0	0
2.34021	-0.0667587	0	1.9711	0.435867	0
1	0.412898	0	0.19734	-0.0461045	0.435867
	0.412898	0	0.19734	-0.0461045	0.435867

Table 8: BHR-553 [7]

[15] P. Degond, A. Lozinski, J. Narski, and C. Negulescu. An asymptotic-preserving method for highly anisotropic elliptic equations based on a micro-macro decomposition. *Journal of Computational Physics*, 231:2724–2740, 2012.

[16] P. Degond and M. Tang. All speed scheme for the low Mach number limit of the isentropic Euler equation. *Communications in Computational Physics*, 10:1–31, 2011.

[17] G. Dimarco and L. Pareschi. Asymptotic preserving implicit-explicit Runge–Kutta methods for nonlinear kinetic equations. *SIAM Journal on Numerical Analysis*, 51(2):1064–1087, 2013.

[18] A. J. Gadd. A split explicit integration scheme for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 104(441):569–582, 1978.

[19] F.X. Giraldo, M. Restelli, and M. Läuter. Semi-implicit formulations of the Navier-Stokes equations: Application to nonhydrostatic atmospheric modeling. *SIAM Journal on Scientific Computing*, 32(6):3394–3425, 2010.

[20] E. Godlewski and P.-A. Raviart. *Hyperbolic Systems of Conservation Laws*. Ellipses Paris, 1991.

[21] E. Godlewski and P.-A. Raviart. *Numerical Approximation of Hyperbolic Systems of Conservation Laws*. Springer New York, 1996.

[22] J. Haack, S. Jin, and J.-G. Liu. An all-speed asymptotic-preserving method for the isentropic Euler and Navier-Stokes equations. *Communications in Computational Physics*, 12:955–980, 2012.

[23] E. Hairer, S.P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I*. Springer Series in Computational Mathematics, 1987.

[24] E. Hairer, M. Roche, and C. Lubich. *The Numerical Solution of Differential-Algebraic Systems by Runge-Kutta Methods*, volume 1409 of *Lecture Notes in Mathematics*. Springer Berlin Heidelberg, 1989.

[25] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II*. Springer Series in Computational Mathematics, 1991.

[26] W. Hundsdorfer and S.-J. Ruuth. IMEX extensions of linear multistep methods with general monotonicity and boundedness properties. *Journal of Computational Physics*, 225(2):2016–2042, 2007.

[27] S. Jin. Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations. *SIAM Journal on Scientific Computing*, 21:441–454, 1999.

[28] S. Jin. Asymptotic preserving (AP) schemes for multiscale kinetic and hyperbolic equations: A review. *Rivista di Matematica della Universita Parma*, 3:177–216, 2012.

[29] S. Jin, L. Pareschi, and G. Toscani. Diffusive relaxation schemes for multiscale discrete-velocity kinetic equations. *SIAM Journal on Numerical Analysis*, 35:2405–2439, 1998.

[30] C. A. Kennedy and M. H. Carpenter. Additive Runge-Kutta schemes for convection-diffusion-reaction equations. *Applied Numerical Mathematics*, 44:139–181, 2003.

[31] S. Klainerman and A. Majda. Singular limits of quasilinear hyperbolic systems with large parameters and the incompressible limit of compressible fluids. *Communications on Pure and Applied Mathematics*, 34:481–524, 1981.

[32] R. Klein. Semi-implicit extension of a Godunov-type scheme based on low Mach number asymptotics I: One-dimensional flow. *Journal of Computational Physics*, 121:213–237, 1995.

[33] R. Klein, N. Botta, T. Schneider, C.D. Munz, S. Roller, A. Meister, L. Hoffmann, and T. Sonar. Asymptotic adaptive methods for multi-scale problems in fluid mechanics. *Journal of Engineering Mathematics*, 39(1):261–343, 2001.

[34] D. Kröner. *Numerical Schemes for Conservation Laws*. Wiley Teubner, 1997.

[35] R. J. LeVeque. *Numerical Methods for Conservation Laws*. Birkhäuser Basel, 1990.

[36] A. Müller, J. Behrens, F.X. Giraldo, and V. Wirth. Comparison between adaptive and uniform discontinuous Galerkin simulations in dry 2d bubble experiments. *Journal of Computational Physics*, 235:371–393, 2013.

[37] S. Noelle, G. Bispen, K.R. Arun, M. Lukacova-Medvidova, and C.-D. Munz. An asymptotic preserving all Mach number scheme for the Euler equations of gas dynamics. *SIAM Journal of Scientific Computing*, 2012.

[38] R. E. O'Malley. *Introduction to singular perturbations*. Academic Press, 1974.

[39] L. Pareschi and G. Russo. Implicit-explicit Runge-Kutta schemes for stiff systems of differential equations. *Recent trends in numerical analysis*, 3:269–289, 2000.

[40] L. Pareschi and G. Russo. High order asymptotically strong-stability-preserving methods for hyperbolic systems with stiff relaxation. In *Hyperbolic Problems: Theory, Numerics, Applications*, pages 241–251. Springer, 2003.

[41] M. Restelli. Semi-lagrangian and semi-implicit discontinuous galerkin methods for atmospheric modeling applications. *PhD thesis Politecnico di Milano*, 2007.

[42] G. Russo and S. Boscarino. IMEX Runge-Kutta schemes for hyperbolic systems with diffusive relaxation. *European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS 2012)*, 2012.

[43] S. Schochet. Fast singular limits of hyperbolic PDEs. *Journal of differential equations*, 114(2):476–512, 1994.

[44] J. Schütz. An asymptotic preserving method for linear systems of balance laws based on Galerkin's method. *Journal of Scientific Computing*, 60:438–456, 2014.

[45] J. Schütz, K. Kaiser, and S. Noelle. The RS-IMEX splitting for the isentropic Euler equations. In *Proceedings to YIC GACM*, 2015.

[46] J. Schütz and S. Noelle. Flux splitting for stiff equations: A notion on stability. *Journal of Scientific Computing*, 64(2):522–540, 2015.

[47] P. Wesseling. *Principles of Computational Fluid Dynamics*, volume 29 of *Springer Series in Computational Mechanics*. Springer Verlag, 2001.

[48] L. Yelash, A. Müller, M. Lukáčová-Medvid'ová, F. X. Giraldo, and V. Wirth. Adaptive discontinuous evolution Galerkin method for dry atmospheric flow. *Journal of Computational Physics*, 268:106–133, 2014.