An asymptotic preserving semi-implicit multiderivative solver

J. Schütz and D. Seal

UHasselt Computational Mathematics Preprint
Nr. UP-19-09

November 8th, 2019

Available online at www.sciencedirect.com
ScienceDirect
[Insert Journal Name Here] 00 (2020) 1–18
Journal ID here]logo
Journal Logo

An asymptotic preserving semi-implicit multiderivative solver

Jochen Schütz†, David C. Seal*

†*Faculty of Sciences, Hasselt University, Agoralaan Gebouw D, BE-3590 Diepenbeek*
**United States Naval Academy, Department of Mathematics, 572C Holloway Road, Annapolis, MD 21402, USA*

Abstract

In this work we construct a multiderivative implicit-explicit (IMEX) scheme for a class of stiff ordinary differential equations. Our solver is high-order accurate and has an asymptotic preserving (AP) property. The proposed method is based upon a two-derivative backward Taylor series base solver, which we show has an AP property. Higher order accuracies are found by iterating the result over a high-order multiderivative interpolant of the right hand side function, which we again prove has an AP property. Theoretical results showcasing the asymptotic consistency as well as the high-order accuracy of the solver are presented. In addition, an extension of the solver to an arbitrarily split right hand side function is also offered. Numerical results for a collection of standard test cases from the literature are presented that support the theoretical findings of the paper.

Keywords: Multiderivative, IMEX, singularly perturbed ODE, asymptotic preserving

1. Introduction

In this work we consider the numerical approximation of the system of differential equations

$$y'(t) = z(t), \qquad z'(t) = \frac{g(y(t), z(t))}{\varepsilon}, \quad 0 \le t \le T, \tag{1}$$

where $g : \mathbb{R}^2 \to \mathbb{R}$ is a smooth function. Our goal is to construct and analyze a high-order, semi-implicit, mulitderivative asymptotic preserving solver for this class of equations. This problem is equipped with initial conditions at time $t = 0$ defined by

$$y(0) = y^0, \qquad z(0) = z^0, \tag{2}$$

and we assume $0 < \varepsilon \ll 1$. The presence of this stiff relaxation parameter turns the problem into a singularly perturbed equation. At this point there is a vast amount of literature on problems of this kind. We refer the interested reader to the classical books [1, 2, 3] and the references therein for an overview on both analysis and applications of this class of problems.

Provided that the initial conditions are appropriately chosen, Eqn. (1) exhibits a multiscale behavior due to the presence of the stiff relaxation parameter. Modern solvers leverage this behavior by splitting the equation into 'stiff' and 'non-stiff' terms for efficient implementations of a numerical discretization. This procedure leads to the now famous implicit-explicit, or IMEX class of methods [4, 5, 6, 7, 8, 9, 10]. IMEX methods can usually be classified as either being an IMEX Runge-Kutta method, or an IMEX multistep method. Recent work has included definitions for IMEX General Linear Methods (GLMs) [11].

A distinct class apart from the aforementioned time integrators include multiderivative methods [12]. These solvers have recently been proven to be promising alternatives to classical Runge-Kutta and multistep schemes, and they are currently experiencing a renaissance with regard to their application to PDEs [13, 14, 15]. Much like Taylor approximations, these solvers work with not only with the first derivative, $y' \equiv z$ and $z' \equiv \frac{g}{\varepsilon}$, respectively, but they also leverage higher time derivatives of the unknowns. In doing so, the formulation of a solver becomes more intricate, but the tradeoff is that it makes the formulation more local, meaning additional information about the ODE can be garnered from each time point or stage value in the solver. This is particularly beneficial for modern high-performance computing architectures, as these solvers have the potential to reduce the memory overhead. To the best of our knowledge, there currently exists no extensions of multiderivative methods to IMEX schemes, which is the subject and aim of this work.

We propose a fourth-order multiderivative IMEX scheme that is based in part on an idea similar to the spectrally deferred correction (SDC) method [16, 17, 18]. Our solver makes use of a second-order two-derivative IMEX Taylor series base solver, which we carefully construct in such a way that it contains an asymptotic preserving property. This solver serves as a 'predictor', upon which iterations are performed on a higher-order scheme that lets us pick up the order of accuracy of the method. In the current work we stop at fourth-order accuracy but our method can be extended to higher orders based upon what is presented here. In this work we analyze the method with respect to its asymptotic properties, we show that it is asymptotically consistent [19, 20] with the continuous asymptotics, and we show numerical results indicating we do indeed find high-order accuracy. Finally, we show that order reduction common to many IMEX solvers for stiff relaxation problems can be overcome in some situations with our proposed method.

The paper is structured as follows: In Sec. 2, we discuss to the necessary theoretical details for equation (1), including restrictions on g and on the initial data. In Sec. 3, the proposed method is presented and consistency in the discretization parameter Δt is shown. This is followed by an *asymptotic* consistency analysis in Sec. 4, i.e., consistency in the singular parameter ε is demonstrated. Numerically, in Sec. 5, we discuss the phenomenon of order reduction and show how the algorithm can be used to address common challenges found with stiff IMEX solvers. As not all singularly perturbed ODEs are of the form defined in (1), we extend the method to more general ordinary differential equation in Sec. 6, where we include stability results for a prototype equation that contains an additive right hand side. Finally, we offers conclusions and an outlook for future work in Sec. 7.

2. The underlying equations: Necessary analytic properties

We consider the equation as given in (1). Although the solution to the equations y and z depend on ε, we only bring this up when necessary in our analysis. In addition, the dependence on the time variable is typically only written explicitly if needed. Furthermore, we need some (frequently used) assumptions on $\partial_z g$, which are related to stable manifolds of the $\varepsilon \to 0$ limit of the solution [1].

Postulate 1. *Assume that g is a smooth function, and that $\partial_z g(y,z) < -c$ for all $y, z \in \mathbb{R}$, with $c > 0$ being a positive constant.*

Remark 1. *Note that many equations only fulfill the assumption locally. This would be enough for our purposes, a local truncation of g would then be sufficient.*

As our solver is a multiderivative solver, we need information about the second derivative of the unknown variables. A straightforward computation gives use the following Lemma.

Lemma 1. *For the solutions y and z to (1), there holds*

$$y'' = \frac{g(y,z)}{\varepsilon}, \qquad \text{and} \qquad z'' = \frac{1}{\varepsilon} \underbrace{\nabla g(y,z) \cdot \begin{pmatrix} z \\ \frac{g(y,z)}{\varepsilon} \end{pmatrix}}_{=:\dot{g}(y,z)}. \tag{3}$$

By ∇g, we denote the vector $\nabla g := \left(\partial_y g, \partial_z g\right)$, i.e., differentiation with respect to y and z.

For the sake of having more explicit error constants, we assume that g and $\dot{g}$ are Lipschitz.

Postulate 2. *We assume that g and $\dot{g}$ are Lipschitz, and that there holds*

$$\|g(y_1, z_1) - g(y_2, z_2)\| \leq L_g \left\| \begin{matrix} y_1 - y_2 \\ z_1 - z_2 \end{matrix} \right\|, \qquad \|\dot{g}(y_1, z_1) - \dot{g}(y_2, z_2)\| \leq \frac{L_{\dot{g}}}{\varepsilon} \left\| \begin{matrix} y_1 - y_2 \\ z_1 - z_2 \end{matrix} \right\|, \tag{4}$$

where L_g and $L_{\dot{g}}$ are constants independent of ε. Note that the ε^{-1} scaling for $\dot{g}$ is the correct one to use here (cf. La. 1).

Problems of the form (1) have an interesting structure in the asymptotic limit as $\varepsilon \to 0$. The starting point for analyzing these problems is to first assume that both y and z can be written in terms of a Hilbert expansion in ε, i.e.,

$$\begin{aligned} y(t) &= y_{(0)}(t) + \varepsilon y_{(1)}(t) + \varepsilon^2 y_{(2)}(t) + O(\varepsilon^3), \\ z(t) &= z_{(0)}(t) + \varepsilon z_{(1)}(t) + \varepsilon^2 z_{(2)}(t) + O(\varepsilon^3). \end{aligned} \tag{5}$$

(For a proof of this identity and more related results, see [1] and the references contained within.) By substituting these expansions into the system of differential equations, equations that each of the $y_{(0)}, z_{(0)}, y_{(1)}, z_{(1)}, \ldots$ must satisfy can be derived. For example, substituting into the first equation in (1), we find that

$$y'_{(0)}(t) + \varepsilon y'_{(1)}(t) + \varepsilon^2 y'_{(2)}(t) + O(\varepsilon^3) = z_{(0)}(t) + \varepsilon z_{(1)}(t) + \varepsilon^2 z_{(2)}(t) + O(\varepsilon^3). \tag{6}$$

As $\varepsilon \to 0$, we have the identity $y'_{(0)} = z_{(0)}$. After substituting the Hilbert expansion (5) into the second equation, we find

$$z'_{(0)}(t) + \varepsilon z'_{(1)}(t) + \varepsilon^2 z'_{(2)}(t) + O(\varepsilon^3) = \frac{1}{\varepsilon} g\Big(y_{(0)} + \varepsilon y_{(1)} + O(\varepsilon^2), z_{(0)} + \varepsilon z_{(1)} + O(\varepsilon^2)\Big). \tag{7}$$

A Taylor expansion of g is given by

$$g\Big(y_{(0)} + \varepsilon y_{(1)} + O(\varepsilon^2), z_{(0)} + \varepsilon z_{(1)} + O(\varepsilon^2)\Big) = g(y_{(0)}, z_{(0)}) + \varepsilon \nabla g(y_{(0)}, z_{(0)}) \cdot \begin{pmatrix} y_{(1)} \\ z_{(1)} \end{pmatrix} + O(\varepsilon^2), \tag{8}$$

which yields with (1)

$$\varepsilon z'_{(0)} + O(\varepsilon^2) = g(y_{(0)}, z_{(0)}) + \varepsilon \nabla g(y_{(0)}, z_{(0)}) \cdot \begin{pmatrix} y_{(1)} \\ z_{(1)} \end{pmatrix} + O(\varepsilon^2), \tag{9}$$

from which one can conclude

$$0 = g(y_{(0)}, z_{(0)}). \tag{10}$$

That is, under the assumption that y and z had Hilbert expansions, we have that the first order terms $y_{(0)}$ and $z_{(0)}$ fulfill the differential-algebraic equation

$$y'_{(0)} = z_{(0)}, \qquad 0 = g(y_{(0)}, z_{(0)}). \tag{11}$$

Eqn. (11) is the limit equation to first order, it is of course possible to extend this procedure. For example, based on the Taylor expansion for g given in (8), we find that to second order there holds

$$y'_{(1)} = z_{(1)}, \qquad z'_{(0)} = \nabla g(y_{(0)}, z_{(0)}) \cdot \begin{pmatrix} y_{(1)} \\ z_{(1)} \end{pmatrix}. \tag{12}$$

These equations can be combined to produce identities for the higher derivatives not only of the original functions y and z, but also of the asymptotic quantities $y_{(0)}$ and the like. For example, we have the following result.

Lemma 2. *The solutions $y_{(0)}$ and $z_{(0)}$ to (11) satisfy*

$$y''_{(0)} = \frac{-\partial_y g(y_{(0)}, z_{(0)}) z_{(0)}}{\partial_z g(y_{(0)}, z_{(0)})}. \tag{13}$$

Proof. Differentiate (11) with respect to time and make use of the chain rule. □

Finally, we further note that thanks to $g(y_{(0)}, z_{(0)}) = 0$ in Eqn. (11), we also have

$$0 = \frac{d}{dt} g(y_{(0)}, z_{(0)}) = \nabla g(y_{(0)}, z_{(0)}) \cdot \begin{pmatrix} y'_{(0)} \\ z'_{(0)} \end{pmatrix} \overset{(11),(12)}{=} \nabla g(y_{(0)}, z_{(0)}) \cdot \begin{pmatrix} z_{(0)} \\ \nabla g(y_{(0)}, z_{(0)}) \cdot \begin{pmatrix} y_{(1)} \\ z_{(1)} \end{pmatrix} \end{pmatrix} \tag{14}$$

This property is important in our asymptotic analysis.

The proposed numerical scheme makes use of not only the right-hand side of (1), but also on the temporal derivative thereof. Intuitively, it is therefore reasonable to extend the concept of *well-preparedness* [21] to cope also with the limit equation to second order.

Definition 1 (Well-preparedness)**.** *We call the initial conditions* (y_0, z_0) well-prepared *if they possess a Hilbert expansion. That is, there exist a collection of unique functions* $y^0_{(0)}, y^0_{(1)}, \dots$ *and* $z^0_{(0)}, z^0_{(1)}, \dots$ *for which the initial conditions can be expanded as*

$$y^0 = y^0_{(0)} + \varepsilon y^0_{(1)} + O(\varepsilon^2), \quad \text{and} \quad z^0 = z^0_{(0)} + \varepsilon z^0_{(1)} + O(\varepsilon^2). \tag{15}$$

Furthermore, we must have

$$g(y^0_{(0)}, z^0_{(0)}) = 0, \qquad \nabla g(y^0_{(0)}, z^0_{(0)}) \cdot \begin{pmatrix} z^0_{(0)} \\ \nabla g(y^0_{(0)}, z^0_{(0)}) \cdot \begin{pmatrix} y^0_{(1)} \\ z^0_{(1)} \end{pmatrix} \end{pmatrix} = 0. \tag{16}$$

Remark 2. *A couple of comments regarding the definition of well-prepared initial conditions are in order.*

- *The well-preparedness property is a necessary condition that the solution to the ODE defined in (1) has a Hilbert expansion given by (5).*
- *Typically, only the first equation in Eqn. (16) is enforced. However, the standard test cases shown in literature, see, e.g., Section 5 in [9], fulfill this property. (In fact, using (12) and higher-version thereof will automatically yield the initial conditions used in [9].)*
- *In [18], Boscarino and collaborators use a more general version of these initial conditions; the initial conditions we are using are called 'well-prepared to order one' in their nomenclature.*

3. The multiderivative implicit-explicit (MD-IMEX) method

We now describe the numerical method proposed in this work, first starting with a definition of the scheme for the class of equations presented in (1). Extensions of this method to larger classes of ODEs are discussed in Sec. 6, but much of the notation that we define here remains the same.

To begin, we start with a mesh spacing

$$0 = t^0 < t^1 < t^2 < \cdots < t^N = T_{end}, \quad t^{n+1} - t^n = \Delta t, \quad n = 0, 1, \dots, N-1, \tag{17}$$

of the time domain $[0, T_{end}]$. We seek discrete numerical approximations $y^n \approx y(t^n)$ and $z^n \approx z(t^n)$ to the exact solutions $y(t)$ and $z(t)$ of (1) at each time point $t = t^n$. For the sake of exposition, we restrict our attention to uniform time steps, but this work can certainly be extended to a non-uniform (or adaptive) time grid.

One well known method for updating the solution to this problem would be to apply the Trapezoidal rule to approximate the integral of the right hand side to produce a second-order solver via:

$$y^{n+1} := y^n + \frac{\Delta t}{2}\left(z^n + z^{n+1}\right) \approx y^n + \int_{t^n}^{t^{n+1}} y'(t)\,\mathrm{d}t = y^n + \int_{t^n}^{t^{n+1}} z(t)\,\mathrm{d}t, \tag{18}$$

$$z^{n+1} := z^n + \frac{\Delta t}{2\varepsilon}\left(g^n + g^{n+1}\right) \approx z^n + \int_{t^n}^{t^{n+1}} z'(t)\,\mathrm{d}t = z^n + \int_{t^n}^{t^{n+1}} \frac{g(y(t), z(t))}{\varepsilon}\,\mathrm{d}t, \tag{19}$$

where $g^n := g(y^n, z^n)$, for $n = 0, 1, \ldots, N$. A lesser well-known strategy is to use a fourth-order integral approximation to the right hand side that makes use of not only the first, but also the second derivative of the right hand side:

$$y^{n+1} := y^n + \frac{\Delta t}{2}\left(z^n + z^{n+1}\right) + \frac{\Delta t^2}{12\varepsilon}\left(g^n - g^{n+1}\right) \approx y^n + \int_{t^n}^{t^{n+1}} z(t)\,\mathrm{d}t, \tag{20}$$

$$z^{n+1} := z^n + \frac{\Delta t}{2\varepsilon}\left(g^n + g^{n+1}\right) + \frac{\Delta t^2}{12\varepsilon}\left(\dot{g}^n - \dot{g}^{n+1}\right) \approx z^n + \int_{t^n}^{t^{n+1}} \frac{g(y(t), z(t))}{\varepsilon}\,\mathrm{d}t, \tag{21}$$

where the total time derivative of the right hand side of z is defined as

$$\dot{g}^n := \nabla g^n \cdot \begin{pmatrix} z^n \\ \frac{1}{\varepsilon} g^n \end{pmatrix}, \quad n = 0, 1, 2, \ldots, N. \tag{22}$$

Of course neither of these methods are semi-implicit. Not only that, but we also seek a high-order method. Therefore, we start with the latter of these two, but we make a modification to the solver so that it becomes a semi-implicit, rather than a fully implicit solver.

Our proposed method is as follows.

Algorithm 1. *For the solution of (1), we propose the following semi-implicit iterative IMEX method to advance the solution from time $t = t^n$ to time $t = t^{n+1}$:*

1. ***Predict.*** *Given the solution (y^n, z^n), we compute a second-order IMEX Taylor approximation*

$$y^{[0]} := y^n + \Delta t z^n + \frac{\Delta t^2}{2\varepsilon} \underbrace{g(y^n, z^n)}_{=:g^n}, \tag{23}$$

$$z^{[0]} := z^n + \frac{\Delta t}{\varepsilon} \underbrace{g(y^{[0]}, z^{[0]})}_{=:g^{[0]}} - \frac{\Delta t^2}{2\varepsilon} \underbrace{\nabla g(y^{[0]}, z^{[0]}) \cdot \begin{pmatrix} z^{[0]} \\ \frac{1}{\varepsilon} g(y^{[0]}, z^{[0]}) \end{pmatrix}}_{:=\dot{g}^{[0]}} \tag{24}$$

   for the unknowns $y^{[0]}$ and $z^{[0]}$ that are our initial guesses for an approximation to $y(t^{n+1})$ and $z(t^{n+1})$. Note that this discretization is based upon a second-order forward *Taylor series in y and a second-order* backward *Taylor series in z. (In due course, we show that the presence of the implicit second order terms is important in the asymptotic analysis of the method.)*
2. ***Correct.*** *Based on this initial step, for $0 \le k \le k_{\max} - 1$ we solve*

$$y^{[k+1]} := y^n + \frac{\Delta t}{2}\left(z^n + z^{[k]}\right) + \frac{\Delta t^2}{12\varepsilon}\left(g^n - g^{[k]}\right), \tag{25}$$

$$z^{[k+1]} := z^n + \frac{\Delta t}{\varepsilon}\left(g^{[k+1]} - g^{[k]}\right) - \frac{\Delta t^2}{2\varepsilon}\left(\dot{g}^{[k+1]} - \dot{g}^{[k]}\right) + \frac{\Delta t}{2\varepsilon}\left(g^n + g^{[k]}\right) + \frac{\Delta t^2}{12\varepsilon}\left(\dot{g}^n - \dot{g}^{[k]}\right), \tag{26}$$

   for $y^{[k+1]}$ and $z^{[k+1]}$. Note that for ease of notation, we define

$$g^{[k]} := g(y^{[k]}, z^{[k]}), \quad \textit{and} \quad \dot{g}^{[k]} := \nabla g^{[k]} \cdot \begin{pmatrix} z^{[k]} \\ \frac{1}{\varepsilon} g^{[k]} \end{pmatrix}, \quad k = 0, 1, \ldots k_{\max}. \tag{27}$$

3. ***Update.*** *The update for the solution is defined as*

$$y^{n+1} := y^{[k_{\max}]}, \qquad z^{n+1} := z^{[k_{\max}]}.$$

Consistency and stability are of central importance for any numerical discretization of a differential equation. Furthermore, for aymptotic-preserving (AP) schemes, the asymptotic stability and accuracy (as $\varepsilon \to 0$) are of paramount import, as these are the defining features of any AP numerical solver. We analyze the latter two central properties in the forthcoming sections, but first we address the consistency of the numerical method by looking at the order of accuracy of the solver (as a fixed function of $\varepsilon > 0$) and letting $\Delta t \to 0$. Stability is investigated in the numerical results section where we consider a prototypical linear case after defining the appropriate extension of this solver to problems with an additive right hand side.

Remark 3. *In every iteration step, $y^{[k]}$ and $z^{[k]}$ are approximations to $y(t^{n+1})$ and $z(t^{n+1})$, respectively, of order $\min\{4, 2+k\}$. That is, the iterates pick up a single order of an order of accuracy with each sweep up the solver, up to a maximal order based on the underlying quadrature rule.*

We formalize the statement of Rmk. 3 in Thm. 1 but we first lay down the foundational ingredients for its proof. As this method is based on the integral formulation of the differential equation, we begin with some lesser well known quadrature identities. Define, for some generic function $f : \mathbb{R}^2 \to \mathbb{R}$,

$$\mathcal{I}\left[f^n, f^{[k]}\right] := \frac{\Delta t}{2}\left(f^n + f^{[k]}\right) + \frac{\Delta t^2}{12}\left(\dot{f}^n - \dot{f}^{[k]}\right), \tag{28}$$

with the obvious notation $f^n := f(y^n, z^n)$ and $f^{[k]} := f(y^{[k]}, z^{[k]})$. Note that $\mathcal{I}$ is a fourth-order accurate quadrature rule, and therefore

$$\mathcal{I}\left[z(t^n), z(t^{n+1})\right] = \int_{t^n}^{t^{n+1}} z(t)\,\mathrm{d}t + \mathcal{O}(\Delta t^5), \quad \text{and} \tag{29}$$

$$\mathcal{I}\left[g(t^n), g(t^{n+1})\right] = \int_{t^n}^{t^{n+1}} g(y(t), z(t))\,\mathrm{d}t + \mathcal{O}(\Delta t^5), \tag{30}$$

assuming enough regularity in the underlying y, z, and g functions that define (1). (The constants in the big-$\mathcal{O}$ estimate do of course depend on $\varepsilon > 0$.) For the sake of readability, we have made the slight abuse of notation and are thinking of $g(t^n) := g(y(t^n), z(t^n))$. Furthermore, observe that the defining equations for $y^{[k]}$, $z^{[k]}$, respectively, with $k > 0$ in the correction step can then be written as

$$y^{[k+1]} = y^n + \mathcal{I}\left[z^n, z^{[k]}\right], \quad \text{and} \tag{31}$$

$$z^{[k+1]} = z^n + \frac{\Delta t}{\varepsilon}\left(g^{[k+1]} - g^{[k]}\right) - \frac{\Delta t^2}{2\varepsilon}\left(\dot{g}^{[k+1]} - \dot{g}^{[k]}\right) + \frac{1}{\varepsilon}\mathcal{I}\left[g^n, g^{[k]}\right], \tag{32}$$

with the understanding that $\dot{z} := \frac{g}{\varepsilon}$, which is required to compute $\mathcal{I}\left[z^n, z^{[k]}\right]$.

As is customary in a consistency analysis, assume that y^n and z^n are the exact solutions evaluated at time t^n. That is, we assume $y^n = y(t^n)$ and $z^n = z(t^n)$. Define

$$\delta_y^{[k]} := y^{[k]} - y(t^{n+1}), \qquad \delta_z^{[k]} := z^{[k]} - z(t^{n+1}), \qquad \text{and} \qquad \delta^{[k]} := \|(\delta_y^{[k]}, \delta_z^{[k]})\|. \tag{33}$$

Note that

$$\begin{aligned}\left|\mathcal{I}\left[z(t^n), z(t^{n+1})\right] - \mathcal{I}\left[z(t^n), z^{[k]}\right]\right| &= \left|\frac{\Delta t}{2}\left(z(t^{n+1}) - z^{[k]}\right) + \frac{\Delta t^2}{12}\left(\dot{z}(t^{n+1}) - \dot{z}^{[k]}\right)\right| \\ &\le \frac{\Delta t}{2}\left|(z(t^{n+1}) - z^{[k]})\right| + \frac{\Delta t^2}{12\varepsilon}\left|g(t^{n+1}) - g^{[k]}\right| \\ &\le \frac{\Delta t}{2}\delta^{[k]} + \frac{\Delta t^2}{12\varepsilon}L_g\delta^{[k]},\end{aligned} \tag{34}$$

where L_g is the Lipschitz constant for g, and similarly

$$\left|\mathcal{I}\left[g(t^n), g(t^{n+1})\right] - \mathcal{I}\left[g(t^n), g^{[k+1]}\right]\right| \le \frac{\Delta t}{2}L_g\delta^{[k]} + \frac{\Delta t^2}{12\varepsilon}L_{\dot{g}}\delta^{[k]}, \tag{35}$$

where $L_{\dot{g}}$ is the Lipschitz constant for $\dot{g}$. Since the exact solution of the differential equation satisfies

$$z(t^{n+1}) = z(t^n) + \frac{1}{\varepsilon}\int_{t^n}^{t^{n+1}} g(y, z)\mathrm{d}t, \tag{36}$$

we have

$$\begin{aligned}\left|\delta_z^{[k+1]}\right| &= \left|z(t^n) + \frac{\Delta t}{\varepsilon}\left(g^{[k+1]} - g^{[k]}\right) - \frac{\Delta t^2}{2\varepsilon}\left(\dot{g}^{[k+1]} - \dot{g}^{[k]}\right) + \frac{1}{\varepsilon}\mathcal{I}\left[g^n, g^{[k]}\right] - z(t^n) - \frac{1}{\varepsilon}\int_{t^n}^{t^{n+1}} g(y,z)\mathrm{d}t\right| \\ &\leq \frac{\Delta t}{\varepsilon}\underbrace{\left|g^{[k+1]} - g^{[k]}\right|}_{\mathbf{I}} + \frac{\Delta t^2}{2\varepsilon}\underbrace{\left|\dot{g}^{[k+1]} - \dot{g}^{[k]}\right|}_{\mathbf{II}} + \frac{1}{\varepsilon}\underbrace{\left|\mathcal{I}\left[g^n, g^{[k]}\right] - \int_{t^n}^{t^{n+1}} g(y,z)\mathrm{d}t\right|}_{\mathbf{III}}.\end{aligned}$$

We estimate each of these terms separately:

$$|\mathbf{I}| = \left|g^{[k+1]} - g^{[k]}\right| \leq \left|g^{[k+1]} - g^{n+1}\right| + \left|g^{n+1} - g^{[k]}\right| \leq L_g\left|\delta^{[k+1]}\right| + L_g\left|\delta^{[k]}\right|, \tag{37}$$

and

$$|\mathbf{II}| = \left|\dot{g}^{[k+1]} - \dot{g}^{[k]}\right| = \left|\dot{g}^{[k+1]} - \dot{g}^{n+1} + \dot{g}^{n+1} - \dot{g}^{[k]}\right| \leq \frac{L_{\dot{g}}}{\varepsilon}\delta^{[k+1]} + \frac{L_{\dot{g}}}{\varepsilon}\delta^{[k]}. \tag{38}$$

Finally, we make use of (35) and (30) to estimate the third term in this inequality:

$$\begin{aligned}|\mathbf{III}| &= \left|\mathcal{I}\left[g^n, g^{[k]}\right] - \int_{t^n}^{t^{n+1}} g(y,z)\mathrm{d}t\right| \leq \left|\mathcal{I}\left[g^n, g^{[k]}\right] - \mathcal{I}\left[g^n, g^{n+1}\right]\right| + \left|\mathcal{I}\left[g^n, g^{n+1}\right] - \int_{t^n}^{t^{n+1}} g(y,z)\mathrm{d}t\right| \\ &\leq \frac{\Delta t}{2}L_g\delta^{[k]} + \frac{\Delta t^2}{12\varepsilon}L_{\dot{g}}\delta^{[k]} + O(\Delta t^5).\end{aligned} \tag{39}$$

All together, we have

$$\begin{aligned}\left|\delta_z^{[k+1]}\right| &\leq \frac{\Delta t}{\varepsilon}|\mathbf{I}| + \frac{\Delta t^2}{2\varepsilon}|\mathbf{II}| + \frac{1}{\varepsilon}|\mathbf{III}| \\ &\leq \frac{L_g\Delta t}{\varepsilon}\delta^{[k+1]} + \frac{L_g\Delta t}{\varepsilon}\delta^{[k]} + \frac{L_{\dot{g}}\Delta t^2}{2\varepsilon^2}\delta^{[k+1]} + \frac{L_{\dot{g}}\Delta t^2}{2\varepsilon^2}\delta^{[k]} + \frac{\Delta t}{2\varepsilon}L_g\delta^{[k]} + \frac{\Delta t^2}{12\varepsilon^2}L_{\dot{g}}\delta^{[k]} + O(\Delta t^5) \\ &= O(\Delta t\delta^{[k+1]}) + O(\Delta t\delta^{[k]}) + O(\Delta t^5).\end{aligned} \tag{40}$$

Note again that the constants in the O-terms depend on ε. Similar results hold for $\delta_y^{[k+1]}$, which show that

$$\left|\delta_y^{[k+1]}\right| \leq O(\Delta t\delta^{[k+1]}) + O(\Delta t\delta^{[k]}) + O(\Delta t^5). \tag{41}$$

These results indicate that $\delta^{[k+1]}$ is one order (in Δt) better than $\delta^{[k]}$, until it reaches the maximum order of the quadrature rule. We formalize this statement in the following theorem.

Theorem 1. *The errors in the iterated approximations defined in Algorithm 1 satisfy $\delta^{[k]} = O(\Delta t^{\min\{5,2+k\}+1})$, with any $k \in \mathbb{Z}_{\geq 0}$, and therefore when $k_{\max} \geq 2$, the method is fourth-order consistent.*

Proof. The predictor is second-order accurate because $y^{[0]}$ and $z^{[0]}$ are computed by a second-order forward/backward Taylor method. That is, $\delta^{[0]} = O(\Delta t^3)$. Combining (40) and (41) gives

$$\delta^{[k+1]} = O(\Delta t\delta^{[k+1]}) + O(\Delta t\delta^{[k]}) + O(\Delta t^5),$$

which yields the desired result after applying induction on the number of iterates, k.

□

Remark 4. *From the analysis it is evident that once the quadrature operator $\mathcal{I}$ is replaced by another, higher-order quadrature, the method exhibits a higher overall order of accuracy. This route opens the possibility to investigate even higher order semi-implicit multiderivative time integrators.*

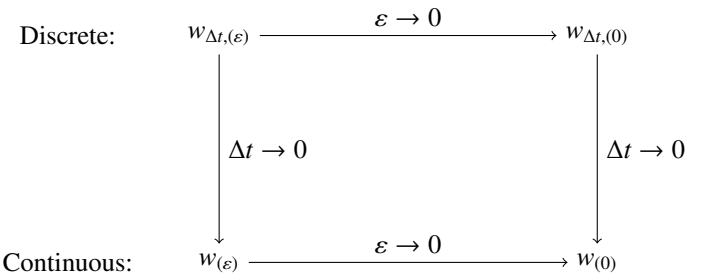

Figure 1. Asymptotic preserving methods. We say a method is asymptotic preserving if the limits in the above diagram commute with each other. That is, $\lim_{\Delta t\to 0}\lim_{\varepsilon\to 0} w_{\Delta t,(\varepsilon)} = \lim_{\varepsilon\to 0}\lim_{\Delta t\to 0} w_{\Delta t,(\varepsilon)} = w_{(0)}$. This property is not automatically preserved with any arbitrary, but consistent numerical method.

4. Asymptotic consistency

Considering the fact that the algorithm should approximate a singularly perturbed equation, it is evident that the behavior of the algorithm in the limiting case $\varepsilon \to 0$ is of utmost importance. Here, we investigate the asymptotic preserving (AP) property of the proposed method. Roughly speaking, an asymptotic preserving scheme means that the discretization found by sending $\varepsilon \to 0$ but holding Δt constant is a consistent discretization of the limit equation, Eqn. (11), found by sending $\varepsilon \to 0$ of the continuous problem. Generic differential equation solvers do not typically have this property.

Formally, if $w_{\Delta t,(\varepsilon)}$ is a discretization of the stiff equations defined in (1), then there are two limits that can be computed. We either send $\Delta t \to 0$ or we can send $\varepsilon \to 0$, from which we send the other variable to zero. On the one hand, if we first send $\Delta t \to 0$, then we end up with a (to be expected) numerical approximation $w_{(\varepsilon)}$ of (1), which we understand relaxes to $w_{(0)}$ as $\varepsilon \to 0$. On the other hand, if we instead first send $\varepsilon \to 0$, then we end up with a discretization $w_{\Delta t,(0)}$, which may or may not converge to the limiting solution $w_{(0)}$ as $\Delta t \to 0$. If it does, we say the numerical method has the *asymptotic preserving property*. This property is summarized in Figure 1.

We begin by showing the well-posedness of $y^{[0]}$, $z^{[0]}$, and the fact that these quantities possess Hilbert expansions.

Lemma 3. *Assume, in addition to Postulate 2, that $\partial_z g$ and $\partial_z \dot{g}$ are Lipschitz in the second argument, i.e., for all $y, z_1, z_2 \in \mathbb{R}$, we have*

$$|\partial_z g(y,z_1) - \partial_z g(y,z_2)| \le L_{\partial_z g}|z_1 - z_2|, \qquad |\partial_z \dot{g}(y,z_1) - \partial_z \dot{g}(y,z_2)| \le \frac{L_{\partial_z \dot{g}}}{\varepsilon}|z_1 - z_2|, \tag{42}$$

and furthermore assume that all occurring derivatives of g are uniformly bounded, and, in the spirit of [1, Sec. VI.3] that

$$g(y^n, z^n) = O(\varepsilon\Delta t), \qquad \dot{g}(y^n, z^n) = O\left(\frac{\Delta t}{\varepsilon}\right). \tag{43}$$

If, in addition to these criteria, we assume y^n and z^n possess Hilbert expansions, then there exists a fixed $\varepsilon_0 > 0$ and $\Delta t_0 > 0$, such that for all $0 < \varepsilon < \varepsilon_0$ and $0 < \Delta t < \Delta t_0$, we have that $y^{[0]}$ and $z^{[0]}$ possess Hilbert expansions.

Remark 5. *Under the assumption that there exists a Hilbert expansion, one can show that the identities in (43) hold with $O(\varepsilon)$. Behind this formulation is hence the implicit assumption that $\varepsilon \ll \Delta t$.*

Proof. Due to the fact that $y^{[0]}$ is computed explicitly, and $g(y^n, z^n) = O(\varepsilon\Delta t)$, it is evident that $y^{[0]}$ possesses a Hilbert expansion, can hence be written as

$$y^{[0]} = y^{[0]}_{(0)} + \varepsilon y^{[0]}_{(1)} + O(\varepsilon^2).$$

More challenging is showing that $z^{[0]}$ has a Hilbert expansion, given that this is nonlinear and implicit. Note that this term is supposed to be a zero of

$$F(z^{[0]}) := z^{[0]} - \frac{\Delta t}{\varepsilon} g(y^{[0]}, z^{[0]}) + \frac{\Delta t^2}{2\varepsilon} \dot{g}(y^{[0]}, z^{[0]}) - z^n = 0. \tag{44}$$

As typically done, see [1], we apply Newton-Kantorovich's theorem to this function F. Direct computation gives

$$F'(z) = 1 - \frac{\Delta t}{\varepsilon} \partial_z g(y^{[0]}, z) + \frac{\Delta t^2}{2\varepsilon} \partial_z \dot{g}(y^{[0]}, z),$$

and hence

$$|F'(z_1) - F'(z_2)| \leq \left(\frac{\Delta t}{\varepsilon} L_{\partial_z g} + \frac{\Delta t^2}{2\varepsilon^2} L_{\partial_z \dot{g}} \right) |z_1 - z_2|.$$

Furthermore, observe that

$$\left| g(y^{[0]}, z^n) \right| = |g\left(y^n + O(\Delta t), z^n\right)| \leq |g(y^n, z^n)| + O(\Delta t) \leq M_g \Delta t,$$

for some constant M_g, because $g(y^n, z^n) = O(\varepsilon \Delta t)$, $\partial_y g$ is bounded and there is some upper bound on ε. Because of our assumption on bounded derivatives of g, we also have

$$\partial_{zz} g(y^{[0]}, z^n) g(y^{[0]}, z^n) \leq M_1 \Delta t$$

for some $M_1 > 0$.

Now, consider Newton's method applied to F, with initial point z^n. Choose an M_2 such that

$$\|\partial_{yz} g\|_\infty |z^n| + \|\partial_y g\|_\infty \leq M_2.$$

Then, taking into account Postulate 1, we have

$$\begin{aligned}
F'(z^n) &= 1 - \frac{\Delta t}{\varepsilon} \partial_z g(y^{[0]}, z^n) + \frac{\Delta t^2}{2\varepsilon} \partial_z \dot{g}(y^{[0]}, z^n) \\
&\geq 1 + \frac{\Delta t}{\varepsilon} c + \frac{\Delta t^2}{2\varepsilon} \left(\partial_{yz} g(y^{[0]}, z^n) z^n + \partial_y g(y^{[0]}, z^n) \right) + \frac{\Delta t^2}{2\varepsilon^2} \left(\partial_{zz} g(y^{[0]}, z^n) g(y^{[0]}, z^n) + (\partial_z g(y^{[0]}, z^n))^2 \right) \\
&\geq 1 + \frac{\Delta t}{\varepsilon} c + \frac{\Delta t^2}{2\varepsilon^2} c^2 + \frac{\Delta t^2}{2\varepsilon} \left(\partial_{yz} g(y^{[0]}, z^n) z^n + \partial_y g(y^{[0]}, z^n) \right) + \frac{\Delta t^2}{2\varepsilon^2} \left(\partial_{zz} g(y^{[0]}, z^n) g(y^{[0]}, z^n) \right) \\
&\geq 1 + \frac{\Delta t}{\varepsilon} \left(c - \frac{\Delta t}{2} M_2 \right) + \frac{\Delta t^2}{2\varepsilon^2} \left(c^2 - \Delta t M_1 \right).
\end{aligned}$$

Choosing Δt_0 small enough (independently of ε!) so that the expressions in brackets are strictly positive (larger than $\alpha > 0$ say) for any $\Delta t < \Delta t_0$, yields that $F'(z^n) \neq 0$ and that

$$F'(z^n) \geq 1 + \frac{\Delta t}{\varepsilon} \alpha + \frac{\Delta t^2}{2\varepsilon^2} \alpha.$$

A similar computation, taking into account $g(y^n, z^n) = O(\Delta t)$ and $\dot{g}(y^n, z^n) = O\left(\frac{\Delta t}{\varepsilon}\right)$, yields

$$\begin{aligned}
|F(z^n)| &= \left| z^n - \frac{\Delta t}{\varepsilon} g(y^{[0]}, z^n) + \frac{\Delta t^2}{2\varepsilon} \dot{g}(y^{[0]}, z^n) - z^n \right| \\
&\leq M_g \frac{\Delta t^2}{\varepsilon} + M_{\dot{g}} \frac{\Delta t^3}{2\varepsilon^2},
\end{aligned}$$

where $M_{\dot{g}}$ is defined similarly to M_g.

The first Newton step would thus have step width

$$\frac{F(z^n)}{F'(z^n)} \leq \Delta t \frac{M_g \frac{\Delta t}{\varepsilon} + M_{\dot{g}} \frac{\Delta t^2}{2\varepsilon^2}}{1 + \frac{\Delta t}{\varepsilon}\alpha + \frac{\Delta t^2}{2\varepsilon^2}\alpha}.$$

This expression can be bounded by Δt times a constant H that does not depend on ε and Δt, i.e.,

$$\frac{F(z^n)}{F'(z^n)} \leq H\Delta t.$$

Hence, there holds

$$\left(\frac{\Delta t}{\varepsilon} L_{\partial_z g} + \frac{\Delta t^2}{2\varepsilon^2} L_{\partial_z \dot{g}}\right) |F'(z^n)^{-1}| \left|\frac{F(z^n)}{F'(z^n)}\right| \leq \frac{\frac{\Delta t}{\varepsilon} L_{\partial_z g} + \frac{\Delta t^2}{2\varepsilon^2} L_{\partial_z \dot{g}}}{1 + \frac{\Delta t}{\varepsilon}\alpha + \frac{\Delta t^2}{2\varepsilon^2}\alpha} H\Delta t.$$

Also this can be bounded by some constant times Δt, hence, choosing Δt sufficiently small makes the expression smaller than one half, and the Newton-Kantorovich theorem can be used. Not only does this imply that $z^{[0]} - z^n = O(1)$, it also implies that $z^{[0]}$ has a Hilbert expansion, because we can repeat the argument for every Newton step. □

Remark 6. *Please note that statement and proof can be extended to the full method.*

To continue, we show the AP-property for the forward/backward starting phase.

Lemma 4. *Assume that $y^{[0]}$ and $z^{[0]}$ possess Hilbert expansions. That is, we are operating under the assumptions presented in Lemma 3. Then, $y^{[0]}$ and $z^{[0]}$ are also well-prepared in the sense of Def. 1, i.e., there holds*

$$g^{[0]}_{(0)} := g\left(y^{[0]}_{(0)}, z^{[0]}_{(0)}\right) = 0, \qquad \nabla g^{[0]}_{(0)} \cdot \begin{pmatrix} z^{[0]}_{(0)} \\ \nabla g^{[0]}_{(0)} \cdot \begin{pmatrix} y^{[0]}_{(1)} \\ z^{[0]}_{(1)} \end{pmatrix} \end{pmatrix} = 0.$$

Proof. The proof starts by considering $z^{[0]}$, given by

$$z^{[0]} := z^n + \frac{\Delta t}{\varepsilon} g^{[0]} - \frac{\Delta t^2}{2\varepsilon} \nabla g^{[0]} \cdot \begin{pmatrix} z^{[0]} \\ \frac{1}{\varepsilon} g^{[0]} \end{pmatrix}.$$

Inserting a Hilbert expansion for all occurring quantities reveals the fact that the highest order is $O(\varepsilon^{-2})$, with corresponding equation being given by

$$\partial_z g^{[0]}_{(0)} g^{[0]}_{(0)} = 0.$$

Due to our assumption on $\partial_z g$, see Postulate 1, there follows $g^{[0]}_{(0)} = 0$.

Now, to $O(\varepsilon^{-1})$, the limiting equations are (be aware that the term scaled with ε^{-2} also contributes, see Eqn. (8)) are given by

$$\Delta t g^{[0]}_{(0)} - \frac{\Delta t^2}{2} \nabla g^{[0]}_{(0)} \cdot \begin{pmatrix} z^{[0]}_{(0)} \\ \nabla g^{[0]}_{(0)} \cdot \begin{pmatrix} y^{[0]}_{(1)} \\ z^{[0]}_{(1)} \end{pmatrix} \end{pmatrix} - \frac{\Delta t^2}{2} \nabla (\partial_z g^{[0]}_{(0)}) \cdot \begin{pmatrix} y^{[0]}_{(1)} \\ z^{[0]}_{(1)} \end{pmatrix} \cdot g^{[0]}_{(0)} = 0.$$

Exploiting the fact that $g^{[0]}_{(0)} = 0$ yields the claim. □

Theorem 2 (The algorithm is AP)**.** *Assume that the discrete solution at time level $t = 0$ is well-prepared in the sense of Def. 1. Assume furthermore that the discrete solution possesses a Hilbert expansion. Then, there holds for all times t^n that*

$$g(y^n_{(0)}, z^n_{(0)}) = 0$$

and

$$\nabla g(y^n_{(0)}, z^n_{(0)}) \cdot \begin{pmatrix} z^n_{(0)} \\ \nabla g(y^n_{(0)}, z^n_{(0)}) \cdot \begin{pmatrix} y^n_{(1)} \\ z^n_{(1)} \end{pmatrix} \end{pmatrix} = 0.$$

This implies that the method is asymptotic preserving.

Proof. The proof is very similar to the one of La. 4 and is hence omitted. Note that due to the $\varepsilon-$dependency of the 'explicit' terms $g^{[k]}$ and $\dot{g}^{[k]}$, terms as in Eqn. (16) at previous time/stage instances do show up. This, contrarily to La. 4, necessitates the need for well-prepared initial conditions as in Def. 1. □

5. Asymptotic accuracy

From a practical point of view, it is not only of interest whether the method is asymptotically *consistent*, but also to what orders the consistency is. This is a more delicate issue than pure consistency; in particular for IMEX Runge-Kutta methods, this leads to rather unwanted results including the stage order of the Runge-Kutta method being a limiting factor, see [8, 1]. In order to investigate this, let us consider van der Pol equation, being in the form (1) with g given by

$$g(y, z) = (1 - y^2)z - y.$$

We put our initial conditions as

$$y(0) = 2, \qquad z(0) = -\frac{2}{3} + \frac{10}{81}\varepsilon - \frac{292}{2187}\varepsilon^2,$$

which is a frequent choice in literature, see, e.g., [9]. Note that these initial conditions are well-prepared in the sense of Def. 1.

In Fig. 2, we plot convergence results for the method presented in Alg. 1. On the top-left, we chose $k_{\max} = 0$ (this means that $y^{n+1} = y^{[0]}$, similarly for z, i.e., only the predictor is taken into account). On the top-right, $k_{\max}$ is set to 2, which is the minimal number of iterations required to produce a fourth-order scheme. We observe that the second-order scheme (which is our second-order base IMEX Taylor solver) exhibits no order reduction. This means that there is second-order convergence uniformly in ε. The fourth-order scheme shows severe order reduction. On the bottom of Fig. 2, we increase $k_{\max}$ to 20 and 100, respectively. It is clearly visible that this enhances convergence. For example, with $k_{\max} = 100$ we observe no order reduction. The reason for this behaviour is that under the assumption that $\left(y^{[k]}, z^{[k]}\right)$ converges as $k \to \infty$, the result of Alg. 1 is equal to the fourth-order quadrature rule:

$$y^{n+1} = y^n + \frac{\Delta t}{2}\left(z^n + z^{n+1}\right) + \frac{\Delta t^2}{12\varepsilon}\left(g^n - g^{n+1}\right),$$
$$z^{n+1} = z^n + \frac{\Delta t}{2\varepsilon}\left(g^n + g^{n+1}\right) + \frac{\Delta t^2}{12\varepsilon}\left(\dot{g}^n - \dot{g}^{n+1}\right),$$

which is apparently less sensitive to order reduction.

To investigate the loss of asymptotic order of accuracy numerically in some closer detail, we define $\delta(\Delta t; \varepsilon)$ to be the Euclidean norm of the error in y and z at end time T_{end} for a given Δt and a given ε, i.e.,

$$\delta(\Delta t; \varepsilon) := \sqrt{\left(y^N - y(T_{end})\right)^2 + \left(z^N - z(T_{end})\right)^2}.$$

(Note that δ does of course also depend on $k_{\max}$, which we have not made explicit.) As for the solution, a Hilbert expansion of δ in terms of ε is assumed, so

$$\delta(\Delta t, \varepsilon) = \delta_0(\Delta t) + \varepsilon\delta_1(\Delta t) + \varepsilon^2\delta_2(\Delta t) + O(\varepsilon^3). \tag{45}$$

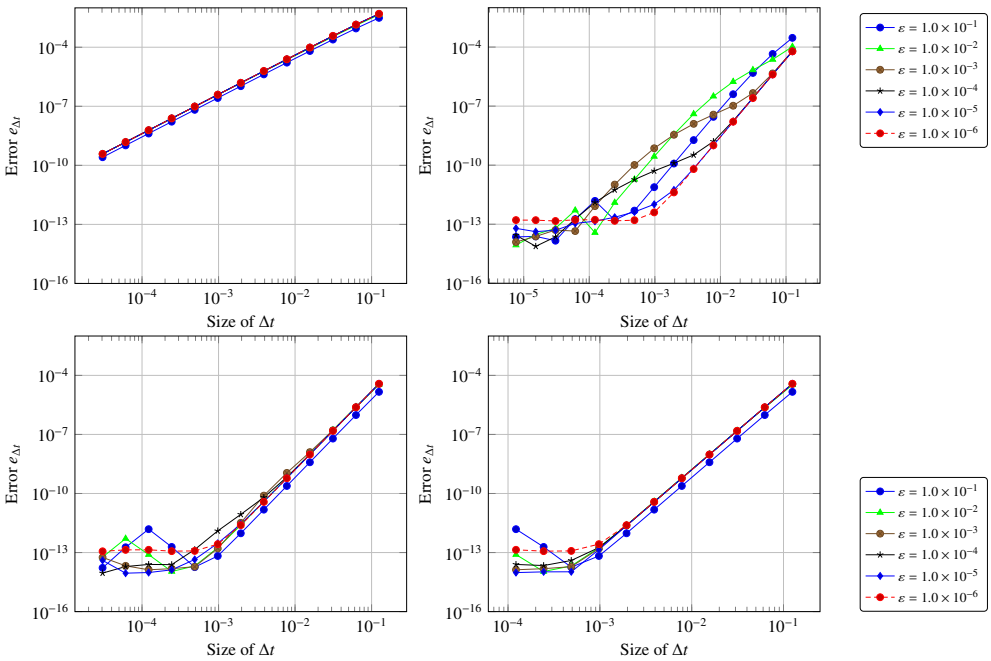

Figure 2. Convergence results for van der Pol equation with different values of ε. Top left: $k_{\max} = 0$, which amounts to only taking the predictor. Top right: $k_{\max} = 2$, with amounts to taking the full fourth-order scheme. Bottom: $k_{\max} = 20$ and $k_{\max} = 100$, respectively. Error measure is defined as the Euclidean norm of the y and z error at end time $T_{end} = 0.5$.

We approximate $\delta_0(\Delta t)$ and $\delta_1(\Delta t)$ numerically through

$$\delta_0 \approx \frac{\delta(\Delta t; \alpha\varepsilon) - \alpha\delta(\Delta t; \varepsilon)}{1-\alpha}, \qquad \varepsilon\delta_1 \approx \omega_1\delta(\Delta t; \varepsilon) + \omega_2\delta(\Delta t; \alpha\varepsilon) + \omega_3\delta(\Delta t; \alpha^2\varepsilon) \tag{46}$$

where we choose the rather arbitrary values $\alpha = \frac{5}{6}$ and $\varepsilon = \alpha^2 \cdot 10^{-5}$. (In our numerical testing, we verified that the results obtained are not influenced to any significant accuracy by this choice of ε.) The weights ω_i are chosen so that

$$\omega_1 + \omega_2 + \omega_3 = 0, \quad \omega_1 + \alpha\omega_2 + \alpha^2\omega_3 = 1, \quad \omega_1 + \omega_2\alpha^2 + \omega_3\alpha^4 = 0.$$

These conditions on the ω_i come out naturally after inserting the expansion (45) into (46).

For the same test case as above (i.e., van der Pol's problem with $k_{max} = 0, 2, 20, 100$, respectively) we plot $\delta_0(\Delta t)$ and $\delta_1(\Delta t)$ in Fig. 3. We remark that it is the contribution of $\delta_1(\Delta t)$ that is responsible for the degredation in the order of the solver. It can be seen that the slope of $\delta_1(\Delta t)$ increases as $k_{\max}$ increases, until machine accuracy issues occur.

6. Extensions and stability

6.1. Extending the method to arbitrary splittings

So far, we have discussed the method for equations of type Eqn. (1), and we have proposed a solver for that class of equations in Algorithm 1. The method developed in this work has a natural extension to a larger class of ODEs, which we now point out. Consider, for example a system of ODEs with a generic additive right hand side:

$$w'(t) = \Phi(w) =: \Phi_{\mathrm{E}}(w) + \Phi_{\mathrm{I}}(w), \tag{47}$$

with $w : \mathbb{R}^{\geq 0} \to \mathbb{R}^n$ and $\Phi_{\mathrm{E}}, \Phi_{\mathrm{I}} : \mathbb{R}^n \to \mathbb{R}^n$ being a splitting of the right hand side function. It is assumed that $\Phi_{\mathrm{I}}(w)$ contains "stiff" terms, and $\Phi_{\mathrm{E}}(w)$ contains non-stiff terms, so $\Phi_{\mathrm{I}}(w)$ should be treated implicitly, while $\Phi_{\mathrm{E}}(w)$ can be

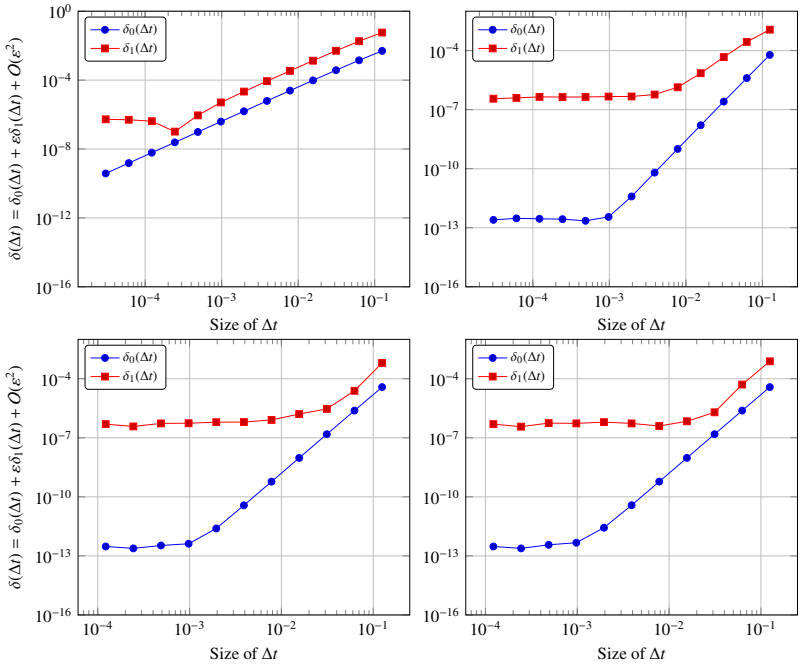

Figure 3. Asymptotic convergence results for van der Pol equation. Top left: $k_{\max} = 0$, which amounts to only taking the predictor. Top right: $k_{\max} = 2$, which is the minimum required number of iterations to obtain the the full fourth-order scheme. Bottom: $k_{\max} = 20$ and $k_{\max} = 100$, respectively. Error measure is defined as the Euclidean norm of the y and z error at end time $T_{end} = 0.5$.

treated explicitly in order to speed up the computations. Note that Eqn. (1) is of this form, with $\Phi_{\mathrm{I}}(w) = (0, \frac{1}{\varepsilon}g(w))$ and $\Phi_{\mathrm{E}}(w) = (z, 0)$. The choice of a suitable splitting is a subtle issue. We refer the reader to [21] for other splittings.

We extend Algorithm 1 in the following fashion to address arbitrary splittings. One of the chief goals is to retain the implicit-explicit (IMEX) type flavor of the underlying ODE and reach higher orders of accuracy all the while keeping the implicit solves as simple as possible. Note that the total time derivative of each piece in the right hand side function is given by

$$\dot{\Phi}_{\mathrm{I}}(w) = \Phi'_{\mathrm{I}}(w)\left(\Phi_{\mathrm{I}}(w) + \Phi_{\mathrm{E}}(w)\right), \qquad \dot{\Phi}_{\mathrm{E}}(w) = \Phi'_{\mathrm{E}}(w)\left(\Phi_{\mathrm{I}}(w) + \Phi_{\mathrm{E}}(w)\right).$$

Algorithm 2. *Consider a differential equation with a split right hand side given by Eqn. (47). To advance the solution from time level $t = t^n$ to $t = t^{n+1}$ we perform the following predictor-corrector strategy:*

1. ***Predict.*** *Given the solution w^n at time level $t = t^n$, we first compute an approximation to $w^{[0]} \approx w^{n+1}$ via*

$$w^{[0]} := w^n + \Delta t\left(\Phi_I(w^{[0]}) + \Phi_E(w^n)\right) + \frac{\Delta t^2}{2}\left(\dot{\Phi}_E(w^n) - \dot{\Phi}_I(w^{[0]})\right). \tag{48}$$

   That is, we perform a forward Taylor expansion on Φ_E and a backward Taylor expansion on Φ_I and integrate the results. This produces a second-order accurate predictor.

2. ***Correct.*** *Based on this initial step, for each $0 \le k \le k_{\max} - 1$ solve*

$$w^{[k+1]} := w^n + \Delta t\left(\Phi_I^{[k+1]} - \Phi_I^{[k]}\right) - \frac{\Delta t^2}{2}\left(\dot{\Phi}_I^{[k+1]} - \dot{\Phi}_I^{[k]}\right) + \frac{\Delta t}{2}\left(\Phi^n + \Phi^{[k]}\right) + \frac{\Delta t^2}{12}\left(\dot{\Phi}^n - \dot{\Phi}^{[k]}\right)$$

   for $w^{[k+1]}$.

3. ***Update.*** *Set $w^{n+1} := w^{[k_{\max}]}$.*

Note that the intermediate iterates need not be stored, and therefore this algorithm needs only the solution at a total of two time levels. This is advantageous when compared to any multistep method, where the solution at each stage needs to be stored, as well as most Runge-Kutta methods (save the methods of the low-storage variety).

Remark 7. *Algorithm 2 is an extension of Algorithm 1. That is, with* $w = (y, z)$, $\Phi_E(w) = (z, 0)$, *and* $\Phi_I(w) = \left(0, \frac{g}{\varepsilon}\right)$, *Algorithm 2 reduces to Algorithm 1.*

We now present the results for this problem on some classical test cases from the literature.

6.2. Kaps Problem

A problem that is not of the form defined in (1) is the so-called Kaps test problem [13]

$$\begin{aligned} y' &= -2y + \frac{1}{\varepsilon}(z^2 - y), \qquad y(0) = 1, \\ z' &= y - z(1 + z), \qquad\quad z(0) = 1, \end{aligned}$$

with exact solution $w := (y, z) = (e^{-2t}, e^{-t})$ for any $\varepsilon > 0$.

We use the most straightforward splitting on this problem given by grouping all of the terms containing ε and putting them into the implicit piece of the right hand side:

$$\Phi_{\mathrm{E}}(w) := (-2y, y - z(1 + y))^T, \qquad \Phi_{\mathrm{I}}(w) := \frac{1}{\varepsilon}\left(z^2 - y, 0\right)^T.$$

We present numerical results in Fig. 4. These results echo the findings of the previous section:

- The second-order scheme does not exhibit order degradation.
- For low $k_{\max}$, we observe order degradation.
- For $k_{\max} \to \infty$, the observed order degradation vanishes.

We thus conclude that the algorithm is capable of also computing solutions to equations that are not given in form (1). This is very important, in particular with respect to an extension of the scheme to singularly perturbed PDEs, where the semi-discretized systems are rarely in the form defined in (1).

6.3. Stability results

In this section, we examine stability of our newly developed scheme. We pay particular attention to the expected behavior of the solver for convection-diffusion equations, in which case the convection terms would be treated explicitely and the diffusive terms would be treated implicitely, as is common in the literature. As pointed out in [6, 7], a suitable prototype equation for the convection diffusion equation is

$$w' = (\lambda + i\mu)w \equiv \lambda w + i\mu w,$$

with $\lambda \leq 0$ and $\mu > 0$. The first part is treated implicitly, as it corresponds to the discretization of a diffusive operator, while the second part is treated explicitly. For a more detailed explanation of the relation between this equation and the convection-diffusion equation, we refer to [6]. To cast this into the framework of Alg. 2, we define

$$\Phi_{\mathrm{I}}(w) := \lambda w, \qquad \Phi_{\mathrm{E}}(w) := i\mu w,$$

which then yields

$$\dot{\Phi}_{\mathrm{I}}(w) = \lambda(\lambda + i\mu)w, \qquad \dot{\Phi}_{\mathrm{E}}(w) = i\mu(\lambda + i\mu)w.$$

We follow the steps done in [7] and define

$$\tilde{\lambda} := \lambda\Delta t \leq 0, \qquad \tilde{\mu} := \mu\Delta t > 0.$$

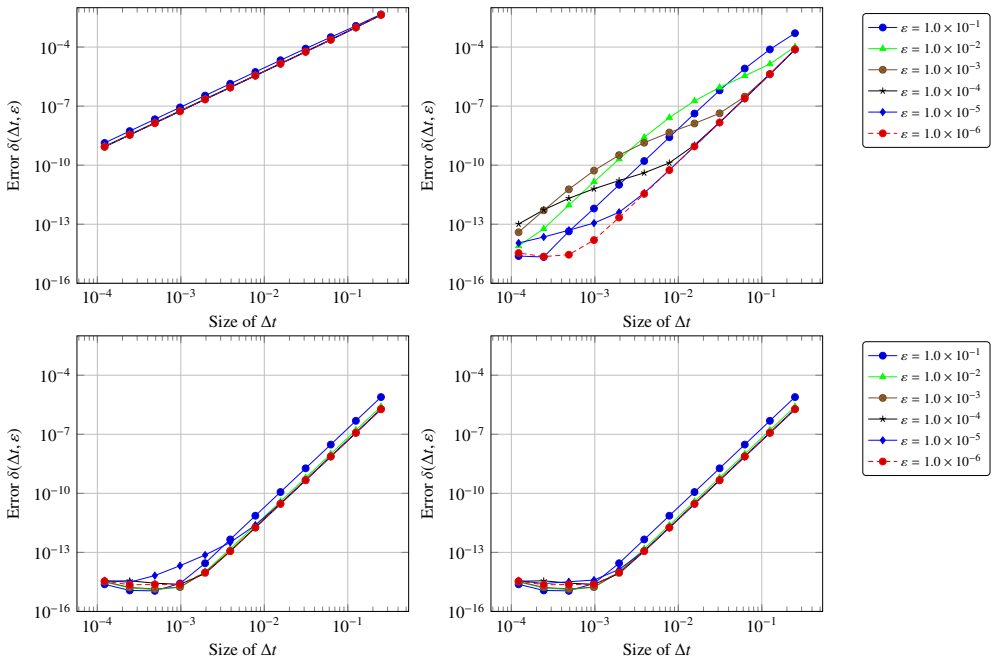

Figure 4. Convergence results for Kaps problem with different values of ε. Top left: $k_{\max} = 0$, which amounts to only taking the predictor. Top right: $k_{\max} = 2$, with amounts to taking the full fourth-order scheme. Bottom: $k_{\max} = 20$ and $k_{\max} = 100$, respectively. Error measure is defined as the Euclidean norm of the y and z error at end time $T_{end} = 1.0$.

The predictor step for Alg. 2 can hence be written as

$$w^{[0]} = \frac{1 + i\tilde{\mu} + \frac{i\tilde{\mu}\tilde{\lambda}}{2} - \frac{\tilde{\mu}^2}{2}}{1 - \tilde{\lambda} + \frac{\tilde{\lambda}^2}{2} + i\frac{\tilde{\lambda}\tilde{\mu}}{2}} w^n =: \Psi(\tilde{\lambda}, \tilde{\mu}) w^n.$$

As already noticed in [7] for IMEX Euler, for $z := \lambda + i\mu$ being on the imaginary axis, i.e., $\lambda = 0$, this can, for $\mu \neq 0$, never yield an unconditionally stable algorithm, as

$$\left| 1 + i\tilde{\mu} - \frac{\tilde{\mu}^2}{2} \right|^2 = \left(1 - \frac{\tilde{\mu}^2}{2} \right)^2 + \tilde{\mu}^2 = 1 + \frac{\tilde{\mu}^4}{4} > 1.$$

This is of course not surprising, as the algorithm reduces to a purely explicit time marching scheme. In the spirit of [7], we keep the ratio of λ and μ constant, i.e., we define

$$\gamma := \frac{\lambda}{\mu} \leq 0,$$

and investigate whether, for a given ratio of the implicit to explicit eigenvalues, γ, the algorithm is stable. This may produces a restrictions on $\tilde{\mu} \equiv \mu \Delta t$, which we can modify by changing the time step size. Any restrictions on $\tilde{\mu} \equiv \mu \Delta t$ will in practice result in a timestep restriction.

Lemma 5. *If $\gamma \leq -1$ then the predictor $w^{[0]}$ for the method is stable. That is, for a single time step, we have*

$$\|w^{[0]}\| \leq \|w^n\|.$$

Proof. Note that $\gamma = \frac{\lambda}{\mu} = \frac{\tilde{\lambda}}{\tilde{\mu}}$ and consider the expression

$$|\Psi(\gamma\tilde{\mu}, \tilde{\mu})| \leq 1,$$

which is equivalent to

$$\left(1-\frac{\tilde{\mu}^2}{2}\right)^2+\left(\tilde{\mu}+\frac{\gamma\tilde{\mu}^2}{2}\right)^2\leq\left(1-\gamma\tilde{\mu}+\frac{(\gamma\tilde{\mu})^2}{2}\right)^2+\left(\frac{\gamma\tilde{\mu}^2}{2}\right)^2.$$

This again reduces to

$$\frac{1-\gamma^4}{4}\tilde{\mu}^4+(\gamma^3+\gamma)\tilde{\mu}^3-2\gamma^2\tilde{\mu}^2+2\gamma\tilde{\mu}\leq 0.$$

As $\tilde{\mu}$ is positive and $\gamma\leq -1$, this proves the claim. □

Similar analysis can of course be made for the full algorithm. Due to the technical difficulties that come with high-order polynomials in λ and μ, we restrict ourselves to a numerical investigation. In Fig. 5, the maximum allowable $\tilde{\mu}$ is shown as a function of γ both for the predictor and the full algorithm, where we have restricted ourselves to $k_{\max}=2$. It can be seen that for $\gamma\to 0$, the maximum allowable timestep for the predictor tends to zero, while for the full algorithm, it tends to a fixed constant. This is due to the fact that the squared absolute value of the iteration function for $k_{\max}=2$ and $\gamma=0$ (hence $\lambda=0$) is given by

$$\frac{\tilde{\mu}^6\left(\tilde{\mu}^6+76\,\tilde{\mu}^4+1392\,\tilde{\mu}^2-7488\right)}{82944}+1,$$

which is smaller than one between $\tilde{\mu}=0$ and $\tilde{\mu}=2.075$. *The full algorithm hence gives a significant improvement in stability compared to just the predictor.*

Finally, we perform our analysis also for the 'limiting' method, i.e., the method defined by

$$w^{n+1}=w^n+\frac{\Delta t}{2}\left(\Phi^n+\Phi^{n+1}\right)+\frac{\Delta t^2}{12}\left(\dot{\Phi}^n-\dot{\Phi}^{n+1}\right). \tag{49}$$

In the case that $w^{[k]}$ converges with $k\to\infty$, the limit exactly coincides with w^{n+1} defined in (49). In terms of $\tilde{\lambda}$ and $\tilde{\mu}$, the iteration is given by

$$w^{n+1}=\frac{1+\frac{1}{2}(\tilde{\lambda}+i\tilde{\mu})+\frac{1}{12}(\tilde{\lambda}+i\tilde{\mu})^2}{1-\frac{1}{2}(\tilde{\lambda}+i\tilde{\mu})+\frac{1}{12}(\tilde{\lambda}+i\tilde{\mu})^2}w^n=:\Theta(\tilde{\lambda},\tilde{\mu})w^n.$$

For this fully implicit method, we have the following result:

Lemma 6. *If $\gamma<0$, the the method defined in (49) has an amplification factor that satisfies*

$$|\Theta(\gamma\tilde{\mu},\tilde{\mu})|\leq 1.$$

Proof. The claim is equivalent to

$$\left(1+\frac{1}{2}\gamma\tilde{\mu}+\frac{1}{12}(\gamma^2\tilde{\mu}^2-\tilde{\mu}^2)\right)^2+\left(\frac{1}{2}\tilde{\mu}+\frac{1}{6}\gamma\tilde{\mu}^2\right)^2\leq\left(1-\frac{1}{2}\gamma\tilde{\mu}+\frac{1}{12}(\gamma^2\tilde{\mu}^2-\tilde{\mu}^2)\right)^2+\left(-\frac{1}{2}\tilde{\mu}+\frac{1}{6}\gamma\tilde{\mu}^2\right)^2,$$

which is then again equivalent to

$$\gamma\tilde{\mu}\left(12+\gamma^2\tilde{\mu}^2+\tilde{\mu}^2\right)\leq 0.$$

This is true for any $\gamma<0$.

□

The finding here is that the limiting solver defined in (49) is numerically stable for any choice of time step.

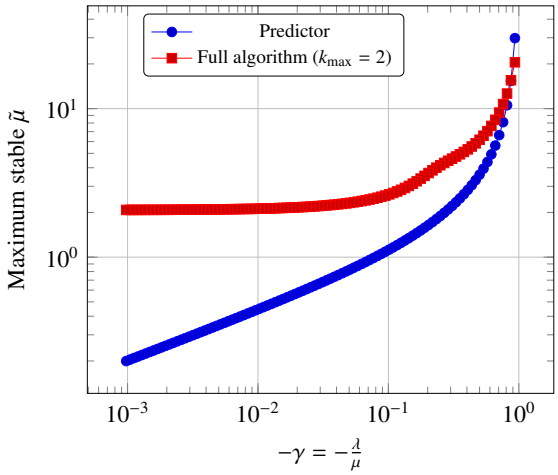

Figure 5. Maximum allowable $\tilde{\mu}$ to guarantee stability depending on γ. Note that a restriction on $\tilde{\mu} \equiv \mu\Delta t$ will in practice yield a restriction on the timestep Δt.

7. Conclusion and Outlook

In this work we have presented a novel time integrator, featuring aspects of the spectrally deferred correction, implicit-explicit (IMEX) time integrators, and the multiderivative class of ODE integrators. We have shown that the method is asymptotically preserving and we have presented results for a class of test cases from the literature. In addition, stability results for this solver have been investigated based on a prototype equation for convection-diffusion PDEs.

There are many extensions of this work that can be found. Next steps include investigating the application of this solver to partial differential equations of the singularly perturbed type. In particular, we are interested in the compressible Navier-Stokes equations at low Mach number [22]. Suitable splittings have already been developed in literature, see, e.g., [23, 24, 25, 26, 27, 28, 29, 30]. Similar to spectral deferred correction (SDC) methods, the proposed algorithm can certainly be parallelized in time, which, again in particular for PDEs, could make for a tremendous benefit in a parallel computing environment. Furthermore, extensions of this method that include variable orders of accuracy introduce the potential to investigate adaptive time stepping which would make for interesting results on their own merit.

Acknowledgements

This study is the outcome of a research stay of D.C. Seal at the University of Hasselt, which was supported by the Special Research Fund (BOF) of Hasselt University. Additional funding came from the Office of Naval Research, grant number N0001419WX01523.

References

[1] E. Hairer, G. Wanner, Solving ordinary differential equations II, Springer Series in Computational Mathematics, 1991.
[2] J. Kevorkian, J. D. Cole, Perturbation Methods in Applied Mathematics, Springer Berlin / Heidelberg / New York, 1981.
[3] R. E. O'Malley, Singular perturbation methods for ordinary differential equations, Vol. 89, Springer Science & Business Media, 2012.
[4] Le Roux, Marie-Noëlle, Semi-discrétisation en temps pour les équations d'évolution paraboliques lorsque l'opérateur dépend du temps, RAIRO. Anal. numér. 13 (2) (1979) 119–137. doi:10.1051/m2an/1979130201191.
URL https://doi.org/10.1051/m2an/1979130201191
[5] M. Crouzeix, Une méthode multipas implicite-explicite pour l'approximation des équations d'évolution paraboliques, Numerische Mathematik 35 (3) (1980) 257–276.
[6] U. M. Ascher, S. Ruuth, B. Wetton, Implicit-Explicit methods for time-dependent partial differential equations, SIAM Journal on Numerical Analysis 32 (1995) 797–823.

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

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

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

[10] S. Boscarino, G. Russo, On a class of uniformly accurate IMEX Runge-Kutta schemes and applications to hyperbolic systems with relaxation, SIAM Journal on Scientific Computing 31 (3) (2009) 1926–1945.

[11] H. Zhang, A. Sandu, S. Blaise, Partitioned and implicit–explicit general linear methods for ordinary differential equations, Journal of Scientific Computing 61 (1) (2014) 119–144.

[12] E. Hairer, G. Wanner, Multistep-multistage-multiderivative methods for ordinary differential equations, Computing (Arch. Elektron. Rechnen) 11 (3) (1973) 287–303.

[13] R. Chan, A. Tsai, On explicit two-derivative runge-kutta methods, Numerical Algorithms 53 (2010) 171–194.

[14] D. Seal, Y. Güçlü, A. Christlieb, High-order multiderivative time integrators for hyperbolic conservation laws, Journal of Scientific Computing 60 (2014) 101–140. doi:DOI10.1007/s10915-013-9787-8.

[15] J. Schütz, D. Seal, A. Jaust, Implicit multiderivative collocation solvers for linear partial differential equations with discontinuous Galerkin spatial discretizations, Journal of Scientific Computing 73 (2017) 1145–1163.

[16] A. Dutt, L. Greengard, V. Rokhlin, Spectral deferred correction methods for ordinary differential equations, BIT 40 (2) (2000) 241–266. doi:10.1023/A:1022338906936.

[17] M. Minion, Semi-implicit spectral deferred correction methods for ordinary differential equations, Communications in Mathematical Sciences 1 (3) (2003) 471–500.

[18] S. Boscarino, J. Qiu, G. Russo, Implicit-explicit integral deferred correction methods for stiff problems, SIAM Journal on Scientific Computing 40 (2018) A787–A816.

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

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

[21] J. Schütz, K. Kaiser, A new stable splitting for singularly perturbed ODEs, Applied Numerical Mathematics 107 (2016) 18–33.

[22] S. Klainerman, 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 (1981) 481–524.

[23] G. Bispen, K. R. Arun, M. Lukáčová-Medvid'ová, S. Noelle, IMEX large time step finite volume methods for low Froude number shallow water flows, Communications in Computational Physics 16 (2014) 307–347.

[24] G. Bispen, M. Lukáčová-Medvid'ová, L. Yelash, Asymptotic preserving IMEX finite volume schemes for low Mach number Euler equations with gravitation, Journal of Computational Physics 335 (2017) 222–248.

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

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

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

[28] 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 (1995) 213–237.

[29] S. Noelle, G. Bispen, K. Arun, M. Lukáčová-Medvid'ová, C.-D. Munz, A weakly asymptotic preserving low Mach number scheme for the Euler equations of gas dynamics, SIAM Journal on Scientific Computing 36 (2014) B989–B1024.

[30] J. Zeifang, J. Schütz, K. Kaiser, A. Beck, M. Lukáčová-Medvid'ová, S. Noelle, A novel full-Euler low Mach number IMEX splitting, Communications in Computational Physics (in press).

More preprints at www.uhasselt.be/cmat.