Asymptotic error analysis of an IMEX Runge–Kutta method

Asymptotic Error Analysis of an IMEX Runge-Kutta method

Klaus Kaiser

IGPM, RWTH Aachen University, Templergraben 55, DE - 52062 Aachen

Jochen Schütz

CMAT, UHasselt, Agoralaan Gebouw D, BE - 3590 Diepenbeek

Abstract

We consider a system of singularly perturbed differential equations with singular parameter $\varepsilon \ll 1$, discretized with an IMEX Runge-Kutta method. The splitting needed for the IMEX method stems from a linearization of the fluxes around the limit solution. We analyze the asymptotic convergence order as $\varepsilon \to 0$. We show that in this setting, the stage order of the implicit part of the scheme is of great importance, thereby explaining earlier numerical results showing a close correlation of errors of the splitting scheme and the fully implicit one.

Keywords: Order reduction, RS-IMEX, IMEX Runge-Kutta, singularly perturbed equation, asymptotic convergence order

1. Introduction

Singularly perturbed equations arise in many applications where a small parameter $\varepsilon \ll 1$ plays an important role. One example is the reaction of a substrate to a product through an enzyme with the concentration of the enzyme being much smaller than the one of the substrate. In this case, the small parameter ε denotes the ratio of enzyme and substrate concentration at initial time. This leads to Michaelis Menten equation [1, 2]. Another typical example are equations associated to low Mach number flow, where the small parameter ε is the ratio of velocity to speed of sound. For $\varepsilon \to 0$, the flow changes type from compressible to incompressible. For details, we refer to, e.g., [3, 4] and the references therein; and in particular to [5] for a short historical discussion. Mathematically speaking, the equations, wich are systems of conservation laws, become extremely stiff, giving rise to the need for efficient solution methodologies.

It is well-known that for systems of conservation laws, implicit time integrators tend to stabilize the solution process but to smear out the solution, while explicit time integrators heavily rely on a rather restrictive CFL-number, but are more accurate than their implicit counterpart [6]. To benefit from both the extended stability and the good approximation properties, implicit/explicit (IMEX) time integrators have been developed, see, e.g., [7, 8, 9, 10]. In particular, we discuss IMEX Runge-Kutta methods in this work [8, 9, 11]. Other IMEX methods, not discussed here, include linear multistep methods [7, 10], general linear methods [12], integral deferred correction methods [13] and many more. All those schemes rely heavily on a splitting of the equation into 'stiff' and 'non-stiff' parts. For the Euler and related equations, several splittings have been developed over the last few decades, e.g. in [14, 15, 16, 17, 18]. Despite the progress made in the last years, some parts of the theory are still lacking. As an example, it is unclear how well *high-order* versions of the methods approximate the equations for small ε as the phenomenon of order reduction can occur [19, 20].

Keeping this motivation in mind, we want to shed some light onto order reduction for one particular set of equations and one particular splitting in a similar spirit as it has been done in [20, 21], hoping that results can later be extended to a more complex setup. Therefore, in this work, we concentrate on the temporal integration part only, i.e., we consider a singularly perturbed ordinary differential equation of form

$$\frac{d}{dt}\begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} f(y,z) \\ \frac{1}{\varepsilon} g(y,z) \end{pmatrix}, \qquad 0 < t < t_{end}. \tag{1}$$

We have already treated this equation in [22] and have applied a new way of splitting the functions f and g into stiff and non-stiff terms to this equation. The splitting was termed RS-IMEX splitting (RS stands for reference solution), it relies on a formal approach and is in principle applicable to many singularly perturbed problems [18, 22, 23, 24, 25, 26].

In [22], we have compared this newly developed splitting to a standard splitting from literature [20], see Fig. 1. There, a convergence plot for two different ε is plotted; the underlying equation is van der Pol equation, see Sec. 5. In Fig. 1, it is clearly visible that the standard splitting exhibits a non-uniform order of convergence, while both the

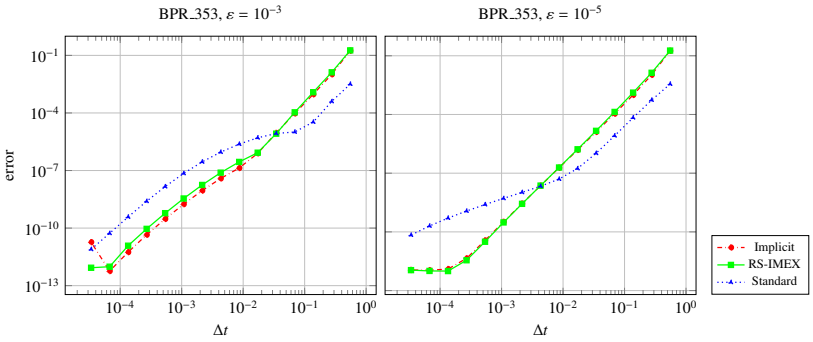

Figure 1: The BPR_353 IMEX Runge-Kutta method [27] applied to van der Pol equation for a standard splitting, the RS-IMEX splitting and a fully implicit method. Two values of ε are treated, it is clearly visible that the standard splitting shows a non-uniform order of convergence. Note that the results for the standard splitting and the RS-IMEX splitting have also been computed – for a different t_{end} – in [22].

RS-IMEX scheme and the fully implicit method converge with the optimal order and uniformly in ε. In [22], we could not explain this behavior; this motivated this work: We investigate the asymptotic convergence properties of the RS-IMEX scheme. It turns out that the stage order of the implicit part of the IMEX Runge-Kutta scheme, see Def. 2, mostly determines the asymptotic behavior of the RS-IMEX scheme. This is surprising, because one would naively assume that the stage order of the method has a greater impact. It also explains why in Fig. 1, there is virtually no difference between the RS-IMEX splitting and the fully implicit solution process. The core theorem, Thm. 1, also gives a-priori information on the expected asymptotic convergence properties of the RS-IMEX scheme.

The paper is structured as follows: In Sec. 2, we introduce the underlying singularly perturbed ordinary differential equation, in Sec. 3, we define IMEX Runge-Kutta schemes and the RS-IMEX splitting. Sec. 4 is the core section of this work where the asymptotic convergence analysis is made. In Sec. 5, we show numerical results and discuss the sharpness of our analytical results. As usual, the last Sec. 6 offers conclusions and outlook.

2. Singularly perturbed ordinary differential equations

In what follows, we consider singularly perturbed ordinary differential equations of form

$$\frac{d}{dt}\begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} f(y,z) \\ \frac{1}{\varepsilon} g(y,z) \end{pmatrix}, \qquad 0 < t < t_{end}, \tag{1}$$

where $f, g : \mathbb{R}^2 \to \mathbb{R}$ are smooth and g fulfills

$$\mu(g_z(y,z)) \leq -1 \tag{2}$$

in an ε independent neighborhood of the solution, with μ being the logarithmic norm. As we assume $g(y,z)$ to be scalar, the logarithmic norm (2) is equivalent to

$$g_z(y,z) \leq -1,$$

i.e. the partial derivative with respect to z is bounded by a negative constant. y and z are the unknowns, and the equation is equipped with suitable initial conditions to be explained in further detail below. We assume that fluxes f, g and end time $t_{end} > 0$ are chosen in such a way that there exist smooth solutions for all $\varepsilon \in (0, 1)$ up to time $t = t_{end}$.

The case $\varepsilon \ll 1$ is of special interest in this work. There, it is obvious that (1) consists of multiple scales in ε. Even more, for the formal limit $\varepsilon = 0$, the equation changes its type from an ordinary (ODE) to an algebraic differential equation (DAE). This can best be understood when considering an asymptotic expansion of the solution, i.e., take a representation of y and z as

$$\begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} y_{(0)} \\ z_{(0)} \end{pmatrix} + \varepsilon \begin{pmatrix} y_{(1)} \\ z_{(1)} \end{pmatrix} + \varepsilon^2 \begin{pmatrix} y_{(2)} \\ z_{(2)} \end{pmatrix} + O(\varepsilon^3). \tag{3}$$

One can show that such an expansion exists, see [19, 20] for details, given that the initial conditions are chosen carefully, see below. Plugging the asymptotic expansion (3) into equation (1), ignoring terms of $O(\varepsilon^3)$, expanding f and g through Taylor, and separating the equations in terms of ε, one obtains the following set of DAEs:

$$\frac{d}{dt} \begin{pmatrix} y_{(0)} \\ 0 \end{pmatrix} = \begin{pmatrix} f(y_{(0)}, z_{(0)}) \\ g(y_{(0)}, z_{(0)}) \end{pmatrix} \tag{4}$$

$$\frac{d}{dt} \begin{pmatrix} y_{(1)} \\ z_{(0)} \end{pmatrix} = \begin{pmatrix} \partial_y f(y_{(0)}, z_{(0)}) y_{(1)} + \partial_z f(y_{(0)}, z_{(0)}) z_{(1)} \\ \partial_y g(y_{(0)}, z_{(0)}) y_{(1)} + \partial_z g(y_{(0)}, z_{(0)}) z_{(1)} \end{pmatrix} \tag{5}$$

$$\begin{aligned} \frac{d}{dt} \begin{pmatrix} y_{(2)} \\ z_{(1)} \end{pmatrix} = & \begin{pmatrix} \partial_y f(y_{(0)}, z_{(0)}) y_{(2)} + \partial_z f(y_{(0)}, z_{(0)}) z_{(2)} \\ \partial_y g(y_{(0)}, z_{(0)}) y_{(2)} + \partial_z g(y_{(0)}, z_{(0)}) z_{(2)} \end{pmatrix} + \frac{1}{2} \begin{pmatrix} \partial_{yy} f(y_{(0)}, z_{(0)}) {y_{(1)}}^2 + \partial_{zz} f(y_{(0)}, z_{(0)}) {z_{(1)}}^2 \\ \partial_{yy} g(y_{(0)}, z_{(0)}) {y_{(1)}}^2 + \partial_{zz} g(y_{(0)}, z_{(0)}) {z_{(1)}}^2 \end{pmatrix} \\ & + \begin{pmatrix} \partial_{yz} f(y_{(0)}, z_{(0)}) y_{(1)} z_{(1)} \\ \partial_{yz} g(y_{(0)}, z_{(0)}) y_{(1)} z_{(1)} \end{pmatrix} \end{aligned}. \tag{6}$$

In this work, we exclusively consider well-prepared initial conditions. Roughly speaking, they prevent the solution from forming an initial layer for $\varepsilon \to 0$ and are consistent with the limit DAEs (4)-(6).

Assumption 1. *[Well prepared initial data] Initial data for equation* (1) *is called* well-prepared*, if it is given by an asymptotic expansion, i.e.*

$$\begin{pmatrix} y(t=0) \\ z(t=0) \end{pmatrix} = \begin{pmatrix} y_{(0)}(t=0) \\ z_{(0)}(t=0) \end{pmatrix} + \varepsilon \begin{pmatrix} y_{(1)}(t=0) \\ z_{(1)}(t=0) \end{pmatrix} + \varepsilon^2 \begin{pmatrix} y_{(2)}(t=0) \\ z_{(2)}(t=0) \end{pmatrix} + O(\varepsilon^3),$$

and $(y_{(i)}(t=0), z_{(i)}(t=0))^T$ *with* $i = 0, 1, 2$ *are valid initial data for equations* (4)-(6).

Equation (4), which can be seen as the $\varepsilon \to 0$ limit, is of special interest. From (2), it follows that $\partial_z g(y, z) \leq -1$ in a neighborhood of the asymptotic solution $y_{(0)}$ and $z_{(0)}$, see also [19]. Then the implicit function theorem guarantees the existence of a function $D(y_{(0)})$ such that

$$g(y_{(0)}, D(y_{(0)})) = 0.$$

The derivative of D can in a straightforward way be computed as

$$D'(y_{(0)}) = -\frac{\partial_y g(y_{(0)}, D(y_{(0)}))}{\partial_z g(y_{(0)}, D(y_{(0)}))}. \tag{7}$$

The equation that $y_{(0)}$ and $z_{(0)}$ are supposed to fulfill can hence be reformulated as

$$\frac{d}{dt} y_{(0)} = f(y_{(0)}, D(y_{(0)})), \qquad z_{(0)} = D(y_{(0)}), \qquad 0 < t < t_{end}. \tag{8}$$

3. IMEX Runge-Kutta and RS-IMEX

IMEX methods are based on a splitting of (1) into stiff $\widetilde{(\cdot)}$ and non-stiff contributions $\widehat{(\cdot)}$. More precisely, we split the equation into

$$\frac{d}{dt} \begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} f(y, z) \\ \frac{1}{\varepsilon} g(y, z) \end{pmatrix} = \begin{pmatrix} \widetilde{f}(y, z) \\ \frac{1}{\varepsilon} \widetilde{g}(y, z) \end{pmatrix} + \begin{pmatrix} \widehat{f}(y, z) \\ \frac{1}{\varepsilon} \widehat{g}(y, z) \end{pmatrix}. \tag{9}$$

$$\begin{array}{c|cc|cc} 0 & 0 & 0 & 0 & 0 \\ \boldsymbol{c} & \widetilde{\boldsymbol{\alpha}} & \widetilde{\boldsymbol{A}} & \widehat{\boldsymbol{\alpha}} & \widehat{\boldsymbol{A}} \end{array}$$

Table 1: The Butcher tableau for an IMEX Runge-Kutta method as given in Def. 1 with $\widetilde{\boldsymbol{\alpha}} = (\widetilde{\alpha}^1, \dots, \widetilde{\alpha}^{s-1})^T$, $\widehat{\boldsymbol{\alpha}} = (\widehat{\alpha}^1, \dots, \widehat{\alpha}^{s-1})^T$ and $\boldsymbol{c} = (c^1, \dots, c^{s-1})^T$, $\widetilde{\boldsymbol{A}} = (\widetilde{A})_{i,j=1}^{s-1,s-1}$ and $\widehat{\boldsymbol{A}} = (\widehat{A})_{i,j=1}^{s-1,s-1}$. We assume that $\widetilde{\boldsymbol{A}}$ is an invertible lower triangular matrix and that $\widehat{\boldsymbol{A}}$ is a strictly lower triangular matrix. Furthermore, the vector of internal time instances $\boldsymbol{c}$ fulfills $\boldsymbol{c} = \widetilde{\boldsymbol{A}} \cdot (1, \dots, 1)^T + \widetilde{\boldsymbol{\alpha}} = \widehat{\boldsymbol{A}} \cdot (1, \dots, 1)^T + \widehat{\boldsymbol{\alpha}}$ by definition.

Tilde terms $\widetilde{(\cdot)}$ are solved through an implicit and hat terms $\widehat{(\cdot)}$ through an explicit method. In this work, we focus on globally stiffly accurate (GSA) IMEX Runge-Kutta methods of type CK [9]:

Definition 1 (IMEX-RK type CK GSA with uniform $\boldsymbol{c}$)**.** *Consider an s-stage IMEX Runge-Kutta method with coefficient vectors $\widetilde{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\alpha}}$ and $\boldsymbol{c} \in \mathbb{R}^{s-1}$; and coefficient matrices $\widetilde{\boldsymbol{A}}$ and $\widehat{\boldsymbol{A}} \in \mathbb{R}^{(s-1)\times(s-1)}$. We assume that $\widehat{\boldsymbol{A}}$ is strictly lower triangular and $\widetilde{\boldsymbol{A}}$ is invertible and lower triangular. For every $t^{n+1} = t^n + \Delta t$ do the following:*

1. *Set $y^{n,1} = y^n$ and $z^{n,1} = z^n$. In the following we use y^n and z^n for the first internal stage.*
2. *For $i = 2, \dots, s$ solve*

$$\begin{aligned} \begin{pmatrix} y^{n,i} \\ z^{n,i} \end{pmatrix} = \begin{pmatrix} y^n \\ z^n \end{pmatrix} &+ \Delta t \left(\widetilde{\alpha}^{i-1} \begin{pmatrix} \widetilde{f}(y^n, z^n) \\ \frac{1}{\varepsilon}\widetilde{g}(y^n, z^n) \end{pmatrix} + \widehat{\alpha}^{i-1} \begin{pmatrix} \widehat{f}(y^n, z^n) \\ \frac{1}{\varepsilon}\widehat{g}(y^n, z^n) \end{pmatrix} \right) \\ &+ \Delta t \left(\sum_{j=2}^{i} \widetilde{A}^{i-1,j-1} \begin{pmatrix} \widetilde{f}(y^{n,j}, z^{n,j}) \\ \frac{1}{\varepsilon}\widetilde{g}(y^{n,j}, z^{n,j}) \end{pmatrix} + \sum_{j=2}^{i-1} \widehat{A}^{i-1,j-1} \begin{pmatrix} \widehat{f}(y^{n,j}, z^{n,j}) \\ \frac{1}{\varepsilon}\widehat{g}(y^{n,j}, z^{n,j}) \end{pmatrix} \right), \end{aligned} \tag{10}$$

   where $(y^{n,i}, z^{n,i})^T$ denotes the solution of the i^{th} internal stage. Note that all splitting functions also depend on the internal time instance

$$t^{n,j} := t^n + c^j \Delta t.$$

3. *Set $(y^{n+1}, z^{n+1})^T = (y^{n,s}, z^{n,s})^T$.*

We can give the coefficients of an IMEX Runge-Kutta method as given in Def. 1 in an extended Butcher tableau, see Tbl. 1 for more details. The Butcher tableaux of the methods used in this contribution, taken from [8] and [27], are given in Appendix B.

Remark 1. *Let us shortly comment on the used abbreviations:*

- *Methods of type CK (Carpenter, Kennedy) have the property that their Butcher tableaux are as in Tbl. 1 with* invertible $\widetilde{\boldsymbol{A}}$*, see [9]; or [20] for an overview.*
- *IMEX-RK methods are termed* globally stiffly accurate *(GSA), if the last stage is equal to the update step, see [27]. Here, this is automatically fulfilled by construction.*
- *We only consider IMEX Runge-Kutta schemes with uniform $\boldsymbol{c}$. This restriction is necessary for the analysis to follow, see also Remark 5 for more details.*

An ubiquitous property for Runge-Kutta methods in the context of singularly perturbed equations is the stage order [19, 29]. We also need what we call the *internal order* in the sequel:

Definition 2 (Stage and interal order)**.** *Assume that a given IMEX Runge-Kutta method has classical order of convergence of p. We say that the i^{th} ($1 \le i \le s$) internal stage of an IMEX Runge-Kutta method has internal order q^i if there holds*

$$y^{n,i} - y(t^n + c^i \Delta t) = O(\Delta t^p) + O(\Delta t^{q^i+1}),$$

for the exact (assumed smooth) solution y. The stage order q of the method is defined to be minimum of the internal orders,

$$q := \min_{1 \le i \le s} q^i.$$

In a similar manner, $\widetilde{q}$ is defined to be the stage order of the implicit part of the IMEX Runge-Kutta method.

The IMEX Runge-Kutta method defined in Def. 1 is formulated stage-wise. Frequently, it can be convenient to rewrite the method in vector notation. Therefore, we introduce the following additional notation:

Remark 2 (Notation)**.** 1. *The solution vectors* $\boldsymbol{y}^\Delta \in \mathbb{R}^{s-1}$ *and* $\boldsymbol{z}^\Delta \in \mathbb{R}^{s-1}$ *are given by*

$$\boldsymbol{y}^\Delta = \left(y^{n,2}, \dots, y^{n,s}\right)^T \qquad \text{and} \qquad \boldsymbol{z}^\Delta = \left(z^{n,2}, \dots, z^{n,s}\right)^T .$$

Note that $y^{n,1} = y^n$ *and* $z^{n,1} = z^n$ *due to the structure of the method, which is why* $\boldsymbol{y}^\Delta$ *and* $\boldsymbol{z}^\Delta$ *are* $s-1-$*dimensional vectors.*

2. *The vector* $\boldsymbol{e} \in \mathbb{R}^{s-1}$ *denotes the* $s-1$ *dimensional vector filled with ones, i.e.* $\boldsymbol{e} = (1, \dots, 1)^T \in \mathbb{R}^{s-1}$.
3. *Considering two vectors of size* $s-1$, $\boldsymbol{y}^\Delta \in \mathbb{R}^{s-1}$ *and* $\boldsymbol{z}^\Delta \in \mathbb{R}^{s-1}$, *then, e.g.,* $g(\boldsymbol{y}^\Delta, \boldsymbol{z}^\Delta)$ *and* $D(\boldsymbol{y}^\Delta)$ *denote element-wise applications of g and D, respectively, thus*

$$g(\boldsymbol{y}^\Delta, \boldsymbol{z}^\Delta) = \begin{pmatrix} g(y^{n,2}, z^{n,2}) \\ \vdots \\ g(y^{n,s}, z^{n,s}) \end{pmatrix} \qquad \text{or} \qquad D(\boldsymbol{y}^\Delta) = \begin{pmatrix} D(y^{n,2}) \\ \vdots \\ D(y^{n,s}) \end{pmatrix}.$$

Based on this remark, the IMEX Runge-Kutta method given in Def. 1 can be written as

$$\begin{aligned} \boldsymbol{y}^\Delta =& y^n \boldsymbol{e} + \Delta t \left(\widetilde{\boldsymbol{\alpha}} \widetilde{f}(y^n, z^n) + \widehat{\boldsymbol{\alpha}} \widehat{f}(y^n, z^n)\right) + \Delta t \left(\widetilde{\boldsymbol{A}} \widetilde{f}(\boldsymbol{y}^\Delta, \boldsymbol{z}^\Delta) + \widehat{\boldsymbol{A}} \widehat{f}(\boldsymbol{y}^\Delta, \boldsymbol{z}^\Delta)\right) \\ \boldsymbol{z}^\Delta =& z^n \boldsymbol{e} + \frac{\Delta t}{\varepsilon} \left(\widetilde{\boldsymbol{\alpha}} \widetilde{g}(y^n, z^n) + \widehat{\boldsymbol{\alpha}} \widehat{g}(y^n, z^n)\right) + \frac{\Delta t}{\varepsilon} \left(\widetilde{\boldsymbol{A}} \widetilde{g}(\boldsymbol{y}^\Delta, \boldsymbol{z}^\Delta) + \widehat{\boldsymbol{A}} \widehat{g}(\boldsymbol{y}^\Delta, \boldsymbol{z}^\Delta)\right). \end{aligned} \tag{11}$$

Obviously, one key part of IMEX methods is the choice of the splitting of f and g into 'implicit' and 'explicit' parts. To obtain a reasonable method, this splitting should have a certain set of properties:

- The splitting should be consistent with the equation.
- The explicit part should be non-stiff, meaning that it can be solved by an explicit method with an ε-independent restriction on the time step.
- The implicit part should be easy to solve, ideally, it should be linear.
- The resulting method should be stable and accurate for $\Delta t < \Delta t_0$ with Δt_0 not depending on ε.
- The resulting method should be consistent with the $\varepsilon \to 0$ limit.

In recent years, several splittings have been developed for different types of equations. The splitting that we have developed over the last few years [18] heavily relies on the $\varepsilon \to 0$ limit of the equation, being called *reference solution*:

Definition 3 (Reference solution)**.** *We call the* $\varepsilon \to 0$ *limiting solutions* reference solutions*, i.e.,*

$$y_{ref} := \lim_{\varepsilon \to 0} y \qquad \text{and} \qquad z_{ref} := \lim_{\varepsilon \to 0} z.$$

Note that y_{ref} *and* z_{ref} *are solutions to equation* (4)*. Following the notation of Rem. 2 we define on the* $n-$*th time slab*

$$\boldsymbol{y}_{ref} := \begin{pmatrix} y_{ref}(t^n + c^1 \Delta t) \\ \vdots \\ y_{ref}(t^n + c^{s-1} \Delta t) \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{z}_{ref} := \begin{pmatrix} z_{ref}(t^n + c^1 \Delta t) \\ \vdots \\ z_{ref}(t^n + c^{s-1} \Delta t) \end{pmatrix},$$

where c^i, $1 \le i \le s-1$, *corresponds to the internal stages of the IMEX Runge-Kutta method given in Def. 1.*

Remark 3. *Unlike in* (3)*, we have chosen to denote the limit solutions by* y_{ref} *and* z_{ref}*, respectively, and not by the* (0) *index. This is due to notational simplicity in the proofs to follow. Equation* (3) *can hence be rewritten as*

$$\begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} y_{ref} \\ z_{ref} \end{pmatrix} + \varepsilon \begin{pmatrix} y_{(1)} \\ z_{(1)} \end{pmatrix} + \varepsilon^2 \begin{pmatrix} y_{(2)} \\ z_{(2)} \end{pmatrix} + O(\varepsilon^3). \tag{3'}$$

The solution $(y,z)^T$ can therefore be considered as a perturbation of the reference solution for a small ε. The idea of the RS-IMEX method is thus to use a linearization of the fluxes f and g around this reference state:

Definition 4 (RS-IMEX splitting)**.** *The RS-IMEX splitting of equation* (9) *is given by*

$$\begin{pmatrix} \widetilde{f}(y,z) \\ \frac{1}{\varepsilon}\widetilde{g}(y,z) \end{pmatrix} := \begin{pmatrix} f(y_{ref},z_{ref}) \\ \frac{1}{\varepsilon}g(y_{ref},z_{ref}) \end{pmatrix} + \partial_y \begin{pmatrix} f(y_{ref},z_{ref}) \\ \frac{1}{\varepsilon}g(y_{ref},z_{ref}) \end{pmatrix}\Big(y-y_{ref}\Big) + \partial_z \begin{pmatrix} f(y_{ref},z_{ref}) \\ \frac{1}{\varepsilon}g(y_{ref},z_{ref}) \end{pmatrix}\Big(z-z_{ref}\Big) \tag{implicit}$$

and

$$\begin{pmatrix} \widehat{f}(y,z) \\ \frac{1}{\varepsilon}\widehat{g}(y,z) \end{pmatrix} := \begin{pmatrix} f(y,z) \\ \frac{1}{\varepsilon}g(y,z) \end{pmatrix} - \begin{pmatrix} \widetilde{f}(y,z) \\ \frac{1}{\varepsilon}\widetilde{g}(y,z) \end{pmatrix}. \tag{explicit}$$

4. Error analysis

In this section, we perform an error analysis of our method similar to the one in [19] and [20].

4.1. Statement of the main theorem

Theorem 1. *We consider a globally stiffly accurate IMEX Runge-Kutta method of type CK coupled with the RS-IMEX splitting and assume that $\Delta t \gg \varepsilon$, then the error at time instance t^{n+1} is given by*

$$\begin{pmatrix} y^{n+1}-y(t^{n+1}) \\ z^{n+1}-z(t^{n+1}) \end{pmatrix} = \begin{pmatrix} O(\Delta t^{r_1}) \\ O(\Delta t^{r_1}) \end{pmatrix} + \varepsilon \begin{pmatrix} O(\Delta t^{r_2+1}) \\ O(\Delta t^{r_2}) \end{pmatrix} + \varepsilon^2 \begin{pmatrix} O(\Delta t^{r_2}) \\ O(\Delta t^{r_2-1}) \end{pmatrix} + O(\varepsilon^3), \tag{12}$$

where the constants r_1 and r_2 are given by

$$r_1 := \min(p, 2(q+1)) \qquad \textit{and} \qquad r_2 := \min(q+1, \widetilde{q}).$$

p denotes the order of convergence of the overall method, q the stage order and $\widetilde{q}$ the implicit stage order.

Proof. We will prove this theorem in multiple steps. Our point of departure is the IMEX Runge-Kutta method in vector notation, see (11). Following the steps in [19] and [20] we assume that all quantities can be represented by their asymptotic expansions, i.e.

$$\boldsymbol{y}^\Delta = \boldsymbol{y}^\Delta_{(0)} + \varepsilon \boldsymbol{y}^\Delta_{(1)} + \varepsilon^2 \boldsymbol{y}^\Delta_{(2)} + O(\varepsilon^3) \qquad \text{and} \qquad z^\Delta = z^\Delta_{(0)} + \varepsilon \boldsymbol{z}^\Delta_{(1)} + \varepsilon^2 \boldsymbol{z}^\Delta_{(2)} + O(\varepsilon^3). \tag{13}$$

Consequently, we can use the asymptotic expansion of the numerical solution (13) and the asymptotic expansion of the exact solution (3) to split the error in orders of ε, i.e.

$$\begin{pmatrix} y^{n+1}-y(t^{n+1}) \\ z^{n+1}-z(t^{n+1}) \end{pmatrix} = \begin{pmatrix} y^{n+1}_{(0)}-y_{(0)}(t^{n+1}) \\ z^{n+1}_{(0)}-z_{(0)}(t^{n+1}) \end{pmatrix} + \varepsilon \begin{pmatrix} y^{n+1}_{(1)}-y_{(1)}(t^{n+1}) \\ z^{n+1}_{(1)}-z_{(1)}(t^{n+1}) \end{pmatrix} + \varepsilon^2 \begin{pmatrix} y^{n+1}_{(2)}-y_{(2)}(t^{n+1}) \\ z^{n+1}_{(2)}-z_{(2)}(t^{n+1}) \end{pmatrix} + O(\varepsilon^3). \tag{14}$$

Similarly as before, see (4)-(6), we obtain the following defining equation for $\boldsymbol{y}^\Delta_{(i)}$ and $\boldsymbol{z}^\Delta_{(i)}$, $0 \le i \le 2$:

$$\begin{aligned}
\boldsymbol{y}^\Delta_{(0)} &= y^n_{(0)}\boldsymbol{e} + \Delta t\left(\widetilde{\boldsymbol{\alpha}}\widetilde{f}_{(0)}(y^n_{(0)}, z^n_{(0)}) + \widehat{\boldsymbol{\alpha}}\widehat{f}_{(0)}(y^n_{(0)}, z^n_{(0)})\right) + \Delta t\left(\widetilde{\boldsymbol{A}}\widetilde{f}_{(0)}(\boldsymbol{y}^\Delta_{(0)}, \boldsymbol{z}^\Delta_{(0)}) + \widehat{\boldsymbol{A}}\widehat{f}_{(0)}(\boldsymbol{y}^\Delta_{(0)}, \boldsymbol{z}^\Delta_{(0)})\right) \\
0 &= \widetilde{\boldsymbol{\alpha}}\widetilde{g}_{(0)}(y^n_{(0)}, z^n_{(0)}) + \widehat{\boldsymbol{\alpha}}\widehat{g}_{(0)}(y^n_{(0)}, z^n_{(0)}) + \widetilde{\boldsymbol{A}}\widetilde{g}_{(0)}(\boldsymbol{y}^\Delta_{(0)}, \boldsymbol{z}^\Delta_{(0)}) + \widehat{\boldsymbol{A}}\widehat{g}_{(0)}(\boldsymbol{y}^\Delta_{(0)}, \boldsymbol{z}^\Delta_{(0)}),
\end{aligned} \tag{15}$$

$$\begin{aligned}
\boldsymbol{y}^\Delta_{(1)} &= y^n_{(1)}\boldsymbol{e} + \Delta t\left(\widetilde{\boldsymbol{\alpha}}\widetilde{f}_{(1)}(y^n_{(0)}, z^n_{(0)}, y^n_{(1)}, z^n_{(1)}) + \widehat{\boldsymbol{\alpha}}\widehat{f}_{(1)}(y^n_{(0)}, z^n_{(0)}, y^n_{(1)}, z^n_{(1)})\right) \\
&\quad + \Delta t\left(\widetilde{\boldsymbol{A}}\widetilde{f}_{(1)}(\boldsymbol{y}^\Delta_{(0)}, \boldsymbol{z}^\Delta_{(0)}, \boldsymbol{y}^\Delta_{(1)}, \boldsymbol{z}^\Delta_{(1)}) + \widehat{\boldsymbol{A}}\widehat{f}_{(1)}(\boldsymbol{y}^\Delta_{(0)}, \boldsymbol{z}^\Delta_{(0)}, \boldsymbol{y}^\Delta_{(1)}, \boldsymbol{z}^\Delta_{(1)})\right) \\
\boldsymbol{z}^\Delta_{(0)} &= z^n_{(0)}\boldsymbol{e} + \Delta t\left(\widetilde{\boldsymbol{\alpha}}\widetilde{g}_{(1)}(y^n_{(0)}, z^n_{(0)}, y^n_{(1)}, z^n_{(1)}) + \widehat{\boldsymbol{\alpha}}\widehat{g}_{(1)}(y^n_{(0)}, z^n_{(0)}, y^n_{(1)}, z^n_{(1)})\right) \\
&\quad + \Delta t\left(\widetilde{\boldsymbol{A}}\widetilde{g}_{(1)}(\boldsymbol{y}^\Delta_{(0)}, \boldsymbol{z}^\Delta_{(0)}, \boldsymbol{y}^\Delta_{(1)}, \boldsymbol{z}^\Delta_{(1)}) + \widehat{\boldsymbol{A}}\widehat{g}_{(1)}(\boldsymbol{y}^\Delta_{(0)}, \boldsymbol{z}^\Delta_{(0)}, \boldsymbol{y}^\Delta_{(1)}, \boldsymbol{z}^\Delta_{(1)})\right),
\end{aligned} \tag{16}$$

$$\begin{aligned}
\boldsymbol{y}^\Delta_{(2)} &= y^n_{(2)}\boldsymbol{e} + \Delta t\left(\widetilde{\boldsymbol{\alpha}}\widetilde{f}_{(2)}(y^n_{(0)}, z^n_{(0)}, y^n_{(1)}, z^n_{(1)}, y^n_{(2)}, z^n_{(2)}) + \widehat{\boldsymbol{\alpha}}\widehat{f}_{(2)}(y^n_{(0)}, z^n_{(0)}, y^n_{(1)}, z^n_{(1)}, y^n_{(2)}, z^n_{(2)})\right) \\
&\quad + \Delta t\left(\widetilde{\boldsymbol{A}}\widetilde{f}_{(2)}(\boldsymbol{y}^\Delta_{(0)}, \boldsymbol{z}^\Delta_{(0)}, \boldsymbol{y}^\Delta_{(1)}, \boldsymbol{z}^\Delta_{(1)}, \boldsymbol{y}^\Delta_{(2)}, \boldsymbol{z}^\Delta_{(2)}) + \widehat{\boldsymbol{A}}\widehat{f}_{(2)}(\boldsymbol{y}^\Delta_{(0)}, \boldsymbol{z}^\Delta_{(0)}, \boldsymbol{y}^\Delta_{(1)}, \boldsymbol{z}^\Delta_{(1)}, \boldsymbol{y}^\Delta_{(2)}, \boldsymbol{z}^\Delta_{(2)})\right) \\
\boldsymbol{z}^\Delta_{(1)} &= z^n_{(1)}\boldsymbol{e} + \Delta t\left(\widetilde{\boldsymbol{\alpha}}\widetilde{g}_{(2)}(y^n_{(0)}, z^n_{(0)}, y^n_{(1)}, z^n_{(1)}, y^n_{(2)}, z^n_{(2)}) + \widehat{\boldsymbol{\alpha}}\widehat{g}_{(2)}(y^n_{(0)}, z^n_{(0)}, y^n_{(1)}, z^n_{(1)}, y^n_{(2)}, z^n_{(2)})\right) \\
&\quad + \Delta t\left(\widetilde{\boldsymbol{A}}\widetilde{g}_{(2)}(\boldsymbol{y}^\Delta_{(0)}, \boldsymbol{z}^\Delta_{(0)}, \boldsymbol{y}^\Delta_{(1)}, \boldsymbol{z}^\Delta_{(1)}, \boldsymbol{y}^\Delta_{(2)}, \boldsymbol{z}^\Delta_{(2)}) + \widehat{\boldsymbol{A}}\widehat{g}_{(2)}(\boldsymbol{y}^\Delta_{(0)}, \boldsymbol{z}^\Delta_{(0)}, \boldsymbol{y}^\Delta_{(1)}, \boldsymbol{z}^\Delta_{(1)}, \boldsymbol{y}^\Delta_{(2)}, \boldsymbol{z}^\Delta_{(2)})\right).
\end{aligned} \tag{17}$$

The quantities $\widetilde{f_{(0)}}$, $\widehat{f_{(0)}}$, ... will be introduced in subsequent sections, they are the terms of the straightforward expansion of $\widetilde{f}$, $\widehat{f}$, ... with respect to ε. Due to Thms. 2, 3 and 4 to be proved later, there holds

$$\begin{pmatrix} y_{(0)}^{n+1} - y_{(0)}(t^{n+1}) \\ z_{(0)}^{n+1} - z_{(0)}(t^{n+1}) \end{pmatrix} = \begin{pmatrix} O(\Delta t^{r_1}) \\ O(\Delta t^{r_1}) \end{pmatrix}, \qquad \begin{pmatrix} y_{(1)}^{n+1} - y_{(1)}(t^{n+1}) \\ z_{(1)}^{n+1} - z_{(1)}(t^{n+1}) \end{pmatrix} = \begin{pmatrix} O(\Delta t^{r_2+1}) \\ O(\Delta t^{r_2}) \end{pmatrix}, \quad \text{and} \quad \begin{pmatrix} y_{(2)}^{n+1} - y_{(2)}(t^{n+1}) \\ z_{(2)}^{n+1} - z_{(2)}(t^{n+1}) \end{pmatrix} = \begin{pmatrix} O(\Delta t^{r_2}) \\ O(\Delta t^{r_2-1}) \end{pmatrix}.$$

Together with (14), this concludes the proof. □

The only thing that remains is to formulate and prove the aforementioned theorems. Because the equations (15), (16) and (17) depend hierarchically on each other, we start with the lowest order one, namely (15).

Remark 4. *In the analysis to come, we always work with exact reference solutions y_{ref} and z_{ref}. In practice, one would replace these functions by a suitable approximation. We have seen in [22] that the influence of this approximation is rather limited.*

4.2. Differential algebraic equation: The error in $\boldsymbol{y}_{(0)}^{\Delta}$ and $\boldsymbol{z}_{(0)}^{\Delta}$

Before proceeding any further, we remind the reader of Assumption 1, guaranteeing that the initial conditions for $y_{(0)}$ and $z_{(0)}$ are consistent with the DAE (4). In this subsection, we analyze the discretization error of both $\boldsymbol{y}_{(0)}^{\Delta}$ and $\boldsymbol{z}_{(0)}^{\Delta}$. Those quantities are defined by (15), with the expansion of the splitted flux functions (see also Def. 4) given by

$$\begin{aligned}
\widetilde{f_{(0)}} =& f(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}}) + \text{Diag}\left\{\partial_y f(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right\}(\boldsymbol{y}_{(0)}^{\Delta} - \boldsymbol{y}_{\text{ref}}) + \text{Diag}\left\{\partial_z f(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right\}(\boldsymbol{z}_{(0)}^{\Delta} - \boldsymbol{z}_{\text{ref}}) \\
\widetilde{g_{(0)}} =& g(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}}) + \text{Diag}\left\{\partial_y g(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right\}(\boldsymbol{y}_{(0)}^{\Delta} - \boldsymbol{y}_{\text{ref}}) + \text{Diag}\left\{\partial_z g(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right\}(\boldsymbol{z}_{(0)}^{\Delta} - \boldsymbol{z}_{\text{ref}}) \\
\widehat{f_{(0)}} =& f(\boldsymbol{y}_{(0)}^{\Delta}, \boldsymbol{z}_{(0)}^{\Delta}) - f(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}}) - \text{Diag}\left\{\partial_y f(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right\}(\boldsymbol{y}_{(0)}^{\Delta} - \boldsymbol{y}_{\text{ref}}) - \text{Diag}\left\{\partial_z f(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right\}(\boldsymbol{z}_{(0)}^{\Delta} - \boldsymbol{z}_{\text{ref}}) \\
\widehat{g_{(0)}} =& g(\boldsymbol{y}_{(0)}^{\Delta}, \boldsymbol{z}_{(0)}^{\Delta}) - g(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}}) - \text{Diag}\left\{\partial_y g(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right\}(\boldsymbol{y}_{(0)}^{\Delta} - \boldsymbol{y}_{\text{ref}}) - \text{Diag}\left\{\partial_z g(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right\}(\boldsymbol{z}_{(0)}^{\Delta} - \boldsymbol{z}_{\text{ref}})
\end{aligned} \tag{18}$$

and similar for $y_{(0)}^{n}$ and $z_{(0)}^{n}$.

Theorem 2. *The numerical solution of* (15) *fulfills the error estimate*

$$\begin{pmatrix} y_{(0)}^{n+1} - y_{(0)}(t^{n+1}) \\ z_{(0)}^{n+1} - z_{(0)}(t^{n+1}) \end{pmatrix} = \begin{pmatrix} O(\Delta t^{r_1}) \\ O(\Delta t^{r_1}) \end{pmatrix}. \tag{19}$$

Proof. The roadmap for the proof is the following:

- We consider an RS-IMEX discretization of the differential equation for $y_{(0)}$ in (8), see La. 1. It is obvious that this discretization is $p-$th order convergent if the IMEX Runge-Kutta method is. Note that unfortunately, this method is not equivalent to (15).
- We then prove that $\boldsymbol{y}_{(0)}^{\Delta}$ can be viewed as a perturbation to the aforementioned IMEX discretization of the ODE in (8) with perturbation error of $O(\Delta t^{r_1})$, see La. 5. This then yields the desired order of convergence in $y_{(0)}$.
- Together with a representation lemma for $\boldsymbol{z}_{(0)}^{\Delta}$, see La. 2, we can show that also $z_{(0)}$ converges with order r_1, see La. 3.

□

As mentioned before, we begin by discretizing the ODE for $y_{(0)}$ in (8) by its own IMEX discretization.

Lemma 1 (RS-IMEX method for the ODE in (8))**.** *The RS-IMEX discretization coupled with an IMEX Runge-Kutta method as given in Def. 1 for the ODE in* (8) *is given by*

$$\begin{aligned}
\mathfrak{y}^{\Delta} =& \mathfrak{y}^{n}\boldsymbol{e} + \Delta t\widehat{\alpha}\left(f(\mathfrak{y}^{n}, D(\mathfrak{y}^{n})) - f(y_{ref}, D(y_{ref})) - \left(\partial_y f(y_{ref}, D(y_{ref})) + \partial_z f(y_{ref}, D(y_{ref}))D'(y_{ref})\right)(\mathfrak{y}^{n} - y_{ref})\right) \\
&+ \Delta t\widetilde{\alpha}\left(f(y_{ref}, D(y_{ref})) + \left(\partial_y f(y_{ref}, D(y_{ref})) + \partial_z f(y_{ref}, D(y_{ref}))D'(y_{ref})\right)(\mathfrak{y}^{n} - y_{ref})\right) \\
&+ \Delta t\widehat{A}\left(f(\mathfrak{y}^{\Delta}, D(\mathfrak{y}^{\Delta})) - f(\boldsymbol{y}_{ref}, D(\boldsymbol{y}_{ref})) - \text{Diag}\left\{\partial_y f(\boldsymbol{y}_{ref}, D(\boldsymbol{y}_{ref}))\right\} + \text{Diag}\left\{\partial_z f(\boldsymbol{y}_{ref}, D(\boldsymbol{y}_{ref}))\right\}D'(\boldsymbol{y}_{ref})\right\}(\mathfrak{y}^{\Delta} - \boldsymbol{y}_{ref})\right) \\
&+ \Delta t\widetilde{A}\left(f(\boldsymbol{y}_{ref}, D(\boldsymbol{y}_{ref})) + \text{Diag}\left\{\partial_y f(\boldsymbol{y}_{ref}, D(\boldsymbol{y}_{ref})) + \text{Diag}\left\{\partial_z f(\boldsymbol{y}_{ref}, D(\boldsymbol{y}_{ref}))\right\}D'(\boldsymbol{y}_{ref})\right\}(\mathfrak{y}^{\Delta} - \boldsymbol{y}_{ref})\right),
\end{aligned}$$

where $\mathfrak{y}^{n}$ denotes the solution at the previous time instance and $\mathfrak{y}^{\Delta}$ the solution vector.

Proof. The flux of the ODE is given by $y \mapsto f(y, D(y))$. Linearizing around y_{ref} yields

$$f(y_{\text{ref}}) + \left(\partial_y f(y_{\text{ref}}, D(y_{\text{ref}})) + \partial_z f(y_{\text{ref}}, D(y_{\text{ref}}))D'(y_{\text{ref}})\right)(y - y_{\text{ref}}).$$

Treating this term implicitly and the remaining terms explicitly (so applying the RS-IMEX technique from Def. 4) and using the IMEX Runge-Kutta method from Def. 1 yields the desired result. □

We continue first with a more precise investigation of $\boldsymbol{z}^{\Delta}_{(0)}$, subsequently, we treat $\boldsymbol{y}^{\Delta}_{(0)}$.

4.2.1. Representation and error of $\boldsymbol{z}^{\Delta}_{(0)}$

Lemma 2. *Assume that $\boldsymbol{y}^{\Delta} - \boldsymbol{y}_{ref} = O(\Delta t^{q+1}\boldsymbol{e})$ and that the errors in both $y^n_{(0)}$ and $z^n_{(0)}$ are in $O(\Delta t^{r_1})$. Then, $\boldsymbol{z}^{\Delta}_{(0)}$ can be represented by*

$$\boldsymbol{z}^{\Delta}_{(0)} = D(\boldsymbol{y}_{ref}) + \operatorname{Diag}\left\{D'(\boldsymbol{y}_{ref})\right\}(\boldsymbol{y}^{\Delta}_{(0)} - \boldsymbol{y}_{ref}) + O(\Delta t^{r_1}\boldsymbol{e}). \tag{20}$$

Proof. Due to the assumption that the errors in $y^n_{(0)}$ and $z^n_{(0)}$ are in $O(\Delta t^{r_1})$, one can compute

$$\widetilde{g}_{(0)}(y^n_{(0)}, z^n_{(0)}) = g(y_{\text{ref}}, z_{\text{ref}}) + \operatorname{Diag}\left\{\partial_y g(y_{\text{ref}}, z_{\text{ref}})\right\}(y^n_{(0)} - y_{\text{ref}}) + \operatorname{Diag}\left\{\partial_z g(y_{\text{ref}}, z_{\text{ref}})\right\}(z^n_{(0)} - z_{\text{ref}}) = O(\Delta t^{r_1}),$$

and similarly for $\widehat{g}(y^n_{(0)}, z^n_{(0)})$. Note that $g(y_{\text{ref}}, z_{\text{ref}}) = 0$ due to the fact that y_{ref} and z_{ref} denote the exact solutions to the limit equation (4). We can then rewrite (15) with fluxes in (18) as

$$\begin{aligned} 0 =& \widehat{\boldsymbol{A}}\left(g(\boldsymbol{y}^{\Delta}_{(0)}, \boldsymbol{z}^{\Delta}_{(0)}) - \operatorname{Diag}\left\{\partial_y g(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right\}(\boldsymbol{y}^{\Delta}_{(0)} - \boldsymbol{y}_{\text{ref}}) - \operatorname{Diag}\left\{\partial_z g(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right\}(\boldsymbol{z}^{\Delta}_{(0)} - \boldsymbol{z}_{\text{ref}})\right) \\ &+ \widetilde{\boldsymbol{A}}\left(\operatorname{Diag}\left\{\partial_y g(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right\}(\boldsymbol{y}^{\Delta}_{(0)} - \boldsymbol{y}_{\text{ref}}) + \operatorname{Diag}\left\{\partial_z g(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right\}(\boldsymbol{z}^{\Delta}_{(0)} - \boldsymbol{z}_{\text{ref}})\right) + O(\Delta t^{r_1}). \end{aligned} \tag{21}$$

Note that due to the construction of our IMEX Runge-Kutta scheme, the temporal instances at which the reference solutions are evaluated, are the same. (There is no $\widehat{\boldsymbol{c}}$ or $\widetilde{\boldsymbol{c}}$ in Def. 1, only a $\boldsymbol{c}$.) Therefore we can collect all $\boldsymbol{z}^{\Delta}_{(0)} - \boldsymbol{z}_{\text{ref}}$ terms and multiply with $\operatorname{Diag}\left\{\partial_z g(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right\}^{-1}\left(\widetilde{\boldsymbol{A}} - \widehat{\boldsymbol{A}}\right)^{-1}$ (note that this inverse exists due to our restrictions on the Runge-Kutta method and on $\partial_z g$) to obtain

$$\begin{aligned} \boldsymbol{z}^{\Delta}_{(0)} =& \boldsymbol{z}_{\text{ref}} - \operatorname{Diag}\left\{\partial_z g(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right\}^{-1}\left(\widetilde{\boldsymbol{A}} - \widehat{\boldsymbol{A}}\right)^{-1}\widehat{\boldsymbol{A}}\left(g(\boldsymbol{y}^{\Delta}_{(0)}, \boldsymbol{z}^{\Delta}_{(0)})\right) \\ &- \operatorname{Diag}\left\{\partial_z g(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right\}^{-1}\operatorname{Diag}\left\{\partial_y g(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right\}(\boldsymbol{y}^{\Delta}_{(0)} - \boldsymbol{y}_{\text{ref}}) + O(\Delta t^{r_1}) \end{aligned} \tag{22}$$

Due to (7) we know that

$$\operatorname{Diag}\left\{D'(\boldsymbol{y}_{\text{ref}})\right\} = -\operatorname{Diag}\left\{\partial_z g(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right\}^{-1}\operatorname{Diag}\left\{\partial_y g(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right\}. \tag{23}$$

The matrix $\left(\widetilde{\boldsymbol{A}} - \widehat{\boldsymbol{A}}\right)^{-1}\widehat{\boldsymbol{A}}$ is a strictly lower triangular matrix, which means that in the i^{th} stage g is only evaluated at *previous* stages. Using a Taylor expansion in g and recursively plugging in (22) shows that $g(\boldsymbol{y}^{\Delta}_{(0)}, \boldsymbol{z}^{\Delta}_{(0)})$ is in $O(\Delta t^{2(q+1)})$, consult Las. 6 and 7 in the appendix for details. Altogether , this yields the desired result

$$\boldsymbol{z}^{\Delta}_{(0)} = D(\boldsymbol{y}_{\text{ref}}) + \operatorname{Diag}\left\{D'(\boldsymbol{y}_{\text{ref}})\right\}(\boldsymbol{y}^{\Delta}_{(0)} - \boldsymbol{y}_{\text{ref}}) + O(\Delta t^{r_1}), \tag{24}$$

with $D(\boldsymbol{y}_{\text{ref}}) = \boldsymbol{z}_{\text{ref}}$. □

Remark 5. *One of the severe restrictions on the chosen IMEX Runge-Kutta methods is the restriction on a uniform $\boldsymbol{c}$, i.e., $\widetilde{\boldsymbol{c}} = \widehat{\boldsymbol{c}}$. In the proof of the previous lemma, the importance of this restriction can be seen, because otherwise the temporal instance at which $\boldsymbol{z}_{ref}$ is evaluated would differ for the explicit and the implicit part, see also Def. 3. Therefore, in* (22) *we would not be able to collect terms in $\boldsymbol{z}_{ref}$.*

With this, we immediately get an error order for $z_{(0)}$:

Lemma 3. *Under the assumptions of La. 2 and the assumption* $y_{(0)}^{n,i} - y_{ref}(t^{n,i}) = O(\Delta t^{q^i+1}) + O(\Delta t^{r_1})$ *for* $i = 2, \dots, s$, *we obtain*

$$z_{(0)}^{n,s} - z_{ref}(t^{n+1}) = z_{(0)}^{n+1} - z_{ref}(t^{n+1}) = O(\Delta t^{r_1}). \tag{25}$$

Proof. Note that $D(\boldsymbol{y}_{\text{ref}}) = \boldsymbol{z}_{\text{ref}}$ and consider the last element of (20). Noting that $p = q^s$ because the method is globally stiffly accurate, this immediately proves the lemma. □

The following lemma gives an estimate on how accurate $\boldsymbol{z}_{(0)}^{\Delta}$ approximates $D(\boldsymbol{y}_{(0)}^{\Delta})$.

Lemma 4. *Under the assumptions of La. 2, there holds:*

$$z_{(0)}^{\Delta} - D(\boldsymbol{y}_{(0)}^{\Delta}) = O((\boldsymbol{y}_{(0)}^{\Delta} - \boldsymbol{y}_{ref})^2) + O(\Delta t^{r_1}). \tag{26}$$

Proof. We compute the Taylor expansion of $D(\boldsymbol{y}^{\Delta})$ around the reference solution $\boldsymbol{y}_{\text{ref}}$ up to terms in $O((\boldsymbol{y}_{(0)}^{\Delta} - \boldsymbol{y}_{\text{ref}})^2)$ and directly obtain the result by using La. 2. □

4.2.2. Error of $\boldsymbol{y}_{(0)}^{\Delta}$

After having obtained error orders for $z_{(0)}$ – always under the assumption that those of $y_{(0)}$ are 'good' enough, which we haven't proved yet – we now turn to the approximation quality of $y_{(0)}$. The core statement is the following:

Lemma 5. *The method for* $\boldsymbol{y}_{(0)}^{\Delta}$, *given in* (15), *is an* $O(\Delta t^{r_1})$ *perturbation of the method given in La. 1.*

Proof. The method (15) for $\boldsymbol{y}_{(0)}^{\Delta}$ can be written as

$$\begin{aligned}
\boldsymbol{y}^{\Delta}{}_{(0)} &= y_{(0)}^{n}\boldsymbol{e} \\
&+ \widehat{\boldsymbol{\alpha}}\Delta t\left(f(y_{(0)}^{n}, z_{(0)}^{n}) - f(y_{\text{ref}}^{0}, z_{\text{ref}}^{0}) - \partial_y f(y_{\text{ref}}^{0}, z_{\text{ref}}^{0})(y_{(0)}^{n} - y_{\text{ref}}^{0}) - \partial_z f(y_{\text{ref}}^{0}, z_{\text{ref}}^{0})(z_{(0)}^{n} - z_{\text{ref}}^{0})\right) \\
&+ \widetilde{\boldsymbol{\alpha}}\Delta t\left(f(y_{\text{ref}}^{0}, z_{\text{ref}}^{0}) + \partial_y f(y_{\text{ref}}^{0}, z_{\text{ref}}^{0})(y_{(0)}^{n} - y_{\text{ref}}^{0}) + \partial_z f(y_{\text{ref}}^{0}, z_{\text{ref}}^{0})(z_{(0)}^{n} - z_{\text{ref}}^{0})\right) \\
&+ \Delta t\widehat{\boldsymbol{A}}\left(f(\boldsymbol{y}_{(0)}^{\Delta}, \boldsymbol{z}_{(0)}^{\Delta}) - f(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}}) - \text{Diag}\left\{\partial_y f(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right\}(\boldsymbol{y}_{(0)}^{\Delta} - \boldsymbol{y}_{\text{ref}}) - \text{Diag}\left\{\partial_z f(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right\}(\boldsymbol{z}_{(0)}^{\Delta} - \boldsymbol{z}_{\text{ref}})\right) \\
&+ \Delta t\widetilde{\boldsymbol{A}}\left(f(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}}) + \text{Diag}\left\{\partial_y f(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right\}(\boldsymbol{y}_{(0)}^{\Delta} - \boldsymbol{y}_{\text{ref}}) + \text{Diag}\left\{\partial_z f(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right\}(\boldsymbol{z}_{(0)}^{\Delta} - \boldsymbol{z}_{\text{ref}})\right)
\end{aligned} \tag{27}$$

The trick is now to apply Las. 2 and 4 elementwise. The core assumption to these lemmas is $\boldsymbol{y}_{(0)}^{\Delta} - \boldsymbol{y}_{\text{ref}} = O(\Delta t^{q+1})$. However, by looking carefully at the proofs, due to the fact that the matrix $\left(\widetilde{\boldsymbol{A}} - \widehat{\boldsymbol{A}}\right)^{-1}\widehat{\boldsymbol{A}}$ is strictly lower triangular, for the $i-$th internal stage, the lemmas are valid given that only $y_{(0)}^{n,j} - y_{\text{ref}}(t^{n,j}) = O(\Delta t^{q+1})$ for $j < i$ is fulfilled. Therefore we can use the results from Las 2 and 4; the following calculation should be considered stage-wise. We plug the representation (20) into the above and rearrange terms to obtain

$$\begin{aligned}
\boldsymbol{y}^{\Delta}{}_{(0)} &= y_{(0)}^{n}\boldsymbol{e} \\
&+ \widehat{\boldsymbol{\alpha}}\Delta t\left(f(y_{(0)}^{n}, D(y_{(0)}^{n})) - f(y_{\text{ref}}^{0}, D(y_{\text{ref}}^{0})) - \left(\partial_y f(y_{\text{ref}}^{0}, D(y_{\text{ref}}^{0})) + \partial_z f(y_{\text{ref}}^{0}, D(y_{\text{ref}}^{0}))D'(y_{\text{ref}}^{0})\right)(y_{(0)}^{n} - y_{\text{ref}}^{0})\right) \\
&+ \widetilde{\boldsymbol{\alpha}}\Delta t\left(f(y_{\text{ref}}^{0}, D(y_{\text{ref}}^{0})) + \left(\partial_y f(y_{\text{ref}}^{0}, D(y_{\text{ref}}^{0})) + \partial_z f(y_{\text{ref}}^{0}, D(y_{\text{ref}}^{0}))D'(y_{\text{ref}}^{0})\right)(y_{(0)}^{n} - y_{\text{ref}}^{0})\right) \\
&+ \Delta t\widehat{\boldsymbol{A}}\left(f(\boldsymbol{y}_{(0)}^{\Delta}, D(\boldsymbol{y}_{(0)}^{\Delta})) - f(\boldsymbol{y}_{\text{ref}}, D(\boldsymbol{y}_{\text{ref}}))\right) \\
&+ \Delta t\widehat{\boldsymbol{A}}\left(-\text{Diag}\left\{\partial_y f(\boldsymbol{y}_{\text{ref}}, D(\boldsymbol{y}_{\text{ref}}))\right\} - \text{Diag}\left\{\partial_z f(\boldsymbol{y}_{\text{ref}}, D(\boldsymbol{y}_{\text{ref}}))\right\}\text{Diag}\left\{D'(\boldsymbol{y}_{\text{ref}})\right\}\right)(\boldsymbol{y}_{(0)}^{\Delta} - \boldsymbol{y}_{\text{ref}}) \\
&+ \Delta t\widetilde{\boldsymbol{A}}\left(f(\boldsymbol{y}_{\text{ref}}, D(\boldsymbol{y}_{\text{ref}})) + \left(\text{Diag}\left\{\partial_y f(\boldsymbol{y}_{\text{ref}}, D(\boldsymbol{y}_{\text{ref}}))\right\} - \text{Diag}\left\{\partial_z f(\boldsymbol{y}_{\text{ref}}, D(\boldsymbol{y}_{\text{ref}}))\right\}\text{Diag}\left\{D'(\boldsymbol{y}_{\text{ref}})\right\}\right)(\boldsymbol{y}_{(0)}^{\Delta} - \boldsymbol{y}_{\text{ref}})\right) \\
&+ O(\Delta t^{r_1+1}).
\end{aligned} \tag{28}$$

This formulation is equal to the RS-IMEX discretization given in Lemma 1 up to terms of order $O(\Delta t^{r_1})$. Thus we can write the error in the i$^{\text{th}}$ stage as

$$y_{(0)}^{n,i} - y_{\text{ref}}(t^{n,i}) = O(\Delta t^{q^i+1}) + O(\Delta t^{r_1}).$$

□

The following corollarys are now easy to show:

Corollary 1. *$\boldsymbol{y}^{\Delta}_{(0)}$ fulfills $\boldsymbol{y}^{\Delta}_{(0)} - \boldsymbol{y}_{ref} = O(\Delta t^{q+1})$.*

Corollary 2. *The error in the $y_{(0)}$ component is given by*

$$y^{n,s}_{(0)} - y_{ref}(t^{n+1}) = y^{n+1}_{(0)} - y_{ref}(t^{n+1}) = O(\Delta t^{r_1}). \tag{29}$$

Together with La. 3, this proves Thm. 2.

4.3. Higher order terms

In this section, we treat the higher-order expansions of the IMEX method. The analysis is a slight modification of the standard work of Hairer and Wanner [19, Theorem 3.4].

4.3.1. Error analysis of $y^{n+1}_{(1)}$ and $z^{n+1}_{(1)}$

The first order expansion is given in (16) together with the expansions of the splitting functions defined by

$$\begin{aligned}
\widetilde{f}_{(1)} &:= \mathrm{Diag}\left\{\partial_y f(\boldsymbol{y}_{\mathrm{ref}}, \boldsymbol{z}_{\mathrm{ref}})\right\} \boldsymbol{y}^{\Delta}_{(1)} + \mathrm{Diag}\left\{\partial_z f(\boldsymbol{y}_{\mathrm{ref}}, z_{\mathrm{ref}})\right\} z^{\Delta}_{(1)} \\
\widetilde{g}_{(1)} &:= \mathrm{Diag}\left\{\partial_y g(\boldsymbol{y}_{\mathrm{ref}}, \boldsymbol{z}_{\mathrm{ref}})\right\} \boldsymbol{y}^{\Delta}_{(1)} + \mathrm{Diag}\left\{\partial_z g(\boldsymbol{y}_{\mathrm{ref}}, z_{\mathrm{ref}})\right\} z^{\Delta}_{(1)} \\
\widehat{f}_{(1)} &:= \mathrm{Diag}\left\{\partial_y f(\boldsymbol{y}^{\Delta}_{(0)}, z^{\Delta}_{(0)})\right\} \boldsymbol{y}^{\Delta}_{(1)} + \mathrm{Diag}\left\{\partial_z f(\boldsymbol{y}^{\Delta}_{(0)}, z^{\Delta}_{(0)})\right\} z^{\Delta}_{(1)} - \mathrm{Diag}\left\{\partial_y f(\boldsymbol{y}_{\mathrm{ref}}, \boldsymbol{z}_{\mathrm{ref}})\right\} \boldsymbol{y}^{\Delta}_{(1)} - \mathrm{Diag}\left\{\partial_z f(\boldsymbol{y}_{\mathrm{ref}}, z_{\mathrm{ref}})\right\} z^{\Delta}_{(1)} \\
\widehat{g}_{(1)} &:= \mathrm{Diag}\left\{\partial_y g(\boldsymbol{y}^{\Delta}_{(0)}, z^{\Delta}_{(0)})\right\} \boldsymbol{y}^{\Delta}_{(1)} + \mathrm{Diag}\left\{\partial_z g(\boldsymbol{y}^{\Delta}_{(0)}, z^{\Delta}_{(0)})\right\} z^{\Delta}_{(1)} - \mathrm{Diag}\left\{\partial_y g(\boldsymbol{y}_{\mathrm{ref}}, \boldsymbol{z}_{\mathrm{ref}})\right\} \boldsymbol{y}^{\Delta}_{(1)} - \mathrm{Diag}\left\{\partial_z g(\boldsymbol{y}_{\mathrm{ref}}, z_{\mathrm{ref}})\right\} z^{\Delta}_{(1)}.
\end{aligned} \tag{30}$$

Theorem 3. *Let $y^{n+1}_{(1)}$ and $z^{n+1}_{(1)}$ be the quantities computed with* (16)*, and let $y_{(1)}$ and $z_{(1)}$ be solutions to* (5)*. There holds:*

$$y^{n+1}_{(1)} - y_{(1)}(t^{n+1}) = O(\Delta t^{r_2+1}) \qquad \textit{and} \qquad z^{n+1}_{(1)} - z_{(1)}(t^{n+1}) = O(\Delta t^{r_2}), \tag{31}$$

where

$$r_2 := \min(q+1, \widetilde{q}).$$

Proof. We start by considering the explicit splitting function $\widehat{f}$ defined in (30) for the numerical solution $\boldsymbol{y}^{\Delta}_{(1)}$ and $\boldsymbol{z}^{\Delta}_{(1)}$. $\boldsymbol{y}^{\Delta}_{(0)}$ and $\boldsymbol{z}^{\Delta}_{(0)}$ are quantities computed with method (15). Therefore, they obey Thm. 2. It is straightforward to compute

$$\widehat{f}_{(1)} = \mathrm{Diag}\left\{\partial_y f(\boldsymbol{y}^{\Delta}_{(0)}, z^{\Delta}_{(0)}) - \partial_y f(\boldsymbol{y}_{\mathrm{ref}}, \boldsymbol{z}_{\mathrm{ref}})\right\} \boldsymbol{y}^{\Delta}_{(1)} + \mathrm{Diag}\left\{\partial_z f(\boldsymbol{y}^{\Delta}_{(0)}, z^{\Delta}_{(0)}) - \partial_z f(\boldsymbol{y}_{\mathrm{ref}}, \boldsymbol{z}_{\mathrm{ref}})\right\} z^{\Delta}_{(1)}. \tag{32}$$

Note that for an exact solution the explicit part sums up to zero. The error in $\boldsymbol{y}^{\Delta}_{(0)}$ and $\boldsymbol{z}^{\Delta}_{(0)}$ is known due to Thm. 2, it is given by the interal orders of the IMEX Runge-Kutta method plus terms in $O(\Delta t^{r_1})$, hence

$$\partial_y f(y^{n,i}_{(0)}, z^{n,i}_{(0)}) - \partial_y f(y_{\mathrm{ref}}, z_{\mathrm{ref}}) = O(\Delta t^{q^i+1}) + O(\Delta t^{r_1}) \qquad \text{and} \qquad \partial_z f(y^{n,i}_{(0)}, z^{n,i}_{(0)}) - \partial_z f(y^{i}_{\mathrm{ref}}, z^{i}_{\mathrm{ref}}) = O(\Delta t^{q^i+1}) + O(\Delta t^{r_1}).$$

The same result can be obtained for $\widehat{g}_{(1)}$. Consequently, the explicit part (note that in (16), $\widehat{f}_{(1)}$ and $\widehat{g}_{(1)}$ are multiplied by Δt) reduces to terms in $O(\Delta t^{q^i+2})$. Even more, for the update step we obtain that the explicit part is given with an accuracy in $O(\Delta t^p) + O(\Delta t^{r_1})$ since the explicit Runge-Kutta scheme is a standard quadrature rule integrating an approximation of the zero given with the desired internal orders. The implicit function $\widetilde{f}_{(1)}$ is given by

$$\widetilde{f}_{(1)} = \mathrm{Diag}\left\{\partial_y f(\boldsymbol{y}_{\mathrm{ref}}, \boldsymbol{z}_{\mathrm{ref}})\right\} \boldsymbol{y}^{\Delta}_{(1)} + \mathrm{Diag}\left\{\partial_z f(\boldsymbol{y}_{\mathrm{ref}}, \boldsymbol{z}_{\mathrm{ref}})\right\} z^{\Delta}_{(1)}. \tag{33}$$

Therefore, the method for computing approximations to $y_{(1)}$ and $z_{(1)}$ can be seen as a perturbation of an implicit Runge-Kutta discretizaion of (5), where the implicit Runge-Kutta scheme is given by the implicit Butcher tableau of

the IMEX scheme. The behavior is hence the same as it would be for this implicit method up to $O(\Delta t^{q+2})$ terms due to the explicit part.

The rest of the proof follows by arguments given in [19], note that the proof is slightly simpler, because we exactly know the function $\partial_y f(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})$: It is shown in [19] that the error for the y component is given by $O(\Delta t^{\widetilde{q}+1})$, in our case, one has to add terms of order $O(\Delta t^{q+1})$ to account for the perturbation due to the explicit part. Therefore, it can be concluded that the error in the y component is given by $O(\Delta t^{r_2+1})$. Similarly, we obtain that the z component has an error of $O(\Delta t^{r_2})$. This proves the theorem. □

Remark 6. *For diagonally-implicit methods, which we are using in this work, $\widetilde{q}$ can be at most* 2 *and in general $q = 1$ due to the explicit term. Therefore* $\min(q+1, \widetilde{q}) = \widetilde{q}$.

4.3.2. Error analysis of $y^{n+1}_{(2)}$ and $z^{n+1}_{(2)}$

There remains to investigate the error in $y^{n+1}_{(2)}$ and $z^{n+1}_{(2)}$. As before, we follow the analysis in [19]. The method is defined in equation (17). For simplicity we only consider $\widehat{f}_{(2)}$ and $\widetilde{f}_{(2)}$ since the steps and ideas are the same for $\widehat{g}_{(2)}$ and $\widetilde{g}_{(2)}$. The splitting functions $\widehat{f}_{(2)}$ and $\widetilde{f}_{(2)}$ are given by

$$\begin{aligned}
\widehat{f}_{(2)} &= \operatorname{Diag}\left\{\partial_y f(\boldsymbol{y}^\Delta_{(0)}, \boldsymbol{z}^\Delta_{(0)})\right\} \boldsymbol{y}^\Delta_{(2)} + \operatorname{Diag}\left\{\partial_z f(\boldsymbol{y}^\Delta_{(0)}, \boldsymbol{z}^\Delta_{(0)})\right\} \boldsymbol{z}^\Delta_{(2)} \\
&\quad + \frac{1}{2}\left(\operatorname{Diag}\left\{\partial_{yy} f(\boldsymbol{y}^\Delta_{(0)}, \boldsymbol{z}^\Delta_{(0)})\right\} \operatorname{Diag}\left\{\boldsymbol{y}^\Delta_{(1)}\right\} \boldsymbol{y}^\Delta_{(1)} + \operatorname{Diag}\left\{\partial_{zz} f(\boldsymbol{y}^\Delta_{(0)}, \boldsymbol{z}^\Delta_{(0)})\right\} \operatorname{Diag}\left\{\boldsymbol{z}^\Delta_{(1)}\right\} \boldsymbol{z}^\Delta_{(1)}\right) \\
&\quad + \operatorname{Diag}\left\{\partial_{yz} f(\boldsymbol{y}^\Delta_{(0)}, \boldsymbol{z}^\Delta_{(0)})\right\} \operatorname{Diag}\left\{\boldsymbol{y}^\Delta_{(1)}\right\} \boldsymbol{z}^\Delta_{(1)} - \operatorname{Diag}\left\{\partial_y f(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right\} \boldsymbol{y}^\Delta_{(2)} - \operatorname{Diag}\left\{\partial_z f(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right\} \boldsymbol{z}^\Delta_{(2)} \\
\widetilde{f}_{(2)} &= \operatorname{Diag}\left\{\partial_y f(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right\} \boldsymbol{y}^\Delta_{(2)} + \operatorname{Diag}\left\{\partial_z f(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right\} \boldsymbol{z}^\Delta_{(2)}
\end{aligned} \tag{34}$$

Again, we start with the explicit part $\widehat{f}_{(2)}$ and rearrange the terms such that

$$\begin{aligned}
\widehat{f}_{(2)} &= \operatorname{Diag}\left\{\partial_y f(\boldsymbol{y}^\Delta_{(0)}, \boldsymbol{z}^\Delta_{(0)}) - \partial_y f(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right\} \boldsymbol{y}^\Delta_{(2)} + \left(\partial_z f(\boldsymbol{y}^\Delta_{(0)}, \boldsymbol{z}^\Delta_{(0)}) - \partial_z f(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right) \boldsymbol{z}^\Delta_{(2)} \\
&\quad + \frac{1}{2}\left(\operatorname{Diag}\left\{\partial_{yy} f(\boldsymbol{y}^\Delta_{(0)}, \boldsymbol{z}^\Delta_{(0)})\right\} \operatorname{Diag}\left\{\boldsymbol{y}^\Delta_{(1)}\right\} \boldsymbol{y}^\Delta_{(1)} + \operatorname{Diag}\left\{\partial_{zz} f(\boldsymbol{y}^\Delta_{(0)}, \boldsymbol{z}^\Delta_{(0)})\right\} \operatorname{Diag}\left\{\boldsymbol{z}^\Delta_{(1)}\right\} \boldsymbol{z}^\Delta_{(1)}\right) \\
&\quad + \operatorname{Diag}\left\{\partial_{yz} f(\boldsymbol{y}^\Delta_{(0)}, \boldsymbol{z}^\Delta_{(0)})\right\} \operatorname{Diag}\left\{\boldsymbol{y}^\Delta_{(1)}\right\} \boldsymbol{z}^\Delta_{(1)}.
\end{aligned}$$

As before, we can conclude that the explicit part is only a perturbation of an implicit method: Following the work in [19] and [20], estimating quadratic terms with the help of a Lipschitz argument which adds an additional term of order $O(\Delta t^{r_2})$. Working similarly as in the proof of Thm. 3, one obtains the following theorem:

Theorem 4. *Let $y^{n+1}_{(2)}$ and $z^{n+1}_{(2)}$ be the quantities computed with* (17)*, and let $y_{(2)}$ and $z_{(2)}$ be solutions to* (6)*. There holds:*

$$y^{n+1}_{(2)} - y_{(2)}(t^{n+1}) = O(\Delta t^{r_2}) \qquad \textit{and} \qquad z^{n+1}_{(2)} - z_{(2)}(t^{n+1}) = O(\Delta t^{r_2 - 1}). \tag{35}$$

5. Numerical verification

In this section, we verify and discuss our analytical results numerically. We compare the RS-IMEX method against the more standard IMEX-splitting, discussed, e.g., in [20].

Definition 5 (Standard splitting [20])**.** *The more standard splitting of the equation* (1) *into stiff and non-stiff terms, discussed, e.g., in [20] is given by*

$$\begin{pmatrix} \widetilde{f}(y,z) \\ \frac{1}{\varepsilon}\widetilde{g}(y,z) \end{pmatrix} := \begin{pmatrix} 0 \\ \frac{1}{\varepsilon} g(y,z) \end{pmatrix} \qquad \textit{and} \qquad \begin{pmatrix} \widehat{f}(y,z) \\ \frac{1}{\varepsilon}\widehat{g}(y,z) \end{pmatrix} := \begin{pmatrix} f(y,z) \\ 0 \end{pmatrix}.$$

	type	$\widetilde{q}$	q	p	r_1	r_2	Tbl.
ARS_443	CK	1	1	3	3	1	B.3
BPR_353	CK	2	1	3	3	2	B.4

Table 2: Classification of IMEX Runge-Kutta methods considered in this work.

The limit of this method serves as an approximation to the reference solution needed in the RS-IMEX splitting.

Two different equations are considered, namely van der Pol equation [12, 19, 20, 21] and Michaelis Menten equation [1, 2]. Those equations are given, respectively, by

$$\frac{d}{dt}\begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} f(y,z) \\ \frac{1}{\varepsilon} g(y,z) \end{pmatrix} := \begin{pmatrix} z \\ \frac{1}{\varepsilon}\left((1-y^2)z - y\right) \end{pmatrix} \tag{vdP}$$

and

$$\frac{d}{dt}\begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} f(y,z) \\ \frac{1}{\varepsilon} g(y,z) \end{pmatrix} := \begin{pmatrix} -y + \left(y + \frac{1}{2}\right) z \\ \frac{1}{\varepsilon}\left(y - (y+1)z\right) \end{pmatrix}. \tag{MM}$$

The last equation describes the transformation of a substrate to a product by an enzyme, see [1, 2]. Both examples are equipped with well prepared initial data. The final time instance is chosen in such a way that there is no singularity for $\varepsilon \to 0$. This happens for van der Pol equation with those initial conditions as given below for $t_{end} \geq 0.8$. We choose for van der Pol

$$t_{end} = 0.55139, \qquad \begin{pmatrix} y(t=0) \\ z(t=0) \end{pmatrix} = \begin{pmatrix} 2 \\ -\frac{2}{3} \end{pmatrix} + \begin{pmatrix} 0 \\ \frac{10}{81} \end{pmatrix} \varepsilon + \begin{pmatrix} 0 \\ -\frac{292}{2187} \end{pmatrix} \varepsilon^2,$$

and for Michaelis Menten

$$t_{end} = 1, \qquad \begin{pmatrix} y(t=0) \\ z(t=0) \end{pmatrix} = \begin{pmatrix} 1 \\ \frac{1}{2} \end{pmatrix} + \begin{pmatrix} 0 \\ \frac{1}{32} \end{pmatrix} \varepsilon + \begin{pmatrix} 0 \\ -\frac{5}{512} \end{pmatrix} \varepsilon^2.$$

Two different IMEX Runge-Kutta methods are considered, tabulated in Appendix B. The properties of the used schemes are summarized in Tbl. 2. From our analysis in Sec. 4, it is known that stage orders of the complete and the implicit part of the IMEX Runge-Kutta scheme are important. Note that the BPR_353 scheme fulfills $\widetilde{q} = 2$ and $q = 1$, while the ARS_443 scheme has $\widetilde{q} = 1$ and $q = 1$. Therefore, we expect that all splittings perform similarly for the ARS_443 scheme and that the RS-IMEX and implicit splitting shows an improved behavior for the BPR_353 scheme.

Remark 7 (Computational cost)**.** *Obviously, the use of the RS-IMEX splitting necessitates the computation of the reference solution. An additional equation (*(4) *or* (8)*) has thus to be solved, rendering the method on first sight less effective. The analysis of computational cost is not part of this research. In the framework of more realistic problems, we have made these investigations in [18, 24].*

In the following we summarize the computations for all different splittings, different IMEX Runge-Kutta schemes, different examples and also different values of ε, i.e. $\varepsilon = 10^{-i}$ for $i = 1, 3, 5, 7$. Please note that the computations have been done with the help of Matlab 2017a [30]; machine accuracy is $\approx 2 \cdot 10^{-16}$.

5.1. Order reduction

In this section, we investigate the phenomenon of order reduction. We compute numerical approximations to van der Pol equation (see Figs. 2 with BPR_353, and 3 with ARS_443) and Michaelis Menten equation (see Figs. 4 with BPR_353 and 5 with ARS_443). Error is defined as error at t_{end}, i.e.,

$$e_{\Delta t} := \sqrt{\left|y(t_{end}) - y^N\right|^2 + \left|z(t_{end}) - z^N\right|^2}, \tag{36}$$

where y^N and z^N are results of the method under consideration. All computations show the expected convergence behavior. For the ARS_443 scheme there is no difference between the splittings concerning order reduction, which we expected since $q = \widetilde{q}$ for the ARS_443 scheme. For the BPR_353 scheme we also obtain the expected results. The RS-IMEX splitting and the implicit method behave very similarly and show an improved convergence compared to the standard splitting. This is what we expected since $\widetilde{q} = 2$ and $q = 1$ for the BPR_353 scheme.

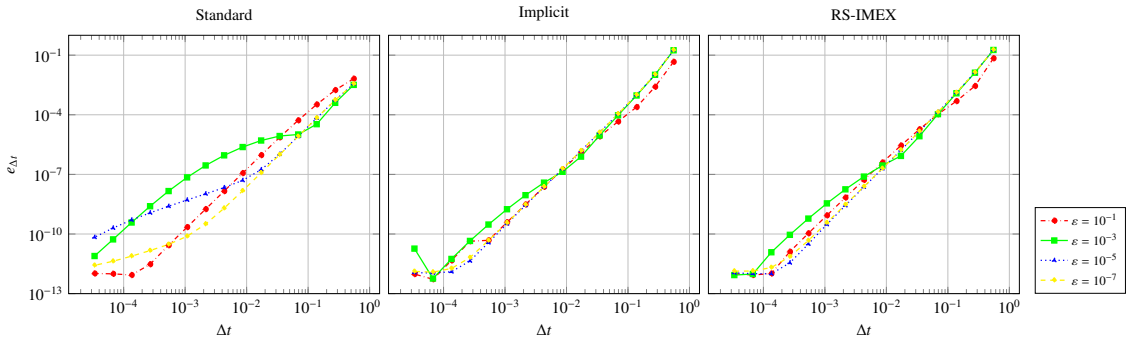

Figure 2: Approximation error $e_{\Delta t}$ of van der Pol equation for different values of ε. Numerical discretization: BPR_353 method coupled with the standard (left), implicit (middle) and RS-IMEX (right) splitting. Note that the results for the standard splitting and the RS-IMEX splitting have also been computed – for a different t_{end} – in [22].

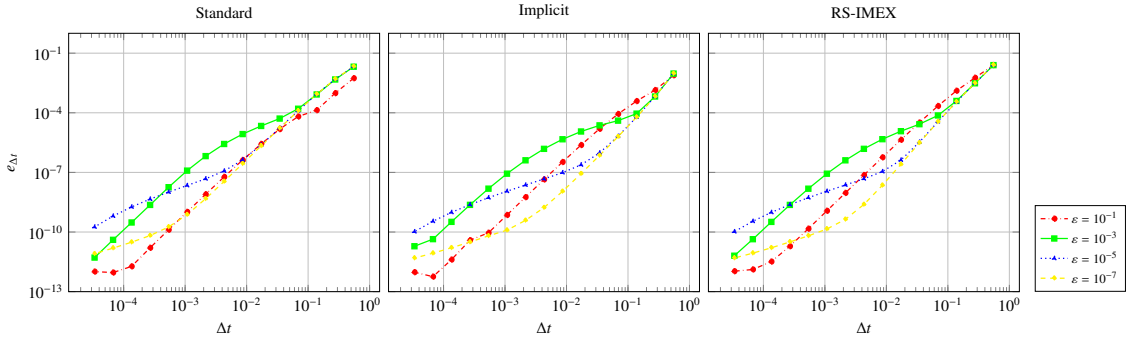

Figure 3: Approximation error $e_{\Delta t}$ of van der Pol equation for different values of ε. Numerical discretization: ARS_443 method coupled with the standard (left), implicit (middle) and RS-IMEX (right) splitting. Note that the results for the standard splitting and the RS-IMEX splitting have also been computed – for a different t_{end} – in [22].

5.2. Convergence in the asymptotic expansion

In the analysis section Sec. 4, we have derived error estimates for the different components of the asymptotic expansion, i.e. we investigated

$$\begin{pmatrix} y^{n+1} - y(t^{n+1}) \\ z^{n+1} - z(t^{n+1}) \end{pmatrix} = \begin{pmatrix} y^{n+1}_{(0)} - y_{(0)}(t^{n+1}) \\ z^{n+1}_{(0)} - z_{(0)}(t^{n+1}) \end{pmatrix} + \varepsilon \begin{pmatrix} y^{n+1}_{(1)} - y_{(1)}(t^{n+1}) \\ z^{n+1}_{(1)} - z_{(1)}(t^{n+1}) \end{pmatrix} + \varepsilon^2 \begin{pmatrix} y^{n+1}_{(2)} - y_{(2)}(t^{n+1}) \\ z^{n+1}_{(2)} - z_{(2)}(t^{n+1}) \end{pmatrix} + O(\varepsilon^3)$$

componentwise. In this section, we implement the methods given in (15)-(17) and compare the resulting numerical approximation to the solution of (4)-(6). For van der Pol equation, the limit solutions at time $t_{end} = 0.55139$ have been computed by a very accurate numerical integration, there holds

$$\begin{pmatrix} y_{(0)}(t_{end}) \\ z_{(0)}(t_{end}) \end{pmatrix} = \begin{pmatrix} 1.54162058100305 \\ -1.11988034477856 \end{pmatrix}, \begin{pmatrix} y_{(1)}(t_{end}) \\ z_{(1)}(t_{end}) \end{pmatrix} = \begin{pmatrix} 0.295547928806445 \\ 1.976165953776134 \end{pmatrix}, \begin{pmatrix} y_{(2)}(t_{end}) \\ z_{(2)}(t_{end}) \end{pmatrix} = \begin{pmatrix} -1.34783260991136 \\ -16.4475455160741 \end{pmatrix}.$$

Numerical results can be found in Fig. 6 for the BPR_353 and in Fig. 7 for the ARS_443 scheme. Error is computed as error at t_{end}. All figures show that the RS-IMEX splitting has the same convergence behavior in every component as the implicit method. The standard splitting is less accurate for the BPR_353 scheme, but it is also less accurate for the ARS_443 scheme in the y component. Since the z component is dominating the error this order reduction cannot be seen in Fig. 3.

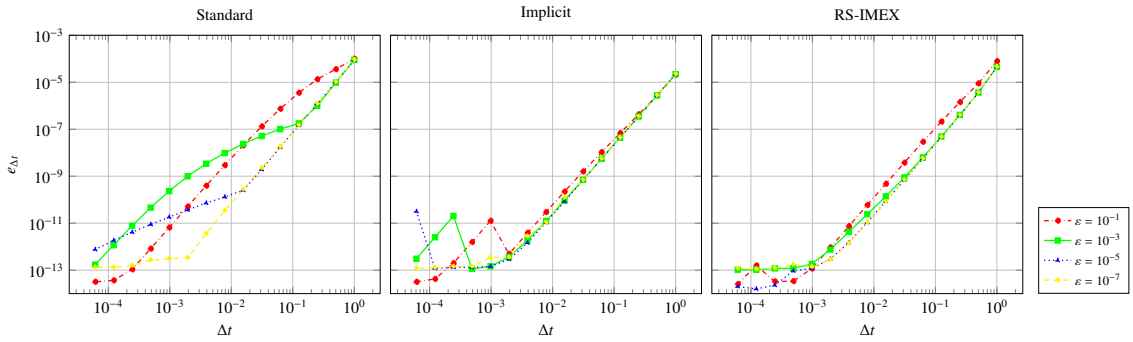

Figure 4: Approximation error $e_{\Delta t}$ of Michaelis Menten equation for different values of ε. Numerical discretization: BPR_353 method coupled with the standard (left), implicit (middle) and RS-IMEX (right) splitting.

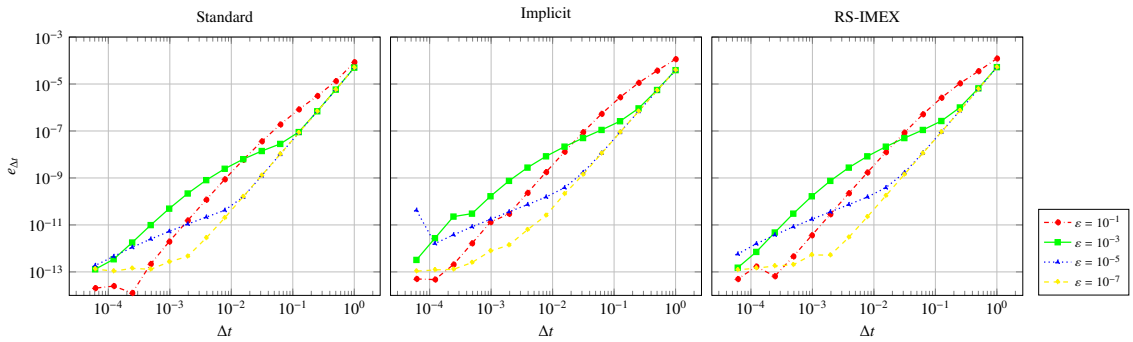

Figure 5: Approximation error $e_{\Delta t}$ of Michaelis Menten equation for different values of ε. Numerical discretization: ARS_443 method coupled with the standard (left), implicit (middle) and RS-IMEX (right) splitting.

5.3. Numerical evidence for very high-order schemes

It is not clear that the bounds given by Thm. 2 are sharp, and we indicate in this subsection that this is not the case. Due to the explicit part, there always holds $q = 1$, and as we are using diagonally-implicit methods, there also holds $\widetilde{q} \leq 2$, hence $r_1 = \min(p, 4)$. To investigate the sharpness of the theorem, we therefore have to chose a method that is at least fifth-order convergent. Good candidates are the IMEX schemes based on integral deferred correction methods (InDC) as presented in [13]. Those methods can be constructed to have any desired order of convergence; they can be rewritten as IMEX Runge-Kutta shemes, see also [21]. Fig. 8 shows numerical results for the computation of the limit differential algebraic equation for van der Pol and Michaelis Menten equation, for different InDC methods based on the IMEX Euler discretization. Methods are termed IMEX_InDC_p, where p denotes the order of the method. It can be seen that the error is decaying faster than $O(\Delta t^{r_1})$, see the dashed reference line. Therefore, the bound derived in Thm. 2 is not sharp.

Remark 8. *Note that the use of InDC methods in Runge-Kutta formulation is here made only for simplicity. In general, this is not recommendable, as the number of stages grows drastically. As an example, IMEX_InDC_6 has 37 stages in this formulation.*

6. Conclusion & Outlook

In this work, we have given a detailed analysis of the asymptotic convergence error of an IMEX Runge-Kutta method applied to a singularly perturbed ODE splitted with the RS-IMEX approach. We have improved and extended earlier results on the topic. The mathematical analysis has been backed up by numerical investigations; also rather

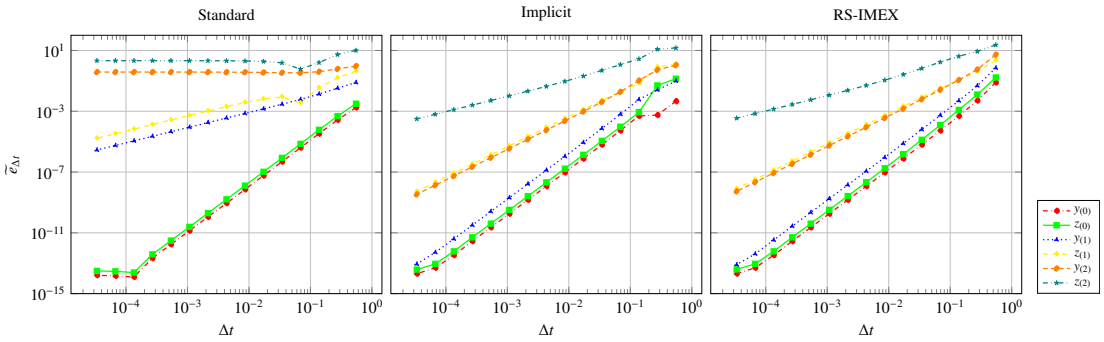

Figure 6: The approximation error in each component of the separated van der Pol oscillator as given in (4)-(6) at the final time instance t_{end}. The used method is given in (15)-(17) coupled with the BPR_353 scheme and different splittings. From left to right: standard splitting, implicit method and RS-IMEX splitting.

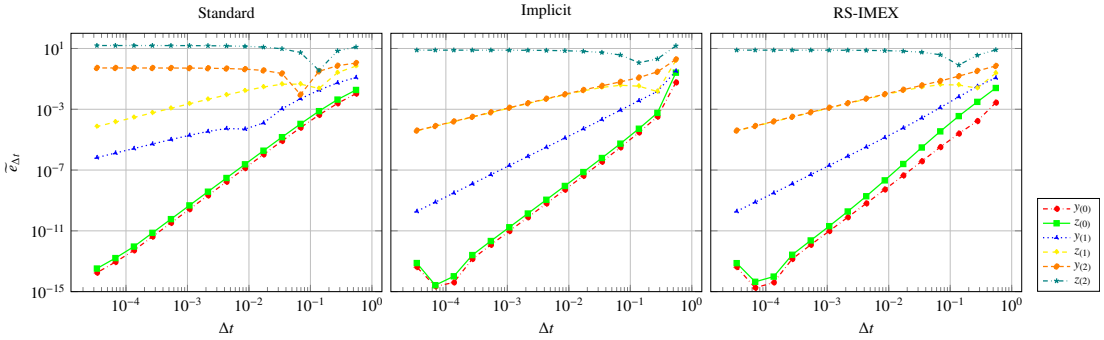

Figure 7: The approximation error in each component of the separated van der Pol oscillator as given in (4)-(6) at the final time instance t_{end}. The used method is given in (15)-(17) coupled with the ARS_443 scheme and different splittings. From left to right: standard splitting, implicit method and RS-IMEX splitting.

high-order discretizations have been considered. A comparison to a more established standard splitting revealed that order reduction is not only a structural phenomenon of a given IMEX Runge-Kutta method, but depends on a subtle interplay between splitting and Runge-Kutta method.

Obviously, a couple of issues still remain. The restrictions on the Runge-Kutta method (type CK, GSA, uniform $\boldsymbol{c}$) are rather strict. In [20, 31], a class of uniformly accurate (for the standard splitting!) IMEX Runge-Kutta methods has been developed. Those methods are not GSA, and we have already observed in [22] that for 'large' Δt, they have stability issues if combined with the RS-IMEX splitting. Currently, the limit differential algebraic equation and its numerical discretization is under investigation. It would be desirable to obtain general conditions on the splitting function – independent of the RS-IMEX approach – and on the Butcher tableaux such that an IMEX Runge-Kutta discretization of this limit equation has optimal order. This is a non-trivial task, as the straightforward application of an IMEX Runge-Kutta method to a general splitting is usually of first order only. Closely related to this issue is the extension of the results in this paper to other singularly perturbed equations, favorably to the (semi-)discretizations of the Euler and Navier-Stokes equations.

Acknowledgment

The authors would like to thank Sebastian Noelle for fruitful discussions. The first author has been partially supported by the German Research Foundation (DFG) through project NO 361/6-1; his study was supported by the Special Research Fund (BOF) of Hasselt University.

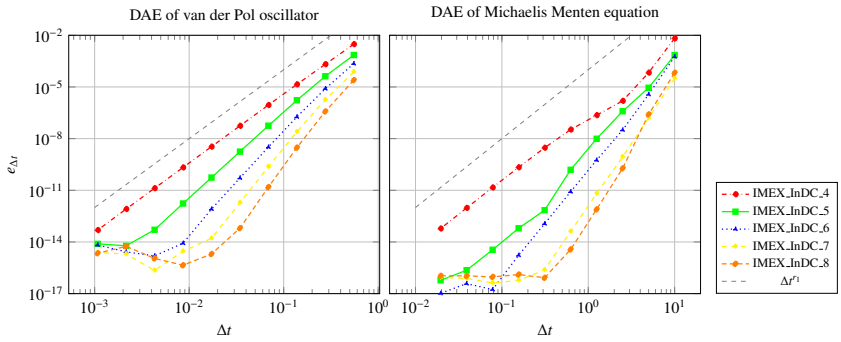

Figure 8: Left: approximation error of van der Pol DAE, Right: approximation error of Michaelis Menten DAE. Approximation is done via different InDC [21] methods with the RS-IMEX splitting. In contrast to before, $t_{end} = 10$ has been chosen for the Michaelis Menten equation, because otherwise, error levels reach machine accuracy really fast.

Appendix A. Postponed lemmas and proofs

In this section, we state and prove some of the lemmas that we used earlier. All of them are of rather technical nature and do not add new techniques to this work, which is why we have sourced them out to the appendix section. The point of departure is the following formula, derived in La. 2:

$$z_{(0)}^{\Delta} = z_{\text{ref}} - \text{Diag}\left\{\partial_z g(\boldsymbol{y}_{\text{ref}}, z_{\text{ref}})\right\}^{-1} \left(\left(\widetilde{\boldsymbol{A}} - \widehat{\boldsymbol{A}}\right)^{-1} \widehat{\boldsymbol{A}} \left(g(\boldsymbol{y}_{(0)}^{\Delta}, z_{(0)}^{\Delta})\right) + \text{Diag}\left\{\partial_y g(\boldsymbol{y}_{\text{ref}}, z_{\text{ref}})\right\} (\boldsymbol{y}_{(0)}^{\Delta} - \boldsymbol{y}_{\text{ref}})\right) + O(\Delta t^{r_1}). \tag{22}$$

Lemma 6. *Assume that* $\boldsymbol{y}_{(0)}^{\Delta} - \boldsymbol{y}_{ref} = O(\Delta t^{q+1}\boldsymbol{e})$ *holds and that* (22) *is valid. Then,*

$$z_{(0)}^{\Delta} - z_{ref} = O(\Delta t^{q+1}\boldsymbol{e}).$$

Proof. For convenience, let us define

$$\boldsymbol{C} := (\widetilde{\boldsymbol{A}} - \widehat{\boldsymbol{A}})^{-1}\widehat{\boldsymbol{A}},$$

which is a strictly lower triangular matrix. We consider (22) element-wise. For the first equation - because $\boldsymbol{C}$ is strictly lower triangular - one gets

$$z_{(0)}^{n,2} - z_{\text{ref}}(t^{n,2}) = -\text{Diag}\left\{\partial_z g(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right\}^{-1} \text{Diag}\left\{\partial_y g(\boldsymbol{y}_{\text{ref}}, \boldsymbol{z}_{\text{ref}})\right\} (\boldsymbol{y}_{(0)}^{\Delta} - \boldsymbol{y}_{\text{ref}}) + O(\Delta t^{r_1}) = O(\Delta t^{q+1}) + O(\Delta t^{r_1}).$$

Let us consider the i^{th} internal stage and assume that $z_{(0)}^{n,k} - z_{\text{ref}}(t^{n,k}) = O(\Delta t^{q+1})$ for $j < i$. Then we can deduce in the same fashion as in La. 7, that $g(y_{(0)}^{n,k}, z_{(0)}^{n,k}) = O(\Delta t^{r_1})$. Therefore,

$$\begin{aligned} z_{(0)}^{n,i} - z_{\text{ref}}(t^{n,i}) =& -\text{Diag}\left\{\partial_z g(y_{\text{ref}}(t^{n,i}), z_{\text{ref}}(t^{n,i}))\right\}^{-1} \text{Diag}\left\{\partial_y g(y_{\text{ref}}(t^{n,i}), z_{\text{ref}}(t^{n,i}))\right\} (\boldsymbol{y}_{(0)}^{\Delta} - \boldsymbol{y}_{\text{ref}}) \\ &+ \partial_z g(y_{\text{ref}}(t^{n,i}), z_{\text{ref}}(t^{n,i}))^{-1} \sum_{j=1}^{i-2} C^{i,j} O(\Delta t^{2(q+1)}) + O(\Delta t^{p}) \\ =& O(\Delta t^{q+1}) + O(\Delta t^{p}). \end{aligned}$$

This proves the lemma. □

Lemma 7. *Assume that* $\boldsymbol{y}_{(0)}^{\Delta} - \boldsymbol{y}_{ref} = O(\Delta t^{q+1}\boldsymbol{e})$ *and* $z_{(0)}^{\Delta} - z_{ref} = O(\Delta t^{q+1}\boldsymbol{e})$ *holds and that* (22) *is valid. Then,*

$$g(\boldsymbol{y}_{(0)}^{\Delta}, z_{(0)}^{\Delta}) = O(\Delta t^{r_1}\boldsymbol{e}).$$

Proof. We can expand g with the help of a Taylor expansion up to second order terms, i.e.

$$g(\boldsymbol{y}^{\Delta}_{(0)}, \boldsymbol{z}^{\Delta}_{(0)}) = \mathrm{Diag}\left\{\partial_y g(\boldsymbol{y}_{\mathrm{ref}}, \boldsymbol{z}_{\mathrm{ref}})\right\}(\boldsymbol{y}^{\Delta}_{(0)} - \boldsymbol{y}_{\mathrm{ref}}) + \mathrm{Diag}\left\{\partial_z g(\boldsymbol{y}_{\mathrm{ref}}, \boldsymbol{z}_{\mathrm{ref}})\right\}(\boldsymbol{z}^{\Delta}_{(0)} - \boldsymbol{z}_{\mathrm{ref}}) + O(\Delta t^{r_1})$$

and plug the representation of $\boldsymbol{z}^{\Delta}$, (22), in

$$g(\boldsymbol{y}^{\Delta}_{(0)}, \boldsymbol{z}^{\Delta}_{(0)}) = -\boldsymbol{C}\left(g(\boldsymbol{y}^{\Delta}_{(0)}, \boldsymbol{z}^{\Delta}_{(0)})\right) + O(\Delta t^{r_1})$$

This can be done several times recursively and we obtain

$$g(\boldsymbol{y}^{\Delta}_{(0)}, \boldsymbol{z}^{\Delta}_{(0)}) = (-\boldsymbol{C})^s\left(g(\boldsymbol{y}^{\Delta}_{(0)}, \boldsymbol{z}^{\Delta}_{(0)})\right) + \sum_{k=1}^{s}(-\boldsymbol{C})^{k-1}\, O(\Delta t^{r_1})$$

Since $\boldsymbol{C} \in \mathbb{R}^{(s-1)\times(s-1)}$ is a lower triangular matrix with 0 on the diagonal, we know that the s^{th} power of this matrix is equal to 0. Therefore we obtain

$$g(\boldsymbol{y}^{\Delta}_{(0)}, \boldsymbol{z}^{\Delta}_{(0)}) = \sum_{k=1}^{s}(-\boldsymbol{C})^{k-1}\, O(\Delta t^{r_1}) + O(\Delta t^{p}).$$

□

Appendix B. IMEX Runge-Kutta methods

In Tbls. B.3 and B.4, we show for convenience the Butcher tableaux of the schemes used in the numerical results section.

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

Table B.3: A third order IMEX RK method called ARS_443 [8]. Left: implicit, right: explicit.

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

Table B.4: A third order IMEX RK method called BPR_353 [27]. Left: implicit, right: explicit.

References

[1] R. E. O'Malley, Singular perturbation methods for ordinary differential equations, Vol. 89, Springer Science & Business Media, 2012.

[2] C. W. Gear, T. J. Kaper, I. G. Kevrekidis, A. Zagaris, Projecting to a slow manifold: Singularly perturbed systems and legacy codes, SIAM Journal on Applied Dynamical Systems 4 (3) (2005) 711–732.

[3] 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.

[4] W.-A. Yong, A note on the zero Mach number limit of compressible Euler equations, Proceedings of the American Mathematical Society 133 (10) (2005) 3079–3085.

[5] S. Schochet, The mathematical theory of low Mach number flows, ESAIM: Mathematical Modelling and Numerical Analysis 39 (03) (2005) 441–458.

[6] D. Kröner, Numerical Schemes for Conservation Laws, Wiley Teubner, 1997.
[7] 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.
[8] 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.
[9] C. A. Kennedy, M. H. Carpenter, Additive Runge-Kutta schemes for convection-diffusion-reaction equations, Applied Numerical Mathematics 44 (2003) 139–181.
[10] W. Hundsdorfer, S.-J. Ruuth, IMEX extensions of linear multistep methods with general monotonicity and boundedness properties, Journal of Computational Physics 225 (2) (2007) 2016–2042.
[11] S. Boscarino, L. Pareschi, On the asymptotic properties of IMEX Runge–Kutta schemes for hyperbolic balance laws, Journal of Computational and Applied Mathematics 316 (2017) 60 – 73.
[12] 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.
[13] A. Christlieb, M. Morton, B. Ong, J.-M. Qiu, Semi-implicit integral deferred correction constructed with additive Runge-Kutta methods, Communications in Mathematical Sciences 9 (2011) 879–902.
[14] 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.
[15] 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.
[16] 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.
[17] 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.
[18] K. Kaiser, J. Schütz, R. Schöbel, S. Noelle, A new stable splitting for the isentropic Euler equations, Journal of Scientific Computing 70 (2017) 1390–1407.
[19] E. Hairer, G. Wanner, Solving Ordinary Differential Equations II, Springer Series in Computational Mathematics, 1991.
[20] S. Boscarino, Error analysis of IMEX Runge-Kutta methods derived from differential-algebraic systems, SIAM Journal on Numerical Analysis 45 (2007) 1600–1621.
[21] S. Boscarino, J. Qiu, G. Russo, Implicit-explicit integral deferred correction methods for stiff problems, arXiv preprint arXiv:1701.04750.
[22] J. Schütz, K. Kaiser, A new stable splitting for singularly perturbed ODEs, Applied Numerical Mathematics 107 (2016) 18–33.
[23] K. Kaiser, J. Schütz, A high-order method for weakly compressible flows, Communications in Computational Physics 22 (4) (2017) 1150–1174.
[24] J. Zeifang, K. Kaiser, A. Beck, J. Schütz, C.-D. Munz, Efficient high-order discontinuous Galerkin computations of low Mach number flows, UHasselt CMAT Preprint UP-17-04.
[25] H. Zakerzadeh, Asymptotic analysis of the RS-IMEX scheme for the shallow water equations in one space dimension, IGPM Preprint Nr. 455.
[26] H. Zakerzadeh, S. Noelle, A note on the stability of implicit-explicit flux splittings for stiff hyperbolic systems, IGPM Preprint Nr. 449.
[27] S. Boscarino, L. Pareschi, G. Russo, Implicit-explicit Runge–Kutta schemes for hyperbolic systems and kinetic equations in the diffusion limit, SIAM Journal on Scientific Computing 35 (1) (2013) A22–A51.
[28] E. Hairer, S. Nørsett, G. Wanner, Solving Ordinary Differential Equations I, Springer Series in Computational Mathematics, 1987.
[29] L. R. Petzold, Order results for implicit Runge–Kutta methods applied to differential / algebraic systems, SIAM 23 (4) (1986) 837–852.
[30] MATLAB, MATLAB and Statistics Toolbox Release 2017a, The MathWorks Inc., Natick, Massachusetts, 2017.
[31] 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.