SIAM J. NUMER. ANAL.
Vol. 45, No. 4, pp. 1600–1621

ERROR ANALYSIS OF IMEX RUNGE–KUTTA METHODS DERIVED FROM DIFFERENTIAL-ALGEBRAIC SYSTEMS*

SEBASTIANO BOSCARINO†

Abstract. In this paper we present an error analysis of the IMEX Runge–Kutta methods when applied to stiff problems containing a nonstiff term and a stiff term, characterized by a small stiffness parameter ε. In this analysis we expand the global error in powers of ε and show that the coefficients of the error are the global errors of the IMEX Runge–Kutta method applied to a differential-algebraic system. Interesting convergence results of these errors and of the remainder of the expansion allow us to determine sharp error bounds for stiff problems. As a representative example of stiff problems we have chosen the van der Pol equation. We illustrate that the theoretical prediction is confirmed by the numerical test. Specifically, an order reduction phenomenon is observed when the problem becomes increasingly stiff. In particular, making several assumptions, we try to improve global error estimates of several IMEX Runge–Kutta methods existing in the literature.

Key words. Runge–Kutta methods, stiff problems, differential-algebraic systems

AMS subject classification. 34E05, 65L06, 65L80

DOI. 10.1137/060656929

1. Introduction.

Several physical phenomena of great importance for applications are described by stiff systems of differential equations in the form

$$U' = F(U) + \frac{1}{\varepsilon}G(U), \tag{1}$$

where $U = U(t) \in R^m$, F,$G : R^m \to R^m$, and $\varepsilon > 0$ is the stiffness parameter.

Systems of such form, with a large number of equations, often arise from the discretization of partial differential equations, such as convection-diffusion problems and hyperbolic systems with relaxation (i.e., discrete kinetic theory of rarefied gases, hydrodynamical models for semiconductors, etc., see [8], [17], [19], [18], [15], [6], [9]), where a method of lines approach is usually used.

In order to be able to treat problems of the form (1), it is important to develop suitable numerical schemes that work in an accurate and efficient way. A general approach to the solution of problem (1) is based on implicit-explicit (IMEX) multistep methods [14], [10], [3] or IMEX Runge–Kutta (R-K) methods [8], [17], [19], [18], [1], [2].

We consider here IMEX R-K methods. An IMEX R-K method consists of applying an implicit discretization for G and an explicit one for F. In general, in order to guarantee simplicity and efficiency in solving the algebraic equations corresponding to the implicit part of the discretization at each step of problem (1), we will consider diagonally implicit R-K (DIRK) methods.

In this paper we show that most of the popular IMEX R-K methods presented in the literature suffer from the phenomenon of order reduction in the stiff regime

*Received by the editors April 10, 2006; accepted for publication (in revised form) February 22, 2007; published electronically August 15, 2007. This research was partially supported by INDAM project "Metodi numerici per lo studio di problemi evolutivi multiscala" and Italian PRIN 2004 project Prot. 2004014411_007.

http://www.siam.org/journals/sinum/45-4/65692.html

†Department of Mathematics and Computer Science, University of Catania, viale A. Doria 6, 95125 Catania, Italy (boscarino@dmi.unict.it).

($\Delta t \gg \varepsilon$) when the classical order is greater than two [8], [17], [18], [1]. To this aim, we investigate this phenomenon and give an answer through a theoretical error analysis using typical techniques of differential-algebraic equations (DAEs) [12], [13], [7], [11].

We observe that system (1) can be written as a system of $2m$ equations in the form

$$\begin{aligned} y' &= f(y,z), \\ \varepsilon z' &= g(y,z) \end{aligned} \tag{2}$$

once we set $U = y+z$, $F(U) = f(y,z)$, and $G(U) = g(y,z)$. On the other hand, system (2) is a particular case of system (1) when $F(U) = (f(y,z),0)$, $G(U) = (0,g(y,z))$. Now, restricting our attention to system (2), such a problem is called a *singular perturbation problem* (SPP). Classical books on this subject are [20] and [16]. These SPPs give us the possibility of studying the dependence of the global error of IMEX R-K methods on the stiffness parameter ε. Then in system (2) we suppose that $0 < \varepsilon \ll 1$ and the functions f and g are sufficiently differentiable, with f, g and the initial values $y(0)$, $z(0)$ that may depend smoothly on ε. For simplicity of notation we suppress this dependence.

When the parameter ε in system (2) is small, the corresponding differential equation is stiff, and when ε tends to zero, the differential equation becomes differential algebraic. A sequence of differential-algebraic systems arises in the study of SPPs. Our analysis is based on the assumption of a smooth solution of system (2) and applies to the stiff case ($\Delta t \gg \varepsilon$).

The paper is organized as follows. In the next section we introduce a description and classification of the different types of IMEX R-K methods present in the literature, based on the structure of the matrix of the implicit part. In section 3 we state our main results, presenting convergence proofs which give sharp error bounds for such methods. On the van der Pol equation, moreover, we provide numerical confirmation of the theoretical analysis and compare the performances of several types of IMEX R-K schemes. Also, these numerical results suggest how we can improve error estimates of some IMEX R-K methods through straightforward assumptions. In section 4 we consider the asymptotic expansion of the exact and numerical solution in terms of the stiffness parameter ε. Sections 5 and 6 are devoted to examining the results obtained when we apply IMEX R-K methods to DAEs of index 1 (zeroth-order expansion) and higher (higher-order expansion). In particular, in section 7 we estimate the remainder of the expansion. Finally, in section 8, conclusions are drawn and work in progress is mentioned.

2. Description and classification of IMEX R-K methods. We consider an IMEX R-K method applied to system (2),

$$\begin{pmatrix} y_{n+1} \\ z_{n+1} \end{pmatrix} = \begin{pmatrix} y_n \\ z_n \end{pmatrix} + h \sum_{i=1}^{s} \begin{pmatrix} \tilde{b}_i k_{ni} \\ b_i \ell_{ni} \end{pmatrix}, \tag{3}$$

where

$$\begin{pmatrix} k_{ni} \\ \varepsilon \ell_{ni} \end{pmatrix} = \begin{pmatrix} f(Y_{ni}, Z_{ni}) \\ g(Y_{ni}, Z_{ni}) \end{pmatrix} \tag{4}$$

and the internal stages are given by

$$\begin{pmatrix} Y_{ni} \\ Z_{ni} \end{pmatrix} = \begin{pmatrix} y_n \\ z_n \end{pmatrix} + h \begin{pmatrix} \sum_{j=1}^{i-1} \tilde{a}_{ij} k_{nj} \\ \sum_{j=1}^{i} a_{ij} \ell_{nj} \end{pmatrix}. \tag{5}$$

The matrices $(\tilde{a}_{ij})$, with $\tilde{a}_{ij} = 0$ for $j \geq i$, and (a_{ij}) are $s \times s$ matrices such that the resulting method is explicit in f and implicit in g. We use a diagonally implicit scheme for g, i.e., $a_{ij} = 0$ for $j > i$. This will guarantee that f is always evaluated explicitly.

Such methods are characterized by the coefficient matrices $\tilde{A} = (\tilde{a}_{ij})$, $A = (a_{ij})$ and vectors $\tilde{c} = (\tilde{c}_1, \ldots, \tilde{c}_s)^T$, $\tilde{b} = (\tilde{b}_1, \ldots, \tilde{b}_s)^T$, $c = (c_1, \ldots, c_s)^T$, $b = (b_1, \ldots, b_s)^T$. They can be represented by a double *tableau* in the usual Butcher notation,

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

The coefficients $\tilde{c}$ and c are given by the usual relation,

$$\tilde{c}_i = \sum_{j=1}^{i-1} \tilde{a}_{ij}, \quad c_i = \sum_{j=1}^{i} a_{ij}, \tag{6}$$

which allows the results of our analysis to be extended to nonautonomous systems. We shall use the notation Name(s, σ, p), where this triplet characterizes the number s of the stages of the implicit scheme, the number σ of stages of the explicit scheme and the combined order of the method, p. Now we give some definitions that we will use later.

DEFINITION 2.1. *We call q_i the stage order of the ith stage of an R-K method if and only if for a problem $\dot{y}(t) = f(t, y(t))$, with $0 \leq t \leq T$ and f a smooth function, the intermediate local errors $y(t_n + c_i h) - Y_i = \mathcal{O}(h^{q_i+1})$, where $Y_i = y(t_n) + h \sum_{j=1}^{s} a_{ij} f(t_n + c_j h, Y_j)$ $(1 \leq i \leq s)$.*

Remark. For stiff differential equations the *stage order* q is an essential ingredient. It is defined by the condition $C(q)$ (see [12] and [13, sect. IV.5]), i.e.,

$$\sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{c_i^k}{k} \quad \text{for } k = 1, \ldots, q \quad \text{and all } \ i. \tag{7}$$

For an s-stage DIRK method, the stage order is 1.

DEFINITION 2.2. *Methods that satisfy the condition $a_{sj} = b_j$, $j = 1, \ldots, s$, are called stiffly accurate.*

Remark. In our analysis we indicate $R(\infty) = \lim_{z \to \infty} R(z)$, with $R(z)$ *the stability function* of the implicit scheme, defined by $R(z) = 1 + z b^T (I - zA)^{-1} \mathbf{1}$ (see [13, sect. IV.3]), with $b^T = (b_1, \ldots, b_s)$ and $\mathbf{1} = (1, \ldots, 1)^T$. From the expression of $R(z)$ follows $R(\infty) = 1 - \sum_{i,j=1}^{s} b_i \omega_{ij}$ with ω_{ij} elements of the inverse of (a_{ij}). Moreover, if the implicit method is stiffly accurate and the matrix A is invertible, one has always $R(\infty) = 0$. We shall use the notation $\tilde{q}_i$, with $\tilde{q}_i \geq 1$, to indicate the *stage order* of the ith stage of the explicit part of the IMEX R-K method, and with q_i, $q_i \geq 1$, the *stage order* of the ith stage of the implicit one. IMEX R-K methods present in the literature can be classified in three different types characterized by the structure of the matrix $A = (a_{ij})_{i,j=1}^{s}$ of the implicit scheme.

DEFINITION 2.3. *We call an IMEX R-K method* type A *(see* [18]*) if the matrix $A \in R^{s \times s}$ is invertible.*

DEFINITION 2.4. *We call an IMEX R-K method* type CK *(see* [8]*) if the matrix $A \in R^{s \times s}$ can be written as*

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

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

Remark. IMEX R-K methods of type ARS (see [1]) are a special case of type CK with the vector $a = 0$.

3. Main results. Motivated by the procedure first suggested by Hairer, Lubich, and Roche [12] (see also [13]), we extend this analysis to different types of IMEX R-K methods. The main results of this paper are summarized in this section in the form of theorems. The aim of these theorems is to present convergence results of these methods when applied to SPP (2). We suppose that the initial values lie on a suitable manifold that allows smooth solutions even in the limit of infinite stiffness and the step size $h = \Delta t \gg \varepsilon$. In fact, arbitrary initial values introduce in the solution a fast transient. One possible way to overcome this difficulty is simply to ensure that the numerical method resolves the transient phase by taking time step $h \ll \varepsilon$ in the first few steps. Then the following results are obtained assuming that the transient phase is over.

An essential ingredient to obtaining these results is to assume that the system is dissipative. More precisely, we assume that

$$\mu\left(g_z(y,z)\right) \le -1 \tag{8}$$

in an ε-independent neighborhood of the solution, where μ denotes the logarithmic norm with respect to some inner product. Condition (8) guarantees the existence of an ε-expansion of problem (2) (see [13, p. 390]).

The proof of the theorems below will be a consequence of the results of sections 5 to 7. We start by considering the limit case $\varepsilon = 0$ (*the reduced problem* or *problems of index* 1) for problem (2).

THEOREM 3.1 (type A). *Consider the stiff problem* (2), (8) *with initial values* $y(0)$, $z(0)$ *admitting a smooth solution. Apply the type-A IMEX R-K method* (3)–(5) *and let* p *be the order of explicit scheme. Assume that the method with coefficients* b_i *and* a_{ij} *is A-stable, that the stability function satisfies* $|R(\infty)| < 1$, *and that* $a_{ii} > 0$ *for all* i. *Furthermore, assume that the weights satisfy the condition* $\tilde{b}_i = b_i$ *for* $i = 1, \dots, s$.

Then if $\sum_{i,j=1}^{s} b_i \omega_{ij} \tilde{c}_j = 1$, *with* ω_{ij} *elements of the inverse matrix of* A, *for any fixed constant* $C > 0$, *the global error satisfies*

$$y_n - y(t_n) = \mathcal{O}(h^p) + \mathcal{O}(\varepsilon h^2), \quad z_n - z(t_n) = \mathcal{O}(h^2)$$

for $\varepsilon \le Ch$; *otherwise, we obtain*

$$y_n - y(t_n) = \mathcal{O}(h^p) + \mathcal{O}(\varepsilon h), \quad z_n - z(t_n) = \mathcal{O}(h).$$

If in addition $a_{si} = b_i$ *and* $\tilde{a}_{si} = \tilde{b}_i$ *for all* i, *we have* $z_n - z(x_n) = \mathcal{O}(h^p) + \mathcal{O}(\varepsilon h^2)$. *The estimates hold uniformly for* $h \le h_0$ *and* $nh \le$ Const.

THEOREM 3.2 (type CK). *Consider the stiff problem* (2), (8) *with initial values* $y(0)$, $z(0)$ *admitting a smooth solution. Apply the type-CK IMEX R-K method* (3)–(5) *with invertible matrix* $\hat{A}$ *and let* p *be the order of the explicit scheme. Assume that the method, with coefficients* b_i *and* a_{ij}, *is A-stable, that the stability function satisfies* $|R(\infty)| < 1$, *and that* $a_{ii} > 0$ *for all* i. *Assume that the weights satisfy the condition* $\tilde{b}_i = b_i$ *for* $i = 1, \dots, s$ *and that the method is stiffly accurate. Then, for any fixed constant* $C > 0$, *the global error satisfies, for* $\varepsilon \le Ch$,

$$y_n - y(x_n) = \mathcal{O}(h^{\tilde{q}+2} + h^p) + \mathcal{O}(\varepsilon h^2), \quad z_n - z(x_n) = \mathcal{O}(h^{\tilde{q}+1} + h^p) + \mathcal{O}(\varepsilon h) \tag{9}$$

with $\tilde{q} = \min\{\tilde{q}_s,\ \tilde{q}_i + 1$ *for all* $i = 2, \dots, s-1\}$. *The estimates hold uniformly for* $h \le h_0$ *and* $nh \le$ Const.

COROLLARY 3.1 (type ARS). *Under the same assumptions of Theorem* 3.2 *and with* $b_1 = 0$, *the global error satisfies* (9). *These estimates hold uniformly for* $h \le h_0$ *and* $nh \le$ Const.

Remark. Next, we shall show that if the method of type ARS is not *stiffly accurate*, one obtains the following estimates:

$$y_n - y(x_n) = \mathcal{O}(h^p + h^3) + \mathcal{O}(\varepsilon h^2), \quad z_n - z(x_n) = \mathcal{O}(h^p + h^2) + \mathcal{O}(\varepsilon h).$$

3.1. Numerical evidence. Before we provide proof of the main theorems, we present numerical results for the different types of IMEX R-K methods developed in the literature (see, e.g., [8], [1], [18], [19]), which confirm the theoretical prediction. Specifically, we will conduct convergence tests to compare the performance of different types of methods. As an example of a stiff problem (2) we consider one of the simplest nonlinear equations (describing nonlinear oscillations) in the stiff literature, the *van der Pol equation*

$$y' = z, \quad \varepsilon z' = (1 - y^2)z - y \tag{10}$$

with $0 \le \varepsilon \ll 1$. When the stiffness parameter ε is sufficiently small, numerical results confirm order reduction especially for the algebraic z-component. In our experiment, errors are computed by choosing initial values

$$y(0) = 2, \qquad z(0) = -\frac{2}{3} + \frac{10}{81}\varepsilon - \frac{292}{2187}\varepsilon^2 - \frac{1814}{19683}\varepsilon^3 + \mathcal{O}(\varepsilon^4) \tag{11}$$

such that the solution is smooth, and $\varepsilon = 10^{-6}$. In the following figures we have plotted the relative global error at $t_{end} = 0.55139$ as a function of the step size h, which was taken to be a constant over the considered interval $[0, t_{end}]$. We use logarithmic scales in both directions. The relative global error behaves like $C \cdot h^r$, where r is the slope of the straight line and C is a constant. We have indicated this behavior in all figures.

Table 1 shows the different types of IMEX R-K methods together with the global errors predicted by Theorems 3.1 and 3.2 and Corollary 3.1. Several conclusions are drawn from the numerical tests.

3.2. Discussion. (a) In Figures 1–7 we see that whenever p is small or h is very large the $\mathcal{O}(h^p)$ term is dominant in the z-component, whereas the other terms can be seen behaving otherwise. Furthermore the estimates in Table 1 demonstrate order reduction for the algebraic component in every type of method for a sufficiently stiff parameter ($\varepsilon = 10^{-6}$).

(b) An important ingredient, suggested by the analysis, is the condition $\tilde{b}_i = b_i$ for all i. Such a choice provides a significant benefit for the differential y-component. In fact the ARS$(4, 4, 3)$ method does not satisfy this condition, and for the y-component the global error drops to first order for a range of the step h. Note, however, that in Theorems 6.1 and 6.2 a satisfactory theoretical explanation of this fact is given.

In particular, the ARS$(4, 4, 3)$ method satisfies the conditions $\tilde{a}_{si} = \tilde{b}_i$ $\tilde{a}_{si} = \tilde{b}_i$ for all i, and in the next sections we shall observe that as a consequence of the above the z-component has the same estimate of the convergence rate as the y-component, justifying the behavior shown in Figure 3.

TABLE 1
Global errors predicted by theorems for the van der Pol equation.

Method	Stiffly accurate	y-comp.	z-comp.
ARS$(3,4,3)$, [1]	yes	$h^3+\varepsilon h^2$	h^2
MARS$(3,4,3)$	yes	$h^3+\varepsilon h^2$	$h^3+\varepsilon h$
ARS$(4,4,3)$, [1]	yes	$h^3+\varepsilon h$	$h^3+\varepsilon h$
ARK3(2)4L[2]SA, [8]	yes	h^3	h^2
ARK5(4)8L[2]SA, [8]	yes	$h^4+\varepsilon h^2$	$h^3+\varepsilon h$
ARK4(3)6L[2]SA, [8]	yes	h^4	$h^3+\varepsilon h$
MARK3(2)4L[2]SA	yes	h^3	$h^3+\varepsilon h$
IMEX-SSP2$(3,3,2)$, [18]	yes	h^2	h^2
IMEX-SSP3$(3,3,2)$, [18]	no	$h^3+\varepsilon h$	h
IMEX-SSP3$(4,3,3)$, [18]	no	$h^3+\varepsilon h$	h

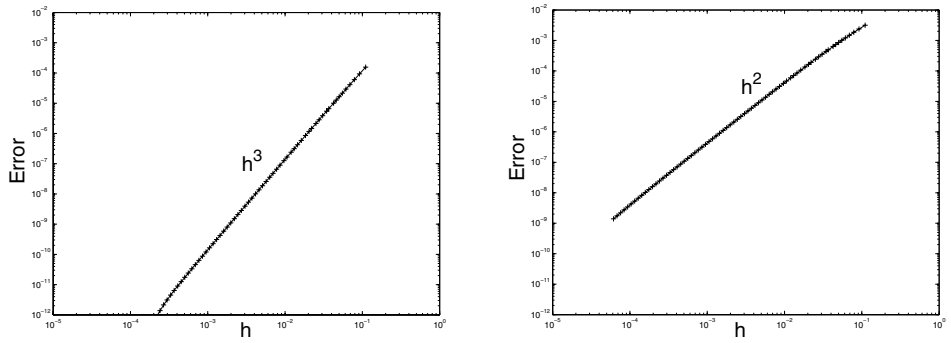

FIG. 1. *Global error versus the step size h for the ARS$(3,4,3)$-IMEX method using the van der Pol equation with $\varepsilon = 10^{-6}$. On the left-hand side is the y-component; on the right-hand side is the z-component.*

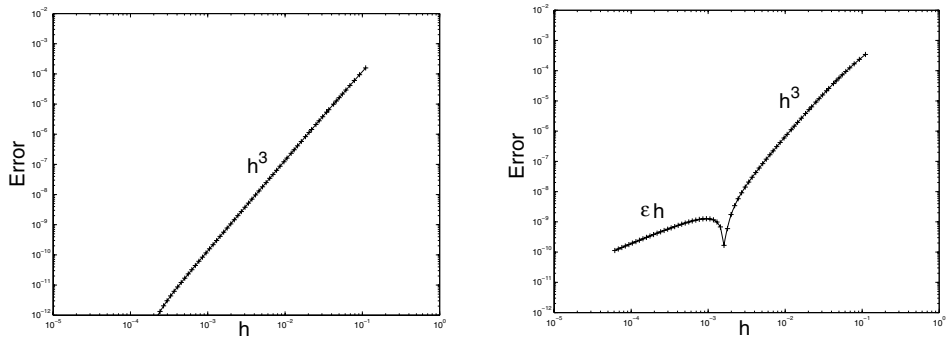

FIG. 2. *Global error versus the step size h for the MARS$(3,4,3)$-IMEX method using the van der Pol equation with $\varepsilon = 10^{-6}$. On the left-hand side is the y-component; on the right-hand side is the z-component.*

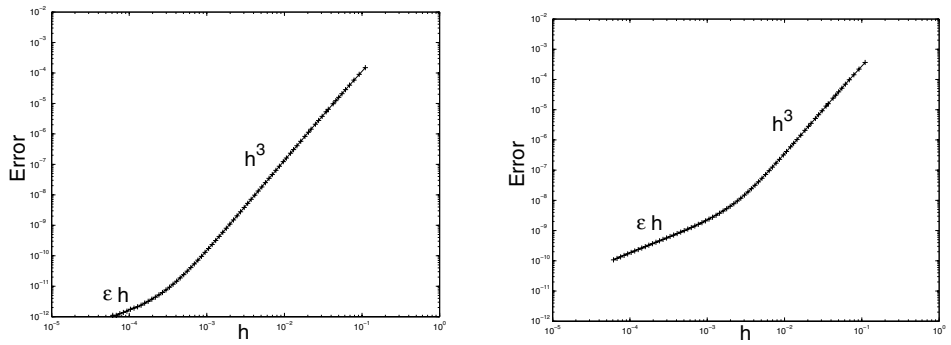

FIG. 3. *Global error versus the step size h for the ARS*$(4,4,3)$*-IMEX method using the van der Pol equation with* $\varepsilon = 10^{-6}$*. On the left-hand side is the y-component; on the right-hand side is the z-component.*

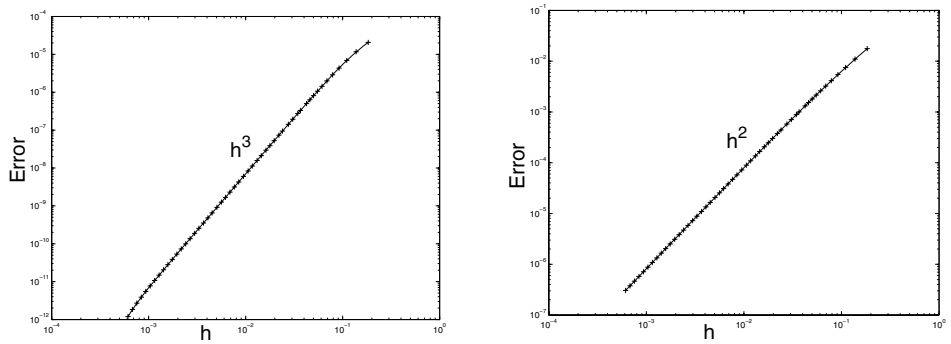

FIG. 4. *Global error versus the step size h for the ARK*$3(2)4L[2]$*SA-IMEX method using the van der Pol equation with* $\varepsilon = 10^{-6}$*. On the left-hand side is the y-component; on the right-hand side is the z-component.*

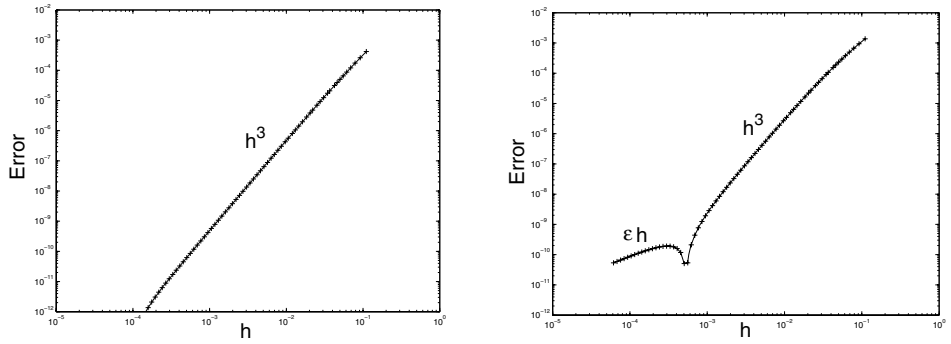

FIG. 5. *Global error versus the step size h for the MARK*$3(2)4L[2]$*SA-IMEX method using the van der Pol equation with* $\varepsilon = 10^{-6}$*. On the left-hand side is the y-component; on the right-hand side is the z-component.*

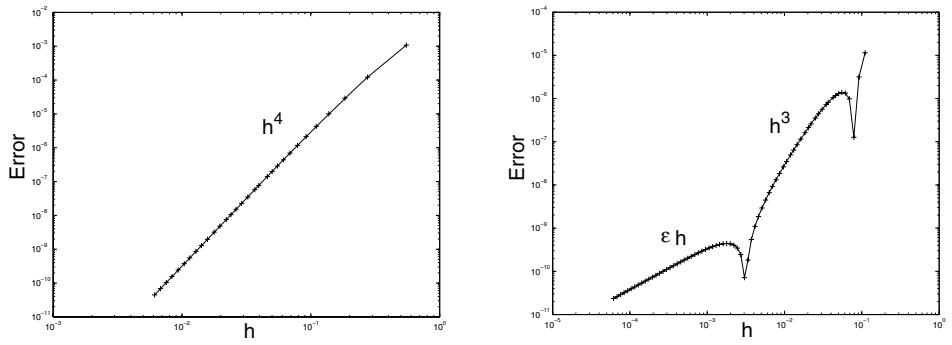

FIG. 6. *Global error versus the step size h for the ARK*4(3)6*L*[2]*SA-IMEX method using the van der Pol equation with* $\varepsilon = 10^{-6}$. *On the left-hand side is the y-component; on the right-hand side is the z-component.*

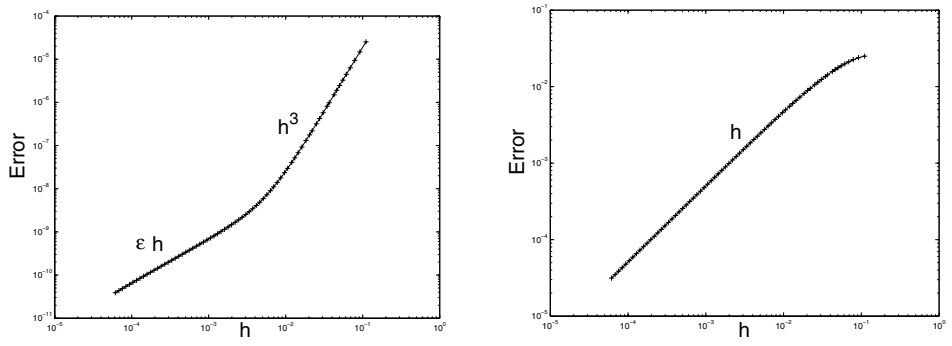

FIG. 7. *Global error versus the step size h for a type A-SSP*$(4,3,3)$*-IMEX method using the van der Pol equation with* $\varepsilon = 10^{-6}$. *On the left-hand side is the y-component; on the right-hand side is the z-component.*

(c) As noted in [8], according to the estimated convergence rates for differential and algebraic variables in [8, Table 12], several IMEX ARK_2 methods confirm the theoretical estimates given in Theorem 3.2. For instance, in order to justify the behavior observed in Figure 6, several pertinent assumptions are satisfied: $b_2 = \tilde{b}_2 = 0$ as well as the formula

$$\sum_{j=1}^{s} \tilde{a}_{ij} c_j = \frac{c_i^2}{2} \tag{12}$$

for $i = 3, \ldots, s$. Thus, using these assumptions, we achieve the estimates in Theorem 3.2.

(d) Finally, it is worth mentioning that the IMEX-SSP3$(4,3,3)$ scheme, as shown in Figure 7, exhibits order reduction both in the differential and algebraic components. Similarly, plots for the IMEX-SSP3$(3,3,2)$ scheme yield similar results. This behavior appears since the IMEX-SSP3$(3,3,2)$ and IMEX-SSP3$(4,3,3)$ schemes don't satisfy the condition $\sum_{ij} b_i \omega_{ij} \tilde{c}_j = 1$ required in Theorem 3.1. On the other hand, the IMEX-SSP2(3,3,2) scheme satisfies this condition so it achieves the anticipated convergence rate.

Improvements of existing schemes. A most relevant point demonstrated by this test is that methods such as *modified* ARK3(2)4L[2]SA (MARK3(2)4L[2]SA) and *modified* ARS$(3,4,3)$ (MARS$(3,4,3)$) produce an estimate for the z-component of the following form:

$$z_n - z(t_n) = \mathcal{O}(h^3) + \mathcal{O}(\varepsilon h) + \mathcal{O}(\varepsilon^2). \tag{13}$$

In this result the term $\mathcal{O}(\varepsilon^2)$ can be neglected since $\varepsilon \ll h$. Furthermore, to illustrate the results shown in Figures 2 and 5, we note that if the step size $h > \varepsilon^{1/2}$, the $\mathcal{O}(h^3)$ term is dominant; otherwise the term $\mathcal{O}(\varepsilon h)$ can be observed. A singularity appears in the neighborhood of $h \approx \varepsilon^{1/2}$ where we have a cancellation of error terms $\mathcal{O}(h^3)$ and $\mathcal{O}(\varepsilon h)$ with error constants of an opposite sign.

Therefore, the modified schemes give an improvement in the error estimate for the z-component when compared to the ARK3(2)4L[2]SA and ARS$(3,4,3)$ methods. In the following sections we will see that the global error estimates of these methods depend on $\tilde{q} = \min\{\tilde{q}_s,\ \tilde{q}_i + 1 \text{ for all } i = 2, \dots, s-1\}$. This fact enables us to construct methods with more accuracy. Notice the following:

(i) For the MARS$(3,4,3)$ method, a natural way to achieve the error estimate (13) is to increase from 1 to 2 the stage order in the sth stage of the explicit scheme so that $\tilde{q} = 2$.

(ii) In order to reach estimate (13) in the case of the MARK3(2)4L[2]SA method, we suggest using formula (12), for $i = 3, \dots, s$, in the explicit scheme accompanied by the assumption $\tilde{b}_2 = 0$. The assumption $\tilde{b}_2 = 0$ is necessary because the assumption (12) cannot be satisfied for $i = 2$; otherwise we would have $c_2 = 0$ and the method would be equivalent to one with fewer stages.

4. Asymptotic expansion. To obtain our main results in a general setting, we start from the ε-expansion of the exact solution of problem (2). Here, in particular, we are interested in smooth solutions which are of the form

$$\begin{aligned} y(t) &= y_0(t) + \varepsilon y_1(t) + \varepsilon^2 y_2(t) + \cdots, \\ z(t) &= z_0(t) + \varepsilon z_1(t) + \varepsilon^2 z_2(t) + \cdots, \end{aligned} \tag{14}$$

where $y_i(t)$ and $z_i(t)$ are ε-independent functions, which are solutions of a sequence of DAEs of arbitrary index.

The aim in this section is to analyze the ε-expansion of the numerical solution for problem (2) and verify how a sequence of differential-algebraic systems arise in the study of such a problem. A general and detailed investigation about the ε-expansion of the exact solution for problem (2) is given in [13] and [16].

We consider the IMEX R-K method (3), (5). We formally expand the quantities Y_{ni}, k_{ni}, y_n, Z_{ni}, ℓ_{ni}, and z_n into powers of ε with ε-independent coefficients:

$$y_n = y_n^0 + \varepsilon y_n^1 + \varepsilon^2 y_n^2 + \cdots, \tag{15a}$$

$$Y_{ni} = Y_{ni}^0 + \varepsilon Y_{ni}^1 + \varepsilon^2 Y_{ni}^2 + \cdots, \tag{15b}$$

$$k_{ni} = k_{ni}^0 + \varepsilon k_{ni}^1 + \varepsilon^2 k_{ni}^2 + \cdots, \tag{15c}$$

$$z_n = z_n^0 + \varepsilon z_n^1 + \varepsilon^2 z_n^2 + \cdots, \tag{15d}$$

$$Z_{ni} = Z_{ni}^0 + \varepsilon Z_{ni}^1 + \varepsilon^2 Z_{ni}^2 + \cdots, \tag{15e}$$

$$\ell_{ni} = \varepsilon^{-1}\ell_{ni}^{-1} + \ell_{ni}^0 + \varepsilon \ell_{ni}^1 + \varepsilon^2 \ell_{ni}^2 + \cdots. \tag{15f}$$

Because of the linearity of relations (3) and (5) we have to order ε^ν, with $\nu = -1$,

$$0 = h\sum_{j=1}^{i} a_{ij}\ell_{nj}^{-1}, \quad 0 = h\sum_{i=1}^{s} b_i\ell_{ni}^{-1} \tag{16}$$

and, for $\nu \geq 0$,

$$\begin{pmatrix} y_{n+1}^\nu \\ z_{n+1}^\nu \end{pmatrix} = \begin{pmatrix} y_n^\nu \\ z_n^\nu \end{pmatrix} + h\sum_{i=1}^{s} \begin{pmatrix} \tilde{b}_i k_{ni}^\nu \\ b_i \ell_{ni}^\nu \end{pmatrix}, \tag{17}$$

$$\begin{pmatrix} Y_{ni}^\nu \\ Z_{ni}^\nu \end{pmatrix} = \begin{pmatrix} y_n^\nu \\ z_n^\nu \end{pmatrix} + h \begin{pmatrix} \sum_{j=1}^{i-1} \tilde{a}_{ij} k_{nj}^\nu \\ \sum_{j=1}^{i} a_{ij} \ell_{nj}^\nu \end{pmatrix}. \tag{18}$$

Inserting (15b), (15c), (15e), and (15f) into (4) and comparing equal powers of ε, we obtain

$$\varepsilon^0 : \begin{cases} k_{ni}^0 = f(Y_{ni}^0, Z_{ni}^0), \\ \ell_{ni}^{-1} = g(Y_{ni}^0, Z_{ni}^0), \end{cases} \tag{19a}$$

$$\varepsilon^1 : \begin{cases} k_{ni}^1 = f_y(Y_{ni}^0, Z_{ni}^0)Y_{ni}^1 + f_z(Y_{ni}^0, Z_{ni}^0)Z_{ni}^1, \\ \ell_{ni}^0 = g_y(Y_{ni}^0, Z_{ni}^0)Y_{ni}^1 + g_z(Y_{ni}^0, Z_{ni}^0)Z_{ni}^1, \end{cases} \tag{19b}$$

$$\cdots\cdots$$

$$\varepsilon^\nu : \begin{cases} k_{ni}^\nu = f_y(Y_{ni}^0, Z_{ni}^0)Y_{ni}^\nu + f_z(Y_{ni}^0, Z_{ni}^0)Z_{ni}^\nu + \varphi_\nu(Y_{ni}^0, Z_{ni}^0, \ldots, Y_{ni}^{\nu-1}, Z_{ni}^{\nu-1}), \\ \ell_{ni}^{\nu-1} = g_y(Y_{ni}^0, Z_{ni}^0)Y_{ni}^\nu + g_z(Y_{ni}^0, Z_{ni}^0)Z_{ni}^\nu + \psi_\nu(Y_{ni}^0, Z_{ni}^0, \ldots, Y_{ni}^{\nu-1}, Z_{ni}^{\nu-1}). \end{cases} \tag{19c}$$

Since (4) has a similar form to (2), the formulas (19a), (19b), and (19c) are exactly the same as those of the expansion in powers of ε for the exact solution (see [13] and [12]). In response to this fact, it follows that the coefficients y_n^0, z_n^0, y_n^1, z_n^1, ... represent the numerical solution of an arbitrary IMEX R-K method applied to DAEs of arbitrary index. Finally, subtracting (15a) and (15d) from (14), we get formally

$$y_n - y(t_n) = \sum_{\nu\geq 0} \varepsilon^\nu (y_n^\nu - y_\nu(t_n)), \quad z_n - z(t_n) = \sum_{\nu\geq 0} \varepsilon^\nu (z_n^\nu - z_\nu(t_n)). \tag{20}$$

Hence, the error of the numerical solution possesses an ε-expansion whose coefficients are the errors of the method applied to the differential-algebraic system. Clearly, in order to study this error, one will investigate only the differences $y_n^\nu - y_\nu(t_n)$, $z_n^\nu - z_\nu(t_n)$.

5. Zeroth-order expansion (index 1). From an arbitrary SPP (2) now we want to study the behavior of the global error of different types of IMEX R-K schemes for $\varepsilon \to 0$. In this section we start by studying the limiting case $\varepsilon = 0$. This gives us the corresponding *reduced* problem

$$\begin{aligned} y' &= f(y, z), \\ 0 &= g(y, z). \end{aligned} \tag{21}$$

We assume that $g_z(y,z)$ is invertible in a neighborhood of the solution of (21). This assumption guarantees the solvability of (21) and that the equation $g(y,z)=0$ possesses a locally unique solution (implicit function theorem). Furthermore, the same assumption guarantees that system (21) is a differential-algebraic one of *index* 1 [13]. Therefore, our first goal is to consider the different types of schemes applied to the reduced problem.

Type A. An IMEX R-K method of type A applied to the reduced problem has the form

$$Y_{ni} = y_n + h\sum_{j=1}^{i-1} \tilde{a}_{ij} f(Y_{nj}, Z_{nj}), \tag{22a}$$

$$0 = g(Y_{ni}, Z_{ni}), \tag{22b}$$

$$y_{n+1} = y_n + h\sum_{i=1}^{s} \tilde{b}_i f(Y_{ni}, Z_{ni}), \tag{22c}$$

$$z_{n+1} = R(\infty) z_n + \sum_{i,j=1}^{s} b_i \omega_{ij} Z_{nj}. \tag{22d}$$

Remarks. (a) By the implicit function theorem applied to (22b), we have $Z_{ni} = G(Y_{ni})$ for $i=1,\dots,s$. Consequently, by $Y_{ni} = y(t_n + \tilde{c}_i h) + \mathcal{O}(h^{\tilde{q}_i+1})$, it follows that the *internal stages* Z_{ni} depend on the coefficients $\tilde{c}_i$ of the explicit scheme.

(b) Concerning system (21), the y-component can be interpreted as the numerical solution of the ordinary differential equation $y' = f(y, H(y))$ with $z = H(y)$ (implicit function theorem). Therefore, for the method (22a)–(22d) we have

$$y_n - y(t_n) = \mathcal{O}(h^p),$$

because the formulas (22a), (22b), and (22c) are independent of z_n with p the order of the explicit scheme. Thus, we have only to prove a convergence result for the z-component.

Type CK. By Definition 2.4, we assume submatrix $\hat{A}$ is invertible. By (16), we have $0 = h a_{i1}\ell_{n1}^{-1} + h\sum_{j=2}^{i} a_{ij}\ell_{nj}^{-1}$ for $i=2,\dots,s$. Now, by the fact that $\ell_{n1}^{-1} = g(y_n, z_n)$, we obtain

$$\ell_{ni}^{-1} = \alpha_i g(y_n, z_n), \tag{23}$$

where $\alpha_i = -\sum_{i,j=2}^{s} \hat{\omega}_{ij} a_{j1}$ for $i=2,\dots,s$, with $\hat{\omega}_{ij}$ elements of the inverse matrix of $\hat{A}$.

Now, looking at (16), Lemma 5.1 follows from (23) and $\ell_{n1}^{-1} = g(y_n, z_n)$.

LEMMA 5.1. *The condition*

$$b_1 + \sum_{i=2}^{s} b_i \alpha_i = 0 \tag{24}$$

is automatically satisfied if the IMEX R-K method of type CK is stiffly accurate.

Therefore we assume that the method is stiffly accurate in the implicit part. This, moreover, yields $z_{n+1} = Z_{ns}$. Next we will use this lemma.

Then for the reduced problem a type-CK IMEX R-K scheme is defined by

$$Y_{ni} = y_n + h\sum_{j=1}^{i-1} \tilde{a}_{ij} f(Y_{nj}, Z_{nj}), \tag{25a}$$

$$y_{n+1} = y_n + h\sum_{i=1}^{s} \tilde{b}_i f(Y_{ni}, Z_{ni}), \tag{25b}$$

$$g(Y_{ni}, Z_{ni}) = \alpha_i g(y_n, z_n), \quad i = 2, \dots, s-1, \tag{25c}$$

$$g(Y_{ns}, z_{n+1}) = \alpha_s g(y_n, z_n). \tag{25d}$$

Type ARS. Since this is a particular case of CK with $a_{i1} = 0$ it follows that $\alpha_i = 0$ and $G(Y_{ni}, Z_{ni}) = 0$ for $i = 2, \dots, s$. As an immediate consequence, we get explicitly $z_{n+1} = R(\infty) z_n + \sum_{i,j=2}^{s} b_j \hat{\omega}_{ij} Z_{nj}$. In particular, if the method is stiffly accurate, $z_{n+1} = Z_{ns}$. In particular more theoretical insight into this type ARS shows that if we have $a_{si} = b_i$ and $\tilde{a}_{si} = \tilde{b}_i$ for $i = 1, \dots, s$, it also follows that $g(y_{n+1}, z_{n+1}) = 0$. Thus if $g_z(y, z)$ is invertible, we may express z_{n+1} as a function of y_{n+1}, and therefore we can declare that the z-component has the same asymptotic error estimate as the y-component.

After having understood the structure of each method, we are now in a position to prove the following results. All the theorems below are built on the assumption that the reduced problem satisfies (8) in a neighborhood of the exact solution $(y(t), z(t))$, and we assume that the initial values are *consistent*, i.e., $g(y_0, z_0) = 0$.

THEOREM 5.1 (type A). *Consider an IMEX R-K method of type A. Let p be the classical order of the explicit R-K method. Assume that the stability function of the implicit scheme satisfies $|R(\infty)| < 1$. Then the numerical solution of* (22a)–(22d) *has global error*

$$z_n - z(t_n) = \begin{cases} \mathcal{O}(h^2) & \text{if } \sum_{i,j=1}^{s} b_i \omega_{ij} \tilde{c}_j = 1, \\ \mathcal{O}(h) & \text{if } \sum_{i,j=1}^{s} b_i \omega_{ij} \tilde{c}_j \neq 1. \end{cases} \tag{26}$$

The estimates (26) *hold uniformly for $t_n - t_0 = nh \leq$ Const.*

Proof. We denote the global error by $\Delta z_n = z_n - z(t_n)$ and $R(\infty) = \rho$. By remark (b), we get $Z_{ni} = z(t_n) + \tilde{c}_i h z'(t_n) + \mathcal{O}(h^2)$. Now, inserting it into (22d) and considering $z(t_{n+1}) = z(t_n) + h z'(t_n) + \mathcal{O}(h^2)$, one obtains

$$\Delta z_{n+1} = \rho \Delta z_n + \left(\sum_{i,j=1}^{s} b_i \omega_{ij} \tilde{c}_j - 1 \right) h z'(t_n) + \mathcal{O}(h^2),$$

which allows us to conclude that

$$\Delta z_{n+1} = \begin{cases} \rho \Delta z_n + \delta_{n+1} & \text{if } \sum_{i,j=1}^{s} b_i \omega_{ij} \tilde{c}_j = 1, \\ \rho \Delta z_n + \delta_{n+1} & \text{if } \sum_{i,j=1}^{s} b_i \omega_{ij} \tilde{c}_j \neq 1, \end{cases} \tag{27}$$

where

$$\delta_{n+1} = \begin{cases} O(h^2) & \text{if } \sum_{i,j=1}^{s} b_i \omega_{ij} \tilde{c}_j = 1, \\ O(h) & \text{if } \sum_{i,j=1}^{s} b_i \omega_{ij} \tilde{c}_j \neq 1. \end{cases} \tag{28}$$

Finally, repeated insertion of these formulas gives

$$\Delta z_n = \begin{cases} \sum_{i=1}^{n} \rho^{n-j}\delta_j & \text{if } \sum_{i,j=1}^{s} b_i\omega_{ij}\tilde{c_j} = 1, \\ \sum_{i=1}^{n} \rho^{n-j}\delta j & \text{if } \sum_{i,j=1}^{s} b_i\omega_{ij}\tilde{c_j} \neq 1 \end{cases} \tag{29}$$

because $\Delta z_0 = 0$. Thus, by the hypothesis $|\rho| < 1$, we obtain

$$\Delta z_n = \begin{cases} \mathcal{O}(h^2) & \text{if } \sum_{i,j=1}^{s} b_i\omega_{ij}\tilde{c_j} = 1, \\ \mathcal{O}(h) & \text{if } \sum_{i,j=1}^{s} b_i\omega_{ij}\tilde{c_j} \neq 1. \end{cases} \tag{30}$$

□

Remark. If the IMEX R-K method is stiffly accurate, it follows by (22b) that $z_{n+1} = Z_{ns} = G(Y_{ns})$. By remark (b), since we get $Z_{ns} - z(t_n + \tilde{c}_s h) = G(Y_{ns}) - G(y(t_n + \tilde{c}_s h)) = \mathcal{O}(h^{\tilde{q}_s+1})$, if $\tilde{c}_s = 1$, this proves the following estimate: $z_n - z(t_n) = \mathcal{O}(h^{\tilde{q}_s+1})$. Moreover, if in the explicit part we also have $\tilde{a}_{si} = \tilde{b}_i$ for $i = 1, \dots, s$, this yields $y_{n+1} = Y_{ns}$. Therefore, by $g(y_{n+1}, z_{n+1}) = 0$ and by the implicit function theorem, it follows that $z_{n+1} = G(y_{n+1})$, and in this situation the estimate is $z_n - z(t_n) = \mathcal{O}(h^p)$.

THEOREM 5.2 (type CK). *Consider an IMEX R-K method of type KC stiffly accurate with invertible matrix $\hat{A}$ and weights $\tilde{b}_i = b_i$ for $i = 1, \dots, s$. Let p be the order of explicit scheme. Assume that the stability function of the implicit scheme satisfies $|R(\infty)| < 1$ and $\delta = |\alpha_s| < 1$. Then the numerical solution of* (25a)–(25d) *has global error*

$$y_n - y(t_n) = \mathcal{O}(h^{\tilde{q}+2}) + \mathcal{O}(h^p), \qquad z_n - z(t_n) = \mathcal{O}(h^{\tilde{q}+1}) + \mathcal{O}(h^p) \tag{31}$$

with $\tilde{q} = \min\{\tilde{q}_s, \tilde{q}_i + 1,\ i = 2, \dots, s-1\}$. These estimates holds uniformly for $nh \leq$ Const.

Proof. By the relation (23), it follows that

$$g(Y_{ni}, Z_{ni}) = \alpha_i g(y_n, z_n), \tag{32}$$

so that Z_{ni} is a function of Y_{ni}, y_n, and z_n for $i = 2, \dots, s$. On the other hand, to provide an optimal estimate for the local error of the y-component we introduce the internal stages U_{ni}, V_{ni} that satisfy the relation

$$g(U_{ni}, V_{ni}) = \alpha_i g(y(t_n), z(t_n)) \tag{33}$$

for $i = 2, \dots, s$. Of course, this implies that V_{ni} is a function of U_{ni}, $y(t_n)$, and $z(t_n)$. Also, the internal stage U_{ni} is defined as

$$U_{ni} = y(t_n) + h\left(\tilde{a}_{i1}y'(t_n) + \sum_{j=2}^{i-1} \tilde{a}_{ij} f(U_{nj}, V_{ni})\right), \tag{34}$$

where $y'(t_n) = f(y(t_n), z(t_n))$ is the exact solution of $y(t)$ in t_n.

Next we shall use the abbreviation $g_z(t_n) = g_z(y(t_n), z(t_n))$, $f_y(t_n) = f_y(y(t_n), z(t_n))$ and denote $\Delta y_n = y_n - y(t_n)$ and $\Delta z_n = z_n - z(t_n)$.

Our proof proceeds in two parts, referred to as (a) and (b).

(a) We first estimate the differences $\|Z_{ni} - V_{ni}\|$, $\|Y_{ni} - U_{ni}\|$ of the internal stages. For this, we subtract $Y_{ni} = y_n + h(\tilde{a}_{i1}f(y_n, z_n) + \sum_{j=2}^{i-1} \tilde{a}_{ij} f(Y_{nj}, Z_{nj}))$ from (34) to obtain

$$\|Y_{ni} - U_{ni}\| \leq \|\Delta y_n\| + \mathcal{O}(h\|y_n\| + h\|\Delta z_n\|) + Ch\sum_{j=2}^{i-1} |\tilde{a}_{ij}| \, \|Z_{nj} - V_{nj}\| \tag{35}$$

for $i = 2, \dots, s$ by the use of a Lipschitz condition for f.

We now linearize (32) and (33). Subtracting the two quantities, by the use of (35) and the condition $g_z^{-1}(t_n + c_i h) g_z(t_n) = I + \mathcal{O}(h)$, we obtain

$$\|Z_{ni} - V_{ni}\| \le |\alpha_i| \, \|\Delta z_n\| + \mathcal{O}(h \, \|\Delta z_n\|) + \mathcal{O}(\|\Delta y_n\|). \tag{36}$$

(b) Our next aim is to prove the recursion

$$\begin{pmatrix} \|\Delta y_{n+1}\| \\ \|\Delta z_{n+1}\| \end{pmatrix} \le \begin{pmatrix} 1 + \mathcal{O}(h) & \mathcal{O}(h^2) \\ \mathcal{O}(1) & \delta + \mathcal{O}(h) \end{pmatrix} \begin{pmatrix} \|\Delta y_n\| \\ \|\Delta z_n\| \end{pmatrix} + \begin{pmatrix} \mathcal{O}(h^{p+1}) \\ \mathcal{O}(h^{\tilde{q}+1}) \end{pmatrix}. \tag{37}$$

For the verification of the first relation in (37) we again linearize the quantities $y(t_n+h)$ and y_{n+1} to obtain

$$\begin{aligned} \Delta y_{n+1} = \Delta y_n + h\tilde{b}_1 (f_y(t_n) \Delta y_n + f_z(t_n) \Delta z_n) + h \sum_{i=2}^{s} \tilde{b}_i (f_y(t_n) \, (Y_{ni} - U_{ni}) \\ + f_z(t_n) \, (Z_{ni} - V_{ni})) + \mathcal{O}(h^2 \, \|y_n\| + h^2 \, \|z_n\|) + \mathcal{O}(h^{p+1}), \end{aligned} \tag{38}$$

and inserting (35) and (36) into (38), we get

$$\|\Delta y_{n+1}\| \le (1 + C_1 h) \, \|\Delta y_n\| + C_2 h^2 \, \|\Delta z_n\| + \mathcal{O}(h^{p+1}). \tag{39}$$

In (39), we applied the statement of Lemma 5.1.

Now we compute the second relation in (37) from (25d) and its exact expression $g(y(t_n + \tilde{c}_s h), z(t_{n+1})) = \alpha_s g(y(t_n), z(t_n))$. Linearizing and subtracting the two quantities, respectively, we obtain

$$\begin{aligned} \Delta z_{n+1} = -g_z^{-1}(t_n + \tilde{c}_s h) g_y(t_n) \Delta Y_{ns} + \alpha_s g_z^{-1}(t_n + \tilde{c}_s h) g_y(t_n) y_n \\ + \alpha_s g_z^{-1}(t_n + \tilde{c}_s h) g_z(t_n) z_n + \mathcal{O}\left(\|\Delta y_n\|^2 + \|\Delta z_n\|^2 \right). \end{aligned}$$

We now assume that

$$\|\Delta y_n\| \le Ch, \quad \|\Delta z_n\| \le Ch, \tag{40}$$

with some fixed constant C.[1] Therefore, by $g_z^{-1}(t_n + h) g_z(t_n) = I + \mathcal{O}(h)$, it follows that

$$\|\Delta z_{n+1}\| \le |\alpha_s| \, \|\Delta z_n\| + \mathcal{O}\left(\|\Delta y_n\| + h \, \|\Delta z_n\| \right) + \mathcal{O}(\|\Delta Y_{ns}\|) \tag{41}$$

as long as assumption (40) is satisfied. Now, using

$$\Delta Y_{ni} = \Delta y_n + h \sum_{j=1}^{i-1} \tilde{a}_{ij} \Delta k_{nj} + \mathcal{O}(h^{\tilde{q}_i + 1}) \tag{42}$$

and a Lipschitz condition for f gives

$$\|\Delta k_{ni}\| \le M \, \|\Delta Y_{ni}\| + N \, \|\Delta Z_{ni}\| \, . \tag{43}$$

[1] This statement should be interpreted to mean that if h is sufficiently small, the numerical solution will never violate the conditions (40).

In order to find an optimal estimate of (41) we proceed as follows. The linearization of (32) and the exact expression $g(y(t_n+\tilde{c}_ih),z(t_n+\tilde{c}_ih))=\alpha_i g(y(t_n),z(t_n))$ yields

$$\|\Delta Z_{ni}\| \le |\alpha_i|\,\|\Delta z_n\| + \mathcal{O}\left(h\,\|\Delta z_n\| + \|\Delta y_n\|\right) + \mathcal{O}(\|\Delta Y_{ni}\|).$$

Inserted into (43), with the help of (42) after repeated insertions of $\|\Delta Y_{ni}\|$, and setting $i=s$, we obtain

$$\|\Delta Y_{ns}\| \le \|\Delta y_n\| + hC_1\left(\|\Delta y_n\| + \|\Delta z_n\|\right) + hC_2(|\alpha_s|\,\|\Delta z_n\| + \|\Delta y_n\|) + \mathcal{O}(h^{\tilde{q}+1})$$

with $\tilde{q}=\min\{\tilde{q}_s,\ \tilde{q}_i+1,\ i=2,\dots,s-1\}$. Now putting the previous formula into (41), it follows that

$$\|\Delta z_{n+1}\| \le C\,\|\Delta y_n\| + (\delta + Ch)\,\|\Delta z_n\| + \mathcal{O}(h^{\tilde{q}+1}), \tag{44}$$

where $\delta=|\alpha_s|$. This completes the proof of formula (37).

Now applying Lemma 5.2 below to (37) gives the estimates (31) for $nh\le$ Const, completing the proof. ∎

LEMMA 5.2. *Let $\{u_n\}$ and $\{v_n\}$ be two sequences of nonnegative numbers satisfying (componentwise)*

$$\begin{pmatrix} u_{n+1} \\ v_{n+1} \end{pmatrix} \le \begin{pmatrix} 1+\mathcal{O}(h) & \mathcal{O}(h^2) \\ \mathcal{O}(1) & \delta+\mathcal{O}(h) \end{pmatrix} \begin{pmatrix} u_n \\ v_n \end{pmatrix} + \begin{pmatrix} \mathcal{O}(h^{p+1}) \\ \mathcal{O}(h^{\tilde{q}+1}) \end{pmatrix} \tag{45}$$

with $0\le\delta<1$. Then the following estimate holds for $nh\le$ Const and $h\le h_0$:

$$\begin{aligned} u_n &\le C(u_0 + h^2v_0 + h^{\tilde{q}+2} + h^p), \\ v_n &\le C(u_0 + (\delta^n + h)\,v_0 + h^{\tilde{q}+1} + h^p). \end{aligned} \tag{46}$$

The proof is similar to that of Lemma 3.9 in [13].

Remarks. It is worth noting that if $b_i=\tilde{b}_i$ for $i=1,\dots,s$, then inserting (35) and (36) into (38) yields the nonzero quantity $\tilde{b}_1+\sum_{i=2}^s\tilde{b}_i\alpha_i$. For the y-component this implies $y_n-y(t_n)=\mathcal{O}(h^p)+\mathcal{O}(h^{\tilde{q}+1})$.

COROLLARY 5.1 (type ARS). *Suppose that the assumptions of Theorem 5.2 are satisfied and $b_1=0$. Then the numerical solution has global error satisfying* (31).

Remarks. We now suppose that the ARS method is not stiffly accurate. In order to obtain an optimal evaluation of Δz_n, we proceed as follows. Since $g(Y_{ni},Z_{ni})=0$, we get $Z_{ni}=G(Y_{ni})$ for $i=2,\dots,s$. By the Lipschitz condition for G, it follows that $\|\Delta Z_{ni}\|\le C\,\|\Delta Y_{ni}\|$. Using (42) and (43), we get

$$\|\Delta Y_{ni}\| \le \|\Delta y_n\| + h\,|\tilde{a}_{i1}|\,(\|\Delta y_n\| + \|\Delta z_n\|) + \mathcal{O}(h^{\tilde{r}_i+1}) \tag{47}$$

with $\tilde{r}_i=\min\{\tilde{q}_i,\tilde{q}_j+1,\ \ j=1,\dots,i-1\}$. It thus follows from the numerical and exact solution that

$$\|\Delta z_{n+1}\| \le |\rho|\,\|\Delta z_n\| + C\sum_{i,j=2}^{s}|b_i\hat{\omega}_{ij}|\,\|\Delta Y_{nj}\| + \mathcal{O}(h^{q+1}), \tag{48}$$

where $\rho=1-\sum_{i,j=2}^s b_i\hat{\omega}_{ij}$ and $q=\min_{i\le s}q_i$. Now, inserting (47) into (48) yields

$$\|\Delta z_{n+1}\| \le (|\rho|+C_2h)\,\|\Delta z_n\| + C_1\,\|\Delta y_n\| + \mathcal{O}(h^{\tilde{r}+1}) + \mathcal{O}(h^{q+1}), \tag{49}$$

where $\tilde{r} = \min\{\tilde{r}_2, \ldots, \tilde{r}_s\}$ with $|\rho| < 1$. We now solve (39) and (49), applying again Lemma 5.2, thus obtaining for the global error

$$y_n - y(t_n) = \mathcal{O}(h^p) + \mathcal{O}(h^3), \quad z_n - z(t_n) = \mathcal{O}(h^p) + \mathcal{O}(h^2).$$

Observe that the estimates obtained above are given since $q = 1$.

It is interesting to note that if $\tilde{b}_1 \neq 0$, in (39) we get $\|\Delta y_{n+1}\| \leq (1+C_1 h)\,\|\Delta y_n\| + C_2 h\,\|\Delta z_n\| + \mathcal{O}(h^{p+1})$ and the proof follows as above.

6. Higher-order expansion (higher index). Now we study the global error of IMEX R-K methods when applied to the SPP (2). To this end, we are interested in studying the differences $y_n^\nu - y_\nu(t_n)$ and $z_n^\nu - z_\nu(t_n)$ from (20). All the theorems below are built on the assumption that the stability function of the implicit scheme satisfies $|R(\infty)| < 1$ and the weights $\tilde{b}_i = b_i$ for all i. In what follows, when we use the superscript 0 in the quantities Y_{ni}, Z_{ni}, k_{ni}, ℓ_{ni}, y_n, z_n, we are treating the behavior of the numerical solution of the *reduced* problem.

THEOREM 6.1 (type A). *Consider an IMEX R-K method of type A such that (a_{ij}) is invertible. Assume* (8) *holds and the initial values of the differential-algebraic system of index $\nu+1$ are consistent. Then if $\sum_{ij}^{s} b_i \omega_{ij} \tilde{c}_j = 1$, the global error of method* (17)–(19c) *satisfies, for $\nu = 1, 2$,*

$$y_n^\nu - y_\nu(t_n) = \mathcal{O}(h^{3-\nu}), \quad z_n^\nu - z_\nu(t_n) = \mathcal{O}(h^{2-\nu}); \tag{50}$$

otherwise

$$y_n^\nu - y_\nu(t_n) = \mathcal{O}(h^{2-\nu}), \quad z_n^\nu - z_\nu(t_n) = \mathcal{O}(h^{1-\nu}). \tag{51}$$

Proof. Here we emphasize some straightforward differences with respect to Theorem 3.4 in [13].

(a) We begin by denoting the differences to the exact solution values:

$$\begin{aligned} \Delta y_n^\nu &= y_n^\nu - y_\nu(t_n), & \Delta z_n^\nu &= z_n^\nu - z_\nu(t_n), \\ \Delta Y_{ni}^\nu &= Y_{ni}^\nu - y_\nu(t_n + \tilde{c}_i h), & \Delta Z_{ni}^\nu &= Z_{ni}^\nu - Z_\nu(t_n + c_i h), \\ \Delta k_{ni}^\nu &= k_{ni}^\nu - y_\nu'(t_n + \tilde{c}_i h), & \Delta \ell_{ni}^\nu &= \ell_{ni}^\nu - z_\nu'(t_n + c_i h). \end{aligned} \tag{52}$$

Furthermore we have for an IMEX R-K method

$$\begin{pmatrix} \Delta Y_{ni}^\nu \\ \Delta Z_{ni}^\nu \end{pmatrix} = \begin{pmatrix} \Delta y_n^\nu \\ \Delta z_n^\nu \end{pmatrix} + h \begin{pmatrix} \sum_{j=1}^{i-1} \tilde{a}_{ij} \Delta k_{nj}^\nu \\ \sum_{j=1}^{i} a_{ij} \Delta \ell_{nj}^\nu \end{pmatrix} + \begin{pmatrix} \mathcal{O}(h^{\tilde{q}_i+1}) \\ \mathcal{O}(h^{q_i+1}) \end{pmatrix}. \tag{53}$$

From Theorem 5.1 it follows that

$$\begin{aligned} \Delta y_n^0 &= \mathcal{O}(h^p), & \Delta Y_{ni}^0 &= \mathcal{O}(h^{\tilde{q}_i+1}), \\ \Delta k_{ni}^0 &= \mathcal{O}(h^{\tilde{q}_i+1}), & \Delta Z_{ni}^0 &= \mathcal{O}(h^{\tilde{q}_i+1}), \end{aligned} \tag{54}$$

and

$$\Delta z_n^0 = \begin{cases} \mathcal{O}(h^2) & \text{if} \quad \sum_{i,j=1}^{s} b_i \omega_{ij} \tilde{c}_j = 1, \\ \mathcal{O}(h) & \text{otherwise.} \end{cases} \tag{55}$$

We also have

$$\Delta \ell_{ni}^0 = \begin{cases} \mathcal{O}(h) & \text{if} \quad \sum_{i,j=1}^{s} \omega_{ij} \tilde{c}_j = 1, \\ \mathcal{O}(1) & \text{otherwise.} \end{cases} \tag{56}$$

Here ω_{ij} are the elements of the inverse of matrix A.

(b) We first consider the case $\nu = 1$. In analogy to the proof of Theorem 3.4 in [13], using the estimates (54), we deduce the following expressions:

$$
\begin{aligned}
\Delta k^1_{ni} &= f_y(t_n + \tilde{c}_i h)\Delta Y^1_{ni} + f_z(t_n + \tilde{c}_i h)\Delta Z^1_{ni} \\
&\qquad + \mathcal{O}(h^{\tilde{q}_i+1} + h^{\tilde{q}_i+1}\left\|\Delta Y^1_{ni}\right\| + h^{\tilde{q}_i+1}\left\|\Delta Z^1_{ni}\right\|), \\
\Delta \ell^0_{ni} &= g_y(t_n + \tilde{c}_i h)\Delta Y^1_{ni} + g_z(t_n + \tilde{c}_i h)\Delta Z^1_{ni} \\
&\qquad + \mathcal{O}(h^{\tilde{q}_i+1} + h^{\tilde{q}_i+1}\left\|\Delta Y^1_{ni}\right\| + h^{\tilde{q}_i+1}\left\|\Delta Z^1_{ni}\right\|).
\end{aligned} \tag{57}
$$

Here we have used the abbreviations $f_y(t) = f_y(y_0(t), z_0(t))$, $g_y(t) = g_y(y_0(t), z_0(t))$. Now, we compute ΔZ^1_{ni} from the second relation in (57). Therefore, inserting it into the first one and using (53), we can eliminate ΔY^1_{ni} and obtain

$$
\begin{aligned}
&\Delta k^1_{ni} - (f_z g_z^{-1})(t_n + \tilde{c}_i h)\Delta \ell^0_{ni} = \mathcal{O}(\left\|\Delta y^1_n\right\|) \\
&+ (f_y - f_z g_z^{-1} g_y)(t_n + \tilde{c}_i h) h \sum_{j=1}^{i-1} \left((f_z g_z^{-1})(t_n + \tilde{c}_j h)\tilde{a}_{ij}\Delta \ell^0_{nj} + \mathcal{O}(h^{\tilde{q}_j+1})\right) + \mathcal{O}(h^{\tilde{q}_i+1}).
\end{aligned}
$$

By (56), it follows that $\Delta k_{ni} = \mathcal{O}(\left\|\Delta y^1_n\right\|) + \mathcal{O}(h)$ if $\sum_{i,j=1}^{i} b_i \omega_{ij} \tilde{c}_j = 1$; otherwise $\Delta k_{ni} = \mathcal{O}(\left\|\Delta y^1_n\right\|) + \mathcal{O}(1)$. A direct estimation of Δy^1_n proves that $\Delta y^1_n = \mathcal{O}(h)$ if $\sum_{i,j=1}^{s} b_i \omega_{ij} \tilde{c}_j = 1$; otherwise $\Delta y^1_n = \mathcal{O}(1)$. However, these estimations are not optimal.

Following the argument in Theorem 3.4 in [13], we now introduce the new variable

$$
\Delta u^1_n = \Delta y^1_n - (f_z g_z^{-1})(t_n)\Delta z^0_n. \tag{58}
$$

At this point the only difference is that we have to treat more carefully the quantity $\Delta k^1_{ni} - f_z g_z^{-1}(t_n)\Delta \ell^0_{ni}$ for all i. For details we refer to [4]. Using the hypothesis $\tilde{b}_i = b_i$ for $i = 1, \ldots, s$, we obtain

$$
\begin{aligned}
\Delta u^1_{n+1} &= \Delta u^1_n + h \sum_{i=1}^{s} b_i \Bigg(\mathcal{O}(\left\|\Delta y^1_n\right\|) + \mathcal{O}(h\left\|\Delta \ell^0_{ni}\right\|) \\
&+ (f_y - f_z g_z^{-1} g_y)(t_n + \tilde{c}_i h) h \sum_{j=1}^{i-1} \left((f_y - f_z g_z^{-1} g_y)(t_n + \tilde{c}_j h)\tilde{a}_{ij}\Delta \ell^0_{nj} + \mathcal{O}(h^{\tilde{q}_j+1})\right) \\
&\qquad + \mathcal{O}(h^{\tilde{q}_i+1}) \Bigg) - ((f_z g_z^{-1})(t_n + h) - (f_z g_z^{-1})(t_n))\Delta z^0_{n+1} + \mathcal{O}(h^{p+1}),
\end{aligned}
$$

where $\mathcal{O}(h\left\|\Delta \ell^0_{ni}\right\|) = ((f_z g_z^{-1})(t_n + \tilde{c}_i h) - (f_z g_z^{-1})(t_n))\Delta \ell^0_{ni}$. Consequently, the first relations in (56) and (55) and the fact that $((f_z g_z^{-1})(t_n + h) - (f_z g_z^{-1})(t_n)) = \mathcal{O}(h)$ imply that

$$
\left\|\Delta u^1_{n+1}\right\| \le (1 + Ch)\left\|\Delta u^1_n\right\| + \mathcal{O}(h^3). \tag{59}
$$

Then we have $\Delta u^1_n = \mathcal{O}(h^2)$ for $nh \le \text{Const}$ (observe that the initial values are assumed to be consistent, i.e., $\Delta u^1_0 = 0$), so that by (58) and (55) we also have $\Delta y^1_n = \mathcal{O}(h^2)$. This implies $\Delta k^1_{ni} = \mathcal{O}(h)$ and $\Delta Y^1_{ni} = \mathcal{O}(h^2)$. The second relation in (57) proves that $\Delta Z^1_{ni} = \mathcal{O}(h)$.

In order to estimate Δz_n^1 we proceed as in Theorem 3.4 in [13] and, because $|R(\infty)| < 1$, we thus obtain $\Delta z_n^1 = \mathcal{O}(h)$. In particular, we emphasize that if we consider the second relation in (56) and (55) in a similar way, we get $\Delta y_n^1 = \mathcal{O}(h)$ with $\Delta k_{ni}^1 = \mathcal{O}(1)$ and, in addition, it follows from $\Delta Z_{ni}^1 = \mathcal{O}(1)$ that $\Delta z_n^1 = \mathcal{O}(1)$.

(c) The proof for general ν is similar to that of Theorem 3.4 in [13]. It is worth commenting that the only difference arises in the quantity $\Delta \ell_{ni}^{\nu-1} = \mathcal{O}(h^{2-\nu})$ if $\sum_{ij}^{s} b_i \omega_{ij} \tilde{c}_j = 1$; otherwise $\Delta \ell_{ni}^{\nu-1} = \mathcal{O}(h^{1-\nu})$. Thus the statement follows with

$$\begin{aligned} &\text{if } \sum_{ij}^{s} b_i \omega_{ij} \tilde{c}_j = 1, \ \Delta Y_{ni}^{\nu} = \mathcal{O}(h^{3-\nu}), \quad \Delta Z_{ni}^{\nu} = \mathcal{O}(h^{2-\nu}); \\ &\text{otherwise} \qquad\qquad \Delta Y_{ni}^{\nu} = \mathcal{O}(h^{2-\nu}), \quad \Delta Z_{ni}^{\nu} = \mathcal{O}(h^{1-\nu}). \end{aligned} \tag{60}$$

∎

THEOREM 6.2 (type CK). *Consider an IMEX R-K method of type CK which is stiffly accurate and such that* $(\hat{a}_{ij})$ *is invertible. If* (8) *holds and if the initial values of the differential-algebraic system of index* $\nu + 1$ *are consistent, then the global error of method* (17)–(19c) *satisfies, for* $\nu = 1, 2$,

$$y_n^{\nu} - y_{\nu}(t_n) = \mathcal{O}(h^{3-\nu}), \qquad z_n^{\nu} - z_{\nu}(t_n) = \mathcal{O}(h^{2-\nu}). \tag{61}$$

Proof. From Theorem 5.2 it follows that

$$\begin{aligned} &\Delta y_n^0 = \mathcal{O}(h^{\tilde{q}+2} + h^p), \quad \Delta Y_{ni}^0 = \mathcal{O}(h^{\tilde{q}_i+1}), \quad \Delta k_{ni}^0 = \mathcal{O}(h^{\tilde{q}_i+1}), \\ &\Delta z_n^0 = \mathcal{O}(h^{\tilde{q}+1} + h^p), \quad \Delta Z_{ni}^0 = \mathcal{O}(h^{\tilde{q}_i+1}), \end{aligned} \tag{62}$$

with $\tilde{q} = \min\{\tilde{q}_s, \tilde{q}_i + 1, i = 2, \ldots, s-1\}$.

Again we consider the case $\nu = 1$. Here the study of convergence needs further investigation. We start by computing the difference $\Delta \ell_{n1}^0$. From (19b), we have $\ell_{n1}^0 = g_y(y_n^0, z_n^0) y_n^1 + g_z(y_n^0, z_n^0) z_n^1$, and this implies $\|\Delta \ell_{n1}^0\| \le C(\|\Delta y_n^0\| + \|\Delta z_n^0\| + \|\Delta y_n^1\| + \|\Delta z_n^1\|)$. Consequently, using (53), we have

$$\Delta \ell_{ni}^0 = \mathcal{O}(\|\Delta y_n^1\| + \|\Delta z_n^1\|) + h^{-1} \hat{\omega}_{i2} (\Delta Z_{n2}^0 - \Delta z_n^0) + \mathcal{O}(h^{\tilde{q}_i}) + \mathcal{O}(h^{q_i}), \tag{63}$$

where $\hat{\omega}_{ij}$ are the elements of the inverse matrix of $\hat{A}$. Therefore, inserting (63) into the quantity $\Delta k_{ni}^1 - (f_z g_z^{-1})(t_n + \tilde{c}_i h) \Delta \ell_{ni}^0$ computed in the previous theorem, we obtain

$$\begin{aligned} \Delta k_{ni}^1 = {} & h^{-1} \hat{\omega}_{i2} (f_z g_z^{-1})(t_n + \tilde{c}_i h)(\Delta Z_{n2}^0 - \Delta z_n^0) \\ & + \mathcal{O}(\|\Delta y_n^1\| + \|\Delta z_n^1\|) + \mathcal{O}(h^{\tilde{q}_i}) + \mathcal{O}(h^{q_i}). \end{aligned}$$

By (62) and $\tilde{q}_2 = 1$, we have $\Delta Y_{n2}^0 = \mathcal{O}(h^2)$ and $\Delta Z_{n2}^0 = \mathcal{O}(h^2)$. Hence, this implies $\Delta k_{ni}^1 = \mathcal{O}(\|\Delta y_n^1\| + \|\Delta z_n^1\|) + \mathcal{O}(h)$, and a direct estimation of Δy_n^1 leads to

$$\|\Delta y_n^1\| \le (1 + Ch) \|\Delta y_n^1\| + Ch \|\Delta z_n^1\| + \mathcal{O}(h^2). \tag{64}$$

Now, using (53), this gives $\Delta \ell_{ni}^1 = \alpha_i \Delta \ell_{n1}^1 + h^{-1} \sum_{j \ge 2} \hat{\omega}_{ij} (\Delta Z_{nj}^1 - \Delta z_n^1) + \mathcal{O}(h^{q_i})$ for $i = 2, \ldots, s$. Since the method is stiffly accurate and $\tilde{b}_i = b_i$ for $i = 1, \ldots, s$, the statement of Lemma 5.1 is satisfied, and from $\Delta z_{n+1}^1 = \Delta z_n^1 + h(b_1 + \sum_{i \ge 2} b_i \alpha_i) \Delta \ell_{n1}^1 + \sum_{i,j \ge 2} b_i \hat{\omega}_{ij} (\Delta Z_{nj}^1 - \Delta z_n^1) + \mathcal{O}(h^{q_i+1})$ we obtain

$$\|\Delta z_{n+1}^1\| \le |\rho| \, \|\Delta z_n^1\| + \sum_{i,j \ge 2} |b_i \hat{\omega}_{ij}| \, \|\Delta Z_{nj}^1\| + \mathcal{O}(h^{q_i+1}) \tag{65}$$

with $|\rho| = |R(\infty)| < 1$. By (63) and $\Delta Z_{ni}^0 = \mathcal{O}(h^2)$, it follows that $\Delta \ell_{ni}^0 = \mathcal{O}(\|\Delta y_n^1\| + \|\Delta z_n^1\|) + \mathcal{O}(h)$. Thus, the second relation of (57) proves that $\Delta Z_{ni}^1 = \mathcal{O}(\|\Delta y_n^1\| + \|\Delta z_n^1\|) + \mathcal{O}(h(\|\Delta y_n^1\| + \|\Delta z_n^1\|)) + \mathcal{O}(h)$. Inserting (65), we obtain

$$\left\|\Delta z_{n+1}^1\right\| \le \left\|\Delta y_n^1\right\| + (|\rho| + Ch)\left\|\Delta z_n^1\right\| + \mathcal{O}(h). \tag{66}$$

Now applying Lemma 5.2 to inequalities (64) and (66) gives $\Delta y_n^1 = \mathcal{O}(h)$, $\Delta z_n^1 = \mathcal{O}(h)$. Again, we can conclude that the estimate about Δy_n^1 is not optimal. Therefore, introducing the new variable (58), we obtain

$$\left\|\Delta u_{n+1}^1\right\| \le (1 + Ch)\left\|\Delta u_n^1\right\| + Ch^2\left\|\Delta z_n^1\right\| + \mathcal{O}(h^3). \tag{67}$$

We now apply Lemma 5.2 again, replacing the inequality (64) with (67). Then by (58) and (62) we have $\Delta y_n^1 = \mathcal{O}(h^2)$. Obviously, the proof for general ν is similar to the one presented in Theorem 6.1. This completes the proof of the theorem. ∎

Remark. Of course, concerning type ARS, under the same assumptions as Theorem 6.2 (with also $b_1 = 0$), we again deduce the estimates (61).

7. Estimates on the remainder. In order to estimate the remainder in the expansion (20), we require the same detailed analysis previously developed by Hairer, Lubich, and Roche in [12] (see also [13, sect. VI.3]). The main purpose in this section is to extend the same results presented in [13, sect. VI.3] to the different types of IMEX R-K methods.

Let us introduce existence and local uniqueness of the numerical solution of (4), (5). Next we shall discuss the influence of perturbations in (5) to the numerical solution.

We shall consider two steps in succession. First, we suppose that (y_n, z_n) are known, denoted by (η, ζ), and prove the existence and uniqueness of (y_{n+1}, z_{n+1}). We assume that $g(\eta, \zeta) = \mathcal{O}(h)$, $\mu(g_z(\eta, \zeta)) \le 1$ and that $a_{ii} > 0$ for all i. Thus we have the nonlinear system for the stage values

$$\begin{pmatrix} Y_i - \eta \\ \varepsilon\,(Z_i - \zeta) \end{pmatrix} = h \begin{pmatrix} \sum_{j=1}^{i-1} \tilde{a}_{ij} f(Y_j, Z_j) \\ \sum_{j=1}^{i} a_{ij} g(Y_j, Z_j) \end{pmatrix}. \tag{68}$$

It is significant to note that if we restrict ourselves to the use of a particular type of IMEX R-K method, for instance, type A, where the matrix A is invertible, we immediately obtain the statement of Theorem 3.5 in [13]. Instead, for type CK, it is worth commenting that the second equation in (68) becomes

$$\frac{\varepsilon}{h}(Z_i - \zeta) - a_{i1} g(\eta, \zeta) - \sum_{j=2}^{i} \hat{a}_{ij} g(Y_j, Z_j) = 0,$$

whereas for type ARS we have $a_{i1} = 0$ for all i. Therefore, we easily find again the statement of Theorem 3.5 in [13].

We now study the influence of perturbations in (68) to the numerical solution. For the perturbed IMEX R-K method

$$\begin{pmatrix} \widehat{Y}_i - \widehat{\eta} \\ \varepsilon\left(\widehat{Z}_i - \widehat{\zeta}\right) \end{pmatrix} = h \begin{pmatrix} \sum_{j=1}^{i-1} \tilde{a}_{ij} f(\widehat{Y}_j, \widehat{Z}_j) \\ \sum_{j=1}^{i} a_{ij} g(\widehat{Y}_j, \widehat{Z}_j) \end{pmatrix} + h \begin{pmatrix} \delta_i \\ \theta_i \end{pmatrix} \tag{69}$$

we allow the following remarks.

Remarks. For an IMEX R-K scheme of type A the statement and the proof is similar to that of Theorem 3.6 in [13]. Extra care must to be taken to properly handle type CK. First observe that in addition to the assumptions of Theorem 3.5 in [13] we suppose that $\widehat{\eta} - \eta = \mathcal{O}(h)$, $\widehat{\zeta} - \zeta = \mathcal{O}(h)$, $\delta_i = \mathcal{O}(1)$, and $\theta_i = \mathcal{O}(h)$ for $i = 2, \dots, s$ with $\delta_1 = 0$ and $\theta_1 = 0$. Then we have for $h \le h_0$ the following estimates:

$$\begin{aligned} &\|\widehat{Y}_i - Y_i\| \le C(\|\widehat{\eta} - \eta\| + h\|\widehat{\zeta} - \zeta\|) + hC(\|\delta\| + \|\theta\|), \\ &\|\widehat{Z}_i - Z_i\| \le C\left(\|\widehat{\eta} - \eta\| + \left(\frac{\varepsilon}{h} + h\right)\|\widehat{\zeta} - \zeta\|\right) + C(h\,\|\delta\| + \|\theta\|), \end{aligned} \tag{70}$$

where $\delta = (\delta_1, \dots, \delta_s)^T$ and $\theta = (\theta_1, \dots, \theta_s)^T$. Later, we note that we have to treat the following homotopy more carefully:

$$\begin{pmatrix} Y_i - \eta \\ \varepsilon\,(Z_i - \zeta) \end{pmatrix} - h \begin{pmatrix} \sum_{j=1}^{i-1} \tilde{a}_{ij} f(Y_j, Z_j) \\ \sum_{j=1}^{i} \hat{a}_{ij} g(Y_j, Z_j) \end{pmatrix} = \tau \begin{pmatrix} \widehat{\eta} - \eta + h\delta_i \\ \varepsilon(\widehat{\zeta} - \zeta) + h a_{i1}(g(\widehat{\eta}, \widehat{\zeta}) - g(\eta, \zeta)) + h\theta_i \end{pmatrix},$$

which relates system (68) for $\tau = 0$ to the perturbed system (71) for $\tau = 1$. Furthermore, we denote by $\hat{a}_{ij}$ the elements of the submatrix $\hat{A}$, and, by the Lipschitz condition for g, we have the inequality

$$\|(g(\widehat{\eta}, \widehat{\zeta}) - g(\eta, \zeta))\| \le L\|\widehat{\eta} - \eta\| + L\|\widehat{\zeta} - \zeta\|.$$

Then, in this situation, the same conclusions of Theorem 3.6 in [13] hold. In particular, if $a_{i1} = 0$ for all i, the same also follows for type ARS.

Following [13], we finally estimate the remainder of the expansion (20).

THEOREM 7.1 (type A). *Under the same hypotheses as those of Theorem* 3.1*, for any fixed constant $c > 0$ and $\varepsilon \le ch$, the global error satisfies*

$$\begin{aligned} &y_n - y(t_n) = \Delta y_n^0 + \varepsilon \Delta y_n^1 + \varepsilon^2 \Delta y_n^2 + \mathcal{O}(\varepsilon^3), \\ &z_n - z(t_n) = \Delta z_n^0 + \varepsilon \Delta z_n^1 + \varepsilon^2 \Delta z_n^2 + \mathcal{O}(\varepsilon^3/h), \end{aligned} \tag{71}$$

where $\Delta y_n^0 = y_n^0 - y_0(t_n)$, $\Delta z_n^0 = z_n^0 - z_0(t_n), \dots$ are the global errors of the method applied to differential-algebraic system. The estimates (71) *hold uniformly for $h \le h_0$ and $nh \le$ Const.*

Remark. In order to enable a direct comparison with Theorem 3.8 in [13] (see also [12]), by Theorem 6.1, and by (50) and (60), if $\sum_{ij}^{s} b_i \omega_{ij} \tilde{c}_j = 1$, it suffices to prove the result for $\nu = 2$; otherwise it must be proven for $\nu = 1$. Therefore, the result follows directly by applying Theorem 3.8 in [13].

THEOREM 7.2 (type CK). *Under the same hypotheses as those of Theorem* 3.2*, then, for any fixed constant $c > 0$ and $\varepsilon \le ch$, the global error satisfies the estimates* (71) *uniformly for $h \le h_0$ and $nh \le$ Const.*

Remark. It is interesting, of course, to know how in the proof of Theorem 7.2 several formulas are related to those of Theorem 3.8 in [13]. For instance, by (19a)–(19c) it follows from (60) and $\nu = 2$ that

$$\begin{aligned} &\widehat{k}_{ni} = f(\widehat{Y}_{ni}, \widehat{Z}_{ni}) + \mathcal{O}(\varepsilon^3), \\ &\varepsilon\widehat{\ell}_{ni} = g(\widehat{Y}_{ni}, \widehat{Z}_{ni}) + \varepsilon^3 \ell_{ni}^2 + \mathcal{O}(\varepsilon^3). \end{aligned} \tag{72}$$

Using $Z_{ni}^{\nu} = z_n^{\nu} + ha_{i1}\ell_{n1}^{\nu} + h\sum_{j=2}^{i}\ell_{nj}^{\nu}$, from (61) and

$$\ell_{n1}^{\nu} = g_y(y_n^0, z_n^0)y_n^{\nu+1} + g_z(y_n^0, z_n^0)z_n^{\nu+1} + \psi_{\nu+1}(y_n^0, z_n^0, \dots, y_n^{\nu}, z_n^{\nu}),$$

we get $\ell_{ni}^2 = \mathcal{O}(h^{-1})$. Together with (18), and by (72), it follows that we obtain a perturbed IMEX R-K method which is of the form (69). Therefore, in the case of Theorem 3.8 in [13], this yields

$$\begin{aligned} &\|\Delta Y_{ni}\| \le C(\|\Delta y_n\| + h\|\Delta z_n\|) + \mathcal{O}(\varepsilon^3), \\ &\|\Delta Z_{ni}\| \le C\left(\|\Delta y_n\| + \left(\frac{\varepsilon}{h} + h\right)\|\Delta z_n\|\right) + \mathcal{O}(\varepsilon^3/h), \end{aligned} \tag{73}$$

provided that Δy_n and Δz_n are of size $\mathcal{O}(h)$. The justification of these assumptions follows by induction on n where $\Delta y_0 = \mathcal{O}(\varepsilon^3)$ and $\Delta z_0 = \mathcal{O}(\varepsilon^3)$ and from $\Delta y_n = \mathcal{O}(\varepsilon^3/h)$, $\Delta z_n = \mathcal{O}(\varepsilon^3/h)$, because $\nu = 2$.

Moreover, we prove the recursion

$$\begin{pmatrix} \|\Delta y_{n+1}\| \\ \|\Delta z_{n+1}\| \end{pmatrix} \le \begin{pmatrix} 1 + \mathcal{O}(h) & \mathcal{O}(\varepsilon + h^2) \\ \mathcal{O}(1) & \alpha + \mathcal{O}(h) \end{pmatrix} \begin{pmatrix} \|\Delta y_n\| \\ \|\Delta z_n\| \end{pmatrix} + \begin{pmatrix} \mathcal{O}(\varepsilon^3) \\ \mathcal{O}(\varepsilon^3/h) \end{pmatrix}. \tag{74}$$

The value $\alpha < 1$ is justified in [4] (see also [13]).

Second, in solving the second relation in (74) we use the result of Lemma 5.1 where we emphasize that the method is stiffly accurate and $\tilde{b}_i = b_i$ for all i. Of course, for type ARS, we have again the estimates (71).

Now by combining Theorems 5.1, 6.1, and 7.1, Theorem 3.1 follows. Theorem 3.2 follows from Theorems 5.2, 6.2, and 7.2. Finally, Corollary 3.1 follows from Corollary 5.1 and from the remarks of Theorems 6.2 and 7.2.

8. Conclusions. A study of the global error for different types of IMEX R-K methods has been investigated for a class of singular perturbation problems (SPPs). This asymptotic analysis enables us to obtain convergence results, based on the smoothness of the solution, giving error bounds for several classes of IMEX R-K methods. In particular, the use of DAE techniques, when applied to the stiff case $\Delta t \gg \varepsilon$, was found to give optimal estimates describing the structure of the solutions of SPPs. Concerning the van der Pol equation, numerical results reveal order reduction for all methods in the second (algebraic) component of the solution for small values of the stiffness parameter ε and likewise an order reduction in the first (differential) component when $\tilde{b}_i$ is not equal to b_i for all i. In fact, the hypothesis $\tilde{b}_i = b_i$ represents the only remedy for preserving the classical order for the differential component of the solution. Also, when ε is sufficiently small, and for a given set of suitable assumptions, we obtain numerical results which display improved error estimates in the algebraic component for some IMEX R-K methods appearing in the literature. These results lead us to develop new IMEX R-K methods that work uniformly for a wide range of values of the stiffness parameter ε. In future work we shall introduce new order conditions for the construction of these IMEX R-K methods (see [5]) and study their stability properties.

Acknowledgments. The author wishes to express his gratitude and appreciation to Prof. Ernst Hairer, who generously spent his time to guide him in this work through numerous suggestions and enlightening discussions, during the author's stay at the mathematics department of the University of Geneva. He also thanks the two unknown referees for their critical remarks on the first draft of this paper.

REFERENCES

[1] U. Ascher, S. Ruuth, and R. J. Spiteri, *Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations*, Appl. Numer. Math., 25 (1997), pp. 151–167.

[2] U. Ascher, S. Ruuth, and R. J. Wetton, *Implicit-explicit methods for time dependent PDE's*, Appl. Numer. Math., 32 (1995), pp. 797–823.

[3] J. G. Blom, W. Hundsdorfer, and J. G. Verwer, *An implicit-explicit approach for atmospheric transport-chemistry problems*, Appl. Numer. Math., 20 (1996), pp. 191–209.

[4] S. Boscarino, *On the Uniform Accuracy of Implicit-Explicit Runge-Kutta Methods*, Ph.D. Thesis, Mathematics for the Technology, Department of Mathematics and Computer Science, University of Catania, Italy, 2005.

[5] S. Boscarino, *Uniformly accurate implicit-explicit (IMEX) Runge-Kutta schemes*, submitted.

[6] R. E. Caflisch, S. Jin, and G. Russo, *Uniformly accurate schemes for hyperbolic systems with relaxation*, SIAM J. Numer. Anal., 34 (1997), pp. 246–281.

[7] S. L. Campbell and C. W. Gear, *The index of general nonlinear DAEs*, Numer. Math., 72 (1995), pp. 173–196.

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

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

[10] J. Frank, W. Hundsdorfer, and J. G. Verwer, *On the stability of implicit-explicit linear multistep methods*, Appl. Numer. Math., 25 (1997), pp. 193–205.

[11] C. W. Gear, *Differential algebraic equations, indices, and integral algebraic equation*, SIAM J. Numer. Anal., 27 (1990), pp. 1527–1534.

[12] E. Hairer, Ch. Lubich, and M. Roche, *Error of Runge Kutta methods for stiff problems via differential algebraic equations*, BIT, 28 (1988), pp. 678–700.

[13] E. Hairer and G. Wanner, *Solving Ordinary Differential Equation* II: *Stiff and Differential Algebraic Problems*, 2nd ed., Springer Ser. Comput. Math. 14, Springer-Verlag, New York, 1991, 1996.

[14] W. Hundsdorfer and J. Jaffré, *Implicit-explicit time stepping with spatial discontinuous finite elements*, Appl. Numer. Math., 45 (2003), pp. 231–254.

[15] S. F. Liotta, V. Romano, and G. Russo, *Central schemes for balance laws of relaxation type*, SIAM J. Numer. Anal., 38 (2000), pp. 1337–1356.

[16] R. E. O'Malley, Jr., *Introduction to Singular Perturbations*, Academic Press, New York, 1974.

[17] L. Pareschi and G. Russo, *Implicit-explicit Runge-Kutta schemes for stiff systems of differential equations*, in Recent Trends in Numerical Analysis, Adv. Theory Comput. Math. 3, Nova Sci. Publ., Huntington, NY, 2001, pp. 269–288.

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

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

[20] A. N. Tikhonov, A. B. Vasl'eva, and A. G. Sveshnikov, *Differential Equations*, translated from the Russian by A. B. Sossinskij, Springer-Verlag, Berlin, 1985.