Formal series and numerical integrators. Part II: Application to index 2 differential-algebraic systems

A. Murua[1]

Konputazio Zientziak eta A. A. saila,
Informatika Fakultatea EHU/UPV,
Donostia - San Sebastián, Spain
email ander@si.ehu.es

Abstract

The present work has been divided in two parts: In the first part, we presented a unified approach to derive and deal with the formal series that arise when studying the convergence of numerical integrators for systems of ODEs and DAEs. This general approach was applied for systems of ODEs. The particular case of Hamiltonian systems was studied in some detail.

In this second part, we study the convergence of a general class of Runge-Kutta type methods (the *partitioned Runge-Kutta* methods) for systems of index 2 DAEs in Hessenberg form. To do that, we use the tools developped in the first part to deal with the relevant formal expansions, the *DA2-series*. The convergence of such partitioned Runge-Kutta methods is studied in a unified way. Our approach allows us to obtain sharp estimates for the behaviour of the global errors, even in the case where the method does not satisfy the algebraic constraints, the errors for the algebraic variables affect the approximations of the differential variables, and non-consistent initial values are used.

1 Introduction

The purpose of this two-part paper is to present a unified approach to derive and handle the formal expansions that arise when studying the order conditions of one-step integrators for ODEs and DAEs. These formal series can be typically written in the form

[1] Supported by the University of the Basque Country (UPV140.226-EA105/96)

$$\mathcal{S}(\mathbf{d}) = \mathrm{id} + \sum_{u\in\mathcal{T}} h^{\rho(u)}\mathbf{d}(u)\, F(u), \tag{1}$$

where the symbols in these formal expressions have the following meaning: (i) $\mathcal{T}$ is a countable set of indices, for instance, a set of rooted trees, (ii) the *order* $\rho(u)$ of each $u \in \mathcal{T}$ is an integer, (iii) the *elementary differentials* $F(u)$ are vector functions that depend on the system of ODEs or DAEs to be integrated, (iv) the *coefficients* $\mathbf{d}(u)$ are real numbers that do not depend on the system.

The aim of this second part of the article is to show how the formalism presented in the first part [13] of our article can be used for the study of the convergence behaviour of one-step methods for semi-explicit systems of differential-algebraic equations [4,8]. In order to illustrate that, we study the convergence of a general class of one-step methods for index 2 DAEs in Hessenberg form (with possibly non-consistent initial values).

Different Runge-Kutta type methods for semi-explicit differential-algebraic systems in Hessenberg form of index 1, 2, and 3 have been considered in the literature, including implicit Runge-Kutta methods, Rosenbrock methods, and half-explicit methods [6,8,4]. Recently, more Runge-Kutta type families of methods have been proposed [1,11,12] for index 2 DAEs. Detailed study of the convergence of these families of methods leads to three classes of series expansions of the form (1). Depending on the index of the system of DAEs, they are sometimes called DA1-series, DA2-series, and DA3-series. In each case, the indices $u \in \mathcal{T}$ are rooted trees of different types of vertices. However, unlike in the case of Runge-Kutta methods for ODEs, the order of each u does not coincide with its number of vertices. Furthermore, in certain cases it is useful to consider expansions with terms of order 0. This was first considered when studying the behavior of Rosenbrock methods applied to index 1 systems with inconsistent initial values (see [8], Section V.I). Recently, DA2-series that include terms of negative powers of h have been considered by Chan and Chartier [5]. The composition of formal series for systems of DAEs was first considered by Jay in [9], who studied the case of DA3-series for consistent initial values (that is, with only positive powers of h. In [5] the composition of DA2-series originated from implicit Runge-Kutta methods with invertible RK matrix for the general case of non-consistent initial values is studied.

We consider index 2 systems of differential-algebraic equations in Hessenberg form

$$y' = f(y,z), \quad 0 = g(y), \tag{2}$$

where f and g are assumed to be sufficiently differentiable and

$$g_y(y) f_z(y,z) \text{ is invertible} \tag{3}$$

in a neighborhood of the solution. Given initial conditions $(y(t_0), z(t_0)) = (y_0, z_0)$ satisfying the consistency equations

$$g(y_0) = 0, \tag{4}$$
$$g_y(y_0)f(y_0, z_0) = 0, \tag{5}$$

the system (2) has a unique solution $(y(t), z(t))$.

We are concerned with a general classs of one-step methods to solve numerically the system (2), namely, the *partitioned Runge-Kutta* (PRK) methods. We proposed this family of methods in [11] and studied it in more detail in [12]. The family of PRK methods includes the application of implicit Runge-Kutta methods to the system (2) [6,4,8], as well as the different families of half-explicit methods of Runge-Kutta type studied in the literature [6,3,1,11]. Fully implicit new PRK methods (the *Gauss-Lobatto* methods) are also proposed in [12]. The later methods are s-stage symmetric methods that give approximations of order $2s$ for the differential variable y.

One step of the partitioned Runge-Kutta method gives an approximation $(y_1, z_1) \approx (y(t_0 + \ h), z(t_0 + \ h))$ defined as follows:

$$y_1 = y_0 + \sum_{i=0}^{s} b_i f(Y_i, Z_i), \quad z_1 \ = \ \sum_{i=0}^{s} d_i Z_i, \tag{6}$$

where $Y_0 = y_0$, $Z_0 = z_0$, and for $i = 1, \ldots, s$,

$$Y_i = y_0 + h\sum_{j=0}^{s} a_{ij} f(Y_j, Z_j), \tag{7}$$
$$\bar{Y}_i = y_0 + h\sum_{j=0}^{s} \bar{a}_{ij} f(Y_j, Z_j), \quad g(\bar{Y}_i) = 0.$$

We denote by A and respectively $\bar{A}$ the $(s+1)\times(s+1)$ matrices with entries a_{ij} and $\bar{a}_{ij}$ ($a_{0j} = \bar{a}_{0j} = 0$), and $b = (b_0, \ldots, b_s)^T$, $d = (d_0, \ldots, d_s)^T$. The coefficients d_i must satisfy the equation $\sum d_i = 1$. We denote by $\tilde{A}$ the $s \times s$ matrix obtained from removing the first row and column of $\bar{A}$.

Note that, in general, the numerical solution (y_1, z_1) depend on both y_0 and z_0. It will depend only on y_0 if

$$d_0 = 0, \quad b_0 = 0, \quad a_{i0} = \bar{a}_{i0} = 0 \quad \text{for all } i. \tag{8}$$

In that case, although the PRK method (6)-(7) has with our notation $s+1$ stages, they effectively have s stages, as the 0th stage derivative $f(Y_0, Z_0) =$

$f(y_0, z_0)$ and Z_0 are not present in the definition of the numerical solution (y_1, z_1). Implicit Runge-Kutta methods ($A = \bar{A}$) with invertible Runge-Kutta matrix, and the half-explicit Runge-Kutta methods proposed in [6] fall within the class of PRK methods satisfying (8). Of particular interest are PRK methods (6)-(7) satisfying

$$b_i = \bar{a}_{si}, \quad 0 \leq i \leq s, \tag{9}$$

so that $y_1 = \bar{Y}_s$, and therefore the numerical solution y_1 satisfies the algebraic equations of (2).

This second part of the paper is organized as follows: Section 2 is devoted to the study of the existence of solution and the influence of perturbations of the PRK scheme (6)-(7). In Section 3, we will summarize some definitions and basic results presented in the first part [13] of our article which will be used later. In Section 4, we see that the numerical solution given by the PRK method (6)-(7) can be expanded as a DA2-series. The DA2-series were introduced by Hairer, Lubich, and Roche [6] for the case of the application of Runge-Kutta methods to the system (2) with consistent initial values, and generalized for the case of inconsistent initial values by Chan and Chartier [5]. In Section 5, we derive recursive formulae to obtain the coefficients of the expansion of the composition of two DA2-series. Rigorous truncated versions of the DA2-series expansions derived in a formal way in Section 4 and Section 5 are obtained in Section 6. Finally, a new procedure to study the convergence of the application of partitioned Runge-Kutta methods to index 2 systems of DAEs of the form (2) is presented in Section 7. This new approach allows us to obtain sharp estimates for the global errors, even in the case where the method does not satisfy the algebraic constraint of (2), the errors for the algebraic variables z affect the values for the differential variables y, and non-consistent initial values are used.

2 Existence and influence of perturbations

In the more general case (where (8) or (9) is not satisfied), the repeated application of more than one step of the PRK method leads to the application of the scheme (6)-(7) with inconsistent (y_0, z_0). The following result on the existence of the numerical solution of (6)-(7) is a particular case of the Lemma 2 below.

Lemma 1 *Assume that $\tilde{A}$ is invertible. Let us consider the system (2) and (y_0, z_0) such that $g_y(y) f_z(y, z)$ is invertible in a neighborhood of (y_0, z_0). If*

$$\epsilon = \max\left(|h|, ||g(y_0)||/|h|, ||g_y(y_0)f(y_0, z_0)||\right) \tag{10}$$

is sufficiently small, then the PRK scheme (6)-(7) has a locally unique solution such that

$$Y_i = y_0 + O(h), \quad Z_i = z_0 + O(\epsilon), \quad y_1 = y_0 + O(h), \quad z_1 = z_0 + O(\epsilon).$$

Our aim now is to study the influence of perturbations of the PRK solution. The equations (7) can be rearranged in such a way that it consists of a nonlinear system of equations with $k_i = f(Y_i, Z_i)$, Z_i $(1 \leq i \leq s)$ as unknowns. Let us consider the following perturbed version of that nonlinear system:

$$k_i = f(y_0 + h\sum_{j=0}^{s} a_{ij}k_j, Z_i) + \gamma_i, \; \theta_i = g(y_0 + h\sum_{j=0}^{s} \bar{a}_{ij}k_j), \; 1 \leq i \leq s, \tag{11}$$

where $k_0 = f(y_0, z_0) + \gamma_0$. In general, the existence and local uniqueness of the solution of (11) is not guaranteed for fixed values of the perturbations θ_i as h tends to 0. Thus, we consider special perturbations of the form

$$\theta_i = g(y_0) + h\,\omega_i, \quad 1 \leq i \leq s. \tag{12}$$

Let us denote $\gamma = (\gamma_0, \ldots, \gamma_s)$, and $\omega = (\omega_1, \ldots, \omega_s)$.

Lemma 2 *Let us consider the system (2) and consistent initial values (y_0^0, z_0^0) such that (3) is satisfied in a neighborhood of (y_0^0, z_0^0). Let us assume that the $s \times s$ matrix $\tilde{A}$ is invertible. Then, the perturbed scheme (11)-(12) has, for $(h, y_0, z_0, \gamma, \omega)$ in a certain neighborhood $\mathcal{U}$ of $(0, y_0^0, z_0^0, 0, 0)$ a locally unique solution $U = (k_1, \ldots, k_s, Z_1, \ldots, Z_s)$ that smoothly depends on $(h, y_0, z_0, \gamma, \omega)$.*

PROOF. Proceeding in a similar way to the proof of Theorem VII.4.1 of [8] when studying the existence of Runge-Kutta methods with invertible A matrix, the equations $\theta_i = g(y_0 + h\sum \bar{a}_{ij}k_j)$ in (11) can be replaced by

$$\omega_i = \int_0^1 \mathrm{d}\tau\, g_y(y_0 + \tau h\sum_{j=0}^{s} \bar{a}_{ij}k_j)\sum_{j=0}^{s} \bar{a}_{ij}k_j.$$

Thus, (11)-(12) is equivalent to a system of the form $\mathcal{F}(h, y_0, z_0, \gamma, \omega, U) = 0$, where $U = (k_1, \ldots, k_s, Z_1, \ldots, Z_s)$ and F is smooth. Next, we will show that the implicit function theorem can be applied to this nonlinear system: First, $\mathcal{F}(0, y_0^0, z_0^0, 0, 0, U^0) = 0$, where $U^0 = (k_0^0, \ldots, k_0^0, z_0^0, \ldots, z_0^0)$, $k_0^0 = f(y_0^0, z_0^0)$. Second, the Jacobian of $\mathcal{F}$ with respect to U at $(0, y_0^0, z_0^0, 0, 0, U^0)$ is

$$\begin{pmatrix} I_{ns} & -I_s \otimes f_z(y_0^0, z_0^0) \\ 0 & \tilde{A} \otimes g_y(y_0^0, z_0^0) \end{pmatrix},$$

where n is the dimension of the differential variable y. Clearly, this matrix is invertible under the hypothesis of the theorem. □

Now, the proof of Lemma 1 can be obtained considering Lemma 2 with $\gamma_i = 0$, $\omega_i = -\frac{1}{h} g(y_0)$ $(1 \leq i \leq s)$.

3 Abstract framework

Here, we summarize the definitions and basic results presented in the first part of our article that are needed in subsequent sections.

Let us consider a countable set $\mathcal{T}$ of mathematical objects such that each $u \in \mathcal{T}$ has attached an integer $\rho(u)$, the *order* of u. Let us assume that each $u \in \mathcal{T}$ has associated a function, called *elementary differential* $F(u) : R^n \to R^n$.

For each $\mathbf{d} : \mathcal{T} \to R$, we denote by $\mathcal{S}(\mathbf{d})$ the formal series (1), and by $\overline{\mathcal{S}}(\mathbf{d})$ the series

$$\overline{\mathcal{S}}(\mathbf{d}) = \mathcal{S}(\mathbf{d}) - \mathrm{id} = \sum_{u \in \mathcal{T}} h^{\rho(u)} \mathbf{d}(u)\, F(u). \tag{13}$$

Given $u_1, \ldots, u_m \in \mathcal{T}$, we denote by $X[u_1, \ldots, u_m]$ the following linear differential operator on functions of n variables which take values in a real vector space (R, R^m, or some matrix space): Given a sufficiently smooth function k of n variables, for each $y \in R^n$,

$$(X[u_1, \ldots, u_m]k)(y) = \frac{1}{\mu_1! \cdots \mu_\nu!} k^{(m)}(y)\, (F(u_1)(y), \ldots, F(u_m)(y)), \tag{14}$$

where ν is the number of distinct elements $u_1^*, \ldots, u_\nu^*$ among $u_1, \ldots, u_m$, and each μ_i is the number of elements u_j that coincide with u_i^*. Clearly, the operator $X[u_1, \ldots, u_m]$ is invariant under permutations of $u_1, \ldots, u_m$.

Definition 3 *Given the set $\mathcal{T}$, we denote by $\widehat{\mathcal{T}}$ the following set of unordered m-tuples $[u_1, \ldots, u_m]$ of elements of $\mathcal{T}$,*

$$\widehat{\mathcal{T}} = \{[\emptyset]\} \cup \{[u_1, \ldots, u_m] \ : \ u_1, \ldots, u_m \in \mathcal{T}\}. \tag{15}$$

We define for each $\hat{w} = [u_1, \ldots, u_m] \in \widehat{\mathcal{T}} - \{[\emptyset]\}$ *the* elementary (differential) operator *corresponding to* $\hat{w}$ *as* $X(\hat{w}) = X[u_1, \ldots, u_m]$ *given by (14). We define* $X[\emptyset]$ *as the identity operator. The* order $\widehat{\rho}(\hat{w})$ *of* $\hat{w}$ *is defined by* $\widehat{\rho}(\hat{w}) = \rho(u_1) + \cdots + \rho(u_m)$ *and* $\widehat{\rho}([\emptyset]) = 0$.

Definition 4 *Given* $\hat{\mathbf{a}} : \widehat{\mathcal{T}} \to R$*, we denote by* $\widehat{\mathcal{S}}(\hat{\mathbf{a}})$ *the formal series of elementary operators*

$$\widehat{\mathcal{S}}(\hat{\mathbf{a}}) = \sum_{\hat{w} \in \widehat{\mathcal{T}}} h^{\widehat{\rho}(\hat{w})} \hat{\mathbf{a}}(\hat{w}) \, X(\hat{w}).$$

The following result allows us to expand in a convenient and compact way the composition $k \circ \mathcal{S}(\mathbf{d})$ of a formal series (1) with a given function k. This will be the main tool for the derivation of the expansions of the PRK solution that we present in Section 4.

Theorem 5 *Given* $\mathbf{d} : \mathcal{T} \to R$*, for each smooth function* k *of* n *variables*

$$k \circ \mathcal{S}(\mathbf{d}) = \widehat{\mathcal{S}}(\mathbf{d}')k,$$

where $\mathbf{d}'([\emptyset]) = 1$*, and* $\mathbf{d}'([u_1, \ldots, u_m]) = \mathbf{d}(u_1) \cdots \mathbf{d}(u_m)$.

Definition 6 *Given* $\hat{w} \in \widehat{\mathcal{T}}$ *and* $v \in \mathcal{T}$*, the* product $\hat{w} \cdot v \in \widehat{\mathcal{T}}$ *is defined as follows:*

- $[\emptyset] \cdot v = [v]$ *if* $v \in \mathcal{T}$,
- $[u_1, \ldots, u_m] \cdot v = [u_1, \ldots, u_m, v]$ *if* $u_1, \ldots, u_m, v \in \mathcal{T}$,

Note that this product has the property that $(\hat{w} \cdot u) \cdot v = (\hat{w} \cdot v) \cdot u$, for $\hat{w} \in \widehat{\mathcal{T}}$, $u, v \in \mathcal{T}$.

Clearly, each element of $\hat{w} \in \widehat{\mathcal{T}} - \{[\emptyset]\}$ can be decomposed as a product $\hat{w}' \cdot u$ where $\widehat{\rho}(\hat{w}') < \widehat{\rho}(\hat{w})$ and $\rho(u) = \widehat{\rho}(\hat{w}) - \widehat{\rho}(\hat{w}')$. This allows a convenient recursive way of representing the elementary operators and computing the coefficients $\mathbf{d}'$ of Theorem 5 using the formula

$$\mathbf{d}'(\hat{w} \cdot u) = \mathbf{d}'(\hat{w})\mathbf{d}(u), \quad \hat{w} \in \widehat{\mathcal{T}}, \quad u \in \mathcal{T}. \tag{16}$$

Note that each $\hat{w} \in \widehat{\mathcal{T}}$ can typically be obtained as the result of different products $\hat{w}' \cdot u$.

The following result will be usefull in Section 5.

Lemma 7 *Given* $\mathbf{d} : \mathcal{T} \to R$ *and* $\hat{w}_1 = [u_1, \ldots, u_m] \in \widehat{\mathcal{T}}$*, let us assume that for each* u_l *(*$l = 1, \ldots, m$*) there exists* $\mathbf{d}^{(u_l)} : \mathcal{T} \to R$ *such that*

$$h^{\rho(u_l)} F(u_l) \circ \mathcal{S}(\mathbf{d}) = \overline{\mathcal{S}}(\mathbf{d}^{(u_l)}), \tag{17}$$

then, for every smooth function k *of* n *variables,*

$$h^{\hat{\rho}(\hat{w}_1)}(X(\hat{w}_1)k) \circ \mathcal{S}(\mathbf{d}) = \widehat{\mathcal{S}}(\mathbf{d}^{(\hat{w}_1)})k \tag{18}$$

where for each $\hat{w}_2 \in \widehat{\mathcal{T}}$ *with* $\hat{w}_2$ *decomposed as* $\hat{w}_2 = \hat{w}_2' \cdot v_2$*,*

$$\mathbf{d}^{(\hat{w}_1)}(\hat{w}_2) = \mathbf{d}^{(\hat{w}_1)}(\hat{w}_2')\mathbf{d}(v_2) + \sum_{\substack{\hat{w}_1' \in \widehat{\mathcal{T}}, v_1 \in \mathcal{T} \\ \hat{w}_1' \cdot v_1 = \hat{w}_1}} \mathbf{d}^{(\hat{w}_1')}(\hat{w}_2')\mathbf{d}^{(v_1)}(v_2). \tag{19}$$

Note that Theorem 5 can be considered as the particular case of Lemma 7 where $\hat{w}_1 = [\emptyset]$ and $\mathbf{d}^{([\emptyset])} = \mathbf{d}'$. In that case, the recursive formula (19) reduces to (16).

4 DA2-series expansion of the PRK solution

In order to illustrate the use of the tools of our abstract framework, the formal expansion of additive Runge-Kutta methods (or equivalently of partitioned Runge-Kutta methods) for ODEs were considered in [13]. Our aim was to show how our approach could be used to derive a suitable class of series (1) (that is, to construct the set $\mathcal{T}$ of indices u and to associate an elementary differential to each $u \in \mathcal{T}$) and at the same time to obtain the actual coefficients $\mathbf{c}(u)$ (depending on the parameters of the method) of each ARK scheme. However, when the problem of finding the expansion of a certain family of numerical integrators is addressed, it is often the case that a suitable class of formal series (1) is guessed from the expansions of similar families of methods, or from the direct derivation of the first terms of the series expansion of the numerical approximation and the exact solution. In that case, it is sufficient to find a mapping $\mathbf{c} : \mathcal{T} \to R$ and prove that the numerical approximation can be expanded as the series $\mathcal{S}(\mathbf{c})$.

The family of PRK methods (6)-(7), it generalizes both the class of implicit Runge-Kutta methods ($\bar{a}_{ij} = a_{ij}$) and the class of half-explicit methods ($\bar{a}_{ij} = a_{i+1,j}$) for systems of DAEs of the form (2), whose series expansions (for consistent initial values) where derived in [6]. It seams reasonable to expect that the series expansion in powers of h of the PRK scheme (6)-(7) also fits in the same class of formal series. The approach adopted to study the convergence of the application of Runge-Kutta type methods for systems of the form (2) in [6] (see also [8]) only requires expanding the numerical solution for consistent initial values. In [5], the expansion of implicit Runge-Kutta methods for inconsistent initial values is considered. Such expansions (the DA2-series) also are of the form (1), but unlike the series that arise when studying methods for ODEs (for instance, the B-series [7]), they include (in the case of inconsistent initial values) terms with negative powers of h.

In Section 7, we will find usefull to expand the numerical approximation given by the application of the PRK scheme (6)-(7) with inconsistent initial values (y_0, z_0). Next, we will show, using the abstract framework of Section 3, that the PRK solution can be formally expanded as a DA2-series.

A DA2-series is a formal expression (1), where $\mathcal{T}$ is the set of *DA2-trees*. Next, we will define $\mathcal{T}$, together with the order $\rho(u)$ and the elementary differential $F(u)$ associated to each DA2-tree.

Recall that, according to the terminology used in the first part [13] of our paper, a 2-tree is a rooted tree with vertices of two different types or colors (black for type 1, and white for type 2). The single 2-tree with a vertex of type ν ($\nu = 1, 2$) is denoted by $[\emptyset]_\nu$, and the 2-trees with more than a vertex can be recursively represented as $u = [u_1, \ldots, u_m]_\nu$, where ν is the color of the root of u, and $u_1, \ldots, u_m$ are the 2-trees obtained from chopping off the root of u.

A DA2-tree is a 2-tree that does not have any subtree either of the form $[v_1, \ldots, v_m]_2$ with some $v_l \in \mathcal{T}_2$ or of the form $[[v]_1]_2$ with $v \in \mathcal{T}_2$. We will alternatively denote the DA2-trees with a single vertex $[\emptyset]_1$ and $[\emptyset]_2$ respectively as $\bullet$ and $\circ$. We denote by $\mathcal{T}_1$ (resp. $\mathcal{T}_2$) the set of DA2-trees with root of type 1 (resp. of type 2). Now, the set $\mathcal{T}$ can be defined using the notation of Definition 3 as follows.

Definition 8 *The set $\mathcal{T} = \mathcal{T}_1 \cup \mathcal{T}_2$ (and $\widehat{\mathcal{T}}^*$) is recursively defined by*

$$\begin{aligned}
\widehat{\mathcal{T}}^* &= \{[\emptyset]\} \cup \{[u_1, \ldots, u_m] : \ u_1, \ldots, u_m \in \mathcal{T}_1\} - \{[[u]_1] : \ u \in \mathcal{T}_2\}, \\
\mathcal{T}_1 &= \{[u_1, \ldots, u_m]_1 : \ [u_1, \ldots, u_m] \in \widehat{\mathcal{T}}\}, \\
\mathcal{T}_2 &= \{[u_1, \ldots, u_m]_2 : \ [u_1, \ldots, u_m] \in \widehat{\mathcal{T}}^*\}.
\end{aligned}$$

Given a DA2-tree $u = [u_1, \ldots, u_m]_\nu$ ($\nu = 1, 2$), we say that $\mathrm{op}(u) = [u_1, \ldots, u_m]$ is the *forest of operands* of the DA2-tree u and that $\mathrm{tp}(u) = \nu$ is the *type* of u.

Definition 9 *The order $\rho(u)$ and the elementary differential $F(u)$ of each DA2-tree u can be defined using the notation introduced in Definition 3 by*

$$F(u) = \begin{pmatrix} X(\mathrm{op}(u))f \\ 0 \end{pmatrix}, \quad \rho(u) \;=\; \widehat{\rho}(\mathrm{op}(u)) + 1, \quad \mathit{if}\; u \in \mathcal{T}_1,$$
$$F(u) = \begin{pmatrix} 0 \\ (-g_y f_z)^{-1} X(\mathrm{op}(u))g \end{pmatrix}, \quad \rho(u) \;=\; \widehat{\rho}(\mathrm{op}(u)) - 1, \quad \mathit{if}\; u \in \mathcal{T}_2.$$

In the context of Section 3 of [13], the order of a 2-tree was the total number of its vertices. However, it is clear from the recursive definition above that the order $\rho(u)$ of a DA2-tree of $\mathcal{T}$ is the number of its vertices of type 1 minus the number of vertices of type 2. In particular, $\rho(\circ) = -1$, and DA2-trees with negative order of arbitrarily high absolute value exist, for instance, those of the form $[\circ, \ldots, \circ]_1$. If (y_0, z_0) are assumed to be consistent, the elementary differentials $F(u)$ corresponding to DA2-trees of order $\rho(u) \leq 0$ vanish, and if only (4) is fulfilled by (y_0, z_0), only the elementary differentials corresponding to DA2-trees of non-negative order remain.

Our aim now is to prove that, under the conditions of Lemma 1, the numerical solution (y_1, z_1) given by (6)-(7) can be expanded as a DA2-series. In order to do that, we will try to determine suitable coefficients $\mathbf{c}(u)$ that only depend on the parameters a_{ij}, $\bar{a}_{ij}$, b_i, d_i of the method such that formally,

$$\begin{pmatrix} y_1 \\ z_1 \end{pmatrix} = \mathcal{S}(\mathbf{c})(y_0, z_0). \tag{20}$$

Given a vector $x = (y^T, x^T)^T$, we will use the notation $x^1 = y$, $x^2 = z$, and similarly,

$$F(u) = \begin{pmatrix} F(u)^1 \\ F(u)^2 \end{pmatrix}, \quad S(\mathbf{c}, h) = \begin{pmatrix} S(\mathbf{c}, h)^1 \\ S(\mathbf{c}, h)^2 \end{pmatrix}.$$

First, we will focus our attention on the expansions of $Y_i, \bar{Y}_i, Z_i$, and try to find $\mathbf{C}_i : \mathcal{T} \to R$ and $\bar{\mathbf{C}}_i : \mathcal{T}_1 \to R$ such that

$$Y_i = \mathcal{S}(\mathbf{C}_i)^1(y_0, z_0), \quad \bar{Y}_i = \mathcal{S}(\bar{\mathbf{C}}_i)^1(y_0, z_0), \quad Z_i = \mathcal{S}(\mathbf{C}_i)^2(y_0, z_0). \tag{21}$$

From (6)-(7), we see that in order to accomplish that it is convenient to know how to expand compositions of the form $f \circ \mathcal{S}(\mathbf{d})$ and $g \circ \mathcal{S}(\mathbf{d})$. But according to Theorem 5, their expansions can be obtained as the action of the series of elementary operators $\widehat{\mathcal{S}}(\mathbf{d}')$ respectively on f and g. By definition, $F(u)^1 = X(\mathrm{op}(u))f$ if $u \in \mathcal{T}_1$, which directly implies that

$$hf \circ \mathcal{S}(\mathbf{d}) = h\widehat{\mathcal{S}}(\mathbf{d}')f = \bar{\mathcal{S}}(\mathbf{d}')^1, \quad \text{with } \; \mathbf{d}'(u) = \mathbf{d}'(\mathrm{op}(u)). \tag{22}$$

As for the expansion of $g \circ \mathcal{S}(\mathbf{d}) = \widehat{\mathcal{S}}(\mathbf{d}')g$, since g does not depend on z, the terms of the series $\widehat{\mathcal{S}}(\mathbf{d}')g$ corresponding to $w = [u_1, \ldots, u_m] \in \widehat{\mathcal{T}}$ with some $u_l \in \mathcal{T}_1$ are identically null. The remaining terms correspond to $\hat{w} \in \widehat{\mathcal{T}}$ such that $\hat{w} = \mathrm{op}(u)$ or $\hat{w} = [[u]_1]$ with $u \in \mathcal{T}_2$. However, given $u \in \mathcal{T}_2$, taking into account the definition of $F(u)$ and the definition of the elementary operators,

$$X[[u]_1]g = -X(\mathrm{op}(u))g = (g_y f_z)F(u)^2, \quad \text{if } u \in \mathcal{T}_2. \tag{23}$$

From that, one obtains

$$g \circ \mathcal{S}(\mathbf{d}) \;=\; \widehat{\mathcal{S}}(\mathbf{d}')g = h(-g_y f_z)\bar{\mathcal{S}}(\widetilde{\mathbf{d}'})^2, \tag{24}$$

where $\widetilde{\mathbf{d}'}(u) = \mathbf{d}'(\mathrm{op}(u)) - \mathbf{d}'([[u]_1]) = \mathbf{d}'(\mathrm{op}(u)) - \mathbf{d}([u]_1)$.

Taking into account (22) and (24), the formal equations obtained from replacing $Y_i, \bar{Y}_i, Z_i$ by (21) in (7) read

$$\begin{aligned}
0 &= \sum_{u \in \mathcal{T}_1} h^{\rho(u)} \left(\mathbf{C}_i(u) - \sum_{j=0}^{s} a_{ij} \mathbf{C}'_j(\mathrm{op}(u)) \right) F(u)^1, \\
0 &= \sum_{u \in \mathcal{T}_1} h^{\rho(u)} \left(\bar{\mathbf{C}}_i(u) - \sum_{j=0}^{s} \bar{a}_{ij} \mathbf{C}'_j(\mathrm{op}(u)) \right) F(u)^1, \\
0 &= (-g_y f_z) \sum_{u \in \mathcal{T}_1} h^{\rho(u)+1} \left(\bar{\mathbf{C}}'_i(\mathrm{op}(u)) - \bar{\mathbf{C}}_i([u]_1) \right) F(u)^2.
\end{aligned}$$

Clearly, these equalities hold if the coefficients $\mathbf{C}_i(u), \bar{\mathbf{C}}_i(u)$ are such that the expressions in the inner parenthesis are null, that is,

$$\mathbf{C}_i(u) = \sum_{j=0}^{s} a_{ij} \mathbf{C}'_j(\mathrm{op}(u)), \quad \bar{\mathbf{C}}_i(u) = \sum_{j=0}^{s} \bar{a}_{ij} \mathbf{C}'_j(\mathrm{op}(u)), \quad u \in \mathcal{T}_1, \tag{25}$$

and $\bar{\mathbf{C}}'_i(\mathrm{op}(u)) = \bar{\mathbf{C}}_i([u]_1)$ if $u \in \mathcal{T}_2$. Now, (25) implies that $\mathbf{C}_i([u]_1) = \sum \bar{a}_{ij} \mathbf{C}_i(u)$, which together with $\mathbf{C}_0(u) = 0$ (as $Y_0 = y_0$ and $Z_0 = z_0$) leads to

$$\mathbf{C}_i(u) = \sum_{j=1}^{s} w_{ij} \bar{\mathbf{C}}'_j(\mathrm{op}(u)), \quad u \in \mathcal{T}_2, \tag{26}$$

where the coefficients w_{ij} are the entries of the inverse of the $s \times s$ matrix $\tilde{A}$.

The equalities (25)-(26) determine the coefficients of all the DA2-trees recursively with respect to the number of vertices. Furthermore, we have that formally $hf(Y_i, Z_i) = \overline{\mathcal{S}}(\mathbf{C}'_i)(y_0, z_0)$. Hence, from (6) it is clear that the numerical solution (y_1, z_1) can be expanded as a DA2-series (20), where the coefficients $\mathbf{c}(u)$ satisfy

$$\mathbf{c}(u) = \sum_{i=1}^{s} b_i \mathbf{C}'_i(u), \quad \text{if } u \in \mathcal{T}_1, \quad \mathbf{c}(u) = \sum_{i=1}^{s} d_i \mathbf{C}_i(u), \quad \text{if } u \in \mathcal{T}_2. \tag{27}$$

It is interesting to note that the elementary differentials $F(u)$ of DA2-trees are independent, so that the coefficients of the expansions (20) and (21) are unique. In fact, the following result can be proven:

Lemma 10 *For each DA2-tree $u \in \mathcal{T}_1$ (resp. $u \in \mathcal{T}_2$) there exists a system (2) with polynomial vector functions f (of dimension n, the number of vertices of u) and g (of dimension m, the number of vertices of type 2 of u), such that $(g_y f_z)(0,0)$ is the identity matrix, and the first component (resp. the $(m+1)$th component) of $F(v)(0,0)$ is non-null if and only if $u = v$.*

We hint the proof of this lemma with an example: For $u = [\bullet, \circ, [\bullet, \bullet]_2]_1$, we take $n = 6$, $m = 2$, and

$$f(y,z) = (y^2 z^1 z^2, 1, 1, 1, z^1, z^2)^T, \quad g(y) = (y^5 + y^3 y^4, y^6 + 1)^T.$$

5 Composition of DA2-series

We are now concerned with the formal expansion of compositions $\mathcal{S}(\mathbf{b}) \circ \mathcal{S}(\mathbf{d})$ of DA2-series. Clearly, it is sufficient to expand compositions of the form $h^{\rho(v)} F(v) \circ \mathcal{S}(\mathbf{d})$. In order to do that, it will be useful to study in some detail the series of the form $\widehat{\mathcal{S}}(\mathbf{a})f$ and $\widehat{\mathcal{S}}(\mathbf{a})g$ for $\mathbf{a} : \widehat{\mathcal{T}} \to R$. First note that the arguments that led us to (22) and (24) are also valid with the mapping $\mathbf{d}' : \widehat{\mathcal{T}} \to R$ replaced by an arbitrary mapping $\hat{\mathbf{a}} : \widehat{\mathcal{T}} \to R$, that is,

$$h\widehat{\mathcal{S}}(\hat{\mathbf{a}})f = \bar{\mathcal{S}}(\hat{\mathbf{a}})^1, \quad \widehat{\mathcal{S}}(\hat{\mathbf{a}})g = h(-g_y f_z)\bar{\mathcal{S}}(\tilde{\hat{\mathbf{a}}})^2, \tag{28}$$

where $\hat{\mathbf{a}}(u) = \hat{\mathbf{a}}(\mathrm{op}(u))$ and $\tilde{\hat{\mathbf{a}}}(u) = \hat{\mathbf{a}}(\mathrm{op}(u)) - \hat{\mathbf{a}}([[u]_1])$. This leads, taking into account Lemma 7, to the following:

Lemma 11 *Given $\mathbf{d} : \mathcal{T} \to R$, and $\hat{w} \in \widehat{\mathcal{T}}^*$,*

$$h^{\hat{\rho}(\hat{w})} X(\hat{w}) g \circ \mathcal{S}(\mathbf{d}) = \widehat{\mathcal{S}}(\mathbf{d}^{(\hat{w})})g = h(-g_y f_z)\overline{\mathcal{S}}(\widetilde{\mathbf{d}^{(\hat{w})}})^2,$$

where $\widetilde{\mathbf{d}^{(\hat{w})}}(u) = \mathbf{d}^{(\hat{w})}(\mathrm{op}(u)) - \mathbf{d}^{(\hat{w})}([[u]_1])$ for $u \in \mathcal{T}_2$.

The following lemma is the basic ingredient of the main result of this section.

Lemma 12 *Given a DA2-series $\mathcal{S}(\mathbf{d})$, the following identity formally holds for each $v \in \mathcal{T}$*

$$h^{\rho(v)} F(v) \circ \mathcal{S}(\mathbf{d}) = \mathcal{S}(\mathbf{d}^{(v)}), \tag{29}$$

where the coefficients $\mathbf{d}^{(v)}(u)$ $(u, v \in \mathcal{T})$ can be recursively obtained with the help of the coefficients $\mathbf{d}^{(w_2)}(w_1)$ $(w_1, w_2 \in \widehat{\mathcal{T}})$ of Lemma 7 as follows:

(1) $\mathbf{d}^{(v)}(u) = 0$ *if* $\mathrm{tp}(u) \neq \mathrm{tp}(v)$,
(2) $\mathbf{d}^{(v)}(u) = \mathbf{d}^{(\mathrm{op}(v))}(\mathrm{op}(u))$ *if* $u, v \in \mathcal{T}_1$,
(3) $\mathbf{d}^{(v)}(u) = \mathbf{d}^{(\mathrm{op}(v))}(\mathrm{op}(u)) + \mathbf{d}^{([[v]_1])}(\mathrm{op}(u)) - \mathbf{d}^{(\mathrm{op}(v))}([[u]_1])$ *if* $u, v \in \mathcal{T}_2$.

Observe that $\mathbf{d}^{(\mathrm{op}(v))}([[u]_1]) = 0$ unless $\mathrm{op}(v) = [\emptyset]$ or $\mathrm{op}(v) = [\bullet]$. From that and the recursive definition of the coefficients $\mathbf{d}^{(w_2)}(w_1)$ in Lemma 7, it can be seen that the formulae above uniquely determines the coefficients $\mathbf{d}^{(v)}(u)$ for $u, v \in \mathcal{T}$.

PROOF. Item 1 immediately follows from the definition of elementary differentials, since $F(u)^\nu = 0$ if $\mathrm{tp}(u) \neq \nu$. Item 2 follows from Lemma 7 and (28). As for item 3, we compose both sides of the first equality of (23) with the DA2-series $\mathcal{S}(\mathbf{d})$, and apply Lemma 11 to obtain

$$(g_y f_z)\overline{\mathcal{S}}(\mathbf{d}^{\widetilde{([[v]_1])}})^2 = -(g_y f_z)\overline{\mathcal{S}}(\mathbf{d}^{\widetilde{(\mathrm{op}(v))}})^2.$$

Item 3 is finally obtained, taking into account the definition of the coefficients $\widetilde{\mathbf{d}^{(\hat{w})}}$ and that $\mathbf{d}^{([[v]_1])}([[u]_1]) = \mathbf{d}^{(v)}(u)$. □

Theorem 13 *Given $\mathbf{d}, \mathbf{b} : \mathcal{T} \to R$, then*

$$\overline{\mathcal{S}}(\mathbf{b}) \circ \mathcal{S}(\mathbf{d}) = \overline{\mathcal{S}}(\mathbf{db}),$$

where for each $u \in \mathcal{T}$

$$\mathbf{db}(u) = \sum_{v \in \mathcal{T}} \mathbf{b}(u)\mathbf{d}^{(v)}(u), \tag{30}$$

and the coefficients $\mathbf{d}^{(v)}(u)$ are defined as in Lemma 12.

In [5], the composition of the particular case of DA2-series corresponding to implicit Runge-Kutta schemes with invertible matrix is studied. They give a systematic way of obtaining the coefficients of the composed DA2-series using *pairs of embedded DA2-trees.* Theorem 13 generalizes this result to the case of composition of general DA2-series, and their formulae in terms of embedded pairs of DA2-trees can be proven for the general case from Lemma 12.

6 Truncated DA2-series expansions

Though the expansions we have obtained so far are merely formal, we will show that they can have a rigorous meaning when truncated in a suitable way.

Let us assume that we are under the hypothesis of Lemma 1. According to Definition 9, the independent terms in the DA2-series satisfy

$$h^{\rho(u)}F(u)^1 = O(h\epsilon^{\eta(u)-1}), \quad h^{\rho(u)}F(u)^2 = O(\epsilon^{\eta(u)}),$$

where ϵ is given by (10), and $\eta(\bullet) = \eta(\circ) = \eta([\bullet]_2) = 1$ and for the rest of the DA2-trees,

$$\begin{aligned} \eta([u_1,\ldots,u_m]_1) &= \eta(u_1) + \ldots + \eta(u_m) + 1, \\ \eta([u_1,\ldots,u_m]_2) &= \eta(u_1) + \ldots + \eta(u_m) - 1. \end{aligned}$$

It can be shown, by induction on the number of vertices, that $\eta(u) \geq 1$ for every $u \in \mathcal{T}$. Moreover, the following sets of DA2-trees

$$\mathcal{T}_1^{[r]} = \{u \in \mathcal{T}_1 \;/\; \eta(u) \leq r\}, \quad \mathcal{T}_2^{[r]} = \{u \in \mathcal{T}_2 \;/\; \eta(u) \leq r\},$$

are finite for each $r \geq 1$.

Given $\mathbf{d} : \mathcal{T} \to R$, $r \geq 1$, we denote by $\mathcal{S}^{[r]}(\mathbf{d})$ the truncated series

$$\mathcal{S}^{[r]}(\mathbf{d}) = \begin{pmatrix} \mathcal{S}^{[r]}(\mathbf{d})^1 \\ \mathcal{S}^{[r]}(\mathbf{d})^2 \end{pmatrix} = \begin{pmatrix} id^1 + \sum\limits_{u \in \mathcal{T}_1^{[r]}} h^{\rho(u)}\mathbf{d}(u)F(u)^1 \\ \mathrm{id}^2 + \sum\limits_{u \in \mathcal{T}_2^{[r]}} h^{\rho(u)}\mathbf{d}(u)F(u)^2 \end{pmatrix}$$

It is important to note that the formal results of Section 5 admit rigorous versions in terms of truncated DA2-series of the form $\mathcal{S}^{[r]}(\mathbf{d})$ and $\overline{\mathcal{S}}^{[r]}(\mathbf{d}) = \mathcal{S}^{[r]}(\mathbf{d}) - id$. In particular,

Lemma 14 *Given* $\mathbf{d} : \mathcal{T} \to R$, *for each* $r \geq 1$,

$$\begin{aligned} h\, f \circ \mathcal{S}^{[r-1]}(\mathbf{d}) &= \overline{\mathcal{S}}^{[r]}(\mathbf{d}')^1 + hO(\epsilon^r), \\ g \circ \mathcal{S}^{[r]}(\mathbf{d}) &= h\,(-g_y f_z)\, \bar{\mathcal{S}}^{[r-1]}(\widetilde{\mathbf{d}'})^2 + hO(\epsilon^r), \end{aligned}$$

where $\mathbf{d}'(u) = \mathbf{d}'(\mathrm{op}(u))$ *and* $\widetilde{\mathbf{d}'(u)} = \mathbf{d}'(\mathrm{op}(u)) - \mathbf{d}([u]_1)$ *for each* $u \in \mathcal{T}$.

Theorem 15 *Consider the coefficients* $\mathbf{c}(u)$ *given by (27), (25), (26). The numerical solution* (y_1, z_1) *given by the PRK scheme (6)-(7) satisfy*

$$y_1 = \mathcal{S}^{[r]}(\mathbf{c})^1(y_0, z_0) + O(h\epsilon^r), \quad z_1 = \mathcal{S}^{[r]}(\mathbf{c})^2(y_0, z_0) + O(\epsilon^{r+1}),$$

for sufficiently small $\epsilon = \max\left(|h|, ||g(y_0)||/|h|, ||g_y(y_0) f(y_0, z_0)||\right)$.

PROOF. It is sufficient to prove that

$$k_i = k_i^{[r]} + O(\epsilon^r), \quad Z_i = Z_i^{[r]} + O(\epsilon^r), \tag{31}$$

where Z_i and $k_i = f(Y_i, Z_i)$ are the intermediate values given by (7), and for each $r \geq 1$ and $1 \leq i \leq s$

$$\begin{aligned} k_i^{[r]} &= \sum_{u \in \mathcal{T}_1^{[r]}} h^{\rho(u)-1} \mathbf{C}_i'(u) F(u)^1(y_0, z_0), \\ Z_i^{[r]} &= z_0 + \sum_{u \in \mathcal{T}_2^{[r-1]}} h^{\rho(u)} \mathbf{C}_i(u) F(u)^2(y_0, z_0). \end{aligned}$$

Clearly, $U^{[r]} = (k_1^{[r]}, \ldots, k_s^{[r]}, Z_1^{[r]}, \ldots, Z_s^{[r]})$ can be considered as the solution of the perturbed PRK scheme (11) with $\gamma_i^{[r]}$ and $\omega_i^{[r]}$ defined by

$$\gamma_i^{[r]} = k_i^{[r]} - f\left(y_0 + h \sum_{j=0}^{s} a_{ij} k_j^{[r]}, Z_i^{[r]}\right), \quad \omega_i^{[r]} = \frac{1}{h} g\left(y_0 + h \sum_{j=0}^{s} \bar{a}_{ij} k_j^{[r]}\right),$$

that is, $\mathcal{F}(h, y_0, z_0, \gamma^{[r]}, \omega^{[r]}, U^{[r]}) = 0$ (where $\mathcal{F}$ is the smooth function considered in the proof of Lemma 2). If the coefficients $\mathbf{C}_i'(u)$ and $\mathbf{C}_i(u)$ satisfy (25) and (26), applying Lemma 14 one obtains that $\gamma_i^{[r]} = O(\epsilon^r)$ and $\omega_i^{[r]} = -\frac{1}{h} g(y_0) + O(\epsilon^r)$. Finally, taking into account that the exact intermediate values $U = (k_1, \ldots, k_s, Z_1, \ldots, Z_s)$ of (6)-(7) satisfy $\mathcal{F}(h, y_0, z_0, \gamma, \omega, U) = 0$ with $\gamma_i = 0$ and $\omega_i = -\frac{1}{h} g(y_0)$, (31) follows from the implicit function Theorem. □

In next section, we will be interested in the truncated expansion of compositions of the form $(g_y f) \circ \mathcal{S}^{[r]}(\mathbf{d})$. Since this composition is $X[\bullet] g \circ \mathcal{S}^{[r]}(\mathbf{d})$, its expansion can be obtained considering the truncated version of Lemma 11 for $\hat{w} = [\bullet]$:

Lemma 16 *Given $\mathbf{d} : \mathcal{T} \to R$, for each $r \geq 1$,*

$$(g_y f) \circ \mathcal{S}^{[r]}(\mathbf{d}) = (-g_y f_z) \bar{\mathcal{S}}^{[r]}(\widetilde{\mathbf{d}^{([\bullet])}})^2 + O(\epsilon^{r+1}), \tag{32}$$

where $\widetilde{\mathbf{d}^{([\bullet])}}(u) = \mathbf{d}^{([\bullet])}(\mathrm{op}(u)) - \mathbf{d}(u)$ for each $u \in \mathcal{T}_2$.

7 Characterization of convergence of PRK methods

A characterization of the convergence of PRK methods satisfying (8) can be obtained using the techniques developed in [6] (see also [8]). To do that, it is sufficient to expand, for each consistent (y_0, z_0), the numerical solution (y_1, z_1) of (6)-(7) in powers of h. However, if (8) is not fulfilled, so that the numerical solution y_1 does depend on z_0, further study is needed to get optimal convergence results. The convergence of different classes of partitioned Runge-Kutta methods for which (8) does not hold but (9) does, is studied in the literature: In [10], a class of implicit Runge-Kutta methods with these properties is studied (in particular, optimal convergence results for the family of Lobatto IIIA methods are obtained). Half-explicit methods belonging to the class of PRK methods with these properties are considered in [1] and [11]. Inspired by the ideas presented in [8] (Section V.I) on the local error for inconsistent initial values of Rosenbrock methods applied to index 1 DAEs, we observe that sharper convergence results for PRK methods that does not satisfy (8) can be obtained studying the local error of the method for inconsistent initial values (y_0, z_0). Our approach gives in general sharper convergence results than other approaches [10,1,11] to study the convergence of the application of a (6)-(7) scheme that satisfy (9) and do not satisfy (8) (so that the numerical solution depend on both y_0 and z_0). Furthermore, we will show that these ideas can be extended to the case of PRK methods that violate both (9) and (8).

First, we will study the simpler case where (9) is satisfied, and the procedure for the general case will be outlined at the end of this section.

Let us consider a certain neighborhood $\mathcal{U}$ of the solution of (2) we are aiming to approximate such that, for each $(y_0, z_0) \in \mathcal{U}$, the equation $g_y(y_0) f(y_0, z) = 0$ has a unique solution $z = G(y_0)$. We also assume without loss of generality that for each $(y_0, z_0) \in \mathcal{U}$ the estimates of Theorem 15 hold with the same constants.

Given $(y_0, z_0) \in \mathcal{U}$ such that (4) holds (but (5) does not necessarily hold), let $(y(t), z(t))$ be the solution of the system (2) with initial values $y(t_0) = y_0$, $z(t_0) = G(y_0)$. Equivalently, $y(t)$ is the exact solution of the ODE

$$y' = f(y, G(y)), \quad y(t_0) = y_0, \tag{33}$$

and $z(t) = G(y(t))$. We denote the local error of the scheme (6)-(7) at (y_0, z_0) by $\delta_y(y_0, z_0, h) = y_1 - y(t)$, $\delta_z(y_0, z_0, h) = z_1 - z(t)$.

We already have the expansion of (y_1, z_1) as a truncated DA2-series (Theorem 15) and we seek for a similar expansion $\mathcal{S}(\mathbf{e})(y_0, z_0)$ of the exact solution $(y(t_0 + h), z(t_0 + h))$. Recall that the initial values are now assumed to satisfy (4), and therefore, the elementary differentials corresponding to DA2-trees with some terminal vertex of type 2 (that is, which has the single DA2-tree $[\emptyset]_2 = \circ$ as subtree) vanish. In order to derive the coefficients $\mathbf{e}(u)$, we replace the expansion $\mathcal{S}(\mathbf{e})(y_0, z_0)$ in (2), apply Lemma 14, and choose the coefficients $\mathbf{e}(u)$ in such a way that the residuals are as small as possible.

Theorem 17 *The exact solution of the system (2) with initial values $y(t_0) = y_0$ and $z(t_0) = G(y_0)$ satisfy for each $r \geq 1$,*

$$y(t_0 + h) = \mathcal{S}^{[r]}(\mathbf{e})^1(y_0, z_0) + hO(\epsilon^r), \tag{34}$$

$$z(t_0 + h) = \mathcal{S}^{[r]}(\mathbf{e})^2(y_0, z_0) + O(\epsilon^{r+1}), \tag{35}$$

where the coefficients $\mathbf{e}(u)$ for DA2-trees without terminal vertices of type 2 are recursively determined by $\mathbf{e}(\bullet) = 1$ and

$$\mathbf{e}(u) = \frac{\mathbf{e}'(u)}{\rho(u)}, \quad \textit{if } u \in \mathcal{T}_1, \quad \mathbf{e}(u) = (\rho(u) + 1)\mathbf{e}'(u), \quad \textit{if } u \in \mathcal{T}_2. \tag{36}$$

PROOF. Let us denote $(y^{[r]}(t_0+h), z^{[r]}(t_0+h)) = \mathcal{S}^{[r]}(\mathbf{e})(y_0, z_0)$. With the coefficients $\mathbf{e}(u)$ chosen in this way, Lemma 14 implies that $(y^{[r]}(t_0+h), z^{[r-1]}(t_0+h))$ is the solution of the perturbed system $y' = f(y, z) + \theta_1(h)$, $g(y) = \theta_2(h)$, where h plays the role of the time, and the residuals satisfy $\theta_1(h) = O(\epsilon^r)$ and $\theta_2(h) = hO(\epsilon^r)$. From that, (34) can be obtained using a continuous analog of Lemma 2, for instance (VI.5.22) in [8]. □

Now, the convergence of the PRK method can be studied in a similar way to one-step methods for ODEs as follows: Assume that y_0 satisfy (4). Let ϕ_h be the flow of the ODE system (33), so that its exact solution satisfy $y(t + h) = \phi_h(y(t))$. Let $(y_n, z_n) \approx (y(t_n), z(t_n))$ the numerical solution y_n

that result from the repeated application ($t_n = t_{n-1} + h$) of the PRK method (6)-(7) satisfying (9). Then, $y_{n+1} = \phi_h(y_n) + \delta_n$, where $\delta_n = \delta_y(y_n, z_n, h)$ and therefore, the errors $\Delta y_n = y_n - y(t_n)$ satisfy the recursion

$$\Delta y_{n+1} = (I + O(h))\Delta y_n + \delta_n. \tag{37}$$

Thus, the convergence of the method can be studied proceeding as follows:

(1) Study the propagation of the residuals in the hidden constraint $\varepsilon_n = g_y(y_n)f(y_n, z_n)$. Although this can be done using the expansion $\mathcal{S}(\mathbf{c} - \mathbf{e})^2$ of the local error for the z component, a more direct procedure is to apply Lemma 16 to obtain the truncated expansion of the composition $(g_y f) \circ \mathcal{S}^{[r]}(\mathbf{c})$. In particular this shows that,

$$\varepsilon_{n+1} = \alpha\,\varepsilon_n + O(h) + O(||\varepsilon_n||^2), \tag{38}$$

where $\alpha = \widetilde{\mathbf{c}^{([\bullet])}}([\bullet]_2) = \mathbf{c}^{([\bullet])}([\bullet]) - \mathbf{c}([\bullet]_2) = 1 - \sum_{i,j,l} d_i w_{ij} \bar{a}_{jl}$. Clearly, the method will not converge to the desired solution if $|\alpha| > 1$. It can be proven that, if $z_1 = Z_j$ for some j, then the $O(||\varepsilon_n||^2)$ term can be removed from (38). Sharper estimates of the $O(h) + O(||\varepsilon_n||^2))$ term in (38) can be obtained inspecting the values of the coefficients $\widetilde{\mathbf{c}^{([\bullet])}}(u)$ for $u \in \mathcal{T}_2$.

(2) Estimate the local errors $\delta_n = \delta_y(y_n, z_n, h)$, taking into account Theorems 15 and 17 and the estimates obtained for the values $\varepsilon_n = g_y(y_n)f(y_n, z_n)$ at the previous step.

(3) Once the local errors δ_n have been estimated, the convergence behavior of the differential components can be studied adapting to (37) the standard techniques for one-step methods for ODEs. The global errors $\Delta z_n = z_n - z(t_n)$ for the algebraic components can be estimated from the estimates for Δy_n and ε_n using the implicit function theorem.

To illustrate our approach, let us consider the two half-explicit methods of order 5 proposed in [1] and [11]. Both methods can be considered as PRK methods, and they satisfy (9) but do not satisfy (8). Also, $z_1 = Z_s$, so that $\varepsilon_{n+1} = \alpha\,\varepsilon_n + O(h)$. Studying their coefficients $\mathbf{c}(u)$ and $\widetilde{\mathbf{c}^{([\bullet])}}(u)$ one observe that in both cases:

(1) $\alpha = 1 - \mathbf{c}([\bullet]_2) = 0$ and $\varepsilon_{n+1} = O(h^4 + h^2||\varepsilon_n|| + h||\varepsilon_n||^2)$. This provides, in the particular case of $n = 0$, an estimate of ε_1 in terms of h and the initial error ε_0. Furthermore, this implies that $\varepsilon_2 = O(h^4 + h^3||\varepsilon_0||^2)$, and $\varepsilon_n = O(h^4)$ for $n \geq 3$.

(2) $\delta_n = O(h^6 + h^3||\varepsilon_n|| + h||\varepsilon_n||^2)$, and therefore, $\delta_0 = O(h^6 + h^3||\varepsilon_0|| + h||\varepsilon_0||^2)$, $\delta_1 = O(h^6 + h^5||\varepsilon_0|| + h^4||\varepsilon_0||^2 + h^3||\varepsilon_0||^4)$ and $\delta_n = O(h^6)$ for $n \geq 2$.

(3) Thus, from (37) one obtains that, for $nh \leq Constant$

$$\Delta y_n = O(h^5 + h^3||\varepsilon_0|| + h||\varepsilon_0||^2), \quad \Delta z_n = O(h^4 + h^3||\varepsilon_0|| + h||\varepsilon_0||^2).$$

Consequently, for both methods it is sufficient to have $\varepsilon_0 = g_y(y_0)f(y_0, z_0) = O(h^2)$ to achieve convergence of order 5 for the differential component and of order 4 for the algebraic component.

General case: The convergence of the general case of PRK methods that satisfy neither (9) nor (8), with possibly inconsistent initial values, can also be studied with a more general approach. The key to do this is to consider the standard projection

$$\tilde{y} = y + f_z(y, z)\nu, \quad g(\tilde{y}) = 0.$$

For each $(y, z) \in \mathcal{U}$, we denote the projected value as $\tilde{y} = \mathcal{P}(y, z)$. Now, the local error of the scheme (6)-(7) at $(y_0, z_0) \in \mathcal{U}$ is defined as

$$\delta_y(y_0, z_0, h) = y_1 - \phi_h(\mathcal{P}(y_0, z_0)), \quad \delta_z(y_0, z_0, h) = z_1 - G(\phi_h(\mathcal{P}(y_0, z_0))).$$

It is not difficult to see that $\mathcal{P}(y_0, z_0)$ can be expanded as a DA2-series $\mathcal{S}(\pi)^1(y_0, z_0)$, where $\pi(u) = 1$ if all the subtrees of $u \in \mathcal{T}_1$ with root of type 1 are of the form $[v]_1$ with $v \in \mathcal{T}_2$, and $\pi(u) = 0$ otherwise. In particular, $\pi(u) = 1$ implies that $\rho(u) = 0$, as it may be expected, since h does not appear in the definition of the projection $\mathcal{P}$. The composition $\phi_h(\mathcal{P}(y_0, z_0))$ can also be expanded as a DA2-series, using the techniques of Section 5, as the composition of two DA2-series. From that, it can be obtained that $\phi_h(\mathcal{P}(y_0, z_0))$ admits a truncated expansion of the form $\mathcal{S}^{[r]}(\mathbf{e})^1$, where (i) $\mathbf{e}(u) = 0$ if $\rho(u) < 0$, (ii) $\mathbf{e}(u) = \pi(u)$ if $\rho(u) = 0$ and $u \in \mathcal{T}_1$, (iii) and for the rest of DA2-trees u such that $\rho(u) \geq 0$ the coefficients are determined by (36).

Then, a possible procedure to study the convergence behavior of the PRK scheme is the following:

(1) Study the propagation of $\varepsilon_n = (\varepsilon_n^1, \varepsilon_n^2)$, where $\varepsilon_n^1 = g(y_n)/h$ and $\varepsilon_n^2 = g_y(y_n)f(y_n, z_n)$. In order to do that, expand $g \circ \mathcal{S}(\mathbf{c})$ and $(g_y f) \circ \mathcal{S}(\mathbf{c})$ with the help of Lemma 14 and Lemma 16. Thus, one obtains

$$\begin{pmatrix} \varepsilon_{n+1}^1 \\ \varepsilon_{n+1}^2 \end{pmatrix} = \begin{pmatrix} 1 - \mathbf{c}([\circ]_1) & \mathbf{c}(\bullet) - \mathbf{c}([[\bullet]_2]_1) \\ -\mathbf{c}(\circ) & 1 - \mathbf{c}([\bullet]_2) \end{pmatrix} \begin{pmatrix} \varepsilon_n^1 \\ \varepsilon_n^2 \end{pmatrix} + O(h) + O(||\varepsilon_n||^2).$$

The contractivity condition $|\alpha| \leq 1$ needed for the convergence of methods satisfying (9) is here replaced by a condition on the eigenvalues of the 2×2 matrix above.

(2) Estimate the local errors $\delta_n = \delta_y(y_n, z_n, h)$ using its expansion $\mathcal{S}(\mathbf{c} - \mathbf{e})^1(y_n, z_n)$ and the estimates obtained for $g(y_n)/h$ and $g_y(y_n)f(y_n, z_n)$.

(3) Compare the numerical solution to $y(t_{n+1}) = \phi_h(y(t_n)) = \phi_h(\mathcal{P}(y(t_n), z_n))$. To do that, observe that

$$\frac{\partial \mathcal{P}(y,z)}{\partial y} = P(y,z) + O(||g(y)||),$$

where $P(y,z) = I - (f_z(-g_y, f_z)^{-1} g_y)(y,z)$. This leads to the following

$$\Delta y_{n+1} = (I + O(h))(P(y_n, z_n) + O(||\Delta y_n||))\Delta y_n + \delta_y(y_n, z_n, h),$$

which allow us to study the convergence behavior of the numerical solution for the differential component.

Summary

In this second part of the present work, we have shown that the general framework of Part I [13] can also be applied for the study of the series expansions that arise when studying the convergence of one-step integrators for semi-explicit index 2 DAEs. In this sense, we want to stress the unifying character of our approach, since the same basic results (proven in [13] and summarized in Section 3) can be used to derive and deal with series corresponding to one-step integrators for: (a) General systems of ODEs, (b) Hamiltonian systems of ODEs, (c) semi-explicit index 2 DAEs. This approach could also be applied to deal with other classes of series (DA1-series, DA3-series) for structured DAEs in a similar way. We find that unifying approach specially advantageous when studying the composition of the different classes of series.

In Section 7, a new procedure for studying the convergence of one-step methods for index 2 DAEs in Hessemberg form is proposed, which makes use of DA2-series for inconsistent initial values. The basic results on DA2-series needed to apply that approach are obtained using our general framework in Sections 4 and 5, and stated in terms of truncated series in Section 6. Needless to say, these basic results (more specifically Lemmas 14 and 16), could also be obtained with a different procedure.

The results presented in Section 7 allow us to obtain sharp estimates for the behaviour of the global errors, even in the case where the method does not satisfy the algebraic constraints, the errors for the algebraic variables affect the approximations of the differential variables, and non-consistent initial values are used. In particular, the approach adopted for the general case could be used to prove in an alternative way the results presented recently by Aubry *et al.* in [2] on the convergence of implicit RK methods that satisfy (8) but do not satisfy (9) (and therefore the local error for z does not affect the global error of y). In [2], series with partial derivatives of the elementary differentials corresponding to DA2-trees of positive order are employed to study the propagation of the errors. Our new approach, which uses series of elementary differentials, is more general, and allows us to study the convergence of methods that satisfy neither (9) nor (8) in a unified way. Moreover, that approach can in principle give sharper convergence estimates (in particular, for non-consistent initial values). This is due to the fact that the approach of [2] is equivalent to linearizing the DA2-series $\mathcal{S}(\mathbf{c})(y_0, z_0)$ with inconsistent initial values with respect to $g(y_0)$ and $g_y(y_0)f(y_0, z_0)$.

References

[1] M. Arnold, *Half-explicit Runge-Kutta methods with explicit stages for differential-algebraic systems of index 2*, submitted to BIT (1995).

[2] A. Aubry, P. Chartier, *On the structure of errors for Radau IA methods applied to index-2 DAEs*, Appl. Num. Math, 22 (1996), 23-34.

[3] V. Brasey and E. Hairer, *Half-explicit Runge-Kutta methods for differential-algebraic systems of index 2*, Siam J. Numer. Anal., Vol. 30, No. 2 (1993), 538-552.

[4] K. E. Brenan, S. L. Campbell, and L. R. Petzold, *Numerical solution of initial-value problems in differential-algebraic equations*, North-Holland, New-York, (1989).

[5] R. P. K. Chan and P. Chartier, *A composition law for Runge-Kutta methods applied to index-2 differential-algebraic equations*, BIT 36:2 (1996), 229-246.

[6] E. Hairer, Ch. Lubich, M. Roche, *The numerical solution of differential-algebraic systems by Runge-Kutta methods*, Lecture-Notes in Mathematics, Vol 1409, Springer-Verlag (1989).

[7] E. Hairer, S.P. Nørset, G. Wanner: *Solving ordinary differential equations I. Non-stiff problems*, Second Edition, Springer-Verlag (1993).

[8] E. Hairer, G. Wanner, *Solving ordinary differential equations II. Stiff and differential-algebraic problems*, Second Edition, Springer-Verlag, (1996).

[9] L. Jay, *Runge-Kutta methods for differential-algebraic systems of index 3 with applications to Hamiltonian systems*, Ph.D. Thesis, Université de Genève (1994).

[10] L. Jay, *Convergence of a class of Runge-Kutta methods for differential-algebraic systems of index 2*, BIT (1993), 137-150.

[11] A. Murua, *Partitioned half-explicit Runge-Kutta methods for differential-algebraic systems of index 2*, Accepted in Computing (1996).

[12] A. Murua, *Partitioned Runge-Kutta methods for semi-explicit differential-algebraic systems of index 2*, Technical Report EHU-KZAA-IKT-196 (1996).

[13] A. Murua, *Formal series and numerical integrators. Part I: Systems of ODEs and Symplectic integrators*, submitted (1997).