MATHEMATICS OF COMPUTATION
Volume 75, Number 254, Pages 641–654
S 0025-5718(05)01809-0
Article electronically published on December 19, 2005

SPECIALIZED RUNGE-KUTTA METHODS FOR INDEX 2 DIFFERENTIAL-ALGEBRAIC EQUATIONS

LAURENT O. JAY

Abstract. We consider the numerical solution of systems of semi-explicit index 2 differential-algebraic equations (DAEs) by methods based on Runge-Kutta (RK) coefficients. For nonstiffly accurate RK coefficients, such as Gauss and Radau IA coefficients, the standard application of implicit RK methods is generally not superconvergent. To reestablish superconvergence projected RK methods and partitioned RK methods have been proposed. In this paper we propose a simple alternative which does not require any extra projection step and does not use any additional internal stage. Moreover, symmetry of Gauss methods is preserved. The main idea is to replace the satisfaction of the constraints at the internal stages in the standard definition by enforcing specific linear combinations of the constraints at the numerical solution and at the internal stages to vanish. We call these methods *specialized Runge-Kutta methods for index* 2 *DAEs (SRK-DAE*2*).*

1. Introduction

We consider the following class of semi-explicit differential-algebraic equations (DAEs)

$$\frac{d}{dt}y = f(y,z), \tag{1.1a}$$

$$0 = g(y) \tag{1.1b}$$

where $y \in \mathbb{R}^n$ are the *differential* variables, $z \in \mathbb{R}^m$ are the *algebraic* variables, and the functions $f : \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}^n$ and $g : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ are assumed to be sufficiently differentiable. The initial values y_0, z_0 at t_0 are supposed to be given and to be consistent, i.e., to satisfy $g(y_0) = 0$ and $g_y(y_0)f(y_0, z_0) = 0$. We make the usual assumption that in a neighborhood of the solution the square matrix $g_y(y)f_z(y,z)$ exists and is invertible. The system of DAEs (1.1) is thus of index 2 [2, 6, 7].

For such index 2 DAEs convergence results have been obtained for some classes of implicit RK (IRK) methods [5, 6, 7, 8]. In this paper, when we discuss convergence and order of convergence it will only concern the differential variables. The same order of convergence for the algebraic variables can be reached by obtaining the z-component through the equation $g_y(y)f(y,z) = 0$. Stiffly accurate IRK methods, such as Radau IIA, Lobatto IIIA, and Lobatto IIIC methods, preserve their high

Received by the editor January 15, 2004 and, in revised form, January 26, 2005.
2000 *Mathematics Subject Classification.* Primary 65L05, 65L06, 65L80.
Key words and phrases. Differential-algebraic equations, index 2, Runge-Kutta methods.
This material is based upon work supported by the National Science Foundation under Grant No. 9983708.

order of convergence, whereas nonstiffly accurate IRK methods, such as Gauss and Radau IA methods, do not. The main motivation of this paper is to show how the standard application of nonstiffly accurate IRK methods to semi-explicit index 2 DAEs can be simply modified in order to reestablish superconvergence. For this purpose projected RK methods [1] and partitioned RK methods [11] have already been proposed. In this paper we propose a simple alternative to projected and partitioned Runge-Kutta methods which does not require any extra projection step and does not use any additional internal stage. Moreover, symmetry of Gauss methods is preserved. The proposed modification is simple in essence. The main idea is to replace the satisfaction of the constraints at the internal stages in the standard definition by enforcing specific linear combinations of the constraints at the numerical solution and at the internal stages to vanish. We call these methods *specialized Runge-Kutta methods for index* 2 *DAEs (SRK-DAE2).*

In this paper we consider RK coefficients having an invertible RK matrix. For RK coefficients with a singular RK matrix, more particularly for methods based on Lobatto coefficients, we refer the reader to [9]. Results obtained in this paper can be generalized directly to the following class of index 2 implicit DAEs:

$$\begin{aligned} \frac{d}{dt}a(t,y) &= f(t,y,z), \\ 0 &= g(t,y) \end{aligned}$$

where it is assumed that $a_y(t,y)$ and $g_y(t,y)a_y^{-1}(t,y)f_z(t,y,z)$ exist and are invertible in a neighborhood of a solution; see [9].

The paper is organized as follows. In section 2 the standard definition of Runge-Kutta methods for index 2 DAEs is given together with projected and partitioned Runge-Kutta methods. In section 3 specialized Runge-Kutta methods for index 2 DAEs (SRK-DAE2) are introduced, and sufficient conditions for symmetry preservation are given. In section 4 we analyze the existence, uniqueness, local error, and global convergence of the numerical solution for the class of SRK-DAE2 methods considered. Finally, a numerical experiment is given in section 5 to illustrate the superconvergence results.

2. Standard, projected, and partitioned Runge-Kutta methods for index 2 DAEs

The coefficients of a Runge-Kutta method are given by its weights $b := (b_i)_{i=1,\dots,s}$, its nodes $c := (c_i)_{i=1,\dots,s}$, and its RK matrix of coefficients $A := (a_{ij})_{i,j=1,\dots,s}$. The nodes generally satisfy $c_i = \sum_{j=1}^{s} a_{ij}$ $(i = 1,\dots,s)$. In this paper we will assume the RK matrix A to be invertible.

The standard definition of an s-stage Runge-Kutta method applied to semi-explicit index 2 DAEs is described as follows [2, 6, 7]. We consider one step with stepsize h_n starting from y_n at t_n. The numerical solution y_{n+1} approximating the exact solution $y(t)$ at $t_{n+1} = t_n + h_n$ is given by

$$y_{n+1} = y_n + h_n \sum_{i=1}^{s} b_i f(Y_{ni}, Z_{ni}) \tag{2.1a}$$

where the s internal stages Y_{ni}, Z_{ni} $(i = 1, \ldots, s)$ are the solution of the system of nonlinear equations

$$Y_{ni} = y_n + h_n \sum_{j=1}^{s} a_{ij} f(Y_{nj}, Z_{nj}), \qquad i = 1, \ldots, s, \tag{2.1b}$$

$$0 = g(Y_{ni}), \qquad i = 1, \ldots, s. \tag{2.1c}$$

Notice that the value z_n does not enter explicitly the above definition (2.1), but it just indicates to which solution branch of $g_y(y)f(y, z) = 0$ the internal stages Z_{ni} $(i = 1, \ldots, s)$ are close. Therefore it is not necessary to obtain an extremely accurate value z_{n+1} in a step-by-step integration. One possibility is to define explicitly z_{n+1} as

$$z_{n+1} = z_n + \sum_{i=1}^{s} \sum_{j=1}^{s} b_i w_{ij} (Z_{nj} - z_n)$$

where $W := (w_{ij})_{i,j=1}^{s}$ denotes the inverse of the RK matrix A, i.e., $W := A^{-1}$ [6, 7]. The possibility to obtain z_{n+1} implicitly as satisfying $g_y(y_{n+1})f(y_{n+1}, z_{n+1}) = 0$ is generally not taken in practice as it requires the solution of an additional system of nonlinear equations.

From the above standard application of IRK methods it is important to notice that from (2.1c) all internal stages Y_{ni} $(i = 1, \ldots, s)$ satisfy the constraint $g(y) = 0$, whereas the numerical solution y_{n+1} generally does not. However, for stiffly accurate RK methods, i.e., for methods satisfying $a_{si} = b_i$ for $i = 1, \ldots, s$ we have $y_{n+1} = Y_{ns}$. Therefore $g(y_{n+1}) = 0$ is automatically satisfied for such methods since from (2.1c) for $i = s$ we have $g(Y_{ns}) = 0$. Superconvergence of stiffly accurate methods has been demonstrated [6, 7, 8]. For example the s-stage Radau IIA method converges with global order $2s - 1$, and the s-stage Lobatto IIIA and Lobatto IIIC methods converge with global order $2s - 2$. Nonstiffly accurate IRK methods generally have a reduced order of convergence. For example the s-stage Gauss and Radau IA methods converge only with global order s [6, 7]. In our opinion, the loss of superconvergence for nonstiffly accurate RK methods is mainly due to the fact that the constraint $g(y) = 0$ is not satisfied at the numerical solution y_{n+1}.

In this paper we are interested in nonstiffly accurate RK coefficients, i.e., where the relations $a_{si} = b_i$ for $i = 1, \ldots, s$ do not hold, e.g., for Gauss coefficients. The corresponding standard RK methods (2.1) generally fail to be superconvergent for index 2 DAEs. Two modifications preserving superconvergence have already been proposed: *projected Runge-Kutta methods* by Ascher and Petzold [1] and *partitioned Runge-Kutta methods* by Murua [11], both described succinctly below.

In a *projected RK method* [1], see also [7, 10], the numerical solution y_{n+1} of a standard Runge-Kutta method (2.1) is projected onto the constraint $g(y) = 0$ for example as follows:

$$\widetilde{y}_{n+1} = y_{n+1} + f_z(\widetilde{y}_{n+1}, z_{n+1})\mu, \qquad 0 = g(\widetilde{y}_{n+1}).$$

This requires the solution of a system of nonlinear equations. Moreover, a drawback of this approach is the fact that the symmetry property of an IRK method with symmetric coefficients is not preserved. A projection procedure preserving symmetry is described in [4]. Notice that if projection is done as a post-processing procedure, the order of a standard RK method can be increased significantly, but it is generally not optimal [3].

In a *partitioned RK method* [11], the standard definition of RK methods (2.1) is modified by introducing s extra internal stages $\widetilde{Y}_{ni}$ and by replacing the equations (2.1c) for the constraints as follows:

$$\widetilde{Y}_{ni} = y_n + h_n \sum_{j=1}^{s} \widetilde{a}_{ij} f(Y_{nj}, Z_{nj}), \quad 0 = g(\widetilde{Y}_{ni}), \quad i = 1, \dots, s.$$

The additional RK coefficients $\widetilde{a}_{ij}$ satisfy in particular $\widetilde{a}_{si} = b_i$ for $i = 1, \dots, s$, i.e., they are stiffly accurate. To Gauss IRK methods correspond the so-called Gauss-Lobatto partitioned RK methods [11]. A drawback of this approach is the fact that for index 2 DAEs of the type (1.2), the s extra internal stages $\widetilde{Y}_i$ are required, thus doubling for the y-component the number of unknowns and corresponding nonlinear equations.

The main objective of this paper is to present a new and simpler approach to retain superconvergence by a proper modification of the standard definition (2.1) of RK methods for index 2 DAEs. This new approach has some advantages over the two modifications described above. First, unlike projected RK methods, symmetry of Gauss methods is preserved and the classical order of convergence of Radau IA methods is preserved. Secondly, unlike partitioned RK methods, the new approach does not use any additional internal stage. In the next section, we describe these new methods in detail.

3. Specialized Runge-Kutta methods for index 2 DAEs

In this section we describe a simple alternative to projected and partitioned Runge-Kutta methods in order to preserve superconvergence of nonstiffly accurate RK methods. The new methods are called *specialized Runge-Kutta methods for index* 2 *DAEs (SRK-DAE2)*. The difference to standard IRK methods (2.1) is in (2.1c). Ideally we would like to satisfy $g(y_{n+1}) = 0$ and $g(Y_{ni}) = 0$ for $i = 1, \dots, s$. This is generally not possible since there are only s algebraic variables $Z_{n1}, \dots, Z_{ns}$ for $s+1$ sets of constraints. The main idea is as follows. Instead of having y_{n+1} and all internal stages Y_{ni} $(i = 1, \dots, s)$ satisfying the constraint $g(y) = 0$, we enforce s linear combinations of $g(y_{n+1})$ and $g(Y_{ni})$ $(i = 1, \dots, s)$ to be equal to zero.

Definition 3.1. One step of an *s-stage specialized Runge-Kutta method* applied to the system of index 2 DAEs (1.1) (SRK-DAE2) with stepsize h_n starting from y_n, z_n at t_n is given by y_{n+1} where y_{n+1} and the s internal stages Y_{ni}, Z_{ni} $(i = 1, \dots, s)$ are the solution of the system of nonlinear equations

$$Y_{ni} = y_n + h_n \sum_{j=1}^{s} a_{ij} f(Y_{nj}, Z_{nj}), \qquad i = 1, \dots, s, \tag{3.1a}$$

$$y_{n+1} = y_n + h_n \sum_{i=1}^{s} b_i f(Y_{ni}, Z_{ni}), \tag{3.1b}$$

$$0 = \sum_{j=1}^{s} \omega_{ij} g(Y_{nj}) + \omega_{i,s+1} g(y_{n+1}), \qquad i = 1, \dots, s, \tag{3.1c}$$

where the coefficients ω_{ij} are those of a matrix $\Omega \in \mathbb{R}^{s \times (s+1)}$.

Using tensor matrix product notation, the equations (3.1c) can be expressed as

$$(\Omega \otimes I_m) \begin{pmatrix} g(Y_{n1}) \\ \vdots \\ g(Y_{ns}) \\ g(y_{n+1}) \end{pmatrix} = 0.$$

Notice that we can multiply the matrix Ω from the left by any invertible $s \times s$ matrix without changing the definition of SRK-DAE2 methods. The major difficulty is to find matrices $\Omega \in \mathbb{R}^{s\times(s+1)}$ such that the order of the SRK-DAE2 method is as high as possible and also such that other properties, such as symmetry, are preserved. To ensure existence and uniqueness of the SRK-DAE2 solution we will assume the square matrix $\Omega\alpha \in \mathbb{R}^{s\times s}$ to be invertible, see Theorem 4.1 below, where the matrix α is defined as

$$\alpha := \begin{pmatrix} A \\ b^T \end{pmatrix} \in \mathbb{R}^{(s+1)\times s}.$$

Moreover, we will assume $b_i \neq 0$, $c_i \neq c_j$ for $i \neq j$, and the matrix A to be invertible. We use the following notation: $e_i \in \mathbb{R}^s$ denotes the i-th unit vector $(i = 1, \ldots, s)$, $e := (1, 1, \ldots, 1)^T \in \mathbb{R}^s$, $0_s := (0, 0, \ldots, 0)^T \in \mathbb{R}^s$, $e_{s+1} := (0, 0, \ldots, 0, 1)^T \in \mathbb{R}^{s+1}$, $I_s := \mathrm{diag}(1, 1, \ldots, 1) \in \mathbb{R}^{s\times s}$, $B := \mathrm{diag}(b_1, b_2, \ldots, b_s) \in \mathbb{R}^{s\times s}$, $C := \mathrm{diag}(c_1, c_2, \ldots, c_s) \in \mathbb{R}^{s\times s}$, and $c^k := C^k e = (c_1^k, c_2^k, \ldots, c_s^k)^T \in \mathbb{R}^s$. Runge-Kutta coefficients generally satisfy some *simplifying assumptions*, expressed here in vector notation

$$\begin{aligned} B(p) \quad &: \quad b^T c^{k-1} = \frac{1}{k} \quad \text{for } k = 1, \ldots, p, \\ C(q) \quad &: \quad Ac^{k-1} = \frac{1}{k} c^k \quad \text{for } k = 1, \ldots, q, \\ D(r) \quad &: \quad b^T C^{k-1} A = \frac{1}{k}(b^T - b^T C^k) \quad \text{for } k = 1, \ldots, r, \end{aligned}$$

for some nonnegative integer values p, q, and r. For given Runge-Kutta coefficients satisfying the simplifying assumption $D(r)$, we define the nonnegative integer value

$$\rho := \min(r, s-1) \leq s - 1.$$

To deal with the constraints (1.1b), for SRK-DAE2 methods we will consider the relations

$$0 = g(y_{n+1}), \qquad 0 = \sum_{i=1}^{s} b_i c_i^{k-1} g(Y_{ni}), \qquad k = 1, \ldots, \rho, \tag{3.2}$$

instead of (2.1c). For $\rho = 0$ we only have the relation $0 = g(y_{n+1})$. An important point is that the second set of relations in (3.2) can be rewritten equivalently as

$$0 = h_n \sum_{i=1}^{s} b_i (c_i h_n)^{k-1} g(Y_{ni}), \qquad k = 1, \ldots, \rho,$$

which can be interpreted as a discretization of

$$0 = \int_{t_n}^{t_n+h_n} (t - t_n)^{k-1} \cdot g(y) dt.$$

There is a total of $\rho+1$ equations in (3.2). This corresponds to having a matrix Ω in (3.1c) with $\rho+1$ rows equal to the first $\rho+1$ rows of the following matrix:

$$\widetilde{\Omega}_0 := \begin{pmatrix} 0_s^T & 1 \\ b^T & 0 \\ \vdots & \vdots \\ b^T C^{s-2} & 0 \end{pmatrix} \in \mathbb{R}^{s\times(s+1)}. \tag{3.3}$$

When $r \leq s-2$, we have $\rho = r \leq s-2$; thus there are at most $s-1$ equations in (3.2). In this situation there is some additional freedom in choosing the remaining $s-\rho-1$ equations. A natural choice is to take (3.2) for $k = 1, \ldots, s-1$, i.e., $\Omega = \widetilde{\Omega}_0$, but this is not a requirement. The choice of $\widetilde{\Omega}_0$ is equivalent to

$$\widehat{\Omega}_0 := \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & -1 & & 0 \\ \vdots & & \ddots & \vdots \\ 1 & 0 & \cdots & -(s-1) \end{pmatrix} \widetilde{\Omega}_0 = \begin{pmatrix} 0_s^T & 1 \\ -b^T & 1 \\ \vdots & \vdots \\ -(s-1)b^T C^{s-2} & 1 \end{pmatrix} \in \mathbb{R}^{s\times(s+1)}.$$

For the local error analysis of SRK-DAE2 methods, see section 4; without loss of generality Ω can be assumed to satisfy $\Omega\alpha = I_s$. We can consider for example

$$\Omega_0 := M^{-1}\widehat{\Omega}_0 = \begin{pmatrix} A^{-1} & 0_s \end{pmatrix} + M^{-1} e \begin{pmatrix} -b^T A^{-1} & 1 \end{pmatrix} \in \mathbb{R}^{s\times(s+1)} \tag{3.4}$$

where M, assumed to be invertible, is given by

$$M := \widehat{\Omega}_0 \alpha = \begin{pmatrix} b^T \\ b^T - b^T A \\ \vdots \\ b^T - (s-1)b^T C^{s-2} A \end{pmatrix}. \tag{3.5}$$

When the simplifying assumption $D(s-1)$ is satisfied this matrix M reduces to

$$M = VB = \begin{pmatrix} b^T \\ b^T C \\ \vdots \\ b^T C^{s-1} \end{pmatrix} \tag{3.6}$$

where

$$V := \begin{pmatrix} e^T \\ c^T \\ \vdots \\ {c^{s-1}}^T \end{pmatrix} = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ c_1 & c_2 & \ldots & c_s \\ \vdots & \vdots & \ldots & \vdots \\ c_1^{s-1} & c_2^{s-1} & \ldots & c_s^{s-1} \end{pmatrix} \in \mathbb{R}^{s\times s}.$$

From our assumptions on the weight vector b and the node vector c this matrix $M = VB$ is invertible since V is a Vandermonde matrix.

More generally we will consider

$$\Omega := \begin{pmatrix} A^{-1} & 0_s \end{pmatrix} + N^{-1} u \begin{pmatrix} -b^T A^{-1} & 1 \end{pmatrix} \in \mathbb{R}^{s\times(s+1)} \tag{3.7}$$

where $u \in \mathbb{R}^s$ satisfies $u_i = 1$ for $i = 1, \ldots, \rho + 1$ and $N \in \mathbb{R}^{s \times s}$ is any invertible matrix of the form

$$N := \begin{pmatrix} b^T \\ b^T C \\ \vdots \\ b^T C^\rho \\ * \end{pmatrix} = \begin{pmatrix} b^T \\ b^T - b^T A \\ \vdots \\ b^T - \rho b^T C^{\rho-1} A \\ * \end{pmatrix}. \tag{3.8}$$

The choice (3.4) corresponds to $u = e$ and $N = M$ of (3.5). Using the matrix Ω directly as given in (3.7) is easier when analyzing the order of the method, see Theorem 4.4, whereas using the equivalent form, see (3.3),

$$\widetilde{\Omega} = \begin{pmatrix} 0_s^T & 1 \\ b^T & 0 \\ \vdots & \vdots \\ b^T C^\rho & 0 \\ * & * \end{pmatrix} \in \mathbb{R}^{s \times (s+1)} \tag{3.9}$$

is easier when analyzing symmetry, see Theorem 3.3, and is also more natural to interpret and to write down; see (3.2).

Theorem 3.2. *Under the notation and assumptions of section* 3 *we have the following relations for the matrix* Ω *in* (3.7)*:*

$$\Omega\alpha = I_s, \tag{3.10}$$

$$b^T \Omega = e_{s+1}^T. \tag{3.11}$$

- *When in addition the simplifying assumptions* $B(p)$ *and* $C(q)$ *are satisfied, then for* $1 \le k \le \min(q, p)$ *we have*

$$\Omega \begin{pmatrix} c^k \\ 1 \end{pmatrix} = k c^{k-1}. \tag{3.12}$$

  Moreover, if the simplifying assumption $D(s-1)$ *is satisfied, then this equality holds for* $1 \le k \le \max(\min(q, p), p - s + 1)$.
- *Alternatively, when in addition the simplifying assumption* $D(r)$ *is satisfied, then for* $1 \le k \le \rho = \min(r, s-1)$ *we have*

$$b^T C^k \Omega = b^T \Omega - k \begin{pmatrix} b^T C^{k-1} & 0 \end{pmatrix}. \tag{3.13}$$

Proof. The first relation (3.10) is obtained by direct calculation as follows:

$$\Omega \begin{pmatrix} A \\ b^T \end{pmatrix} = A^{-1}A + N^{-1}u(-b^T A^{-1} A + b^T) = I_s.$$

Since the first row of N is b^T we have $b^T N^{-1} = e_1^T$; thus $b^T N^{-1} u = 1$. Hence, we obtain the second relation (3.11) as follows:

$$b^T \Omega = \begin{pmatrix} b^T A^{-1} & 0 \end{pmatrix} + b^T N^{-1} u \begin{pmatrix} -b^T A^{-1} & 1 \end{pmatrix} = e_{s+1}^T.$$

For the third relation (3.12) we have

$$\Omega \begin{pmatrix} c^k \\ 1 \end{pmatrix} = A^{-1} c^k + N^{-1} u(-b^T A^{-1} c^k + 1).$$

Using successively the relations $A^{-1}c^k = k c^{k-1}$ for $k = 1, \ldots, q$ coming from $C(q)$ and $b^T c^{k-1} = 1/k$ for $k = 1, \ldots, p$ from $B(p)$ gives the desired result for

$1 \le k \le \min(q,p)$. When in addition $D(s-1)$ is satisfied the matrix Ω can be expressed as $\Omega = \Omega_0$ of (3.4) with $M = VB$ of (3.6). The equality (3.12) is thus equivalent to

$$\widehat{\Omega}_0 \begin{pmatrix} c^k \\ 1 \end{pmatrix} = kVBc^{k-1}.$$

For $1 \le k \le p-s+1$ both sides simplify to

$$k \begin{pmatrix} 1/k \\ 1/(k+1) \\ \vdots \\ 1/(k+s-1) \end{pmatrix}.$$

For the fourth relation (3.13) we have

$$b^T C^k \Omega = \begin{pmatrix} b^T C^k A^{-1} & 0 \end{pmatrix} + b^T C^k N^{-1} u \begin{pmatrix} -b^T A^{-1} & 1 \end{pmatrix}.$$

The result follows from the relations $b^T C^k A^{-1} = b^T A^{-1} - k b^T C^{k-1}$ for $k = 1, \dots, r$ from $D(r)$, from $b^T C^k N^{-1} = e_{k+1}^T$ for $k = 1, \dots, \rho$ coming directly from (3.8), and finally from (3.11). □

In the following theorem we give sufficient conditions for a SRK-DAE method to preserve symmetry.

Theorem 3.3. *Assume symmetric RK coefficients, i.e.,*

$$a_{s+1-i,s+1-j} + a_{ij} = b_j = b_{s+1-j} \quad \text{for} \quad i,j = 1, \dots, s, \tag{3.14a}$$

$$c_i = 1 - c_{s+1-i} \quad \text{for} \quad i = 1, \dots, s. \tag{3.14b}$$

Then the corresponding SRK-DAE2 method (3.1) *applied to the index* 2 *DAEs* (1.1) *with matrix* Ω *given by* (3.3)*, or equivalently* (3.4)*, is symmetric.*

Proof. We can assume y_n to satisfy $g(y_n) = 0$ since the constraint $g(y) = 0$ is satisfied at each timestep. Exchanging $h_n \leftrightarrow -h_n$, $y_n \leftrightarrow y_{n+1}$, $t_n \leftrightarrow t_n + h_n$, we obtain the adjoint SRK-DAE2 method with coefficients $b_i^* = b_{s+1-i}$, $c_i^* = 1 - c_{s+1-i}$, and $a_{ij}^* = b_{s+1-j} - a_{s+1-i,s+1-j}$ in (3.1a)-(3.1b); i.e.,

$$Y_{ni}^* = y_n + h_n \sum_{j=1}^{s} a_{ij}^* f(Y_{nj}^*, Z_{nj}^*), \qquad i = 1, \dots, s, \tag{3.15a}$$

$$y_{n+1}^* = y_n + h_n \sum_{i=1}^{s} b_i^* f(Y_{ni}^*, Z_{ni}^*). \tag{3.15b}$$

From the assumption $g(y_n) = 0$ we obtain $g(y_{n+1}^*) = 0$ for the adjoint method. For the constraints we have

$$0 = \sum_{i=1}^{s} b_i c_i^{k-1} g(Y_{n,s+1-i}^*), \qquad k = 1, \dots, s-1,$$

which can be reexpressed thanks to (3.14) as

$$0 = \sum_{i=1}^{s} b_{s+1-i} c_{s+1-i}^{k-1} g(Y_{ni}^*) = \sum_{i=1}^{s} b_i (1-c_i)^{k-1} g(Y_{ni}^*), \qquad k = 1, \dots, s-1.$$

Developing $(1-c_i)^{k-1}$ we obtain by induction on k,

$$0 = \sum_{i=1}^{s} b_i c_i^{k-1} g(Y_{ni}^*), \qquad k = 1, \ldots, s-1. \tag{3.15c}$$

From (3.14) the equations (3.15a) and (3.15b) are of the same form as (3.1a) and (3.1b) respectively since under the assumptions (3.14) we have $b_i^* = b_i$, $c_i^* = c_i$, and $a_{ij}^* = a_{ij}$. The equations (3.15c) for the constraints are also of the same form as (3.2) for $k = 1, \ldots, s-1$. Therefore, we must have $Y_{ni}^* = Y_{ni}$ and $Z_{ni}^* = Z_{ni}$ for $i = 1, \ldots, s$, and $y_{n+1}^* = y_{n+1}$. Symmetry has thus been proved. □

As a direct corollary to this theorem we have that for Gauss coefficients the corresponding Gauss SRK-DAE2 methods are symmetric. For $s = 1$ the Gauss SRK-DAE2 method of order 2 reads as follows:

$$Y_{n1} = y_n + \frac{h_n}{2} f(Y_{n1}, Z_{n1}), \quad y_{n+1} = y_n + h_n f(Y_{n1}, Z_{n1}), \quad 0 = g(y_{n+1}),$$

which can be reexpressed as a system of implicit equations for y_{n+1} and Z_{n1} as

$$y_{n+1} = y_n + h_n f\left(\frac{1}{2}(y_n + y_{n+1}), Z_{n1}\right), \quad 0 = g(y_{n+1}).$$

With $0 = g(Y_{n1})$ instead of $0 = g(y_{n+1})$ to treat the constraint (1.1b), the standard Gauss IRK method has only order 1. Using the form (3.2), for $s = 2$ the Gauss SRK-DAE2 method of order 4 reads as follows:

$$\begin{aligned}
Y_{n1} &= y_n + h_n \left(\frac{1}{4} f(Y_{n1}, Z_{n1}) + \left(\frac{1}{4} - \frac{\sqrt{3}}{6}\right) f(Y_{n2}, Z_{n2})\right), \\
Y_{n2} &= y_n + h_n \left(\left(\frac{1}{4} + \frac{\sqrt{3}}{6}\right) f(Y_{n1}, Z_{n1}) + \frac{1}{4} f(Y_{n2}, Z_{n2})\right), \\
y_{n+1} &= y_n + h_n \left(\frac{1}{2} f(Y_{n1}, Z_{n1}) + \frac{1}{2} f(Y_{n2}, Z_{n2})\right), \\
0 &= \frac{1}{2} g(Y_{n1}) + \frac{1}{2} g(Y_{n2}), \\
0 &= g(y_{n+1}).
\end{aligned}$$

With $0 = g(Y_{n1})$ and $0 = g(Y_{n2})$ instead of $0 = \frac{1}{2} g(Y_{n1}) + \frac{1}{2} g(Y_{n2})$ and $0 = g(y_{n+1})$ to treat the constraint (1.1b), the standard Gauss IRK method has only order 2.

For a certain class of RK coefficients, including Lobatto coefficients, we have proposed in [9] a similar modification for semi-explicit index 2 DAEs. Instead of a matrix Ω as defined in (3.7), a different matrix Q is used in [9]. It can be shown that for Lobatto coefficients this matrix Q is equivalent to the matrix Ω_0 through the relation $Q = A_3 \Omega_0$ when taking the RK matrix $A = A_3$ in the definition (3.4) of Ω_0 (here we have used the notation of [9], for Lobatto methods the RK matrix A_3 corresponds to Lobatto IIIC coefficients).

4. Analysis of SRK-DAE2 methods

In this section, to simplify the notation we will remove the index n in the definition (3.1) or replace it by 0 when necessary. Existence and uniqueness of the system of nonlinear equations of SRK-DAE2 methods (3.1) is shown in the following theorem.

Theorem 4.1. *Suppose that $y_0 = y_0(h), z_0 = z_0(h)$ satisfy*

$$g(y_0) = O(h^2), \quad g_y(y_0) f(y_0, z_0) = O(h), \tag{4.1}$$

and that $g_y(y) f_z(y, z)$ exists and is invertible in a neighborhood of (y_0, z_0). Assume that for the SRK-DAE2 method (3.1) the matrix $\Omega\alpha \in \mathbb{R}^{s\times s}$ is invertible. Then for $h \le h_0$ there exists a locally unique SRK-DAE2 solution that satisfies

$$Y_i - y_0 = O(h) \quad i = 1, \dots, s, \quad y_1 - y_0 = O(h), \quad Z_i - z_0 = O(h) \quad i = 1, \dots, s.$$

Proof. The proof of this theorem can be done by application of the implicit function theorem, as in the proof of [7, Theorem VII.4.1]. To simplify the notation we denote $Y_{s+1} := y_1$. Dividing the right-hand side of (3.1c) by h we obtain the equations

$$0 = \frac{1}{h} \sum_{j=1}^{s+1} \omega_{ij} g(Y_j), \quad i = 1, \dots, s. \tag{4.2}$$

Inserting the relations

$$Y_j - y_0 = h \sum_{k=1}^{s} \alpha_{jk} f(Y_k, Z_k), \quad j = 1, \dots, s+1,$$

coming from (3.1a)-(3.1b), in the development

$$g(Y_j) = g(y_0) + \int_0^1 g_y(y_0 + \tau(Y_j - y_0)) d\tau \cdot (Y_j - y_0)\,,$$

the equations (4.2) can be reexpressed as

$$0 = \frac{1}{h} \sum_{j=1}^{s+1} \omega_{ij} g(y_0) + \sum_{j=1}^{s+1} \sum_{k=1}^{s} \omega_{ij} \alpha_{jk} \left(\int_0^1 g_y(y_0 + \tau(Y_j - y_0)) d\tau \cdot f(Y_k, Z_k) \right) \tag{4.3}$$

for $i = 1, \dots, s$. By the assumptions (4.1), at $h = 0$ these equations (4.3) are satisfied since we have $Y_i(0) = y_0$ for $i = 1, \dots, s+1$ and $Z_i(0) = z_0$ for $i = 1, \dots, s$.

Using tensor matrix product notation the Jacobian of (3.1a), (3.1b), and (4.3) with respect to $Y_1, \dots, Y_s, Y_{s+1}, Z_1, \dots, Z_s$ is equal at $h = 0$ to

$$\begin{pmatrix} I_{s+1} \otimes I_n & O \\ * & \Omega\alpha \otimes g_y(y_0) f_z(y_0, z_0) \end{pmatrix},$$

and is therefore invertible. □

The goal now is to obtain a local error estimate for the differential variables y_1 compared to the exact solution $y(t)$ at $t_0 + h$ passing through consistent initial values y_0, z_0 at t_0. In this paper we will not define once again the whole tree theory for semi-explicit index 2 DAEs which can be found for example in [6, 7]. Definitions of trees $t \in LDAT2$ and related quantities $\rho(t), \gamma(t)$, etc., are as in [6, Section 5] and [7, Section VII.5]. We only mention the main results, ideas, and differences.

Without loss of generality we can assume $\Omega\alpha = I_s$; see Theorem 4.1 and (3.10). Defining $k_i(h) := h f(Y_i, Z_i)$ for $i = 1, \dots, s$ and using the notation $Y_{s+1} := y_1$ we can rewrite (3.1a)-(3.1b) as

$$Y_i = y_0 + \sum_{j=1}^{s} \alpha_{ij} k_j, \quad i = 1, \dots, s+1.$$

First we give some expressions for the derivatives of k_i and Z_i which are similar to those given in [6, Theorem 5.7] and [7, Theorem VII.5.6].

Theorem 4.2. *Assuming $\Omega\alpha = I_s$, for $i = 1, \dots, s$ we have*

$$\begin{aligned} k_i^{(q)}(0) &= \sum_{\substack{t \in LDAT2_y \\ \rho(t)=q}} \gamma(t)\Phi_i(t)F(t)(y_0, z_0), \\ Z_i^{(q)}(0) &= \sum_{\substack{u \in LDAT2_z \\ \rho(u)=q}} \gamma(u)\Phi_i(u)F(u)(y_0, z_0), \end{aligned}$$

where the coefficients $\Phi_i(t)$ and $\Phi_i(u)$ are given recursively by $\Phi_i(\tau) = 1$ and

$$\begin{aligned} \Phi_i(t) &= \sum_{\mu_1=1}^{s} \cdots \sum_{\mu_m=1}^{s} \alpha_{i\mu_1} \cdots \alpha_{i\mu_m} \Phi_{\mu_1}(t_1) \cdots \Phi_{\mu_m}(t_m)\Phi_i(u_1) \cdots \Phi_i(u_n) \\ &\qquad \text{if } t = [t_1, \dots, t_m, u_1, \dots, u_n]_y, \\ \Phi_i(u) &= \sum_{j=1}^{s+1} \sum_{\mu_1=1}^{s} \cdots \sum_{\mu_m=1}^{s} \omega_{ij}\alpha_{j\mu_1} \cdots \alpha_{j\mu_m} \Phi_{\mu_1}(t_1) \cdots \Phi_{\mu_m}(t_m) \\ &\qquad \text{if } u = [t_1, \dots, t_m]_z. \end{aligned} \tag{4.4}$$

A proof of this theorem can be obtained along the lines of [6, Theorem 5.7] and [7, Theorem VII.5.6]; it is therefore omitted. A direct consequence is:

Theorem 4.3. *The numerical solution $y_1(h)$ of (3.1) satisfies*

$$y_1^{(q)}(0) = \sum_{\substack{t \in LDAT2_y \\ \rho(t)=q}} \gamma(t) \sum_{i=1}^{s} b_i \Phi_i(t) F(t)(y_0, z_0).$$

For the local error we obtain:

Theorem 4.4. *Consider the semi-explicit system of index 2 DAEs (1.1) with consistent initial values (y_0, z_0) at t_0 and such that $g_y(y)f_z(y, z)$ exists and is invertible in a neighborhood of the exact solution. Consider an SRK-DAE2 method (3.1) having an invertible RK matrix A, with RK coefficients satisfying the simplifying assumptions $B(p)$, $C(q)$, and $D(r)$, and with matrix Ω as in (3.7) or equivalently (3.9). Then we have*

$$y_1 - y(t_0 + h) = O(h^{\mu+1})$$

where $\mu := \min(p, 2\sigma, 2q+2, q+\rho+1)$ with $\rho := \min(r, s-1)$, and $\sigma := q$ if $r \le s-2$ or $\sigma := \max(q, p-s+1)$ if $r \ge s-1$. If the function $f(y, z)$ of (1.1a) is linear in z, then the values 2σ and $2q+2$ in μ can be changed respectively to $2\sigma+1$ and $2q+3$.

Proof. A proof of this theorem can be obtained along the lines of [6, Theorem 5.9] and [7, Theorem VII.5.10]; details are therefore omitted. We only mention some major differences.

Denoting $W := A^{-1}$, the simplifying assumptions $C(q)$ and $D(r)$ can be expressed in terms of the inverse W as follows:

$$\begin{aligned} IC(q) &: \quad Wc^k = kc^{k-1} \quad \text{for } k = 1, \dots, q, \\ ID(r) &: \quad b^T C^k W = b^T W - k b^T C^{k-1} \quad \text{for } k = 1, \dots, r. \end{aligned}$$

The stiff accuracy assumption $a_{si} = b_i$ $(i = 1, \dots, s)$ would lead to $b^T W = e_s^T$. The relations $IC(q)$, $ID(r)$, and $b^T W = e_s^T$ are crucial in the proofs of [6, Theorem 5.9] and [7, Theorem VII.5.10]. In our situation, the numerical solution $y_1 = Y_{s+1}$ also plays a role in order to obtain the expansions (4.4). We cannot apply directly the simplifying assumptions $IC(q)$, $ID(r)$, and $b^T W = e_s^T$, but similar ones where the matrix Ω replaces the matrix W and the matrix α replaces the matrix A. To $WA = I_s$ corresponds (3.10), to $IC(q)$ corresponds (3.12), to $ID(r)$ corresponds (3.13), and to $b^T W = e_s^T$ corresponds (3.11). As in the proof of [6, Theorem 5.9] and [7, Theorem VII.5.10] we make repeated application of the simplifying assumptions $B(p)$, $C(q)$, and $D(r)$, and also of the relations (3.11)-(3.12)-(3.13) to obtain the desired result. The values of σ and ρ in μ come as consequences of (3.12) and (3.13) respectively. □

Following for example [7, Theorem VII.4.5], global convergence of SRK-DAE2 methods can be obtained:

Theorem 4.5. *Consider the semi-explicit system of index* 2 *DAEs* (1.1) *with consistent initial values* (y_0, z_0) *at* t_0 *and such that* $g_y(y) f_z(y, z)$ *exists and is invertible in a neighborhood of the exact solution. Consider a SRK-DAE2 method* (3.1) *with matrix* Ω *such that* $\Omega\alpha \in \mathbb{R}^{s \times s}$ *is invertible, and with local error order* μ*, i.e.,* $y_1 - y(t_0 + h) = O(h^{\mu+1})$*. Then the SRK-DAE2 method is convergent of order* μ*, i.e., its global error satisfies*

$$y_n - y(t_n) = O(h^\mu)$$

for $|t_n - t_0| \leq \text{Const}$ *and* $h := \max(|h_1|, \dots, |h_n|)$*.*

A direct consequence of Theorem 4.5 is the superconvergence of Gauss and Radau IA SRK-DAE2 methods:

Corollary 4.6. *Consider the semi-explicit system of index* 2 *DAEs* (1.1) *with consistent initial values* (y_0, z_0) *at* t_0 *and such that* $g_y(y) f_z(y, z)$ *exists and is invertible in a neighborhood of the exact solution. For* $|t_n - t_0| \leq \text{Const}$ *and* $h := \max(|h_1|, \dots, |h_n|)$*, the global error of the* s*-stage symmetric Gauss SRK-DAE2 method with matrix* Ω *as in* (3.3)*, or equivalently* (3.4)*, satisfies*

$$y_n - y(t_n) = O(h^{2s}).$$

For the s*-stage Radau IA SRK-DAE2 method with matrix* Ω *as in* (3.3)*, or equivalently* (3.4)*, the global error satisfies*

$$y_n - y(t_n) = O(h^{2s-1}).$$

Proof. The s-stage Gauss coefficients satisfy the simplifying assumptions $B(2s)$, $C(s)$, and $D(s)$. These coefficients are also known to be symmetric [7]. From Theorem 3.3, the s-stage Gauss SRK-DAE2 method is therefore symmetric. The s-stage Radau IA coefficients satisfy the simplifying assumptions $B(2s-1)$, $C(s-1)$, and $D(s)$. The global superconvergence estimates are a direct consequence of Theorem 4.4 and Theorem 4.5. □

Notice that with our approach the superconvergence of Radau IA SRK-DAE2 methods is one order higher than for projected Radau IA methods [7, 10].

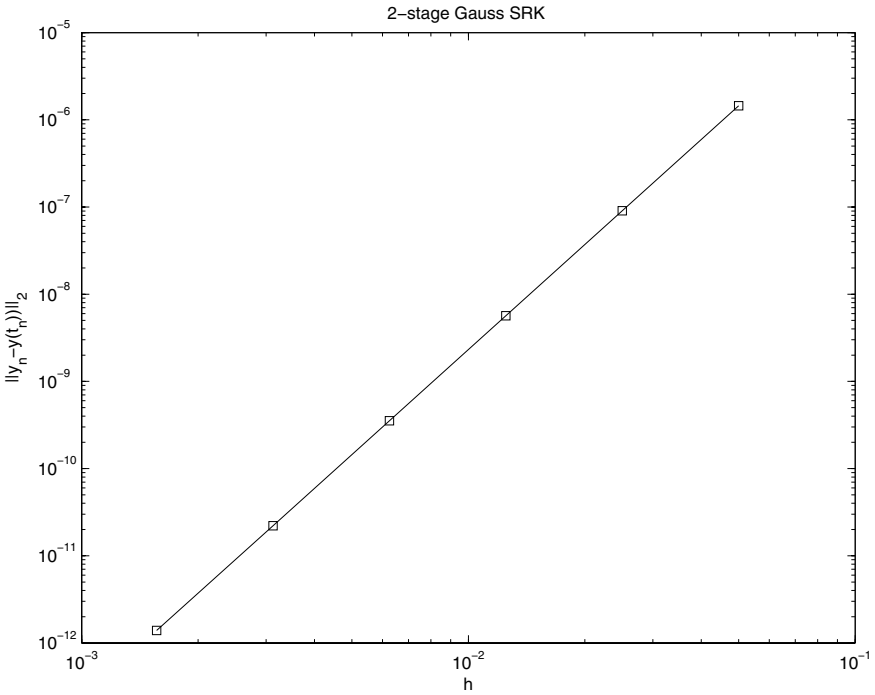

FIGURE 1. Global error of the 2-stage Gauss SRK-DAE2 method of order 4 applied with constant stepsizes to the test problem (5.1).

5. A numerical experiment

To illustrate the superconvergence results, we have applied the 2-stage Gauss SRK-DAE2 method with constant stepsize h to the following semi-explicit system of index 2 DAEs:

$$\begin{pmatrix} y_1' \\ y_2' \end{pmatrix} = \begin{pmatrix} y_1 y_2^2 z_1^2 \\ y_1^2 y_2^2 - 3 y_2^2 z_1 \end{pmatrix}, \tag{5.1a}$$

$$0 = y_1^2 y_2 - 1. \tag{5.1b}$$

For the initial conditions $y_1(0) = 1, y_2(0) = 1$ at $t_0 = 0$ the exact solution to this test problem is given by

$$y_1(t) = e^t, \quad y_2(t) = e^{-2t}, \quad z_1(t) = e^{2t}.$$

In Figure 1 we have plotted the global errors at $t_n = 1$ with respect to different stepsizes h. Logarithmic scales have been used so that a curve appears as a straight line of slope k whenever the leading term of the global error is of order k, i.e., when $\|y_n - y(t_n)\| = O(h^k)$. For the 2-stage Gauss SRK-DAE2 method of order 4 we observe a straight line of slope 4, thus confirming the order of convergence predicted by Corollary 4.6.

References

1. U. Ascher and L. R. Petzold, *Projected implicit Runge-Kutta methods for differential-algebraic equations*, SIAM J. Numer. Anal. **28** (1991), 1097–1120. MR1111456 (92f:65082)
2. K. E. Brenan, S. L. Campbell, and L. R. Petzold, *Numerical solution of initial-value problems in differential-algebraic equations*, SIAM Classics in Appl. Math., SIAM, Philadelphia, Second Edition, 1996. MR1363258 (96h:65083)
3. R. P. K. Chan, P. Chartier, and A. Murua, *Post-projected Runge-Kutta methods for index-2 differential-algebraic equations*, Appl. Numer. Math. **42** (2002), 77–94. MR1921330 (2003f:65124)
4. ———, *Reversible methods of Runge-Kutta type for index-2 differential-algebraic equations*, Numer. Math. **97** (2004), 427–440. MR2059464 (2005b:65088)
5. E. Hairer and L. O. Jay, *Implicit Runge-Kutta methods for higher index differential-algebraic systems*, WSSIAA Contributions in numerical mathematics, 2 (1993), 213–224. MR1299761 (95g:65099)
6. E. Hairer, Ch. Lubich, and M. Roche, *The numerical solution of differential-algebraic systems by Runge-Kutta methods*, Lect. Notes in Math., vol. 1409, Springer, Berlin, 1989. MR1027594 (91a:65178)
7. E. Hairer and G. Wanner, *Solving ordinary differential equations ii. stiff and differential-algebraic problems*, Comput. Math., vol. 14, Springer, Berlin, Second Revised Edition, 1996. MR1439506 (97m:65007)
8. L. O. Jay, *Convergence of a class of Runge-Kutta methods for differential-algebraic systems of index* 2, BIT **33** (1993), 137–150. MR1326008 (96a:65106)
9. ———, *Solution of index* 2 *implicit differential-algebraic equations by Lobatto Runge-Kutta methods*, BIT **43** (2003), 91–104. MR1981642 (2004g:65100)
10. Ch. Lubich, *On projected Runge-Kutta methods for differential-algebraic equations*, BIT **31** (1991), 545–550. MR1127491 (92h:65109)
11. A. Murua, *Partitioned Runge-Kutta methods for semi-explicit differential-algebraic systems of index* 2, Tech. Report EHU-KZAA-IKT-196, Univ. of the Basque country, 1996.

Department of Mathematics, 14 MacLean Hall, The University of Iowa, Iowa City, Iowa 52242-1419

E-mail address: ljay@math.uiowa.edu

E-mail address: na.ljay@na-net.ornl.gov

URL: http://www.math.uiowa.edu/~ljay/