BIT Numerical Mathematics **43:** 93–106, 2003.

SOLUTION OF INDEX 2 IMPLICIT DIFFERENTIAL-ALGEBRAIC EQUATIONS BY LOBATTO RUNGE-KUTTA METHODS *

L. O. JAY †

Department of Mathematics, The University of Iowa, 14 MacLean Hall, Iowa City, IA 52242-1419, USA. email: ljay@math.uiowa.edu and na.ljay@na-net.ornl.gov

Abstract.

We consider the numerical solution of systems of index 2 implicit differential-algebraic equations (DAEs) by a class of super partitioned additive Runge–Kutta (SPARK) methods. The families of Lobatto IIIA-B-C-C*-D methods are included. We show super-convergence of optimal order $2s-2$ for the s-stage Lobatto families provided the constraints are treated in a particular way which strongly relies on specific properties of the SPARK coefficients. Moreover, reversibility properties of the flow can still be preserved provided certain SPARK coefficients are symmetric.

AMS subject classification: 65L05, 65L06, 65L80, 70F25, 70H45.

Key words: Differential-algebraic equations, index 2, Lobatto coefficients, mechanical systems, implicit Runge–Kutta methods.

1 Introduction.

We consider the following class of systems of implicit differential-algebraic equations (DAEs)

$$a'(t,y) = f(t,y,z), \tag{1.1a}$$

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

where $t \in \mathbb{R}$ is the independent variable, $y \in \mathbb{R}^n$ is the *differential* variable, $z \in \mathbb{R}^m$ is the *algebraic* variable, and the functions $a : \mathbb{R} \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$, $f : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}^n$, and $g : \mathbb{R} \times \mathbb{R}^n \longrightarrow \mathbb{R}^m$ are assumed to be sufficiently differentiable. In a neighborhood of a solution we assume that $a_y(t,y)$ and $g_y(t,y)a_y^{-1}(t,y)f_z(t,y,z)$ exist and are invertible. As shown in section 2, the system of DAEs (1.1) is therefore of index 2 and when $a(t,y) = y$ we obtain Hessenberg DAEs of index 2 [2, 5, 7]. The DAEs (1.1) include the formulation of mechanical systems with mixed holonomic, nonholonomic, scleronomic, and rheonomic constraints provided holonomic constraints are differentiated once explicitly with respect to t [8, 16, 17, 18]. The algebraic variable z corresponds to

*Received February 2002. Revised August 2002. Communicated by Christian Lubich.

†This material is based upon work supported by the National Science Foundation under Grant No. 9983708.

Lagrange multipliers when the DAEs can be derived from a variational principle [8, 16].

Solutions to these DAEs (1.1) can be approximated numerically by applying a class of super partitioned additive Runge–Kutta (SPARK) methods, such as the combination of Lobatto IIIA-B-C-C*-D methods [10]. SPARK methods can take advantage of splitting the differential equations into different terms and of partitioning the variables into different classes. Several properties of the SPARK coefficients, satisfied by the Lobatto families, permit to treat the constraints and the algebraic variables properly in order to retain super-convergence properties.

For Hessenberg DAEs of index 2 convergence results have been obtained for some classes of implicit RK methods [5, 7, 9, 15]. Methods preserving their super-convergence order are either stiffly accurate or involve an extra projection step. In this paper we show in particular that neither is necessary for Lobatto SPARK methods. Super-convergence can be obtained even for non-stiffly accurate Lobatto RK methods such as the Lobatto IIIB, Lobatto IIIC*, and Lobatto IIID methods. This is possible provided the constraints are treated in a particular way which only requires a simple linear combination of the constraints evaluated at the internal stages and at the numerical solution as succinctly mentioned in [12] for implementation purposes.

The paper is organized as follows. In section 2, the class of implicit DAEs considered in this article is presented in more details. In section 3 the definition of SPARK methods applied to these DAEs is given. Some properties of the SPARK coefficients are given which are crucial to treat the constraints appropriately in order to obtain super-convergence results, and also in order to preserve reversibility properties of the flow. In section 4 we analyze the existence, uniqueness, local error, and global convergence of the numerical solution for the class of SPARK methods considered. Finally, a numerical experiment is given in section 5 to illustrate the theoretical results.

2 The system of index 2 implicit DAEs.

We consider the system of implicit DAEs (1.1). Applying the chain rule to $a'(t,y)$ in (1.1) and then inverting $a_y(t,y)$, we can obtain an explicit expression for y'

$$y' = a_y^{-1}(t,y)\left(f(t,y,z) - a_t(t,y,z)\right). \tag{2.1a}$$

Taking the total derivative of $g(t,y)$ with respect to t in (1.1) we obtain the additional underlying constraints

$$0 = g_y(t,y)a_y^{-1}(t,y)f(t,y,z) - g_y(t,y)a_y^{-1}(t,y)a_t(t,y,z) + g_t(t,y). \tag{2.1b}$$

The initial values y_0, z_0 at t_0 are supposed to be given and to be consistent, i.e., to satisfy (1.1b) and (2.1b). Recall that we have supposed $g_y(t,y)a_y^{-1}(t,y)f_z(t,y,z)$ to be invertible. All these conditions ensure existence and uniqueness of a solution passing through the initial values. By the implicit function theorem, the algebraic variables can be expressed implicitly from (2.1b) as a function of t and y. Differentiating the constraints (1.1b) a second time with respect to t, we can

obtain an explicit expression for the derivative z'. The system of implicit DAEs (1.1) is therefore of index 2. By defining

$$u := a(t,y), \quad v := z,$$
$$F(t,u,v) := f(t,a^{-1}(t,u),v), \quad G(t,u) := g(t,a^{-1}(t,u)), \tag{2.2}$$

the system of DAEs (1.1) can be expressed equivalently as

$$u' = F(t,u,v), \tag{2.3a}$$
$$0 = G(t,u), \tag{2.3b}$$

with the standard assumption for Hessenberg index 2 DAEs that $G_u(t,u)F_v(t,u,v)$ exists and is invertible.

With the equations of mechanical systems in mind where different types of forces are present, see [8, 10, 17, 18], decompositions of the right-hand side $f(t,y,z)$ of (1.1a) can be considered

$$f(t,y,z) = \sum_{m=1}^{M} f_m(t,y,z). \tag{2.4a}$$

The functions f_m are supposed to have distinct properties and can therefore be numerically treated in a different way. The value of M corresponds to different classes of right-hand side terms. This value must correspond to the number of different methods to be used in conjunction in the SPARK scheme considered and is reasonably small, e.g., $M = 5$. For the applications of the numerical methods considered in this paper, the following additional assumption is made

$$f_1(t,y,z) = f_1(t,y). \tag{2.4b}$$

This is not a restriction on the system (1.1), but rather a restriction on the application of SPARK methods, see section 3.

3 SPARK methods.

In this paper we consider numerical methods applied directly to (1.1), not to (2.1). This has the advantage of not requiring the computation of $a_t(t,y)$ and $a_y^{-1}(t,y)$.

Before giving a precise definition of SPARK methods applied to (1.1), we first consider the application of a Runge–Kutta (RK) method with coefficients $(a_{ij})_{i,j=1,\ldots,s}$ and $(b_j,c_j)_{j=1,\ldots,s}$ to (1.1a). Following [11] it is natural to take

$$a(T_i,Y_i) = a(t_0,y_0) + h\sum_{j=1}^{s} a_{ij} f(T_j,Y_j,Z_j) \quad \text{for} \quad i = 1,\ldots,s,$$
$$a(t_1,y_1) = a(t_0,y_0) + h\sum_{j=1}^{s} b_j f(T_j,Y_j,Z_j),$$

where $T_i := t_0 + c_i h, t_1 := t_0 + h$, and h is the step-size. This definition is equivalent to the standard definition of a RK method applied to (2.3a), then re-expressed in terms of the variables y, z satisfying the relations (2.2). A major difficulty is to define the internal algebraic variables Z_j by dealing properly with the constraints (1.1b). A priori we would like not only $0 = g(T_i, Y_i)$ to be satisfied for $i = 1, \ldots, s$, but also $0 = g(t_1, y_1)$. For stiffly accurate RK methods, i.e., methods satisfying $a_{sj} = b_j$ for $j = 1, \ldots, s$, this last equation is automatically satisfied since $y_1 = Y_s$ and $t_1 = T_s$. However, for non-stiffly accurate RK methods, such as Gauss and Radau IA methods, it is not possible to satisfy all these equations at the same time since there are only sm internal algebraic variables Z_j for $(s+1)m$ constraints. For such methods one way to circumvent this problem is to remove the equation $0 = g(t_1, y_1)$ and to project the solution y_1 obtained onto the constraint (1.1b) in an additional step [1, 3, 7, 14]. A similar approach preserving symmetry is given by the symmetric projection method [4, 6]. In this paper we will show as a particular result that a projection procedure is not necessary for RK methods such as Lobatto IIIB, Lobatto IIIC*, and Lobatto IIID. Another possibility is to replace the conditions $0 = g(T_i, Y_i)$ by $0 = g(T_i, \widehat{Y}_i)$ where $\widehat{Y}_i$ satisfies

$$a(T_i, \widehat{Y}_i) = a(t_0, y_0) + h \sum_{j=1}^{s} \widehat{a}_{ij} f(T_j, Y_j, Z_j)$$

for other RK coefficients $(\widehat{a}_{ij})_{i,j=1,\ldots,s}$ satisfying the stiff accuracy condition $\widehat{a}_{sj} = b_j$ for $j = 1, \ldots, s$. In the case of Lobatto methods we can consider taking the coefficients $\widehat{a}_{ij}$ of Lobatto IIIA or of Lobatto IIIC methods which are both stiffly accurate. Order conditions for such partitioned Runge–Kutta (PRK) methods can be found in [15]. We will see however in what follows that for certain classes of RK methods such as Lobatto IIIB, Lobatto IIIC*, and Lobatto IIID methods, there is an alternative way of dealing with constraints which does not require any projection or the introduction of additional internal stages. The main idea is to add the equation $0 = g(t_1, y_1)$ and to replace the constraints equations $0 = g(T_i, Y_i)$ for $i = 1, \ldots, s$ by a well-chosen linear combination of lower dimension. We obtain the system of sm equations

$$0 = \sum_{j=1}^{s} r_{ij} g(T_j, Y_j) \quad \text{for} \quad i = 1, \ldots, s-1, \qquad 0 = g(t_1, y_1).$$

The main difficulty is to define these linear combinations of constraints in such a way that the method has highest possible order. For the methods under consideration in this paper the coefficients $(r_{ij})_{i=1,\ldots,s-1,j=1,\ldots,s}$ will be chosen as those of $\widetilde{A}_1$ in (3.2).

We state hereafter some assumptions and properties of the coefficients of the SPARK methods that we will consider, according to [12]. In this paper we denote the $m \times m$ identity matrix by I_m, the ith s-dimensional unit basis vector by $e_i := (0, \ldots, 0, 1, 0, \ldots, 0)^T$, the s-dimensional zero vector by $0_s := (0, \ldots, 0)^T$, the weight vector by $b := (b_1, \ldots, b_s)^T$, the node vector by $c := (c_1, \ldots, c_s)^T$, and

the RK coefficients matrices of M distinct RK methods by $A_m := (a_{ij,m})_{i,j=1,\ldots,s}$ for $m = 1, \ldots, M$. It is assumed that the number s of internal stages satisfies $s \geq 2$. SPARK methods satisfying the following assumptions are considered

$$e_1^T A_1 = 0_s^T, \tag{3.1a}$$

$$e_s^T A_1 = b^T, \tag{3.1b}$$

$$A_1 A_m = \begin{pmatrix} 0_s^T \\ N \end{pmatrix} \quad \text{for } m = 2, \ldots, M, \tag{3.1c}$$

$$\begin{pmatrix} N \\ b^T \end{pmatrix} \text{ is invertible,} \tag{3.1d}$$

$$e_s^T A_3 = b^T. \tag{3.1e}$$

These assumptions are satisfied for example by the s-stage Lobatto SPARK families with $M = 5$ and A_1, A_2, A_3, A_4, A_5 being the RK matrices of Lobatto IIIA-B-C-C*-D coefficients respectively [10, 12]. The assumptions (3.1be) are *stiff accuracy* conditions. Let $\widetilde{A}_1$ be the $(s-1) \times s$ sub-matrix of A_1 given by the relation

$$A_1 = \begin{pmatrix} 0_s^T \\ \widetilde{A}_1 \end{pmatrix}. \tag{3.2}$$

We define the $s \times (s+1)$ matrix Q by

$$Q := L \begin{pmatrix} \widetilde{A}_1 & 0_{s-1} \\ 0_s^T & 1 \end{pmatrix} \tag{3.3}$$

where the $s \times s$ matrix L is any invertible matrix. To simplify the analysis hereafter we take

$$L := \begin{pmatrix} \widetilde{A}_1 \\ e_s^T \end{pmatrix}^{-1}.$$

The invertibility of the matrix on the right-hand side and of A_3 follows from the assumption (3.1d) and the relation

$$\begin{pmatrix} N \\ b^T \end{pmatrix} = \begin{pmatrix} \widetilde{A}_1 \\ e_s^T \end{pmatrix} A_3$$

which is a simple consequence of the assumptions (3.1ce) for $m = 3$. Thus, we take

$$Q := \begin{pmatrix} \widetilde{A}_1 \\ e_s^T \end{pmatrix}^{-1} \begin{pmatrix} \widetilde{A}_1 & 0_{s-1} \\ 0_s^T & 1 \end{pmatrix} \tag{3.4}$$

This matrix Q satisfies

$$Q = \begin{pmatrix} I_s & 0_s \end{pmatrix} + \begin{pmatrix} O_{s,s-1} & -p & p \end{pmatrix}, \qquad p := \begin{pmatrix} \widetilde{A}_1 \\ e_s^T \end{pmatrix}^{-1} e_s.$$

Hence, we have $p_s = 1$, $q_{sj} = 0$ for $j = 1, \ldots, s$ and $q_{s,s+1} = 1$. From this result we obtain easily the fact that

$$Q \begin{pmatrix} v \\ d \\ d \end{pmatrix} = \begin{pmatrix} v \\ d \end{pmatrix} \tag{3.5}$$

where v is an arbitrary vector in $\mathbb{R}^{s-1}$ and d an arbitrary constant. To simplify the expression of our results and their proof, we introduce the $(s+1)\times s$ matrices

$$\alpha_m := \begin{pmatrix} A_m \\ b^T \end{pmatrix} \quad \text{for} \quad m = 1, \ldots, M.$$

From the assumptions (3.1) it follows that

$$Q\alpha_1 = A_1, \qquad Q\alpha_m = A_3 \quad \text{for } m = 2, \ldots, M. \tag{3.6}$$

Using the above notations and assumptions, we are now in position to give a precise definition of the application of SPARK methods to (1.1):

Definition 3.1. *One step of an* s-stage super partitioned additive Runge–Kutta (SPARK) method *applied with step-size* h *to the system of index* 2 *implicit DAEs* (1.1) *with decomposition* (2.4)*, initial values* y_0, z_0 *at* t_0*, and step-size* h *reads*

$$0 = a(T_i, Y_i) - a(t_0, y_0) - h \sum_{j=1}^{s} \sum_{m=1}^{M} a_{ij,m} f_m(T_j, Y_j, Z_j) \quad \text{for} \quad i = 1, \ldots, s, \tag{3.7a}$$

$$0 = a(t_1, y_1) - a(t_0, y_0) - h \sum_{j=1}^{s} b_j f(T_j, Y_j, Z_j), \tag{3.7b}$$

$$0 = \sum_{j=1}^{s} q_{ij} g(T_j, Y_j) + q_{i,s+1} g(t_1, y_1) \quad \text{for} \quad i = 1, \ldots, s, \tag{3.7c}$$

where the coefficients q_{ij} *are those of a matrix* Q *satisfying* (3.3) *(the equation* $0 = g(t_1, y_1)$ *is thus satisfied).*

Remark 3.1. In a $o(1)$-neighborhood of y_0 and z_0 the solution of (3.7) does not depend on z_0 (see Theorem 4.1 below). The value z_0 only determines to which branch $z = z(t, y)$ of (2.1b) the solution is close. An accurate value z_1 is not required since the values z_n do not influence the global convergence properties of the differential variable y. For SPARK methods satisfying $c_s = 1$ the approximation given by $z_1 := Z_s$ is generally adequate.

An extension of the above definition including also index 3 constraints is given in [12]. A major difficulty of SPARK methods lies in the numerical solution of the resulting systems of nonlinear equations. Not only the use of matrix Q reestablishes super-convergence properties of non-stiffly accurate Lobatto methods, but it also allows for an efficient implementation of SPARK methods by

application of modified Newton iterations [11, 12, 13]. The structure of the approximate Jacobian becomes greatly simplified and facilitates the construction of efficient preconditioners for stiff problems [12, 13].

In order to preserve reversibility properties of the flow, a question of interest is if a SPARK method based on symmetric RK coefficients still preserves its symmetry. The answer is affirmative:

THEOREM 3.1. *Consider the system of index* 2 *implicit DAEs* (1.1) *with consistent initial values* (y_0, z_0) *at* t_0. *Consider a SPARK method* (3.1) *with coefficients satisfying the assumptions* (3.1), *such that its coefficients* $(b_i, c_i, a_{ij,m})$ *for* $m = 1, \dots, M$ *are symmetric for* $m = 1$ *and when* $f_m \not\equiv 0$ *for* $m \geq 2$, *i.e.,*

$$
\begin{aligned}
&b_i = b_{s+1-i}, \quad \text{for } i = 1, \dots, s,\\
&c_i = 1 - c_{s+1-i}, \quad \text{for } i = 1, \dots, s,\\
&a_{s+1-i,s+1-j,m} + a_{ij,m} = b_j \quad \text{for } i = 1, \dots, s, \text{ and } j = 1, \dots, s.
\end{aligned}
$$

Then the SPARK method is symmetric.

PROOF. Exchanging y_0 and y_1 and h with $-h$ in (3.7), we obtain a method with internal stages $\widetilde{T}_i, \widetilde{Y}_i, \widetilde{Z}_i$ for $i = 1, \dots, s$ and numerical solution $\overline{y}_1$. By defining $\overline{T}_i := \widetilde{T}_{s+1-i}, \overline{Y}_i := \widetilde{Y}_{s+1-i}, \overline{Z}_i := \widetilde{Z}_{s+1-i}$ we obtain the same equations (3.7a) with T_i, Y_i, Z_i, y_1 replaced respectively by $\overline{T}_i, \overline{Y}_i, \overline{Z}_i, \overline{y}_1$. We have assumed Q of the form (3.3) and we can choose $M = I_s$. For the constraints equations (3.7c) we obtain $g(\overline{t}_1, \overline{y}_1) = 0$ and

$$
0 = \sum_{j=1}^{s} a_{ij,1} g(\overline{T}_{s+1-j}, \overline{Y}_{s+1-j}) \quad \text{for } i = 2, \dots, s. \tag{3.8}
$$

Since A_1 is supposed to satisfy (3.1b) we can subtract to the equations (3.8) for $i = 2, \dots, s-1$ the equation (3.8) for $i = s$. We obtain

$$
0 = \sum_{j=1}^{s} (b_j - a_{ij,1}) g(\overline{T}_{s+1-j}, \overline{Y}_{s+1-j}) \quad \text{for } i = 2, \dots, s-1.
$$

By using the symmetry of coefficients $a_{ij,1}$ this leads to

$$
0 = \sum_{j=1}^{s} a_{s+1-i,j,1} g(\overline{T}_j, \overline{Y}_j) \quad \text{for } i = 2, \dots, s-1.
$$

Together with $0 = \sum_{j=1}^{s} b_j g(\overline{T}_j, \overline{Y}_j)$ we obtain again the equations (3.7c) with T_j, Y_j replaced respectively by $\overline{T}_j, \overline{Y}_j$. □

4 Analysis of SPARK methods.

We consider the following *simplifying assumptions*

$$
B(p): \sum_{i=1}^{s} b_i c_i^{k-1} = \frac{1}{k}, \quad \text{for } k = 1, \dots, p,
$$

$$C_m(q_m)\colon \sum_{j=1}^{s} a_{ij,m} c_j^{k-1} = \frac{c_i^k}{k} \quad \text{for } i = 1,\dots,s, \text{ and } k = 1,\dots,q_m,$$

$$D_m(r_m)\colon \sum_{i=1}^{s} b_i c_i^{k-1} a_{ij,m} = \frac{b_j}{k}\big(1 - c_j^k\big) \quad \text{for } j = 1,\dots,s, \text{ and } k = 1,\dots,r_m,$$

where $m \in \{1,\dots,M\}$. If one of the assumptions $C_m(1)$ is satisfied for an index m we can assume the system (1.1) to be autonomous. Notice that the only method of real interest which does not satisfy this condition is the 2-stage Lobatto IIIB method. In this paper we will assume that $C_1(1)$ and $C_3(1)$ hold in any case. To analyze SPARK methods we can also assume that $a(t,y) \equiv y$ since by definition SPARK methods applied to (1.1) are equivalent to their application to (2.3). Therefore, we can assume the system (1.1) to be autonomous and that $a(t,y) \equiv y$, i.e., we can simply consider Hessenberg index 2 DAEs in any part of our analysis

$$y' = f(y,z) = f_1(y) + \sum_{m=2}^{M} f_m(y,z), \tag{4.1a}$$

$$0 = g(y), \tag{4.1b}$$

with the standard assumption that $g_y(y) f_z(y,z)$ exists and is invertible in a neighborhood of a solution. This greatly simplifies the proofs. All results can then be directly reformulated for SPARK methods (3.7) applied to (1.1).

Definition 3.1 of SPARK methods applied to (4.1) can be expressed as follows

$$0 = Y_i - y_0 - h \sum_{j=1}^{s} \sum_{m=1}^{M} \alpha_{ij,m} f_m(Y_j, Z_j) \quad \text{for} \quad i = 1,\dots,s+1, \tag{4.2a}$$

$$0 = \sum_{j=1}^{s+1} q_{ij} g(Y_j) \quad \text{for} \quad i = 1,\dots,s \tag{4.2b}$$

where we have formally introduced the notation $T_{s+1} := t_1$ and $Y_{s+1} := y_1$. In general existence and uniqueness to these nonlinear equations cannot be shown unless some assumptions on the SPARK coefficients are made.

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

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

$g_y(y) f_z(y,z)$ exists and is invertible in a neighborhood of (y_0, z_0). Assume that the SPARK method (4.2) *has coefficients satisfying the assumptions* (3.1) *and that the simplifying assumptions $C_1(1)$ and $C_3(1)$ hold. Then for $h \le h_0$ there exists a locally unique SPARK solution which satisfies*

$$Y_i - y_0 = O(h) \quad \text{for } i = 1,\dots,s+1, \qquad Z_i - z_0 = O(h) \quad \text{for } i = 1,\dots,s.$$

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]. We rewrite the equations (4.2b) for $i = 1, \ldots, s$ as

$$0 = \frac{1}{h}\sum_{j=1}^{s+1} q_{ij} g(Y_j) = \frac{1}{h} g(y_0) + \frac{1}{h}\sum_{j=1}^{s+1} q_{ij} \int_0^1 g_y(y_0 + \tau(Y_j - y_0)) d\tau \cdot (Y_j - y_0)\,. \tag{4.3}$$

By inserting the relation

$$\frac{1}{h}(Y_j - y_0) = \sum_{k=1}^{s} \alpha_{jk,1} f_1(Y_k) + \sum_{m=2}^{M}\sum_{k=1}^{s} \alpha_{jk,m} f_m(Y_k, Z_k),$$

in the right-hand side of (4.3) and by using (3.6), for $h = 0$ and $Y_i(0) = y_0, Z_i(0) = z_0$ we obtain

$$\sum_{k=1}^{s} a_{ik,1} g_y(y_0) f_1(y_0) + \sum_{k=1}^{s} a_{ik,3} g_y(y_0) \left(\sum_{m=2}^{M} f_m(y_0, z_0) \right).$$

By using $C_1(1)$ and $C_3(1)$ this reduces to

$$c_i g_y(y_0) \left(f_1(y_0) + \sum_{m=2}^{M} f_m(y_0, z_0) \right) = c_i g_y(y_0) f(y_0, z_0)$$

which therefore vanishes. Using the tensor matrix product notation $\otimes$, the Jacobian of (4.2a) and (4.3) with respect to $Y_1, \ldots, Y_s, Y_{s+1}, Z_1, \ldots, Z_s$ is equal at $h = 0$ to

$$\begin{pmatrix} I_{s+1} \otimes I_n & O \\ \star & A_3 \otimes g_y(y_0) f_z(y_0, z_0) \end{pmatrix}$$

by (3.6) and is therefore invertible. □

The goal now is to obtain a local error estimate for the differential variable 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 Hessenberg index 2 DAEs which can be found for example in [5, 7]. Definitions of trees t and related quantities $\rho(t), \gamma(t)$, etc., can be given as in [5, Section 5] and [7, Section VII.5]. We only present the main differences. To each meager node we associate in addition a number m with $m \in \{1, \ldots, M\}$, this corresponds to f_m. To a tree with a meager root we associate the quantity $m(t)$ which is the number m associated to its root. Notice that because our assumption $f_1(y, z) = f_1(y)$, each meager node with associated number 1 cannot be directly connected upward by a branch to a fat node. We denote the corresponding sets of trees by $LADAT2_{y,m}$ if the root is meager with associated number m, and by $LADAT2_z$ if the root is fat. We use the notation $LADAT2$ to emphasize that these sets are an extension of $LDAT2$ of [5, 7] with the additional letter A standing for the word additive.

Defining $k_{i,m}(h) := hf_m(Y_i, Z_i)$ for $i = 1, \ldots, s$ we can rewrite (4.2a) as

$$Y_i = y_0 + \sum_{j=1}^{s} \sum_{m=1}^{M} \alpha_{ij,m} k_{j,m} \quad \text{for} \quad i = 1, \ldots, s+1.$$

First we give some expressions for the derivatives of $k_{i,m}$ and Z_i which are similar to those given in [5, Theorem 5.7] and [7, Theorem VII.5.6].

THEOREM 4.2. *For $i = 1, \ldots, s$ and $m = 1, \ldots, M$ we have*

$$k_{i,m}^{(q)}(0) = \sum_{\substack{t \in LADAT2_{y,m} \\ \rho(t)=q}} \gamma(t)\Phi_i(t)F(t)(y_0, z_0),$$

$$Z_i^{(q)}(0) = \sum_{\substack{u \in LADAT2_{z} \\ \rho(t)=q}} \gamma(u)\Phi_i(u)F(u)(y_0, z_0),$$

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

$$\Phi_i(t) = \sum_{\mu_1=1}^{s} \cdots \sum_{\mu_m=1}^{s} \alpha_{i\mu_1,m(t_1)} \cdots \alpha_{i\mu_m,m(t_n)} \Phi_{\mu_1}(t_1) \cdots \Phi_{\mu_m}(t_m)\Phi_i(u_1) \cdots \Phi_i(u_n)$$
$$\text{if } t = [t_1, \ldots, t_m, u_1, \ldots, u_n]_y,$$

$$\Phi_i(u) = \sum_{j=1}^{s} \sum_{k=1}^{s+1} \sum_{\mu_1=1}^{s} \cdots \sum_{\mu_m=1}^{s} \omega_{ij} q_{jk} \alpha_{k\mu_1,m(t_1)} \cdots \alpha_{k\mu_m,m(t_n)} \Phi_{\mu_1}(t_1) \cdots \Phi_{\mu_m}(t_m)$$
$$\text{if } u = [t_1, \ldots, t_m]_z. \tag{4.4}$$

The coefficients q_{jk} are those of matrix Q in (3.4) *and the coefficients ω_{ij} are those of $\omega := A_3^{-1}$.*

A proof of this theorem can be obtained completely similarly to the ones of [5, 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* (4.2) *satisfies*

$$y_1^{(q)}(0) = \sum_{m=1}^{M} \sum_{\substack{t \in LADAT2_{y,m} \\ \rho(t)=q}} \gamma(t) \sum_{i=1}^{s} b_i \Phi_i(t) F(t)(y_0, z_0).$$

For the local error we obtain:

THEOREM 4.4. *Consider the Hessenberg index 2 DAEs* (4.1) *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 SPARK method with coefficients satisfying the assumptions of Theorem* 4.1 *and the simplifying assumptions $B(p)$, $C_m(q_m)$ and $D_m(r_m)$ for $m = 1, \ldots, M$. Then we have*

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

where $\mu := \min(p, 2q + 2, q + r + 2, 2q_3, q_3 + r_3 + 1)$, with $q := \min(q_1, \ldots, q_M)$ and $r := \min(r_1, \ldots, r_M)$. If the function $f(y, z)$ of (4.1a) is linear in z then the value $2q_3$ in μ can be changed to $2q_3 + 1$.

REMARK 4.1. The same local error estimate also holds for SPARK methods (3.1) applied to the system of index 2 implicit DAEs (1.1).

A proof of this theorem can be obtained completely similarly to the ones of [5, Theorem 5.9] and [7, Theorem VII.5.10], it is therefore omitted. After application of the simplifying assumptions $C_m(q_m)$ in (4.4) we simply use the fact that

$$Q \begin{pmatrix} c_1^k \\ \vdots \\ c_s^k \\ 1 \end{pmatrix} = \begin{pmatrix} c_1^k \\ \vdots \\ c_s^k \end{pmatrix}$$

when $c_s = 1$ which directly follows from (3.5).

Following for example [7, Theorem VII.4.5], global convergence of SPARK methods is easily obtained:

THEOREM 4.5. *Consider the system of index* 2 *implicit DAEs* (1.1) *with consistent initial values* (y_0, z_0) *at* t_0 *and such that*

$$a_y(t, y) \quad \textit{and} \quad g_y(t, y)a_y^{-1}(t, y)f_z(t, y, z)$$

exist and are invertible in a neighborhood of the exact solution. Consider a SPARK method (3.1) *with coefficients satisfying the assumptions of Theorem* 4.1 *with local error order* μ*, i.e.,* $y_1 - y(t_0 + h) = O(h^{\mu+1})$*. Then the SPARK method is convergent of order* μ*, i.e., the global error satisfies*

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

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

A direct consequence of Theorem 4.5 is the super-convergence of Lobatto IIIA-B-C-C*-D SPARK methods:

COROLLARY 4.6. *Consider the system of index* 2 *implicit DAEs* (1.1) *with consistent initial values* (y_0, z_0) *at* t_0 *and such that*

$$a_y(t, y) \quad \textit{and} \quad g_y(t, y)a_y^{-1}(t, y)f_z(t, y, z)$$

exist and are invertible in a neighborhood of the exact solution. Then the global error of the s*-stage Lobatto IIIA-B-C-C***-D SPARK method* (3.1) *satisfies*

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

for $t_n - t_0 \leq \text{Const}$ *and* $h := \max(|h_1|, \ldots, |h_n|)$*. Moreover, reversibility properties of the flow are still preserved provided* $f_3 \equiv 0$ *and* $f_4 \equiv 0$.

PROOF. The s-stage Lobatto IIIA-B-C-C*-D coefficients satisfy the simplifying assumptions $B(p)$, $C_m(q_m)$, and $D_m(r_m)$ for $m = 1, \ldots, 5$ with $p = 2s - 2, q_1 = s, r_1 = s - 2, q_2 = s - 2, r_2 = s, q_3 = s - 1, r_3 = s - 1, q_4 = s - 1, r_4 = s - 1, q_5 = s - 1, r_5 = s - 1$. The Lobatto IIIA-B-D coefficients are known to be symmetric [7, 10], hence reversibility preservation is a simple consequence of Theorem 3.1. □

REMARK 4.2. Notice that the above result is not a generalization of [9, Corollary 5.3] for Lobatto IIIA methods since here we have assumed $f_1(y,z) = f_1(y)$. However, it can be seen as a generalization of the super-convergence result of Lobatto IIIC methods for Hessenberg index 2 DAEs which was obtained in [5].

5 A numerical experiment.

To show the relevance of the theoretical results, we have applied the 2-stage and 3-stage Lobatto IIIA-B-C-C*-D SPARK methods with constant stepsize h to the following Hessenberg index 2 DAEs

$$\text{(5.1a)} \qquad \begin{pmatrix} y_1' \\ y_2' \end{pmatrix} = f_1(y_1,y_2) + \sum_{m=2}^{5} f_m(y_1,y_2,z_1), \quad 0 = g(y_1,y_2),$$

where

$$f_1(y_1,y_2) = \begin{pmatrix} y_2 - 2y_1^2 y_2 \\ -y_1^2 \end{pmatrix}, \quad f_2(y_1,y_2,z_1) = \begin{pmatrix} y_1 y_2^2 z_1^2 \\ e^{-t} z_1 - y_1 \end{pmatrix},$$
$$f_3(y_1,y_2,z_1) = \begin{pmatrix} -y_2^2 z_1 \\ -3y_2^2 z_1 \end{pmatrix}, \quad f_4(y_1,y_2,z_1) = \begin{pmatrix} 2y_1 y_2^2 - 2e^{-2t} y_1 y_2 \\ z_1 \end{pmatrix},$$
$$f_5(y_1,y_2,z_1) = \begin{pmatrix} 2y_2^2 z_1^2 \\ y_1^2 y_2^2 \end{pmatrix}, \quad g(y_1,y_2) = y_1^2 y_2 - 1.$$

For the initial conditions $y_{10} = 1, y_{20} = 1$ 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 Fig. 5.1 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 method of order 2 we observe a straight line of slope 2 and for the 3-stage method of order 4 we observe a straight line of slope 4, thus confirming the orders of convergence predicted by Corollary 4.6.

6 Conclusion.

The numerical approximation of the solution of a class of index 2 implicit DAEs by SPARK methods has been considered. We have shown that for certain classes of SPARK methods super-convergence can be achieved even if the SPARK coefficients are not all stiffly accurate. This is possible by taking a specific linear combination of the constraints evaluated at the internal stages and at the numerical solution. Certain properties of SPARK coefficients are essential with that respect. We have proved in particular that the s-stage Lobatto IIIA-B-C-C*-D SPARK methods retain their classical order of super-convergence equal to $2s-2$. Moreover, reversibility properties of the flow can still be preserved provided certain SPARK coefficients are symmetric.

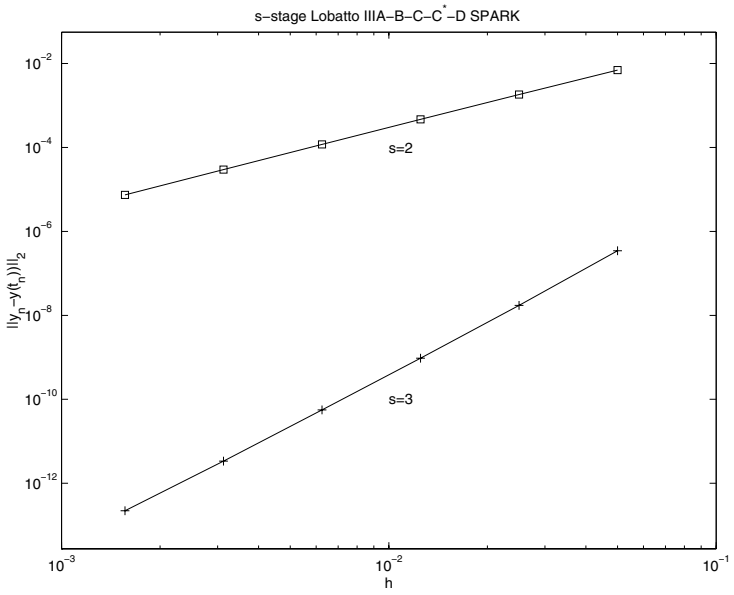

Figure 5.1: Global errors of the s-stage Lobatto IIIA-B-C-C*-D SPARK methods ($s = 2, 3$) with constant stepsizes applied to the test problem (5.1).

7 Acknowledgments.

The author would like to thank Professor Juan I. Montijano from the Department of Applied Mathematics of the University of Zaragoza, Spain, for his invitation and fruitful discussions.

REFERENCES

1. U. Ascher and L. R. Petzold, *Projected implicit Runge–Kutta methods for differential-algebraic equations*, SIAM J. Numer. Anal., 28 (1991), pp. 1097–1120.
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, 2nd ed, 1996.
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), pp. 77–94.
4. E. Hairer, *Geometric integration of ordinary differential equations on manifolds*, BIT, 41 (2001), pp. 996–1007.
5. E. Hairer, C. Lubich, and M. Roche, *The Numerical Solution of Differential-Algebraic Systems by Runge–Kutta Methods*, Vol. 1409, Lecture Notes in Mathematics, Springer-Verlag, Berlin, 1989.
6. E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration*, Vol. 31, Comput. Mathematics, Springer-Verlag, Berlin, 2002.

7. E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems*, Vol. 14 of Comput. Mathematics, Springer-Verlag, Berlin, 2nd ed., 1996.
8. E. J. Haug, *Computer Aided Kinematics and Dynamics of Mechanical Systems. Volume I: Basic Methods*, Allyn & Bacon, Boston, USA, 1989.
9. L. O. Jay, *Convergence of a class of Runge–Kutta methods for differential-algebraic systems of index* 2, BIT, 33 (1993), pp. 137–150.
10. L. O. Jay, *Structure preservation for constrained dynamics with super partitioned additive Runge–Kutta methods*, SIAM J. Sci. Comput., 20 (1998), pp. 416–446.
11. L. O. Jay, *Inexact simplified Newton iterations for implicit Runge–Kutta methods*, SIAM J. Numer. Anal., 38 (2000), pp. 1369–1388.
12. L. O. Jay, *Iterative solution of nonlinear equations for SPARK methods applied to DAEs*, Numer. Algorithms, 31 (2002), pp. 171–191.
13. L. O. Jay, *Preconditioning and parallel implementation of implicit Runge–Kutta methods*, Tech. Rep., Department of Mathematics, University of Iowa, USA, 2002.
14. C. Lubich, *On projected Runge–Kutta methods for differential-algebraic equations*, BIT, 31 (1991), pp. 545–550.
15. A. Murua, *Partitioned Runge–Kutta methods for semi-explicit differential-algebraic systems of index* 2, Tech. Rep., EHU-KZAA-IKT-196, University of the Basque Country, 1996.
16. P. J. Rabier and W. C. Rheinboldt, *Nonholonomic Motion of Rigid Mechanical Systems from a DAE Viewpoint*, SIAM, Philadelphia, 2000.
17. W. O. Schiehlen (ed.), *Multibody Systems Handbook*, Springer-Verlag, Berlin, 1990.
18. W. O. Schiehlen (ed.), *Advanced Multibody System Dynamics, Simulation and Software Tools*, Kluwer Academic Publishers, Dordrecht, 1993.