

RUNGE-KUTTA METHODS APPLIED TO FULLY IMPLICIT DIFFERENTIAL-ALGEBRAIC EQUATIONS OF INDEX 1

ANNE KVAERNØ

ABSTRACT. In this paper we study the order of Runge-Kutta methods applied to differential-algebraic equations of index one. We derive general order conditions for the local order k_L , and give a convergence result, which shows that the order k_G of the global error satisfies $k_G \geq k_L - 1$. We also describe some numerical experiments, which are in agreement with our results.

1. INTRODUCTION

A general differential-algebraic equation (DAE) has the form

$$(1.1) \quad F(v, v', x) = 0$$

with initial values

$$v(x_0) = v_0, \quad v'(x_0) = v'_0,$$

where $v: \mathbf{R} \rightarrow \mathbf{R}^{m-1}$ and $F: \mathbf{R}^{m-1} \times \mathbf{R}^{m-1} \times \mathbf{R} \rightarrow \mathbf{R}^{m-1}$ is a function for which we assume sufficient differentiability. We also assume $\partial F / \partial v'$ to be singular with constant rank, (1.1) to be of index 1 over the whole interval of integration $[x_0, x_{\text{end}}]$, and the initial values to be consistent, i.e.,

$$F(v_0, v'_0, x_0) = 0.$$

The *index* of a DAE is the number of times the algebraic part of the system has to be differentiated to obtain an ODE. The index 1 system (1.1) is supposed to be *solvable* in the sense that for each set of consistent initial values there exists a unique solution of the system. For more precise definitions of index and solvability, see [8].

According to Petzold [13], an s -stage Runge-Kutta method applied to (1.1) is defined by

$$(1.2) \quad F \left(v_n + h \sum_{j=1}^s a_{ij} V'_j, V'_i, x_n + c_i h \right) = 0, \quad i = 1, \dots, s,$$

$$(1.3) \quad v_{n+1} = v_n + h \sum_{i=1}^s b_i V'_i,$$

Received May 27, 1988; revised April 14, 1989.

1980 *Mathematics Subject Classification* (1985 *Revision*). Primary 65L05.

©1990 American Mathematical Society
0025-5718/90 \$1.00 + \$.25 per page

and the stage vectors V_i are given by

$$(1.4) \quad V_i = v_n + h \sum_{j=1}^s a_{ij} V_j', \quad i = 1, \dots, s.$$

An s -stage Runge-Kutta method is described by its Butcher tableau

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \cdots & a_{1s} \\ c_2 & a_{21} & a_{22} & \cdots & a_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array}$$

or by

$$\begin{array}{c|c} c & \mathcal{A} \\ \hline & b^T \end{array},$$

where $c_i = \sum_{j=1}^s a_{ij}$, $i = 1, \dots, s$. The matrix \mathcal{A} has to be nonsingular, and we define

$$\mathcal{D} = (d_{ij}) = \mathcal{A}^{-1}.$$

The *local truncation error* is given by

$$d_{n+1} = \tilde{v}_{n+1} - v(x_n + h),$$

where \tilde{v}_{n+1} is the solution of (1.3) when $v_n = v(x_n)$. The *local order* of the method is k_L if

$$d_{n+1} = \mathcal{O}(h^{k_L}).$$

The *global error* is defined by

$$e_n = v_n - v(x_n),$$

and the order of the method is k_G if

$$e_n = \mathcal{O}(h^{k_G}).$$

The *stability constant* r is given by

$$r = 1 - b^T \mathcal{A}^{-1} \varepsilon_s,$$

where $\varepsilon_s = [1, \dots, 1]^T \in \mathbf{R}^s$.

Recently, the behavior of Runge-Kutta methods applied to differential-algebraic problems has received considerable attention. In [13], Petzold derived a complete set of order conditions for linear constant-coefficient index-1 equations, assuming that $|r| < 1$. Under the same assumption, she also derived a sufficient set of order conditions for nonlinear problems (1.1) linear in v' . Later, Burrage and Petzold [2] extended these results to also include the classes of methods with $|r| = 1$. Kvaernø [9] derived a complete set of order conditions for the local truncation error for this class of problems by comparing the

Taylor expansion of the exact and numerical solution of the equation. Roche [14, 15] derived general order conditions for Runge-Kutta methods applied to semiexplicit index-1 problems, using the theory of Butcher series and rooted trees. Very recently, this theory has been extended to the Hessenberg form index-2 DAE's by Hairer et al. [7].

The DAE (1.1) can (at least in theory) be transformed to an autonomous, partitioned system, by using the following arguments. The system can be written as an autonomous system with no loss of generality. This is done by adding the differential equation

$$v'_m = 1, \quad v(x_0) = x_0,$$

with the solution $v_m(x) = x$, to the system. We then have

$$(1.5) \quad F(v, v') = 0$$

with $v: \mathbf{R} \rightarrow \mathbf{R}^m$ and $F: \mathbf{R}^m \times \mathbf{R}^m \rightarrow \mathbf{R}^m$. Equation (1.5) can be split into a differential and an algebraic part. Gear [6] uses the following argument: Suppose that

$$\text{rank } F_{v'} = r < m.$$

Then there exists a nonsingular $r \times r$ submatrix of $F_{v'}$. Suppose that the equations have been numbered so that $\text{rank } \partial f / \partial v' = r$ over the whole interval of integration, where f represents the first r equations in F . Let g be the last $m - r$ equations in F . Suppose that the variables are numbered such that

$$\frac{\partial f}{\partial v'} = \begin{bmatrix} \frac{\partial f}{\partial v'_1} & \frac{\partial f}{\partial v'_2} \end{bmatrix},$$

where $v = [v_1^T, v_2^T]^T$, $v_1 \in \mathbf{R}^r$, $v_2 \in \mathbf{R}^{m-r}$, and $\partial f / \partial v'_1$ is nonsingular. Then, by the implicit function theorem, $f = 0$ can be solved for v'_1 , that is, $v'_1 = f_1(v_1, v_2, v'_2)$. This can be substituted into the last $m - r$ equations to get an implicit relationship between v_1 and v_2 . The vector v'_2 cannot be involved, or we would be able to solve (1.5) for additional components of v' , contrary to the assumption about the rank of $F_{v'}$. Thus, (1.5) can be written as

$$(1.6) \quad \begin{aligned} f(v, v') &= 0, \\ g(v) &= 0. \end{aligned}$$

The system (1.6) has index 1 if and only if

$$\begin{bmatrix} f_{v'} \\ g_v \end{bmatrix}$$

is nonsingular. In this paper, we are only concerned with solvable index-1 DAE's of the form (1.6). However, the results obtained are valid also for the more general form (1.1), as long as the rank of $\partial F / \partial v'$ is constant over the whole interval of integration.

We have used a model equation for the derivation of the local order conditions.

Theorem 1.1. *The set of order conditions derived for the model equation*

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

is equivalent to the set of order conditions for the fully implicit problem (1.6).

Proof. The fully implicit index-1 DAE is given by (1.6). By the definition of the index, we know that

$$\begin{bmatrix} \partial f / \partial v' \\ \partial g / \partial v \end{bmatrix}$$

is nonsingular. Then, $\text{rank } g_v = m - r$, and there exists a nonsingular $(m-r) \times (m-r)$ submatrix of g_v . Suppose that the variables are now numbered such that g_v can be written as

$$\begin{bmatrix} \partial g / \partial y & \partial g / \partial z \end{bmatrix},$$

where $v = [y^T, z^T]^T$, and g_z is nonsingular. Then g can be solved for z , and (1.6) can be written as

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

or, by inserting the expression for z into the differential equation, as

$$(1.7) \quad f_2(y, y', z') = 0, \quad z = \tilde{g}(y).$$

The numerical solution defined by (1.2) and (1.3), applied to (1.6), is given by

$$(1.8) \quad f(V_i, V'_i) = 0, \quad g(V_i) = 0.$$

The equation (1.8) is the same as (1.6) with v replaced by V_i and v' by V'_i . Using the same arguments as above, (1.8) can be written as

$$(1.9) \quad f_2(Y_i, Y'_i, Z'_i) = 0, \quad Z_i = \tilde{g}(Y_i),$$

where $V_i = [Y_i^T, Z_i^T]^T$. Note that the choice of y and z so that g_z is nonsingular is not necessarily unique, neither will g_z necessarily be constant over the whole interval of integration. But, at least in some neighborhood of the solution at x_n , g_z will be nonsingular, and we assume h to be small enough to keep g_z nonsingular over the whole step. For the first step, (1.4) and (1.3) can be written as

$$(1.10) \quad Y_i = y_n + h \sum_{j=1}^s a_{ij} Y'_j, \quad Z_i = z_n + h \sum_{j=1}^s a_{ij} Z'_j, \quad i = 1, \dots, s,$$

and

$$(1.11) \quad y_{n+1} = y_n + h \sum_{i=1}^s b_i Y'_i, \quad z_{n+1} = z_n + h \sum_{i=1}^s b_i Z'_i.$$

By using (1.10) together with the algebraic part of (1.9) we have

$$z_n + h \sum_{j=1}^s a_{ij} Z'_j = \tilde{g} \left(y_n + h \sum_{j=1}^s a_{ij} Y'_j \right),$$

or, by solving for Z'_i and using the Taylor expansion around (y_n, z_n) ,

$$\begin{aligned} Z'_i &= \frac{1}{h} \sum_{j=1}^s d_{ij} \left(\tilde{g} \left(y_n + h \sum_{k=1}^s a_{jk} Y'_k \right) - z_n \right) \\ &= \tilde{g}_y Y'_i + \frac{h}{2} \sum_{j,k,l=1}^s d_{ij} a_{jk} a_{jl} \tilde{g}_{yy}(Y'_k, Y'_l) + \dots \end{aligned}$$

Inserting this into the differential part of (1.9), we obtain

$$f_2(Y_i, Y'_i, Z'_i) = f_2 \left(Y_i, Y'_i, \tilde{g}_y Y'_i + \frac{h}{2} \sum_{j,k,l=1}^s d_{ij} a_{jk} a_{jl} \tilde{g}_{yy}(Y'_k, Y'_l) + \dots \right) = 0.$$

The term Z'_i involves $\tilde{g}_y Y'_i$, so the term Y'_i in (1.9) will give no additional order conditions. We then have that the equation

$$\tilde{f}(Y_i, Z'_i) = 0, \quad Z_i = \tilde{g}(Y_i), \quad i = 1, \dots, s,$$

together with (1.11) and (1.10), will give all the necessary order conditions. \square

In §2.1 we develop a general scheme for the Taylor expansion of the exact solution of the model equation. In §2.2 we give a complete set of order conditions for the local truncation error when a Runge-Kutta method is applied to the model equation. These order conditions take on a simple form, with the help of the “tree model” derived in §2.1. Convergence results are given in §3, while numerical experiments are described in §4.

2. THE ORDER OF THE LOCAL TRUNCATION ERROR

The aim of this section is to derive a set of necessary and sufficient order conditions for the local truncation error. In §2.1, we expand the solution of the model equation into a Taylor series. This series is expressed in terms of rooted trees. In §2.2, we derive the Taylor expansion of the numerical solution of the model equation. The coefficients of the Taylor series are obtained directly from the trees derived in §2.1. By comparing the Taylor series of the exact and numerical solution, the order conditions are obtained. The main result in this section is given in Theorem 2.2.

Some of the trees derived in §2.1 will correspond to identical order conditions. In §2.3, a reduced set of trees is introduced, so that each of the order conditions is given by one, and only one, tree. In Figure 2, all the order conditions up to order 4 are exhibited, together with their related trees.

2.1. Taylor expansion of the exact solution of a model equation. Consider the index-1 equation

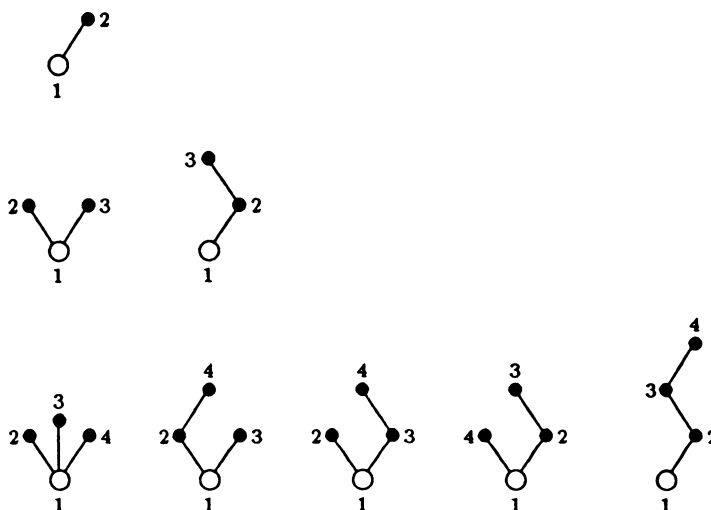
$$(2.1) \quad f(y, z') = 0, \quad z = g(y)$$

with consistent initial values $y(x_0) = y_0$ and $z(x_0) = z_0$, where $y: \mathbf{R} \rightarrow \mathbf{R}^r$, $z: \mathbf{R} \rightarrow \mathbf{R}^{m-r}$, $f: \mathbf{R}^r \times \mathbf{R}^{m-r} \rightarrow \mathbf{R}^r$, and $g: \mathbf{R}^r \rightarrow \mathbf{R}^{m-r}$. The functions f and

g are assumed to be sufficiently differentiable. Repeated differentiation of the algebraic part of (2.1) yields

$$\begin{aligned}
 (2.2) \quad & z' = g_y y', \\
 & z'' = g_{yy}(y', y') + g_y y'', \\
 & z''' = g_{yyy}(y', y', y') + g_{yy}(y'', y') + g_{yy}(y', y'') + g_{yy}(y', y'') + g_y y''', \\
 & \vdots
 \end{aligned}$$

These expressions can be written in terms of trees as follows:



By inserting z' from (2.2) into the differential part of (2.1) we have

$$(2.3) \quad f(y, g_y y') = 0.$$

Since (2.1) is an index 1 equation, (2.3) can be solved for y' , and $f_{z'} g_y$ is nonsingular. Repeated differentiation of the differential part of (2.1) gives

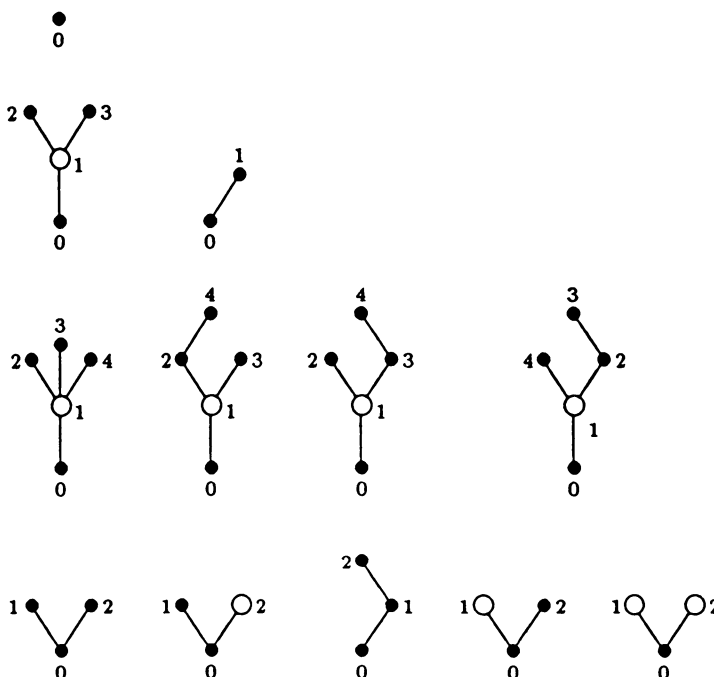
$$\begin{aligned}
 & f_y y' + f_{z'} z'' = 0, \\
 & f_{yy}(y', y') + f_{yz'}(y', z'') + f_y y'' + f_{z'y}(z'', y') + f_{z'z'}(z'', z'') + f_{z'} z''' = 0, \\
 & \vdots
 \end{aligned}$$

By replacing the highest derivative of z with the expression given in (2.2), we

find

$$\begin{aligned}
 y' &= y', \\
 y'' &= (-f_{z'}g_y)^{-1}(f_{z'}g_{yy}(y', y') + f_y y'), \\
 y''' &= (-f_{z'}g_y)^{-1}(f_{z'}g_{yyy}(y', y', y') + f_{z'}g_{yy}(y'', y') \\
 &\quad + f_{z'}g_{yy}(y', y'') + f_{z'}g_{yy}(y', y'') \\
 &\quad + f_{yy}(y', y') + f_{yz'}(y', z'') \\
 &\quad + f_y y'' + f_{z'y}(z'', y') + f_{z'z'}(z'', z'')), \\
 &\vdots
 \end{aligned}
 \tag{2.4}$$

These expressions can also be written in terms of trees:



These graphs motivate us to introduce a set of *special monotonically labelled trees*, t_s , given by the following definitions.

Definition 2.1. The set of special trees, SDA1T , is the set of directed graphs, consisting of light and heavy vertices, with one single root, such that:

- (1) $t_s \in \text{SDA1T}_{yy}$ if the root is light, only the root has ramifications, and each of the branches consist of only light, or only heavy vertices. If the root has no ramification, then the tree consists of only light vertices.
- (2) $t_s \in \text{SDA1T}_{yz}$ if the root is light without ramifications, but followed by a heavy vertex. The heavy vertex has at least two branches with no ramifications, and the branches all consist of only light vertices.

- (3) $t_s \in \text{SDA1T}_z$ if the root is heavy, only the root has ramifications, and the branches consist of only light vertices.
- (4) $\text{SDA1T}_y = \text{SDA1T}_{yy} \cup \text{SDA1T}_{yz}$.
- (5) $\text{SDA1T} = \text{SDA1T}_y \cup \text{SDA1T}_z$.

Let $\omega(t_s)$ be the number of vertices in a tree.

Definition 2.2. Let $t_s \in \text{SDA1T}$. We say that t_s is *monotonically labelled* if every vertex is associated with an integer i satisfying

$$0 \leq i \leq \omega(t_s) - 1 \quad \text{if } t_s \in \text{SDA1T}_y,$$

$$1 \leq i \leq \omega(t_s) \quad \text{if } t_s \in \text{SDA1T}_z,$$

and if, following each branch of t_s , the labels are monotonically increasing.

Definition 2.3. SLDA1T_{yy} , SLDA1T_{yz} , and SLDA1T_z are the sets of monotonically labelled trees satisfying conditions (1)–(3), respectively, in Definition 2.1, and

$$\text{SLDA1T}_y = \text{SLDA1T}_{yy} \cup \text{SLDA1T}_{yz},$$

$$\text{SLDA1T} = \text{SLDA1T}_y \cup \text{SLDA1T}_z.$$

The set of trees, SLDA1T , corresponds to the SLDAT -trees defined by Roche [14]. The differences between the two sets of trees derives from the fact that the trees of Roche are constructed for a semiexplicit index 1 equation, while the trees used in this paper are constructed for the model equation (2.1). Also, in the rest of this paper, we will use similar notations as used by Roche. To distinguish between the two kind of trees, we use the notation DA1T -trees in place of the DAT -trees used by Roche.

Let $\omega_m(t_s)$ be the number of light vertices, and let $\omega_f(t_s)$ be the number of heavy vertices in $t_s \in \text{SLDA1T}$.

Definition 2.4. The *order* $\rho(t_s)$ of a tree $t_s \in \text{SLDA1T}$ is defined by

- (1) $\rho(t_s) = \omega_m(t_s) + \omega_f(t_s)$ if $t_s \in \text{SLDA1T}_{yy}$,
- (2) $\rho(t_s) = \omega_m(t_s) - 1$ if $t_s \in \text{SLDA1T}_{yz}$,
- (3) $\rho(t_s) = \omega_m(t_s)$ if $t_s \in \text{SLDA1T}_z$.

Then t_s is the tree representation for one of the terms in $y^{(\rho(t_s))}$ if $t_s \in \text{SLDA1T}_y$, and one of the terms in $z^{(\rho(t_s))}$ if $t_s \in \text{SLDA1T}_z$. Let

$$\tau_y = \bullet \quad \tau_z = \bigcirc \text{---} \bullet$$

Definition 2.5. For every tree $t_s \in \text{SDA1T}_y$ we define a function $F_s(t_s): \mathbf{R}' \times \mathbf{R}^{m-r} \rightarrow \mathbf{R}'$, and for every tree $u_s \in \text{SDA1T}_z$ we define a function $G_s(u_s): \mathbf{R}' \times \mathbf{R}^{m-r} \rightarrow \mathbf{R}^{m-r}$, by

- (1) $F_s(\tau_y)(y, z) = y'$ and $G_s(\tau_z)(y, z) = g_y y' = z'$,

The number of ways to label a tree $t_s \in \text{SDA1T}$ is the number of times the corresponding derivative appears in the Taylor expansion of the exact solution. We call this number $\beta(t_s)$. We can now state the following lemma.

Lemma 2.1. *For the exact solution of (2.1) we have*

$$\begin{aligned} y^{(p)} &= \sum_{\substack{t_s \in \text{SLDA1T}_y \\ \rho(t_s)=p}} F_s(t_s) = \sum_{\substack{t_s \in \text{SDA1T}_y \\ \rho(t_s)=p}} \beta(t_s) F_s(t_s), \\ z^{(p)} &= \sum_{\substack{u_s \in \text{SLDA1T}_z \\ \rho(u_s)=p}} G_s(u_s) = \sum_{\substack{u_s \in \text{SDA1T}_z \\ \rho(u_s)=p}} \beta(u_s) G_s(u_s). \end{aligned}$$

We now know how to express $y^{(p)}$ and $z^{(p)}$ in terms of partial derivatives of f and g , and of lower derivatives of y and z . What we want is to express $y^{(p)}$ and $z^{(p)}$ in terms of partial derivatives of f and g , and of y' . Such an expression is already given for y'' in (2.4). By inserting this into the expression for z'' in (2.2) we obtain

$$z'' = g_{yy}(y', y') + g_y(-f_{z'} g_y)^{-1} f_y y' + g_y(-f_{z'} g_y)^{-1} f_{z'} g_{yy}(y', y').$$

The expressions for y'' and z'' can be inserted into the expression for y''' , and then for z''' , etc. We now find a new set of trees corresponding to these expressions. This set is defined as follows.

Definition 2.6. We denote by DA1T , DA1T_y , and DA1T_z the set of trees defined recursively by

1. $\tau_y \in \text{DA1T}_y$ and $\tau_z \in \text{DA1T}_{zy}$.
2. (a) If $t_1, \dots, t_k \in \text{DA1T}_y$ and $k > 1$, then $[t_1, \dots, t_k]_z \in \text{DA1T}_{zz}$.
 (b) If $t_1, \dots, t_k \in \text{DA1T}_y$, $u_1, \dots, u_l \in \text{DA1T}_z \setminus \{\tau_z\}$, $k > 0$ or $k = 0$, and $l > 1$, then $[t_1, \dots, t_k, u_1, \dots, u_l]_y \in \text{DA1T}_{yy}$.
 (c) If $u \in \text{DA1T}_{zz}$ then $[u]_y \in \text{DA1T}_{yz}$.
 (d) If $t \in \text{DA1T}_y$ then $[t]_z \in \text{DA1T}_{zy}$.
3. $\text{DA1T}_y = \text{DA1T}_{yy} \cup \text{DA1T}_{yz}$, $\text{DA1T}_z = \text{DA1T}_{zy} \cup \text{DA1T}_{zz}$.
4. $\text{DA1T} = \text{DA1T}_y \cup \text{DA1T}_z$.

Here, $t = [t_1, \dots, t_k, u_1, \dots, u_l]_y$ is the tree obtained by connecting the roots of $t_1, \dots, t_k, u_1, \dots, u_l$ by $k + l$ arcs to a new light vertex which becomes the new root of t . Similarly, $u = [t_1, \dots, t_k]_z$ is the tree obtained in the same manner, but with a new heavy root.

Definition 2.7. The order $\rho(t)$ of a tree $t \in \text{DA1T}$ is given by

- (1) $\rho(t) = \omega_m(t) - \omega_f(t)$ if $t \in \text{DA1T}_y$,
- (2) $\rho(t) = \omega_m(t) - \omega_f(t) + 1$ if $t \in \text{DA1T}_z$,

where $\omega_m(t)$ and $\omega_f(t)$ are the number of light, resp. heavy, vertices in the tree.

There is a one-to-one correspondence between these trees and the terms appearing in the Taylor expansion of the exact solution. This correspondence is given in the following definition.

Definition 2.8. For every tree $t \in \text{DA1T}_y$ we define a function $F(t)(y, z): \mathbf{R}^r \times \mathbf{R}^{m-r} \rightarrow \mathbf{R}^r$, and for every tree $u \in \text{DA1T}_z$ we define a function $G(u)(y, z): \mathbf{R}^r \times \mathbf{R}^{m-r} \rightarrow \mathbf{R}^{m-r}$ recursively by:

- (1) $F(\tau_y)(y, z) = y'$, $G(\tau_z)(y, z) = z' = g_y y'$,
- (2) $F(t)(y, z) = (-f_{z'} g_y)^{-1} f_{ky|z'}(F(t_1), \dots, F(t_k), G(u_1), \dots, G(u_l))$ if $t = [t_1, \dots, t_k, u_1, \dots, u_l]_y$,
- (3) $G(u)(y, z) = g_{ky}(F(t_1), \dots, F(t_k))$ if $u = [t_1, \dots, t_k]_z$,

where $t_1, \dots, t_k \in \text{DA1T}_y$ and $u_1, \dots, u_l \in \text{DA1T}_z$. The expressions $F(t)(y, z)$ and $G(u)(y, z)$ are called the *elementary differentials* associated with the tree t , respectively u .

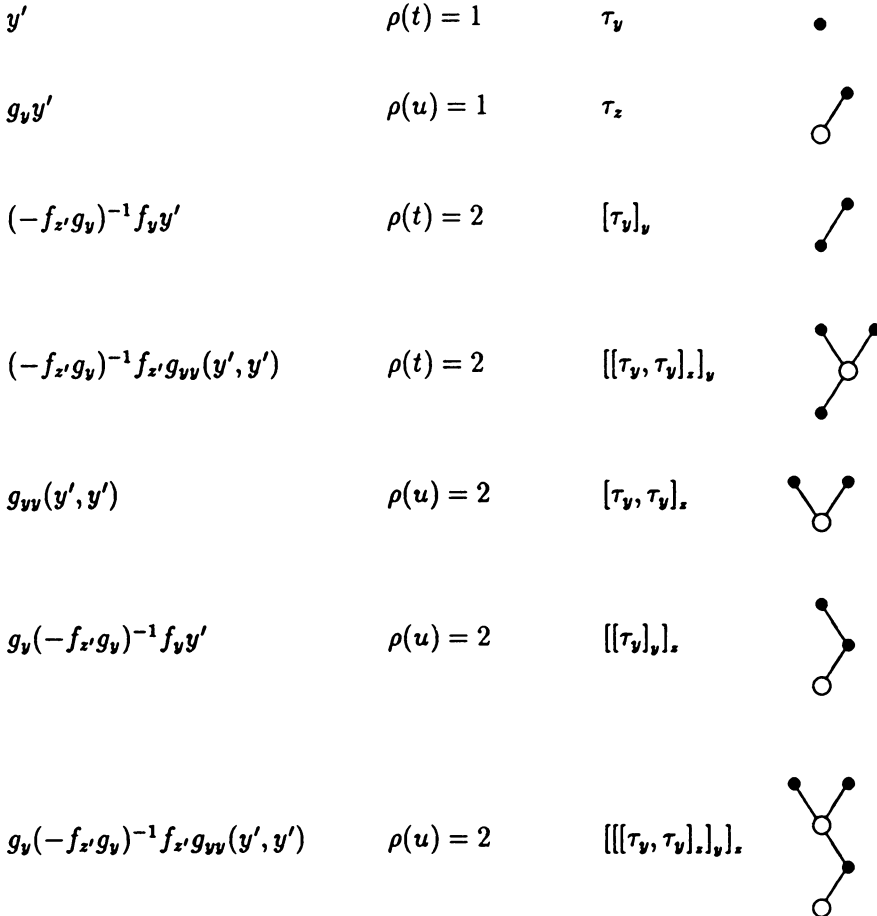


FIGURE 1. Elementary differentials and corresponding trees

Figure 1 shows the elementary differentials and the corresponding trees for y' , z' , y'' , and z'' .

There exists a relation between the trees of SDA1T and those of DA1T. Let $t_s \in \text{SDA1T}_{yy}$ and

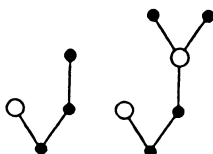
$$F_s(t_s) = (-f_{z'}g_y)^{-1} f_{ky/z'}(y^{(p_1)}, \dots, y^{(p_k)}, z^{(q_1)}, \dots, z^{(q_l)})$$

with $\rho(t_s) = p$. Suppose that all the trees $t \in \text{DA1T}$, $\rho(t) \leq p$, are given. Replace each branch with p_i light vertices once with each of the trees $t \in \text{DA1T}_y$, $\rho(t) = p_i$, and each of the branches with $q_j - 1$ heavy vertices once with each of the trees $u \in \text{DA1T}_z$, $\rho(u) = q_j$. Then we have the set of trees $t \in \text{DA1T}$, $\rho(t) = p$. See Example 2.2.

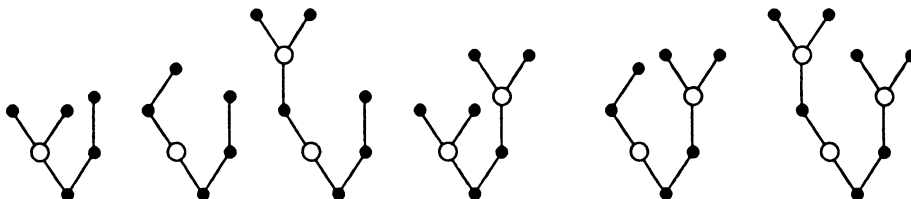
Example 2.2. Let $t_s \in \text{SDA1T}_y$, $\rho(t_s) = 4$, be the tree



with corresponding derivative $(-f_{z'}g_y)^{-1} f_{z'y}(z'', y'')$. Replace the light branch with each of the trees corresponding to y'' , that is all the trees given in Figure 1 with a light root and $\rho(t) = 2$. We then obtain all the trees of DA1T corresponding to t_s :

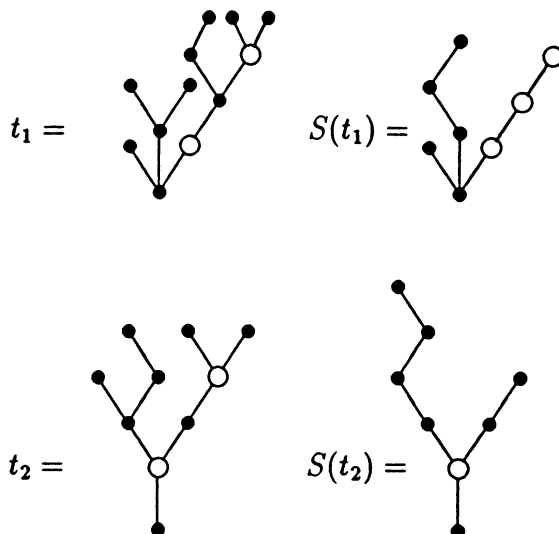


Then replace the heavy branch with each of the trees corresponding to z'' , that is all the trees given in Figure 1 with a heavy root and $\rho(u) = 2$:



Similar transformations can be carried out with all the trees in SDA1T. For each $t_s \in \text{SDA1T}$ there is a corresponding set of trees $t \in \text{DA1T}$. Let $t \in \text{DA1T}$ be one of the trees obtained from a tree $t_s \in \text{SDA1T}$ as described above. Then we call t_s the *special tree corresponding to t* , and denote it by $S(t)$. This is illustrated in the following example.

Example 2.3. Let $t_1 \in \text{SDA1T}_{yy}$ and $t_2 \in \text{SDA1T}_{yz}$. The corresponding special trees are given by



Let $\alpha(t)$ be the number of times the elementary differential corresponding to $t \in \text{DA1T}$ appears in the exact solution of the model equation (2.1). We can then state the following result.

Theorem 2.1. For the exact solution of (2.1) we have

$$\begin{aligned}
 y^{(p)}(x_0) &= \sum_{\substack{t \in \text{DA1T}_y \\ \rho(t)=p}} \alpha(t) F(t)(y_0, z_0), \\
 z^{(p)}(x_0) &= \sum_{\substack{u \in \text{DA1T}_z \\ \rho(u)=p}} \alpha(u) G(u)(y_0, z_0)
 \end{aligned}
 \tag{2.5}$$

and

$$\begin{aligned}
 y(x_0 + h) &= y_0 + \sum_{t \in \text{DA1T}_y} \alpha(t) F(t)(y_0, z_0) \frac{h^{\rho(t)}}{\rho(t)!}, \\
 z(x_0 + h) &= z_0 + \sum_{u \in \text{DA1T}_z} \alpha(u) G(u)(y_0, z_0) \frac{h^{\rho(u)}}{\rho(u)!}.
 \end{aligned}
 \tag{2.6}$$

Lemma 2.2. $\alpha(t)$ is given recursively by

- (1) $\alpha(\tau_y) = \alpha(\tau_z) = 1$.
- (2) If $t = [t_1, \dots, t_k, u_1, \dots, u_l]_y$, then $\alpha(t) = \beta(S(t))\alpha(t_1) \cdots \alpha(t_k)\alpha(u_1) \cdots \alpha(u_l)$, and if $u = [t_1, \dots, t_k]_z$, then $\alpha(u) = \beta(S(u))\alpha(t_1) \cdots \alpha(t_k)$.

Proof. From Lemma 2.1, Definitions 2.5 and 2.8, and Theorem 2.1 we have

$$\begin{aligned}
 y^{(p)} &= \sum_{\substack{t_s \in \text{SDA1T}_y \\ \rho(t_s)=p}} \beta(t_s) F_s(t_s)(y, z) \\
 &= \sum_{\substack{t_s \in \text{SDA1T}_y \\ \rho(t_s)=p}} \beta(t_s) (-f_{z'} g_y)^{-1} f_{ky/z'}(y^{(p_1)}, \dots, y^{(p_k)}, z^{(q_1)}, \dots, z^{(q_l)}) \\
 (2.7) \quad &= \sum_{\substack{t_s \in \text{SDA1T}_y \\ \rho(t_s)=p}} \beta(t_s) (-f_{z'} g_y)^{-1} f_{ky/z'} \\
 &\quad \cdot \left(\sum_{\substack{t_1 \in \text{DA1T}_y \\ \rho(t_1)=p_1}} \alpha(t_1) F(t_1)(y, z), \dots, \sum_{\substack{u_l \in \text{DA1T}_z \\ \rho(u_l)=q_l}} \alpha(u_l) G(u_l)(y, z) \right).
 \end{aligned}$$

Let $S(t) \in \text{SDA1T}_y$, $t = [t_1, \dots, t_k, u_1, \dots, u_l]_y$, and $\rho(t) = p$ be one of the trees in (2.7). The value of $\alpha(t)$ is given directly by comparing (2.7) with (2.5). A similar procedure can be used to prove the lemma for $\alpha(u)$. \square

2.2. Taylor expansion of the numerical solution of the model equation. To be able to find the order conditions for Runge-Kutta methods applied to the model equation (2.1), we have to find the Taylor expansion of the numerical solution. To do this, we introduce the concept of DA1-series. Similar series are given by Roche [14] for semiexplicit index-1 problems.

Definition 2.9. Let $\mathbf{a}: \text{DA1T}_y \rightarrow \mathbf{R}$ and $\mathbf{b}: \text{DA1T}_z \rightarrow \mathbf{R}$ be any mappings. The series

$$\begin{aligned}
 \text{DA1}_y(\mathbf{a}, y_0, z_0) &= y_0 + \sum_{t \in \text{DA1T}_y} \mathbf{a}(t) \alpha(t) F(t) \frac{h^{\rho(t)}}{\rho(t)!}, \\
 \text{DA1}_z(\mathbf{b}, y_0, z_0) &= z_0 + \sum_{u \in \text{DA1T}_z} \mathbf{b}(u) \alpha(u) G(u) \frac{h^{\rho(u)}}{\rho(u)!}
 \end{aligned}$$

are called DA1_y , respectively DA1_z , series.

A Runge-Kutta method applied to (2.1) is given by

$$\begin{aligned}
 f \left(y_0 + h \sum_{j=1}^s a_{ij} Y'_j, Z'_i \right) &= 0, \\
 (2.8) \quad z_0 + h \sum_{j=1}^s a_{ij} Z'_j &= g \left(y_0 + h \sum_{j=1}^s a_{ij} Y'_j \right), \quad i = 1, \dots, s,
 \end{aligned}$$

$$(2.9) \quad y_1 = y_0 + h \sum_{i=1}^s b_i Y'_i, \quad z_1 = z_0 + h \sum_{i=1}^s b_i Z'_i.$$

The stage values are given by

$$(2.10) \quad \begin{aligned} Y_i &= y_0 + h \sum_{j=1}^s a_{ij} Y'_j, \\ Z_i &= z_0 + h \sum_{j=1}^s a_{ij} Z'_j, \end{aligned} \quad i = 1, \dots, s.$$

Now Y_i, y_1 , respectively Z_i and z_1 , $i = 1, \dots, s$, can be written as $DA1_y$, respectively $DA1_z$, series as follows:

$$(2.11) \quad \begin{aligned} Y_i(x_0 + h) &= y_0 + \sum_{t \in DA1T_y} \mathbf{v}_i(t) \alpha(t) F(t) \frac{h^{\rho(t)}}{\rho(t)!}, \\ Z_i(x_0 + h) &= z_0 + \sum_{u \in DA1T_z} \mathbf{w}_i(u) \alpha(u) G(u) \frac{h^{\rho(u)}}{\rho(u)!}, \end{aligned} \quad i = 1, \dots, s,$$

$$(2.12) \quad \begin{aligned} y_1(x_0 + h) &= y_0 + \sum_{t \in DA1T_y} \mathbf{y}_1(t) \alpha(t) F(t) \frac{h^{\rho(t)}}{\rho(t)!}, \\ z_1(x_0 + h) &= z_0 + \sum_{u \in DA1T_z} \mathbf{z}_1(u) \alpha(u) G(u) \frac{h^{\rho(u)}}{\rho(u)!}. \end{aligned}$$

The stage derivatives are written as

$$(2.13) \quad \begin{aligned} Y'_i(x_0 + h) &= \sum_{t \in DA1T_y} \mathbf{l}_i(t) \alpha(t) F(t) \frac{h^{\rho(t)-1}}{\rho(t)!}, \\ Z'_i(x_0 + h) &= \sum_{u \in DA1T_z} \mathbf{k}_i(u) \alpha(u) G(u) \frac{h^{\rho(u)-1}}{\rho(u)!}, \end{aligned} \quad i = 1, \dots, s.$$

By inserting (2.13) into (2.10) we get

$$\begin{aligned} Y_i(x_0 + h) &= y_0 + h \sum_{j=1}^s a_{ij} Y'_j = y_0 + \sum_{t \in DA1T_y} \left(\sum_{j=1}^s a_{ij} \mathbf{l}_j(t) \right) \alpha(t) F(t) \frac{h^{\rho(t)}}{\rho(t)!}, \\ Z_i(x_0 + h) &= z_0 + h \sum_{j=1}^s a_{ij} Z'_j = z_0 + \sum_{u \in DA1T_z} \left(\sum_{j=1}^s a_{ij} \mathbf{k}_j(u) \right) \alpha(u) G(u) \frac{h^{\rho(u)}}{\rho(u)!}. \end{aligned}$$

Comparing this with (2.11), we have

$$(2.14) \quad \mathbf{v}_i(t) = \sum_{j=1}^s a_{ij} \mathbf{l}_j(t), \quad \mathbf{w}_i(t) = \sum_{j=1}^s a_{ij} \mathbf{k}_j(u).$$

Similarly, by inserting (2.13) into (2.9) and comparing with (2.12), we have

$$(2.15) \quad \mathbf{y}_1(t) = \sum_{i=1}^s b_i \mathbf{l}_i(t), \quad \mathbf{z}_1(u) = \sum_{i=1}^s b_i \mathbf{k}_i(u).$$

The Taylor expansion of the numerical solution can be written as

$$(2.16) \quad Y_i' = \sum_{p=1}^{\infty} \tilde{Y}_i^{(p)} \frac{h^{p-1}}{p!}, \quad Z_i' = \sum_{p=1}^{\infty} \tilde{Z}_i^{(p)} \frac{h^{p-1}}{p!},$$

$$(2.17) \quad Y_i = y_0 + \sum_{p=1}^{\infty} \bar{Y}_i^{(p)} \frac{h^p}{p!}, \quad Z_i = z_0 + \sum_{p=1}^{\infty} \bar{Z}_i^{(p)} \frac{h^p}{p!},$$

$$(2.18) \quad y_1 = y_0 + \sum_{p=1}^{\infty} \hat{y}^{(p)} \frac{h^p}{p!}, \quad z_1 = z_0 + \sum_{p=1}^{\infty} \hat{z}^{(p)} \frac{h^p}{p!}.$$

Comparing (2.13) with (2.16), (2.11) with (2.17), and (2.12) with (2.18), we have

$$(2.19) \quad \tilde{Y}_i^{(p)} = \sum_{\substack{t \in \text{DAIT}_y \\ \rho(t)=p}} \mathbf{l}_i(t) \alpha(t) F(t), \quad \tilde{Z}_i^{(p)} = \sum_{\substack{u \in \text{DAIT}_z \\ \rho(u)=p}} \mathbf{k}_i(u) \alpha(u) G(u),$$

$$(2.20) \quad \bar{Y}_i^{(p)} = \sum_{\substack{t \in \text{DAIT}_y \\ \rho(t)=p}} \mathbf{v}_i(t) \alpha(t) F(t), \quad \bar{Z}_i^{(p)} = \sum_{\substack{u \in \text{DAIT}_z \\ \rho(u)=p}} \mathbf{w}_i(u) \alpha(u) G(u),$$

$$(2.21) \quad \hat{y}^{(p)} = \sum_{\substack{t \in \text{DAIT}_y \\ \rho(t)=p}} \mathbf{y}_1(t) \alpha(t) F(t), \quad \hat{z}^{(p)} = \sum_{\substack{u \in \text{DAIT}_z \\ \rho(u)=p}} \mathbf{z}_1(u) \alpha(u) G(u).$$

We use the notation $\tilde{Y}_i^{(p_k)}(t_k)$ for the term $\alpha(t_k) \mathbf{l}_i(t_k) F(t_k)$, and $\tilde{Z}_i^{(q_l)}(u_l)$ for the term $\alpha(u_l) \mathbf{k}_i(u_l) G(u_l)$, where $p_k = \rho(t_k)$ and $q_l = \rho(u_l)$. Similar notations are used for the terms in $\bar{Y}_i^{(p)}$, $\bar{Z}_i^{(p)}$, $\hat{y}^{(p)}$, and $\hat{z}^{(p)}$. Equation (2.8) can be written as

$$(2.22) \quad f \left(y_0 + \sum_{p=1}^{\infty} \bar{Y}_i^{(p)} \frac{h^p}{p!}, \sum_{p=1}^{\infty} \tilde{Z}_i^{(p)} \frac{h^{p-1}}{p!} \right) = 0,$$

$$(2.23) \quad z_0 + \sum_{p=1}^{\infty} \bar{Z}_i^{(p)} \frac{h^p}{p!} = g \left(y_0 + \sum_{p=1}^{\infty} \bar{Y}_i^{(p)} \frac{h^p}{p!} \right).$$

The n th derivative of (2.23), evaluated at $h = 0$, is

$$\bar{Z}_i^{(n)} = \sum_{\substack{S(u) \in \text{SLDAIT}_z \\ u = [t_1, \dots, t_k]_z \\ \rho(u)=n}} g_{ky}(\bar{Y}_i^{(p_1)}(t_1), \dots, \bar{Y}_i^{(p_k)}(t_k)).$$

By use of (2.20), Lemma 2.2, and Definition 2.8 we have

$$\begin{aligned}
 \overline{Z}_i^{(n)} &= \sum_{\substack{S(u) \in \text{SDA1T}_z \\ u=[t_1, \dots, t_k]_z \\ \rho(u)=n}} \beta(S(u)) \alpha(t_1) \cdots \alpha(t_k) \mathbf{v}_i(t_1) \cdots \mathbf{v}_i(t_k) \\
 &\quad \cdot g_{ky}(F(t_1), \dots, F(t_k)) \\
 (2.24) \quad &= \sum_{\substack{S(u) \in \text{SDA1T}_z \\ u=[t_1, \dots, t_k]_z \\ \rho(u)=n}} \alpha(u) \mathbf{v}_i(t_1) \cdots \mathbf{v}_i(t_k) G(u).
 \end{aligned}$$

Comparing this with the expression for $\overline{Z}_i^{(n)}$ in (2.20), we obtain

$$(2.25) \quad \mathbf{w}_i(u) = \mathbf{v}_i(t_1) \cdots \mathbf{v}_i(t_k), \quad i = 1, \dots, s,$$

for all $u = [t_1, \dots, t_k]_z \in \text{DA1T}_z$. Multiplying the $(n-1)$ st derivative of f (given by (2.22)), evaluated at $h = 0$, by $(-f_{z'} g_y)^{-1}$ gives

$$\begin{aligned}
 &(-f_{z'} g_y)^{-1} f^{(n-1)}|_{h=0} \\
 &= \sum_{\substack{S(t) \in \text{SLDA1T}_{yV} \\ t=[t_1, \dots, t_l]_V \\ \rho(t)=n}} (-f_{z'} g_y)^{-1} f_{kylz'} \\
 (2.26) \quad &\cdot \left(\overline{Y}_i^{(p_1)}(t_1), \dots, \overline{Y}_i^{(p_k)}(t_k), \frac{1}{q_1} \tilde{Z}_i^{(q_1)}(u_1), \dots, \frac{1}{q_l} \tilde{Z}_i^{(q_l)}(u_l) \right) \\
 &+ (-f_{z'} g_y)^{-1} f_{z'} \frac{1}{n} \tilde{Z}_i^{(n)} \\
 &= 0.
 \end{aligned}$$

From (2.10), (2.16), and (2.17) we have

$$\tilde{Z}_i^{(n)} = \sum_{j=1}^s d_{ij} \overline{Z}_j^{(n)} = \sum_{j=1}^s d_{ij} \sum_{\substack{S(u) \in \text{SLDA1T}_z \\ u=[t_1, \dots, t_k]_z \\ \rho(u)=n}} g_{ky}(\overline{Y}_j^{(p_1)}(t_1), \dots, \overline{Y}_j^{(p_k)}(t_k)).$$

Thus, from (2.24),

$$\begin{aligned}
 (-f_{z'} g_y)^{-1} f_{z'} \tilde{Z}_i^{(n)} &= \sum_{j=1}^s d_{ij} \sum_{\substack{S(t) \in \text{SLDA1T}_{yz} \\ t=[t_1, \dots, t_k]_z \\ \rho(t)=n}} (-f_{z'} g_y)^{-1} f_{z'} g_{ky}(\overline{Y}_j^{(p_1)}(t_1), \dots, \overline{Y}_j^{(p_k)}(t_k)) \\
 &\quad - \sum_{j=1}^s d_{ij} \overline{Y}_j^{(n)}.
 \end{aligned}$$

Substituting this into (2.26), and using (2.19), (2.20), Lemma 2.2, and Defini-

tion 2.8, we obtain

$$\begin{aligned}
 \overline{Y}_i^{(n)} &= \sum_{\substack{t \in \text{DA1T}_y \\ \rho(t)=n}} \mathbf{v}_i(t) \alpha(t) F(t) \\
 &= n \sum_{j=1}^s a_{ij} \sum_{\substack{S(t) \in \text{SDA1T}_{yy} \\ t=[t_1, \dots, u_l]_y \\ \rho(t)=n}} \alpha(t) \mathbf{v}_j(t_1) \cdots \mathbf{v}_j(t_k) \frac{1}{q_1} \mathbf{k}_j(u_1) \cdots \frac{1}{q_l} \mathbf{k}_j(u_l) F(t) \\
 &\quad + \sum_{\substack{S(t) \in \text{SDA1T}_{yz} \\ t=[[t_1, \dots, t_k]_z]_y \\ \rho(t)=n}} \alpha(t) \mathbf{v}_i(t_1) \cdots \mathbf{v}_i(t_k) F(t).
 \end{aligned}$$

We then have

$$\begin{aligned}
 (2.27) \quad \mathbf{v}_i(t) &= n \sum_{j=1}^s a_{ij} \mathbf{v}_j(t_1) \cdots \mathbf{v}_j(t_k) \frac{1}{\rho(u_1)} \mathbf{k}_j(u_1) \cdots \frac{1}{\rho(u_l)} \mathbf{k}_j(u_l), \\
 &\quad i = 1, \dots, s,
 \end{aligned}$$

if $t = [t_1, \dots, t_k, u_1, \dots, u_l]_y \in \text{DA1T}_{yy}$, $\rho(t) = n$, and

$$(2.28) \quad \mathbf{v}_i(t) = \mathbf{v}_i(t_1) \cdots \mathbf{v}_i(t_k), \quad i = 1, \dots, s,$$

if $t = [[t_1, \dots, t_k]_z]_y \in \text{DA1T}_{yz}$. By using (2.14), (2.15), (2.25), (2.27), and (2.28) we can state the following lemma.

Lemma 2.3. *The quantities Y_i, y_1 , respectively Z_i and z_1 , $i = 1, \dots, s$, are DA1_y , respectively DA1_z , series given by (2.11) and (2.12). The coefficients of these series, and the series of Y'_i and Z'_i , in (2.13), are given recursively by*

$$\begin{aligned}
 (2.29) \quad \mathbf{l}_i(t) &= \rho(t) \mathbf{v}_i(t_1) \cdots \mathbf{v}_i(t_k) \frac{1}{\rho(u_1)} \mathbf{k}_i(u_1), \dots, \frac{1}{\rho(u_l)} \mathbf{k}_i(u_l), \\
 \mathbf{k}_i(u) &= \sum_{j=1}^s d_{ij} \mathbf{v}_j(t_1) \cdots \mathbf{v}_j(t_k)
 \end{aligned}$$

for $t = [t_1, \dots, t_k, u_1, \dots, u_l]_y \in \text{DA1T}_y$, and $u = [t_1, \dots, t_k]_z \in \text{DA1T}_z$, and by

$$(2.30) \quad \mathbf{v}_i(t) = \sum_{j=1}^s a_{ij} \mathbf{l}_j(t), \quad \mathbf{w}_i(u) = \sum_{j=1}^s a_{ij} \mathbf{k}_j(u)$$

and

$$(2.31) \quad \mathbf{y}_1(t) = \sum_{i=1}^s b_i \mathbf{l}_i(t), \quad \mathbf{z}_1(u) = \sum_{i=1}^s b_i \mathbf{k}_i(u),$$

where $\mathbf{l}_i(\emptyset) = 0$, $\mathbf{k}_i(\emptyset) = 0$, and

$$(2.32) \quad \mathbf{l}_i(\tau_y) = 1, \quad \mathbf{k}_i(\tau_z) = 1.$$

Note that the coefficients $\mathbf{l}_i(t)$ and $\mathbf{k}_i(u)$ can be written as

$$(2.33) \quad \mathbf{l}_i(t) = \rho(t) \frac{1}{\rho(u_1)} \cdots \frac{1}{\rho(u_l)} \sum_{n_1, \dots, n_k=1}^s a_{in_1} \cdots a_{in_k} \cdot \mathbf{l}_{n_1}(t_1) \cdots \mathbf{l}_{n_k}(t_k) \mathbf{k}_i(u_1) \cdots \mathbf{k}_i(u_l)$$

for $t = [t_1, \dots, t_k, u_1, \dots, u_l]_y \in \text{DA1T}_y$, and

$$(2.34) \quad \mathbf{k}_i(u) = \sum_{j=1}^s d_{ij} \sum_{n_1, \dots, n_k=1}^s a_{jn_1} \cdots a_{jn_k} \mathbf{l}_{n_1}(t_1) \cdots \mathbf{l}_{n_k}(t_k)$$

for $u = [t_1, \dots, t_k]_z$. Equation (2.32) can be proved by inserting (2.16) into (2.8) and using the Taylor expansions of f and g . We can then express (2.8) as power series in h . The first terms in these expressions will give us $f(y, g_y \tilde{Y}'_i) = 0$ and $\tilde{Z}'_i = g_y \tilde{Y}'_i$, so that

$$(2.35) \quad \tilde{Y}'_i = y' \quad \text{and} \quad \tilde{Z}'_i = z'.$$

Comparing this with (2.19), we have that

$$\tilde{Y}'_i = \mathbf{l}_i(\tau_y) F(\tau_y) = y' \Rightarrow \mathbf{l}_i(\tau_y) = 1.$$

Therefore, $\mathbf{k}_i(\tau_z)$ is given by (2.29) and (2.30).

We observe that $\mathbf{y}_i(t)$ and $\mathbf{z}_i(t)$ can be written as $\gamma(t)\Phi(t)$, where $\gamma(t)$ is a rational number and $\Phi(t)$ is some combination of the method coefficients. The number $\gamma(t)$ is given by the definition below, and $\Phi(t)$ can be read directly from the tree, by the following procedure. Let $t \in \text{DA1T}$. To the root of the tree, attach the label i if the root is light and j if the root is heavy. To the other vertices, attach other labels, say k, l, m, \dots . For each arc write down the factor a_{vw} if the succeeding vertex (labelled w) is light, v, w being the labels at the end of the arc. Similarly, write down the factor d_{vw} if the succeeding vertex is a heavy vertex. Insert a further factor b_i if $t \in \text{DA1T}_y$, and $b_i d_{ij}$ if $t \in \text{DA1T}_z$, and sum over each index i, j, \dots, k in the range from 1 to s . The sum is $\Phi(t)$ and is called *the elementary weight* for the tree t . To each tree we also associate the following rational number:

Definition 2.10. Let $\gamma(t): \text{DA1T} \rightarrow \mathbf{Q}$, where \mathbf{Q} is the set of rational numbers, be defined recursively by

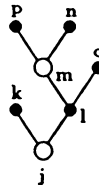
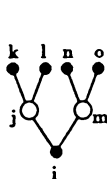
$$\gamma(\tau_y) = 1, \quad \gamma(\tau_z) = 1, \\ \gamma(t) = \rho(t) \frac{1}{\rho(u_1)} \cdots \frac{1}{\rho(u_l)} \cdot \gamma(t_1) \cdots \gamma(t_k) \gamma(u_1) \cdots \gamma(u_l)$$

for $t = [t_1, \dots, t_k, u_1, \dots, u_l]_y \in \text{DA1T}_y$, and

$$\gamma(u) = \gamma(t_1) \cdots \gamma(t_k)$$

for $u = [t_1, \dots, t_k]_z \in \text{DA1T}_z$.

Example 2.4. Let $t \in \text{DA1T}_y$ and $u \in \text{DA1T}_z$. Their elementary weights and rational numbers are given by



$$\Phi(t) = \sum_{ijklmno=1}^s b_i d_{ij} a_{jk} a_{jl} d_{lm} a_{mn} a_{mo} \quad \Phi(u) = \sum_{ijklmnop=1}^s b_i d_{ij} a_{jk} a_{jl} a_{lo} d_{lm} a_{mn} a_{mp}$$

$$\gamma(t) = \frac{3}{4} \quad \gamma(u) = \frac{3}{2}$$

The order conditions for Runge-Kutta methods applied to DAE problems are given by the following theorem.

Theorem 2.2. *If the method (1.2), (1.3) is applied to the problem (2.1), then the order of the local truncation error is $p + 1$ if and only if*

$$\Phi(t) = \frac{1}{\gamma(t)}$$

for all $t \in \text{DA1T}$ with $\rho(t) \leq p$.

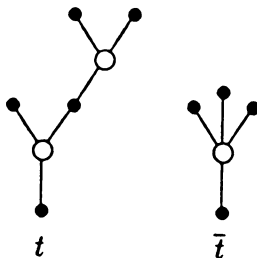
Proof. The Taylor expansion of the exact solution of (2.1) can be written as the DA1 series

$$(2.36) \quad y(x_0 + h) = \text{DA1}_y(\mathbf{p}_y, y_0, z_0), \quad z(x_0 + h) = \text{DA1}_z(\mathbf{p}_z, y_0, z_0)$$

with $\mathbf{p}_y, \mathbf{p}_z = 1$ for all $t \in \text{DA1T}$. By comparing (2.36) term by term with the DA1 series (2.12) for the numerical solution, and by using (2.33), (2.34), and Lemma 2.3, we have proved Theorem 2.2 for the model equation (2.1). Theorem 1.1 shows that this result is also valid for the general index-1 problem. \square

2.3. Simplification of Theorem 2.2. Let $t \in \text{DA1T}_y$; then t can be associated with a simplified tree \bar{t} as follows: If a heavy vertex has no ramifications, and is followed by a light vertex, then the tree can be simplified by removing these two vertices. Similarly, if a light vertex (except the root) with no ramifications is followed by a heavy vertex, the tree can be simplified by removing these two vertices. The simplified tree \bar{t} corresponding to t is the tree which is simplified as much as possible.

Example 2.5. The figure shows a tree t and its corresponding simplified tree \bar{t} .



The set of *simplified trees* \bar{t} is defined recursively by

Definition 2.11. The sets $\overline{\text{DAIT}}_y \subset \text{DAIT}_y$ and $\overline{\text{DAIT}}_z \subset \text{DAIT}_z$ of simplified trees \bar{t} are defined recursively as follows:

1. $\tau_y \in \overline{\text{DAIT}}_{yy}$.
2. (a) If $\bar{t}_1, \dots, \bar{t}_k \in \overline{\text{DAIT}}_{yy}$ and $k > 1$, then $[\bar{t}_1, \dots, \bar{t}_k]_z \in \overline{\text{DAIT}}_z$.
 (b) If $\bar{t}_1, \dots, \bar{t}_k \in \overline{\text{DAIT}}_{yy}$, $\bar{u}_1, \dots, \bar{u}_l \in \overline{\text{DAIT}}_z$, $k > 0$ or $k = 0$, $l > 1$, then $[\bar{t}_1, \dots, \bar{t}_k, \bar{u}_1, \dots, \bar{u}_l]_y \in \overline{\text{DAIT}}_{yy}$.
 (c) If $\bar{u} \in \overline{\text{DAIT}}_z$, then $[\bar{u}]_y \in \overline{\text{DAIT}}_{yz}$.
3. $\overline{\text{DAIT}}_y = \overline{\text{DAIT}}_{yy} \cup \overline{\text{DAIT}}_{yz}$.

Theorem 2.3. To each tree $t \in \text{DAIT}$ there is a corresponding simplified tree $\bar{t} \in \overline{\text{DAIT}}_y$ such that $\Phi(t) = \Phi(\bar{t})$ and $\gamma(t) = \gamma(\bar{t})$.

Before we prove this theorem, we will give an example.

Example 2.6. Consider the tree t and the corresponding simplified tree \bar{t} given by Example 2.5. Their elementary weights are given by

$$\Phi(t) = \sum_{i,j,k,l=1}^s b_i d_{ij} c_j a_{jk} d_{kl} c_l^2, \quad \Phi(\bar{t}) = \sum_{i,j=1}^s b_i d_{ij} c_j^3.$$

By using the fact that

$$\sum_{k=1}^s a_{jk} d_{kl} = \delta_{jl} = \begin{cases} 1 & \text{if } j = l, \\ 0 & \text{otherwise,} \end{cases}$$

we have that $\Phi(t) = \Phi(\bar{t})$. In addition, we have

$$\gamma(t) = 3 \cdot (\tfrac{1}{3} \cdot (1 \cdot 2 \cdot (\tfrac{1}{2} \cdot (1 \cdot 1)))) = 1, \quad \gamma(\bar{t}) = 3 \cdot (\tfrac{1}{3} \cdot (1 \cdot 1 \cdot 1)) = 1,$$

in agreement with the statement of the theorem.

Proof of Theorem 2.3. Let

$$t = [\bar{t}_1, \dots, \bar{t}_k, t_p, \bar{u}_1, \dots, \bar{u}_l, u_q]_y \in \text{DAIT}_y,$$

where

$$\begin{aligned} t_p &= [\bar{u}_p]_y = [[\bar{t}_{p,1}, \dots, \bar{t}_{p,k_p}]_z]_y \in \overline{\text{DAIT}}_{yz}, \\ u_q &= [\bar{t}_q]_z = [[\bar{t}_{q,1}, \dots, \bar{t}_{q,k_q}, \bar{u}_{q,1}, \dots, \bar{u}_{q,l_q}]_y]_z \in \text{DAIT}_{zy}, \\ \bar{t}_1, \dots, \bar{t}_k, \bar{t}_q, \bar{t}_{p,1}, \dots, \bar{t}_{p,k_p}, \bar{t}_{q,1}, \dots, \bar{t}_{q,k_q} &\in \overline{\text{DAIT}}_{yy}, \\ \bar{u}_1, \dots, \bar{u}_l, \bar{u}_p, \bar{u}_{q,1}, \dots, \bar{u}_{q,l_q} &\in \overline{\text{DAIT}}_z. \end{aligned}$$

From (2.33), (2.34), and Definition 2.7 we have

$$\begin{aligned} (2.37) \quad \mathbf{l}_j(t_p) &= \rho(t_p) \frac{1}{\rho(\bar{u}_p)} \mathbf{k}_j(\bar{u}_p) \\ &= \sum_{m=1}^s d_{jm} \sum_{N_1, \dots, N_{k_p}=1}^s a_{mN_1} \cdots a_{mN_{k_p}} \mathbf{l}_{N_1}(t_{p,1}) \cdots \mathbf{l}_{N_{k_p}}(t_{p,k_p}), \end{aligned}$$

$$\begin{aligned}
\mathbf{k}_i(u_q) &= \sum_{j=1}^s d_{ij} \sum_{m=1}^s a_{jm} \mathbf{l}_m(\bar{t}_q) = \mathbf{l}_i(\bar{t}_q) \\
(2.38) \quad &= \rho(\bar{t}_q) \frac{1}{\rho(\bar{u}_{q,1}) \cdots \rho(\bar{u}_{q,l_q})} \\
&\quad \cdot \sum_{M_1, \dots, M_{k_q}=1}^s a_{iM_1} \cdots a_{iM_{k_q}} \mathbf{l}_{M_1}(\bar{t}_{q,1}) \cdots \mathbf{k}_i(\bar{u}_{q,l_q}).
\end{aligned}$$

By using the fact that $\rho(u_q) = \rho(\bar{t}_q)$ we have

$$(2.39) \quad \mathbf{l}_i(t) = \Gamma(t) \sum_{\substack{n_1, \dots, n_k=1 \\ N_1, \dots, N_{k_p}=1 \\ M_1, \dots, M_{k_q}=1}}^s a_{in_1} \cdots a_{in_k} a_{iN_1} \cdots a_{iN_{k_p}} a_{iM_1} \cdots a_{iM_{k_q}} \mathcal{L}(t) \mathcal{K}(t),$$

where

$$\begin{aligned}
\Gamma(t) &= \rho(t) \frac{1}{\rho(\bar{u}_1)} \cdots \frac{1}{\rho(\bar{u}_l)} \frac{1}{\rho(\bar{u}_{q,1})} \cdots \frac{1}{\rho(\bar{u}_{q,l_q})}, \\
\mathcal{L}(t) &= \mathbf{l}_{n_1}(\bar{t}_1) \cdots \mathbf{l}_{n_k}(\bar{t}_k) \mathbf{l}_{N_1}(\bar{t}_{p,1}) \cdots \mathbf{l}_{N_{k_p}}(\bar{t}_{p,k_p}) \mathbf{l}_{M_1}(\bar{t}_{q,1}) \cdots \mathbf{l}_{M_{k_q}}(\bar{t}_{q,k_q}),
\end{aligned}$$

and

$$\mathcal{K}(t) = \mathbf{k}_i(\bar{u}_1) \cdots \mathbf{k}_i(\bar{u}_l) \mathbf{k}_i(\bar{u}_{q,1}) \cdots \mathbf{k}_i(\bar{u}_{q,l_q}).$$

Therefore,

$$(2.40) \quad \mathbf{l}_i(t) = \mathbf{l}_i(\bar{t}),$$

where

$$\begin{aligned}
\bar{t} &= [\bar{t}_1, \dots, \bar{t}_k, \bar{t}_{p,1}, \dots, \bar{t}_{p,k_p}, \bar{t}_{q,1}, \dots, \bar{t}_{q,k_q}, \\
&\quad \bar{u}_1, \dots, \bar{u}_l, \bar{u}_{q,1}, \dots, \bar{u}_{q,l_q}]_y \in \overline{\text{DA1T}}_{yy}.
\end{aligned}$$

It is obvious that this result is valid even if the tree t consists of more (or less) than one subtree $t \in \overline{\text{DA1T}}_{yz}$ and more (or less) than one $u \in \text{DA1T}_{zy}$. Similarly, if $u = [\bar{t}_1, \dots, \bar{t}_k, t_p]_z$, where $t_p = [\bar{u}_p]_y = [[\bar{t}_{p,1}, \dots, \bar{t}_{p,k_p}]_z]_y \in \overline{\text{DA1T}}_{yz}$ and $\bar{t}_1, \dots, \bar{t}_k, \bar{t}_{p,1}, \dots, \bar{t}_{p,k_p} \in \overline{\text{DA1T}}_{yy}$, then

$$(2.41) \quad \mathbf{k}_i(u) = \mathbf{k}_i(\bar{u}),$$

where



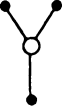


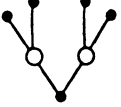




$$\bar{u} = [\bar{t}_1, \dots, \bar{t}_l, \bar{t}_{p,1}, \dots, \bar{t}_{p,l_p}]_z \in \overline{\text{DA1T}}_z.$$

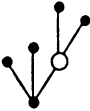

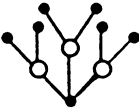

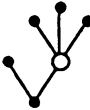

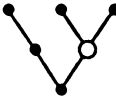
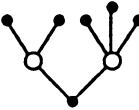
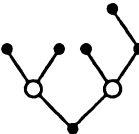
If $t = [u]_y \in \text{DA1T}_{yz}$, then $\bar{t} = [\bar{u}]_{yz}$. By repeated use of (2.38) and (2.41) we can show that for all $t \in \text{DA1T}_y$ and for all $u \in \text{DA1T}_z$ there exist corresponding simplified trees $\bar{t} \in \overline{\text{DA1T}}_y$ and $\bar{u} \in \overline{\text{DA1T}}_z$ such that $\mathbf{l}_i(t) = \mathbf{l}_i(\bar{t})$ and $\mathbf{k}_i(u) = \mathbf{k}_i(\bar{u})$. From (2.37) we have

$$\mathbf{l}_i(\bar{t}) = \mathbf{k}_i(\bar{u}) = \mathbf{k}_i(u) \quad \text{if } t = [u]_y \in \text{DA1T}_{yz}.$$

Inserting this into (2.31), we get $y_1(t) = y_1(\bar{t})$ and $z_1(u) = y_1(\bar{t})$, where $\bar{t} = [\bar{u}]_y \in \overline{\text{DAIT}}_{yz}$. Theorem 2.3 is proved. \square

In Figure 2 we give the order condition related to the trees up to order 4.

$\rho(t)$	t	$\Phi(t) = \frac{1}{\gamma(t)}$
1		$\sum_i b_i = 1$
2		$\sum_i b_i c_i = \frac{1}{2}$
2		$\sum_{ij} b_i d_{ij} c_j^2 = 1$
3		$\sum_i b_i c_i^2 = \frac{1}{3}$
3		$\sum_{ij} b_i c_i d_{ij} c_j^2 = \frac{2}{3}$
3		$\sum_{ijk} b_i d_{ij} c_j^2 d_{ik} c_k^2 = \frac{4}{3}$
3		$\sum_{ij} b_i a_{ij} c_j = \frac{1}{6}$
3		$\sum_{ij} b_i d_{ij} c_j^3 = 1$
3		$\sum_{ijk} b_i d_{ij} c_j a_{jk} c_k = \frac{1}{2}$
4		$\sum_i b_i c_i^3 = \frac{1}{4}$

$\rho(t)$	t	$\Phi(t) = \frac{1}{\gamma(t)}$
4		$\sum_{ij} b_i c_i^2 d_{ij} c_j^2 = \frac{1}{2}$
4		$\sum_{ijk} b_i c_i d_{ij} c_j^2 d_{ik} c_k^2 = 1$
4		$\sum_{ijkl} b_i d_{ij} c_j^2 d_{ik} c_k^2 d_{il} c_l^2 = 2$
4		$\sum_{ij} b_i c_i a_{ij} c_j = \frac{1}{8}$
4		$\sum_{ij} b_i c_i d_{ij} c_j^3 = \frac{3}{4}$
4		$\sum_{ijk} b_i c_i d_{ij} c_j a_{jk} c_k = \frac{3}{8}$
4		$\sum_{ijk} b_i a_{ij} c_j d_{ik} c_k^2 = \frac{1}{4}$
4		$\sum_{ijk} b_i d_{ij} c_j^2 d_{ik} c_k^3 = \frac{3}{2}$
4		$\sum_{ijkl} b_i d_{ij} c_j^2 d_{ik} c_k a_{kl} c_l = \frac{3}{4}$

$\rho(t)$	t	$\Phi(t) = \frac{1}{\gamma(t)}$
4		$\sum_{ij} b_i a_{ij} c_j^2 = \frac{1}{12}$
4		$\sum_{ijk} b_i a_{ij} c_j d_{jk} c_k^2 = \frac{1}{6}$
4		$\sum_{ijkl} b_i a_{ij} d_{jk} c_k^2 d_{jl} c_l^2 = \frac{1}{3}$
4		$\sum_{ijk} b_i a_{ij} a_{jk} c_k = \frac{1}{24}$
4		$\sum_{ij} b_i d_{ij} c_j^4 = 1$
4		$\sum_{ijk} b_i d_{ij} c_j^2 a_{jk} c_k = \frac{1}{2}$
4		$\sum_{ijk} b_i d_{ij} c_j a_{jk} c_k^2 = \frac{1}{3}$
4		$\sum_{ijkl} b_i d_{ij} c_j a_{jk} c_k d_{kl} c_l^2 = \frac{2}{3}$

$\rho(t)$	t	$\Phi(t) = \frac{1}{\gamma(t)}$
4		$\sum_{ijklm} b_i d_{ij} c_j a_{jk} d_{kl} c_l^2 d_{km} c_m^2 = \frac{4}{3}$
4		$\sum_{ijkl} b_i d_{ij} c_j a_{jk} a_{kl} c_l = \frac{1}{6}$
4		$\sum_{ijkl} b_i d_{ij} a_{jk} c_k a_{kl} c_l = \frac{1}{4}$

FIGURE 2. Order conditions related to trees.

3. CONVERGENCE RESULTS

This section deals with the convergence of a Runge-Kutta method applied to the differential-algebraic equation (1.6). The system is assumed to be a uniform index-1 problem. We first give some preliminary results about the existence of a solution of the Runge-Kutta equations, and the influence of perturbations to the solution of these equations. The convergence results are given in §3.2.

3.1. Preliminary results. The results of the next two theorems are essentially the same as those given by Hairer et al. [7] for the index-1 component of a Hessenberg form DAE of size 2.

3.1.1. *The existence of a Runge-Kutta solution.*

Theorem 3.1. *Let (ν, ζ) satisfy*

$$f(\nu, \zeta) = \mathcal{O}(h), \quad g(\nu) = \mathcal{O}(h^2).$$

Let the coefficients of the method satisfy

$$(3.1) \quad f(\tilde{\nu}_i, \zeta) = \mathcal{O}(h), \quad g(\tilde{\nu}_i) = \mathcal{O}(h^2), \quad i = 1, \dots, s,$$

where $\tilde{\nu}_i = \nu + c_i h \zeta$ and $c_i = \sum_{j=1}^s a_{ij}$. Suppose that

$$\left\| \begin{bmatrix} f_{v'} \\ g_v \end{bmatrix}^{-1} \right\| \leq M_1 \quad \text{and} \quad \|f_v\| \leq M_2$$

in an h -independent neighborhood of (ν, ζ) . If the coefficient matrix \mathcal{A} of the method is invertible, then there exists a solution of (1.2), (1.4) which satisfies

$$(3.2) \quad V_i' = \zeta + \mathcal{O}(h), \quad V_i = \nu + \mathcal{O}(h).$$

Proof. Consider the homotopy

$$(3.3) \quad \begin{aligned} f \left(\nu + h \sum_{j=1}^s a_{ij} V_j', V_i' \right) &= (1 - \tau) f(\tilde{\nu}_i, \zeta), \\ g \left(\nu + h \sum_{j=1}^s a_{ij} V_j' \right) &= (1 - \tau) g(\tilde{\nu}_i), \\ V_i &= \nu + h \sum_{j=1}^s a_{ij} V_j'. \end{aligned}$$

For $\tau = 0$, this system has the solution $V_i' = \zeta$, $V_i = \tilde{\nu}_i$. For $\tau = 1$ it is equivalent to (1.2), (1.4). We consider V_i' as functions of τ , and differentiate (3.3) with respect to this parameter:

$$\begin{aligned} f_{v'}(V_i, V_i') \dot{V}_i' + f_v(V_i, V_i') h \sum_{j=1}^s a_{ij} \dot{V}_j' &= -f(\tilde{\nu}_i, \zeta), \\ g_v(V_i) h \sum_{j=1}^s a_{ij} \dot{V}_j' &= -g(\tilde{\nu}_i), \end{aligned} \quad i = 1, \dots, s.$$

Dividing the second equation by h , the system can be written in matrix form as follows:

$$(3.4) \quad \begin{bmatrix} \{f_{v'}\} + h\{f_v\}(\mathcal{A} \otimes I_m) \\ \{g_v\}(\mathcal{A} \otimes I_m) \end{bmatrix} \dot{V}' = \begin{bmatrix} -\tilde{f}(\nu, \zeta) \\ -\frac{1}{h} \tilde{g}(\nu) \end{bmatrix}.$$

Here, $\{f_{v'}\}$, $\{f_v\}$, and $\{g_v\}$ are block-diagonal matrices:

$$\{f_{v'}\} = \text{blockdiag}[f_{v'}(V_1, V_1'), \dots, f_{v'}(V_s, V_s')],$$

etc. Furthermore, $\dot{V}' = [\dot{V}_1^T, \dots, \dot{V}_s^T]^T$, I_m is the $m \times m$ identity matrix, and $\tilde{f}(\nu, \zeta)$ and $\tilde{g}(\nu, \zeta)$ are

$$[f^T(\tilde{\nu}_1, \zeta), \dots, f^T(\tilde{\nu}_s, \zeta)]^T \quad \text{and} \quad [g^T(\tilde{\nu}_1), \dots, g^T(\tilde{\nu}_s)]^T,$$

resp. We have $g_v(V_i) a_{ij} = a_{ij} g_v(V_j) + \mathcal{O}(d)$, provided that $\|V_i - V_j\| \leq d$, d independent of h . Then the coefficient matrix of (3.4) can be written as

$$\begin{bmatrix} \{f_{v'}\} + \mathcal{O}(h) \\ (\mathcal{A} \otimes I_m) \{g_v\} + \mathcal{O}(d) \end{bmatrix},$$

and (3.4) becomes

$$(3.5) \quad \begin{bmatrix} \{f_{v'}\} + \mathcal{O}(h) \\ \{g_v\} + \mathcal{O}(d) \end{bmatrix} \dot{V}' = \begin{bmatrix} -\tilde{f}(\nu, \zeta) \\ -\frac{1}{h}(\mathcal{A} \otimes I_m)^{-1} \tilde{g}(\nu) \end{bmatrix}.$$

The matrix here has a bounded inverse, provided that d and h are sufficiently small. Using the assumption (3.1), we have $\dot{V}' = \mathcal{O}(h)$ and

$$(3.6) \quad V'_i = \zeta + \int_0^\tau \dot{V}'_i(t) dt = \zeta + \mathcal{O}(h).$$

This shows that all $V'_i(\tau)$ remain in a small, h -independent neighborhood of ζ for all $|\tau| \leq 1$ and for sufficiently small h . Hence, the differential equation (3.4) with initial values $V'(0) = \varepsilon_s \otimes \zeta$ possesses a solution at least for $0 \leq \tau \leq 1$. This proves the existence of a solution V'_i of (1.2) and also the first estimate of (3.2). The estimate for V_i follows directly from (1.4). \square

3.1.2. The influence of perturbations.

Theorem 3.2. Let V'_i , V_i be given by (1.2), (1.4) and consider perturbed values \hat{V}'_i , \hat{V}_i satisfying

$$(3.7) \quad \begin{aligned} f(\hat{V}_i, \hat{V}'_i) + \delta_i &= 0, \\ g(\hat{V}_i) + \theta_i &= 0, \\ \hat{V}_i &= \hat{\nu} + h \sum_{j=1}^s a_{ij} \hat{V}'_j, \end{aligned} \quad i = 1, \dots, s.$$

In addition to the hypotheses of Theorem 3.1, assume that

$$(3.8) \quad \hat{\nu} - \nu = \mathcal{O}(h^2), \quad \delta_i = \mathcal{O}(h), \quad \theta_i = \mathcal{O}(h^2).$$

Then, for $h \leq h_0$, we have the estimate

$$(3.9) \quad \|\hat{V}'_i - V'_i\|_\infty \leq C \left(\frac{1}{h} \|g_v(\nu) \cdot (\hat{\nu} - \nu)\|_\infty + \|f_v(\nu, \zeta) \cdot (\hat{\nu} - \nu)\|_\infty + \|\hat{\nu} - \nu\|_\infty + \|\delta\|_\infty + \frac{1}{h} \|\theta\|_\infty \right),$$

where $\delta = [\delta_1^T, \dots, \delta_s^T]^T$ and $\theta = [\theta_1^T, \dots, \theta_s^T]^T$.

Proof. Consider the homotopy

$$(3.10) \quad \begin{aligned} f(V_i, V'_i) + (1 - \tau)\delta_i &= 0, \\ g(V_i) + (1 - \tau)\theta_i &= 0, \\ V_i &= \nu + h \sum_{j=1}^s a_{ij} V'_j + (1 - \tau)(\hat{\nu} - \nu), \end{aligned} \quad i = 1, \dots, s.$$

For $\tau = 1$, this system is equivalent to (1.2), (1.4); for $\tau = 0$, it is equivalent to the perturbed system (3.7). As before, V_i and V'_i are considered as functions

of τ , and (3.10) is differentiated with respect to this parameter. We then obtain

$$\begin{aligned} f_{v'}(V_i, V_i')\dot{V}_i' + f_v(V_i, V_i')\dot{V}_i - \delta_i &= 0, \\ g_v(V_i)\dot{V}_i - \theta_i &= 0, \\ \dot{V}_i &= h \sum_{j=1}^s a_{ij}\dot{V}_j' - (\hat{\nu} - \nu), \end{aligned} \quad i = 1, \dots, s.$$

Inserting the last equation into the other two and using as before

$$g_v(V_i)a_{ij} = a_{ij}g_v(V_j) + \mathcal{O}(d),$$

we have

$$(3.11) \quad \begin{bmatrix} \{f_{v'}\} + \mathcal{O}(h) \\ \{g_v\} + \mathcal{O}(d) \end{bmatrix} \dot{V}' = \begin{bmatrix} \{f_v\}(\varepsilon_s \otimes (\hat{\nu} - \nu)) + \delta \\ \frac{1}{h}(\mathcal{A} \otimes I_m)^{-1}(\{g_v\}(\varepsilon_s \otimes (\hat{\nu} - \nu)) + \theta) \end{bmatrix}.$$

As in the proof of Theorem 3.1, the matrix is nonsingular for h and d sufficiently small, and we have

$$(3.12) \quad \begin{aligned} \|\dot{V}'\|_\infty &\leq C_1(\|f_v(\hat{\nu} - \nu)\|_\infty + \|\delta\|_\infty + h\|(\hat{\nu} - \nu)\|_\infty) \\ &\quad + \frac{C_2}{h}(\|g_v(\hat{\nu} - \nu)\|_\infty + \|\theta\|_\infty + h\|(\hat{\nu} - \nu)\|_\infty). \end{aligned}$$

The term $h\|(\hat{\nu} - \nu)\|_\infty$ comes from the $\mathcal{O}(h)$ term in the matrix of (3.11). Now, by (3.6), the assertion of the theorem is proved. \square

For the special case where $\delta = \theta = 0$ and $\hat{\nu} - \nu = \mathcal{O}(h)$ we have

$$(3.13) \quad \|\hat{V}_i' - V_i'\|_\infty \leq C_3.$$

3.2. The convergence of the methods. We now give a result concerning the growth of a perturbation in the solution computed in a single step of the method.

Lemma 3.1. *Let the assumptions of Theorem 3.1 be satisfied. Suppose that $v_n = v(x_n) + e_n$, where $\|e_n\| = \mathcal{O}(h^{k_G})$. Let v_{n+1} be the solution of*

$$(3.14) \quad \begin{aligned} f\left(v_n + h \sum_{j=1}^s a_{ij}\tilde{V}_j', \tilde{V}_i'\right) &= 0, \\ g\left(v_n + h \sum_{j=1}^s a_{ij}\tilde{V}_j'\right) &= 0, \\ v_{n+1} &= v_n + h \sum_{i=1}^s b_i\tilde{V}_i'. \end{aligned} \quad i = 1, \dots, s.$$

Then

$$e_{n+1} = \left(I_m - (b^T \mathcal{A}^{-1} \varepsilon_s) \begin{bmatrix} f_{v'} \\ g_v \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ g_v \end{bmatrix} \right) e_n + \tilde{\eta} + d_{n+1},$$

where

$$\tilde{\eta} = \begin{cases} \mathcal{O}(h^{k_G+1}) & \text{if } k_G \geq 2, \\ \mathcal{O}(h^{k_G+1}) & \text{if } k_G = 1 \text{ and } f \text{ is linear in } v', \\ \mathcal{O}(h) & \text{if } k_G = 1 \text{ and } f \text{ is nonlinear in } v', \end{cases}$$

and d_{n+1} is the local truncation error.

Proof. Let $\tilde{V}'_i = V'_i + \Delta V'_i$, where V'_i is the solution of (3.14) if $v_n = v(x_n)$. The equation (3.14) can be written as

$$(3.15) \quad \begin{aligned} f \left(V_i + e_n + h \sum_{j=1}^s a_{ij} \Delta V'_j, V'_i + \Delta V'_i \right) &= 0, \\ g \left(V_i + e_n + h \sum_{j=1}^s a_{ij} \Delta V'_j \right) &= 0, \end{aligned} \quad i = 1, \dots, s,$$

where $V_i = v(x_n) + h \sum_{j=1}^s a_{ij} V'_j$. Theorem 3.1 shows that $V_i = v(x_n) + \mathcal{O}(h)$ and $V'_i = v'(x_n) + \mathcal{O}(h)$. We assume that (3.14) is solved exactly. From Theorem 3.2 with $\hat{v} - v = e_n$ we have that $\|\Delta V'_i\| \leq Ch$ for $k_G \geq 2$, while $\|\Delta V'_i\| \leq C_3$ if $k_G = 1$ (see (3.13)). Let $\|\Delta V'_i\| \leq \delta$ and $\|e_n\| \leq \varepsilon$. In the following, if nothing else is said, all functions and their derivatives are evaluated in $(x_n, v(x_n))$. By expanding (3.15) in a Taylor series around (V_i, V'_i) we have

$$(3.16) \quad f(V_i, V'_i) + f_{v'}(V_i, V'_i) \Delta V'_i + \eta_f = 0,$$

where η_f is the sum of higher-order terms, composed of terms of the form

$$\begin{aligned} & f_v \cdot \left(e_n + h \sum_{j=1}^s a_{ij} \Delta V'_j \right), \\ & f_{vv'} \cdot \left(e_n + h \sum_{j=1}^s a_{ij} \Delta V'_j, \Delta V'_i \right), \\ & f_{v'v'} \cdot (\Delta V'_i, \Delta V'_i). \end{aligned}$$

Thus, we find that

$$\eta_f = \mathcal{O}(\varepsilon + h\delta) + \mathcal{O}(\varepsilon\delta + h\delta^2) + \mathcal{O}(\delta^2).$$

Using that $f_{v'}(V_i, V'_i) = f_{v'} + \mathcal{O}(h)$, we have from (3.16)

$$(3.17) \quad f_{v'} \Delta V'_i = \mathcal{O}(\varepsilon + h\delta + \varepsilon\delta + h\delta^2 + \delta^2).$$

Similarly,

$$g(V_i) + (g_v + \mathcal{O}(h)) \left(e_n + h \sum_{j=1}^s a_{ij} \Delta V'_j \right) + \eta_g = 0,$$

where the higher-order term η_g is composed of terms of the form

$$g_{vv} \cdot \left(e_n + h \sum_{j=1}^s a_{ij} \Delta V_j' \right)^2,$$

so that

$$\eta_g = \mathcal{O}(\varepsilon^2 + h\delta\varepsilon + h^2\delta^2).$$

Thus, we find that

$$(3.18) \quad g_v \Delta V_i' = -\frac{1}{h} \left(\sum_{j=1}^s d_{ij} \right) g_v e_n + \mathcal{O} \left(\varepsilon + h\delta + \frac{\varepsilon^2}{h} + \delta\varepsilon + h\delta^2 \right).$$

Equations (3.17) and (3.18) can be solved for $\Delta V_i'$:

$$(3.19) \quad \Delta V_i' = \begin{bmatrix} f_{v'} \\ g_v \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ -\frac{1}{h} \sum_{j=1}^s d_{ij} g_v \end{bmatrix} e_n + \mathcal{O} \left(\varepsilon + h\delta + \varepsilon\delta + h\delta^2 + \frac{\varepsilon^2}{h} + \delta^2 \right).$$

To find a lower bound for $\Delta V_i'$, we use the same strategy as in [1 and 10]. We have

$$\|\Delta V_i'\| \leq K \left(\frac{\varepsilon}{h} + \varepsilon + h\delta + \varepsilon\delta + h\delta^2 + \frac{\varepsilon^2}{h} + \delta^2 \right).$$

Let δ be the solution of the equation

$$(3.20) \quad \delta = K \left(\frac{\varepsilon}{h} + \varepsilon + h\delta + \varepsilon\delta + h\delta^2 + \frac{\varepsilon^2}{h} + \delta^2 \right).$$

Let $\varepsilon = \mathcal{O}(h^{k_G})$, where $k_G \geq 2$, and solve (3.7) by functional iteration $\delta = G(\delta)$ with the initial value $\delta^{(0)} = k_1 h^{k_G-1}$. Then $\delta^{(1)} = G(\delta^{(0)}) = \mathcal{O}(h^{k_G-1})$ and $|\partial G / \partial \delta| = \mathcal{O}(h) \leq 1$ for h sufficiently small. We can now use the Banach Fixed Point Theorem to conclude that the iteration converges to a solution satisfying

$$\delta = \mathcal{O}(h^{k_G-1})$$

for $k_G \geq 2$. Inserting this into (3.19), we have

$$(3.21) \quad \Delta V_i' = \begin{bmatrix} f_{v'} \\ g_v \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ -\frac{1}{h} \sum_{j=1}^s d_{ij} g_v \end{bmatrix} e_n + \mathcal{O}(h^{k_G}) \quad \text{for } k_G \geq 2.$$

If $k_G = 1$, then $\delta = \mathcal{O}(1)$, so that

$$(3.22) \quad \Delta V_i' = \begin{bmatrix} f_{v'} \\ g_v \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ -\frac{1}{h} \sum_{j=1}^s d_{ij} g_v \end{bmatrix} e_n + \mathcal{O}(1) \quad \text{for } k_G = 1.$$

The term $\mathcal{O}(1)$ comes from terms of the form

$$f_{v'v'}(\Delta V_i', \Delta V_i').$$

These terms do not appear in problems linear in v' . We conclude that (3.21) is valid for such problems, even for $k_G = 1$. Inserting (3.21) or (3.22) into (3.14), we have

$$\begin{aligned} v_{n+1} &= v_n + h \sum_{i=1}^s b_i (V'_i + \Delta V'_i) \\ &= v(x_n) + h \sum_{i=1}^s b_i V'_i + e_n - \sum_{i,j=1}^s b_i d_{ij} \begin{bmatrix} f_{v'} \\ g_v \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ g_v \end{bmatrix} e_n + \tilde{\eta}, \end{aligned}$$

where $\tilde{\eta} = \mathcal{O}(h^{k_G+1})$ for $k_G \geq 2$, or if $k_G = 1$ and f is linear in v' . If $k_G = 1$ and f is nonlinear in v' , then $\tilde{\eta} = \mathcal{O}(h)$. From

$$v_{n+1} = v(x_n + h) + d_{n+1} + \left(I_m - (b^T \mathcal{A}^{-1} \varepsilon_s) \begin{bmatrix} f_{v'} \\ g_v \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ g_v \end{bmatrix} \right) e_n + \tilde{\eta}$$

we thus have

$$e_{n+1} = \left(I_m - (b^T \mathcal{A}^{-1} \varepsilon_s) \begin{bmatrix} f_{v'} \\ g_v \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ g_v \end{bmatrix} \right) e_n + \tilde{\eta} + d_{n+1}. \quad \square$$

We now consider the global error resulting from repeated use of the method.

Theorem 3.3. *Suppose the index 1 problem (1.6) is solved numerically by an s -stage Runge-Kutta method. Assume that*

1. f, g are sufficiently differentiable.
2. The initial values satisfy $\|v_0 - v(x_0)\| \leq \mathcal{O}(h^{k_G})$ and $f(v_0, v'_0) = \mathcal{O}(h)$.
3. (a) $k_G \geq 2$ for problems nonlinear in v' ;
(b) $k_G \geq 1$ for problems linear in v' , provided that $f(v_n, v'_n) = \mathcal{O}(h)$ and $g(v_n) = \mathcal{O}(h^2)$.
4. The local truncation error satisfies
(a) $P_n d_{n+1} = \mathcal{O}(h^{k_G+1})$, $S_n d_{n+1} = \mathcal{O}(h^{k_G})$ if $|1 - b^T \mathcal{A}^{-1} \varepsilon_s| < 1$;
(b) $P_n d_{n+1} = \mathcal{O}(h^{k_G+1})$, $S_n d_{n+1} = \mathcal{O}(h^{k_G+1})$ if $|1 - b^T \mathcal{A}^{-1} \varepsilon_s| = 1$,
where

$$S_n = \begin{bmatrix} f_{v'}(v_n, v'_n) \\ g_v(v_n) \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ g_v(v_n) \end{bmatrix}, \quad P_n = I_m - S_n.$$

Then the global error is at least of order k_G .

Before we prove this theorem, we comment on assumptions 2 and 3(b). The relations

$$(3.23) \quad f(v_n, v'_n) = \mathcal{O}(h) \quad \text{and} \quad g(v_n) = \mathcal{O}(h^2)$$

have to be satisfied to ensure a solution of the system (1.2). The relations (3.23) hold for the initial step by assumption 2. For $k_G \geq 2$, (3.23) is satisfied if $v'_n = v'(x) + \mathcal{O}(h)$, which is easily obtained, for example by using $v'_{n+1} = V'_s$.

For $k_G = 1$ we have to make these assumptions explicitly. All Runge-Kutta methods with $b_i = a_{si}$ satisfy assumption (3.23).

Proof of Theorem 3.3. The interval of integration, $[x_0, x_{\text{end}}]$, can be split in N subintervals, $H_k = [X_k, X_{k+1}]$, $k = 0, \dots, N-1$, $X_0 = x_0$, and $X_N = x_{\text{end}}$. We assume each X_k to coincide with some x_n . In each subinterval, there exists a constant permutation matrix $Q(x)$, $Q(x)v = [y^T, z^T]^T$ ensuring g_z to be nonsingular over each subinterval. The subintervals are chosen so that $Q(x)$ has to be changed over two adjacent subintervals. First, we will prove the theorem over the first subinterval, H_0 , and we assume the variables to be numbered such that $Q(x) = I_m$ in H_0 . Thus, we have

$$\begin{aligned} \begin{bmatrix} f_{y'} \\ g_y \end{bmatrix}^{-1} &= \begin{bmatrix} f_{y'} & f_{z'} \\ g_y & g_z \end{bmatrix}^{-1} \\ &= \begin{bmatrix} (f_{y'} - f_{z'} g_z^{-1} g_y)^{-1} & -(f_{y'} - f_{z'} g_z^{-1} g_y)^{-1} f_{z'} g_z^{-1} \\ -g_z^{-1} g_y (f_{y'} - f_{z'} g_z^{-1} g_y)^{-1} & g_z^{-1} g_y (f_{y'} - f_{z'} g_z^{-1} g_y)^{-1} f_{z'} g_z^{-1} + g_z^{-1} \end{bmatrix}. \end{aligned}$$

Inserting this into the expression for the global error in Lemma 3.1, we have

$$(3.24) \quad e_{n+1} = (I_m - (b^T \mathcal{A}^{-1} \varepsilon_s) S(y, z)) e_n + \tilde{\eta}_{n+1} + d_{n+1},$$

where

$$S(y, z) = \begin{bmatrix} -F_P f_{z'} g_z^{-1} g_y & -F_P f_{z'} \\ g_z^{-1} g_y F_P f_{z'} g_z^{-1} g_y + g_z^{-1} g_y & g_z^{-1} g_y F_P f_{z'} + I_{m-r} \end{bmatrix}$$

and $F_P = (f_{y'} - f_{z'} g_z^{-1} g_y)^{-1}$. Thus, (3.24) can be written as

$$(3.25) \quad e_{n+1} = P_n e_n + r S_n e_n + \tilde{d}_{n+1},$$

where $P = I_m - S$, $S_n = S(y_n, z_n)$, $P_n = P(y_n, z_n)$, $\tilde{d}_{n+1} = d_{n+1} + \tilde{\eta}_{n+1}$, and $r = 1 - b^T \mathcal{A}^{-1} \varepsilon_s$. For each subinterval, S_n and P_n are projection operators satisfying $S_n^2 = S_n$ and $P_n^2 = P_n$. In fact, $P(y, z)$ represents a projection into the tangent space of the surface given by $g(y, z) = 0$, because $[g_y, g_z]P \equiv 0$. The components of P and S are smooth, so that

$$\begin{aligned} S_{n+1} S_n &= S_n + \mathcal{O}(h), & P_{n+1} P_n &= P_n + \mathcal{O}(h), \\ S_{n+1} P_n &= \mathcal{O}(h), & P_{n+1} S_n &= \mathcal{O}(h). \end{aligned}$$

Multiplying (3.25) once by P_{n+1} and once by S_{n+1} , we obtain

$$(3.26) \quad \begin{bmatrix} \|P_{n+1} e_{n+1}\| \\ \|S_{n+1} e_{n+1}\| \end{bmatrix} \leq \begin{bmatrix} 1 + \mathcal{O}(h) & \mathcal{O}(h) \\ \mathcal{O}(h) & |r|(1 + \mathcal{O}(h)) \end{bmatrix} \begin{bmatrix} \|P_n e_n\| \\ \|S_n e_n\| \end{bmatrix} + \begin{bmatrix} \|P_{n+1} \tilde{d}_{n+1}\| \\ \|S_{n+1} \tilde{d}_{n+1}\| \end{bmatrix}.$$

Suppose that $\|P_{n+1} \tilde{d}_{n+1}\| \leq D_p$ and $\|S_{n+1} \tilde{d}_{n+1}\| \leq D_s$. Then (3.26) can be written as

$$(3.27) \quad \begin{bmatrix} \|P_n e_n\| \\ \|S_n e_n\| \end{bmatrix} \leq A^n(h) \begin{bmatrix} \|P_0 e_0\| \\ \|S_0 e_0\| \end{bmatrix} + \sum_{i=1}^n A^{n-i}(h) \begin{bmatrix} D_p \\ D_s \end{bmatrix},$$

where

$$A(h) = \begin{bmatrix} 1 + \mathcal{O}(h) & \mathcal{O}(h) \\ \mathcal{O}(h) & |r|(1 + \mathcal{O}(h)) \end{bmatrix}.$$

The matrix $A(h)$ can be diagonalized by a matrix $T(h)$ such that

$$T(h)A(h)T^{-1}(h) = \begin{bmatrix} 1 + \mathcal{O}(h) & 0 \\ 0 & |r|(1 + \mathcal{O}(h)) \end{bmatrix},$$

where

$$T(h) = \begin{bmatrix} 1 + \mathcal{O}(h) & \mathcal{O}(h) \\ \mathcal{O}(h) & 1 + \mathcal{O}(h) \end{bmatrix}.$$

The inequality (3.27) can be written as

$$\begin{aligned} \begin{bmatrix} \|P_n e_n\| \\ \|S_n e_n\| \end{bmatrix} &\leq T^{-1}(h) \begin{bmatrix} (1 + K_1 h)^n & 0 \\ 0 & |r|^n (1 + K_2 h)^n \end{bmatrix} T(h) \begin{bmatrix} \|P_0 e_0\| \\ \|S_0 e_0\| \end{bmatrix} \\ &+ T^{-1}(h) \sum_{i=1}^n \begin{bmatrix} (1 + K_1 h)^{n-i} & 0 \\ 0 & |r|^{n-i} (1 + K_2 h)^{n-i} \end{bmatrix} T(h) \begin{bmatrix} D_p \\ D_s \end{bmatrix}. \end{aligned}$$

By direct computation we have for $|r| < 1$

$$\begin{bmatrix} \|P_n e_n\| \\ \|S_n e_n\| \end{bmatrix} \leq \begin{bmatrix} C_1 \|P_0 e_0\| + \mathcal{O}(h) \|S_0 e_0\| \\ \mathcal{O}(h) \|P_0 e_0\| + C_2 (|r|^n + h) \|S_0 e_0\| \end{bmatrix} + \begin{bmatrix} \frac{1}{h} C_3 D_p + C_4 D_s \\ C_5 D_p + C_6 D_s \end{bmatrix}$$

and for $|r| = 1$

$$\begin{bmatrix} \|P_n e_n\| \\ \|S_n e_n\| \end{bmatrix} \leq \begin{bmatrix} C_1 \|P_0 e_0\| + \mathcal{O}(h) \|S_0 e_0\| \\ \mathcal{O}(h) \|P_0 e_0\| + C_2 \|S_0 e_0\| \end{bmatrix} + \begin{bmatrix} \frac{1}{h} C_3 D_p + C_4 D_s \\ \frac{1}{h} C_5 D_p + C_6 D_s \end{bmatrix}.$$

By using the fact that $\|e_n\| \leq \|P_n e_n\| + \|S_n e_n\|$ we obtain

$$(3.28) \quad \|e_n\| \leq \begin{cases} C(\|P_0 e_0\| + (|r|^n + h)\|S_0 e_0\| + \frac{1}{h} D_p + D_s) & \text{for } |r| < 1, \\ C(\|e_0\| + \frac{1}{h}(D_p + D_s)) & \text{for } |r| = 1. \end{cases}$$

Thus, if $e_0 = \mathcal{O}(h^{k_G})$, $D_s = \mathcal{O}(h^{k_G+1})$, and $D_p = \mathcal{O}(h^{k_G})$ for $|r| < 1$, or $D_s = \mathcal{O}(h^{k_G+1})$ for $|r| = 1$, we have that the global error at the end of the first subinterval is given by

$$(3.29) \quad E_1 = e_n = \mathcal{O}(h^{k_G}).$$

In that case, the global error after $k+1$ subintervals is bounded by

$$\|E_{k+1}\| \leq G\|E_k\| + Dh^{k_G},$$

and the error at the end of the interval of integration is bounded by

$$\|E_N\| \leq C^N \|e_0\| + \left(\sum_{i=1}^N C^{N-i} D \right) h^{k_G}. \quad \square$$

Theorem 3.4. *Let assumptions 1, 2, and 3 of Theorem 3.3 be satisfied. Then, if $|1 - b^T \mathcal{A}^{-1} \varepsilon_s| < 1$, and*

$$\Phi(t) = \frac{1}{\gamma(t)} \quad \forall t \in \begin{cases} \overline{\text{DAIT}}_{yy}, & \rho(t) \leq k_G, \\ \overline{\text{DAIT}}_{yz}, & \rho(t) \leq k_G - 1, \end{cases}$$

the order of the global error is k_G .

Proof. We first prove this theorem for the model equation (2.1). The local truncation error for a method with local order k_L applied to this problem is given by

$$d_n = \left[\begin{array}{c} \sum_{\substack{t \in \text{DA1T}_{yy} \\ \rho(t)=k_L}} (\gamma(t)\Phi(t) - 1)\alpha(t)F(t) \frac{h^{\rho(t)}}{\rho(t)!} \\ \sum_{\substack{u \in \text{DA1T}_{zz} \\ \rho(u)=k_L}} (\gamma(u)\Phi(u) - 1)\alpha(u)G(u) \frac{h^{\rho(u)}}{\rho(u)!} \end{array} \right] + \mathcal{O}(h^{k_L+1}).$$

The first term in this expression is called the *principal error term*. The principal error term can also be written as

$$\left[\begin{array}{cc} (-f_{z'}g_y)^{-1} & -(-f_{z'}g_y)^{-1}f_{z'} \\ g_y(-f_{z'}g_y)^{-1} & -g_y(-f_{z'}g_y)^{-1}f_{z'} + I_{n-r} \end{array} \right] \cdot \left[\begin{array}{c} \sum_{\substack{t \in \text{DA1T}_{yy} \\ \rho(t)=k_L}} (\gamma(t)\Phi(t) - 1)\alpha(t)\tilde{F}(t) \frac{h^{\rho(t)}}{\rho(t)!} \\ \sum_{\substack{u \in \text{DA1T}_{zz} \\ \rho(u)=k_L}} (\gamma(u)\Phi(u) - 1)\alpha(u)G(u) \frac{h^{\rho(u)}}{\rho(u)!} \end{array} \right],$$

where $\tilde{F}(t) = f_{ky|z'}(F(t_1), \dots, F(t_k), G(u_1), \dots, G(u_l))$ for $t \in [t_1, \dots, t_k, u_1, \dots, u_l]_y$, and $F(t)$ and $G(u)$ are given by Definition 2.8. For the model problem, $S(y, z)$ and $P(y, z)$ are given by

$$S = \left[\begin{array}{cc} I_r & -(-f_{z'}g_y)^{-1}f_{z'} \\ 0 & I_{m-r} - g_y(-f_{z'}g_y)^{-1}f_{z'} \end{array} \right], \quad P = \left[\begin{array}{cc} 0 & (-f_{z'}g_y)^{-1}f_{z'} \\ 0 & g_y(-f_{z'}g_y)^{-1}f_{z'} \end{array} \right].$$

Multiplying d_n by P_n , we obtain

$$(3.30) \quad P_n d_n = \left[\begin{array}{cc} I_r & 0 \\ g_y & 0 \end{array} \right] \left[\begin{array}{c} \sum_{\substack{t \in \text{DA1T}_{yy} \\ \rho(t)=k_L}} (\gamma(t)\Phi(t) - 1)\alpha(t)F(t) \frac{h^{\rho(t)}}{\rho(t)!} \\ \sum_{\substack{u \in \text{DA1T}_{zz} \\ \rho(u)=k_L}} (\gamma(u)\Phi(u) - 1)\alpha(u)G(u) \frac{h^{\rho(u)}}{\rho(u)!} \end{array} \right] + \mathcal{O}(h^{k_L+1}).$$

Let $\Phi(t) = 1/\gamma(t)$ for all $t \in \text{DA1T}_{yy}$, $\rho(t) = k_L$, and $\Phi(t) \neq 1/\gamma(t)$ for at least one $t \in \text{DA1T}_{zz}$, $\rho(t) = k_L$. By (3.30) we have that $P_n d_n = \mathcal{O}(h^{k_L+1})$, while $S_n d_n = \mathcal{O}(h^{k_L})$. From Theorem 3.3 we then know that the global order of the method is k_L . Now, the assertion of the theorem follows directly from

the fact that the set of simplified trees associated to DAIT_{yy} , resp. DAIT_{zz} , is $\overline{\text{DAIT}}_{yy}$, resp. $\overline{\text{DAIT}}_{yz}$.

This result can be extended also to include the more general equation (1.6). For this equation, the principal error is composed of terms of the form

$$\begin{bmatrix} f_{y'} & f_{z'} \\ g_y & g_z \end{bmatrix}^{-1} \begin{bmatrix} f_{kylz} \hat{k} y' \hat{l} z' (y^{(p_1)}, \dots, y^{(p_{k+k})}, z^{(q_1)}, \dots, z^{(q_{l+i})}) \\ g_{kylz} (y^{(p_1)}, \dots, y^{(p_k)}, z^{(q_1)}, \dots, z^{(q_k)}) \end{bmatrix} h^{k_L},$$

where $y^{(p)}$ and $z^{(q)}$ are derivatives of y and z of order less than k_L . We can show that

$$P \begin{bmatrix} f_{y'} & f_{z'} \\ g_y & g_z \end{bmatrix}^{-1} = \begin{bmatrix} (f_{y'} - f_{z'} g_z^{-1} g_y)^{-1} & 0 \\ -g_z^{-1} g_y (f_{y'} - f_{z'} g_z^{-1} g_y)^{-1} & 0 \end{bmatrix}.$$

The projection operator P suppresses all the contributions to the local truncation error coming from derivatives of the algebraic equation. Since the algebraic equation in (1.6) and the model problem are equivalent, assuming g_z nonsingular in (1.6), the conclusion for the model equation is also valid for the general problem. \square

4. NUMERICAL EXPERIMENTS

In this section we present the results of some numerical experiments on various index 1 systems. The experiments confirm that the local order predicted in §2 occurs in practice, and in no case is the observed global order less than the lower bound given in Theorems 3.3 and 3.4. The experiments below were performed in single precision on a Cray X-MP computer. The test problems are:

- P1: A linear constant-coefficient system of the form $Av' + Bv = g(x)$.
- P2: A linear system with time-dependent coefficients, $A(x)v' + B(x)v = g(x)$.
- P3: A nonlinear system, linear in v' , $A(v, x)v' = f(v, x)$.
- P4: A system nonlinear also in v' , $f(v, v', x) = 0$.

The problems are described in Appendix A. The test problems were solved by the following Runge-Kutta methods:

- M1: 2-stage, 3rd-order, A -stable SDIRK method (Nørsett [11]).
- M2: 3-stage, 3rd-order, B -stable SDIRK method (Nørsett and Thomsen [12]).
- M3: 5-stage, 4th-order, strongly S -stable SDIRK method (Cash [4]).
- M4: 7-stage, 3rd-order, extrapolation method, based on fully implicit backward Euler, written as a DIRK method.
- M5: 2-stage, 2nd-order, Lobatto IIIC method (Chipman [5]).
- M6: 3-stage, 4th-order, Lobatto IIIC method (Chipman [5]).
- M7: 3-stage, 5th-order, Radau IA method (Butcher [3, p. 228]).
- M8: 2-stage, 4th-order, Kuntzmann-Butcher method (Butcher [3, p. 219]).

M9: 3-stage, 6th-order, Kuntzmann-Butcher method (Butcher [3, p. 220]).

The results of the experiments are given in Tables 4.1–4.4. The following notations have been used:

- k_d : The order of the method when applied to an ODE.
- k_p : The order of the global error, predicted by Burrage and Petzold in [2, Theorem 1].
- k_L : Predicted order of the local error.
- k_l : Observed order of the local error.
- k_G : Predicted order of the global error.
- k_g : Observed order of the global error.

TABLE 4.1. Predicted/Observed Orders for P1.

Method	k_d	k_p	k_L	k_l	k_G	k_g
M1	3	2	2	2	2	2
M2	3	2	2	2	2	2
M3	4	2	3	5	2	4
M4	3	2	4	4	3	3
M5	2	2	3	3	2	2
M6	4	3	5	5	4	4
M7	5	3	3	3	3	3
M8	4	2	3	3	2	2
M9	6	4	4	4	3	4

TABLE 4.2. Predicted/Observed Orders for P2.

Method	k_d	k_p	k_L	k_l	k_G	k_g
M1	3	2	2	2	2	2
M2	3	2	2	2	2	2
M3	4	2	3	5	2	4
M4	3	2	4	4	3	3
M5	2	2	3	3	2	2
M6	4	3	5	5	4	4
M7	5	3	3	3	3	3
M8	4	2	3	3	2	2
M9	6	4	4	4	3	4

TABLE 4.3. Predicted/Observed Orders for P3.

Method	k_d	k_p	k_L	k_l	k_G	k_g
M1	3	2	2	2	2	2
M2	3	2	2	2	2	2
M3	4	2	3	3	2	2
M4	3	2	4	4	3	3
M5	2	2	3	3	2	2
M6	4	3	5	5	4	4
M7	5	3	3	3	3	3
M8	4	2	3	3	2	2
M9	6	4	4	4	3	4

TABLE 4.4. Predicted/Observed Orders for P4.

Method	k_d	k_P^a	k_L	k_l	k_G	k_g
M1	3	-	2	2	2	2
M2	3	-	2	2	2	2
M3	4	-	3	3	2	2
M4	3	-	4	4	3	3
M5	2	-	3	3	2	2
M6	4	-	5	5	4	4
M7	5	-	3	3	3	3
M8	4	-	3	3	2	2
M9	6	-	4	4	3	4

^a Theorem 1 in [2] gives no lower bound for problems nonlinear in v' .

In no case is the observed local and global order lower than the predicted one. However, for P1, the order observed is higher than expected for several methods. The reason is the simplicity of the problem. For linear, constant-coefficients systems, the only order conditions which have to be satisfied, in addition to the classical ODE-conditions, are

$$b^T \mathcal{A}^{-1} c^j = 1, \quad j = 1, \dots, q,$$

where $q = k_G - 1$ if $|r| < 1$, or $q = k_G$ if $|r| = 1$. The Cash method M3 solves problem P2 with a higher than expected order for similar reasons. The elementary differentials causing the method to be reduced to an order-2 method are not present in this problem.

For the rest of the problems, there is agreement between the observed and predicted local order. There is also agreement between the observed and predicted global order, with one exception, the Kuntzmann-Butcher method M9. In this case, the stability constant $r = 1 - b^T \mathcal{A}^{-1} \varepsilon_s = -1$. The contribution to the global error from the local error of two adjacent steps is $d_{n+1} + r d_n = \mathcal{O}(h^{k_L+1})$, that is, the local errors from two adjacent steps cancel each other. See [2] for a better explanation of this phenomenon. M9 is also the only method where the order predicted by Burrage and Petzold is better than the order predicted by our theory.

5. DISCUSSION

In this paper we have derived a set of necessary and sufficient order conditions for the local truncation error when a Runge-Kutta method is applied to a fully implicit differential-algebraic equation of index-1. The numerical experiments, described in §4, confirm our results.

In Appendix B, the results given in this paper are compared with the results given by Burrage and Petzold [2, 13]. We observe that there is no conflict between the two theories. Let $Q(x)$ be a permutation matrix, ensuring

g_z to be nonsingular. The algebraic part of the equation is $g(v) = 0$, and $v = Q^{-1}(x)[y^T, z^T]^T$. The order conditions derived in §2 are only valid if $Q(x)$ does not change over the step. The theory of Petzold does not have this restriction.

If $Q(x)$ does not vary too frequently, it is possible to use embedding for the control of the local error. In practice, one should be careful when choosing a method for solving general classes of index-1 equations. Some methods will attain a higher order for some classes of equations, like semiexplicit equations and linear constant-coefficient equations. This is observed in §4, Table 4.1. For the linear constant-coefficient problem, the two variables were solved with different accuracy. For the methods with $r = 0$, v_1 was solved with order k_d , while v_2 was solved exactly. When choosing a method for solving DAE's, such aspects have to be considered.

ACKNOWLEDGMENTS

The author thanks her thesis adviser, Syvert P. Nørsett, for encouragement and guidance. She is also grateful to Michel Roche and the unknown referees for their suggestions, which improved the original convergence results.

APPENDIX A. TEST PROBLEMS

P1: Linear constant-coefficient problem.

$$\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} v' + \begin{bmatrix} 1 & 2 \\ 2 & 5 \end{bmatrix} v = \begin{bmatrix} 0 \\ \sin x \end{bmatrix}$$

with $x \in [0, 1]$ and initial values

$$v(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{and} \quad v'(0) = \begin{bmatrix} -3 \\ 1 \end{bmatrix}.$$

The exact solution is

$$v_1(x) = e^{-x} - 2 \sin x, \quad v_2(x) = \sin x.$$

The local order was derived at $x = 0.5$.

P2: Linear problem with time-dependent coefficients.

$$\begin{aligned} (x+1)v_1' + (x+1)v_2' + xv_1 - 0.5v_2 &= e^{-x}, \\ (x^2 - 1.3^2)v_1 + (x^2 - 0.3^2)v_2 &= (x^2 - 1.3^2)xe^{-x} + (x^2 - 0.3^2)\sqrt{x+1} \end{aligned}$$

with $x \in [0, 1]$ and initial values

$$v(0) = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \text{and} \quad v'(0) = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}.$$

The exact solution is

$$v_1(x) = xe^{-x}, \quad v_2(x) = \sqrt{x+1}.$$

The local order was derived at $x = 0.28$.

P3: Nonlinear problem, linear in v' .

$$\begin{aligned} v_1' + v_3 v_2' - (v_2 + 1)v_3' &= -v_1 + 1 + \sin x, \\ (v_3 + 1)v_1' + v_1 v_2' &= -e^{-x}, \\ 0 &= v_1 v_2 v_3 - 0.5e^{-x} \sin(2x) \end{aligned}$$

with $x \in [0, 1]$ and initial values

$$v(0) = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \quad \text{and} \quad v'(0) = \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}.$$

The exact solution is

$$v_1(x) = e^{-x}, \quad v_2(x) = \sin x, \quad v_3(x) = \cos x.$$

The local order was derived at $x = 0.5$.

P4: Problem nonlinear in v' .

$$\begin{aligned} (\sin^2 v_1' + \cos^2 v_2')(v_2')^2 - (x-6)^2(x-2)^2 v_1 e^{-x} &= 0, \\ (4-x)(v_2 + v_1)^3 - 64x^2 e^{-x} v_1 v_2 &= 0 \end{aligned}$$

with $x \in [0.5, 1]$. The exact solution is

$$v_1 = x^4 e^{-x}, \quad v_2 = x^3 e^{-x} (4-x).$$

The local order was derived at $x = 0.75$.

APPENDIX B. COMPARISON BETWEEN OUR RESULTS AND THE ORDER RESULTS GIVEN BY PETZOLD

Here we want to explain why there is no contradiction between the order results given by Petzold et al. [2, 13] and the results given in this paper. In fact, we can prove that $k_G \geq k_p$, k_p is the order predicted by Petzold in the theorem cited below, for all methods with $r \neq -1$. The following relations are defined in [2]:

$$\begin{aligned} A(w): \sum_{i,j=1}^s b_i d_{ij} c_j^p &= 1, \quad p = 1, \dots, w, \\ B(w): \sum_{i=1}^s b_i c_i^{p-1} &= \frac{1}{p}, \quad p = 1, \dots, w, \\ C(w): \sum_{j=1}^s a_{ij} c_j^{p-1} &= \frac{c_i^p}{p}, \quad i = 1, \dots, s, p = 1, \dots, w. \end{aligned}$$

Theorem 1 in [2] is as follows.

Theorem 1. Suppose that (1.1) is uniform index 1 and linear in v' , the Runge-Kutta method satisfies the stability condition $|r| \leq 1$, the errors in the initial conditions are $\mathcal{O}(h^{k_p})$, and the errors in terminating the Newton iterations are $\mathcal{O}(h^{k_p+\delta})$ where $\delta = 1$ if $|r| = 1$ and $\delta = 0$ otherwise, and $k_p \geq 2$. Then the global errors satisfy $\|e_n\| = \mathcal{O}(h^{k_p})$, where

$$k_p = \begin{cases} q & \text{if } C(q) \text{ and } B(q), \\ q+1 & \text{if } C(q), B(q+1), \text{ and } -1 \leq r < 1, \\ q+1 & \text{if } C(q), B(q+1), A(q+1), \text{ and } r = 1. \end{cases}$$

This, together with the following lemma and Theorem 3.3 or Theorem 3.4 shows that $k_G \geq k_p$, for all methods with $r \neq -1$.

Lemma B.1. If $C(q)$, $B(q)$, then

$$\Phi(t) = \frac{1}{\gamma(t)} \quad \forall t \in \overline{\text{DAIT}}_y, \rho(t) \leq q.$$

If $C(q)$, $B(q+1)$, then

$$\Phi(t) = \frac{1}{\gamma(t)} \quad \begin{cases} \forall t \in \overline{\text{DAIT}}_{yy}, \rho(t) \leq q+1, \\ \forall t \in \overline{\text{DAIT}}_{yz}, \rho(t) \leq q. \end{cases}$$

If $C(q)$, $B(q+1)$, $A(q+1)$, then

$$\Phi(t) = \frac{1}{\gamma(t)} \quad \forall t \in \overline{\text{DAIT}}_y, \rho(t) \leq q+1.$$

Proof. In the following, we use the notation $\hat{t} \in \overline{\text{DAIT}}_y$ for the bushy trees, that is, the trees where all the vertices are directly connected to the root. For such trees we have

$$(B.1) \quad \mathbf{l}_i(\hat{t}) = \rho(\hat{t})c_i^{\rho(\hat{t})-1}, \quad i = 1, \dots, s.$$

Also, τ_y is a bushy tree. Similarly, we use the notation \hat{u} for trees composed by only bushy trees, that is,

$$\hat{u} = [\hat{t}_1, \dots, \hat{t}_k]_z.$$

Now from (2.34) and (B.1) we have

$$\mathbf{k}_i(\hat{u}) = \sum_{j=1}^s d_{ij} \sum_{n_1, \dots, n_k=1}^s \rho(\hat{t}_1) a_{jn_1} c_{n_1}^{\rho(\hat{t}_1)-1} \dots \rho(\hat{t}_k) a_{jn_k} c_{n_k}^{\rho(\hat{t}_k)-1}.$$

If $C(q)$, where $q+1 \geq \rho(\hat{u})$, this equation can be written as

$$(B.2) \quad \mathbf{k}_i(\hat{u}) = \sum_{j=1}^s d_{ij} c_j^{\rho(\hat{t}_1)} \dots c_j^{\rho(\hat{t}_k)} = \sum_{j=1}^s d_{ij} c_j^{\rho(\hat{u})},$$

where the fact has been used that $\rho(\hat{u}) = \rho(\hat{t}_1) + \dots + \rho(\hat{t}_k)$. If $q \geq \rho(\hat{u})$, then

$$(B.3) \quad \mathbf{k}_i(\hat{u}) = \rho(\hat{u})c_i^{\rho(\hat{u})-1}.$$

Now let $t = [\hat{t}_1, \dots, \hat{t}_k, \hat{u}_1, \dots, \hat{u}_l]_y \in \overline{\text{DAIT}}_{yy}$. Then, from (2.33) and (B.3), we have

$$\mathbf{l}_i(t) = \rho(t) \sum_{n_1, \dots, n_k=1}^s \rho(\hat{t}_1) a_{in_1} c_{n_1}^{\rho(\hat{t}_1)-1} \dots \rho(\hat{t}_k) a_{in_k} c_{n_k}^{\rho(\hat{t}_k)-1} c_i^{\rho(\hat{u}_1)-1} \dots c_i^{\rho(\hat{u}_l)-1}.$$

If $C(q)$, where $q+1 \geq \rho(t)$, then

$$(B.4) \quad \mathbf{l}_i(t) = \rho(t) c_i^{\rho(\hat{t}_1)} \dots c_i^{\rho(\hat{t}_k)} c_i^{\rho(\hat{u}_1)-1} \dots c_i^{\rho(\hat{u}_l)-1} = \rho(t) c_i^{\rho(t)-1},$$

since $\rho(t) = \rho(\hat{t}_1) + \dots + \rho(\hat{t}_k) + \rho(\hat{u}_1) + \dots + \rho(\hat{u}_l) - l + 1$. By repeated use of (B.2) and (B.4) we know that if $C(q)$, then for all $t \in \overline{\text{DAIT}}_{yy}$ with $\rho(t) \leq q+1$ we have $\mathbf{l}_i(t) = \mathbf{l}_i(\hat{t})$, where \hat{t} is a bushy tree, and $\rho(t) = \rho(\hat{t})$. Inserting this into (2.31), we have

$$\mathbf{y}_1(t) = \sum_{i=1}^s b_i \mathbf{l}_i(t) = \rho(t) \sum_{i=1}^s b_i c_i^{\rho(t)-1} \quad \forall t \in \overline{\text{DAIT}}_{yy}, \rho(t) \leq q+1.$$

Similarly, for all $t \in \overline{\text{DAIT}}_{yz}$, $\rho(t) \leq q+1$, there exists a \hat{u} such that $\mathbf{l}_i(t) = \mathbf{k}_i(\hat{u})$ and $\rho(t) = \rho(\hat{u})$. From (B.2), (B.3), and (2.31) we have

$$\mathbf{y}_1(t) = \sum_{i=1}^s b_i \mathbf{k}_i(\hat{u}) = \sum_{i=1}^s b_i d_{ij} c_j^{\rho(t)} \quad \forall t \in \overline{\text{DAIT}}_{yz}, \rho(t) \leq q+1,$$

and

$$\mathbf{y}_1(t) = \rho(t) \sum_{i=1}^s b_i c_i^{\rho(t)-1} \quad \forall t \in \overline{\text{DAIT}}_{yy}, \rho(t) \leq q+1.$$

If $C(q)$, all the order conditions up to order $q+1$ are reduced to $A(q+1)$ and $B(q+1)$. Thus, the lemma is proved. \square

BIBLIOGRAPHY

1. K. E. Brenan and L. R. Petzold, *The numerical solution of higher index differential/algebraic equations by implicit Runge-Kutta methods*, Preprint, UCRL-95906, Lawrence Livermore National Laboratory, 1986.
2. K. Burrage and L. Petzold, *On order reduction for Runge-Kutta methods applied to differential/algebraic systems and to stiff systems of ODEs*, Preprint, UCRL-98046, Lawrence Livermore National Laboratory, 1988.
3. J. C. Butcher, *The numerical analysis of ordinary differential equations, Runge-Kutta and general linear methods*, Wiley, New York, 1987.
4. J. R. Cash, *Diagonally implicit Runge-Kutta formulae with error estimates*, J. Inst. Math. Appl. **24** (1979), 293–301.
5. F. H. Chipman, *A-stable Runge-Kutta processes*, BIT **11** (1971), 384–388.
6. C. W. Gear, *Differential-algebraic equation index transformations*, Preprint, Dept. of Computer Science, University of Illinois at Urbana-Champaign, 1986.
7. E. Hairer, Ch. Lubich, and M. Roche, *The numerical solution of differential-algebraic systems by Runge-Kutta methods*, Report, Dept. de Mathématiques, Université de Genève, 1988.
8. B. Leimkuhler, L. R. Petzold, and C. W. Gear, *On the consistent initialization of differential-algebraic equations*, Dept. of Computer Science, University of Illinois, Urbana, Illinois, 1987.

9. A. Kvaerno, *Order conditions for Runge-Kutta methods applied to differential-algebraic systems of index 1*, Report No. 4/87, Div. of Numerical Mathematics, The Norwegian Institute of Technology, Norway, 1987.
10. P. Lötstedt and L. Petzold, *Numerical solution of nonlinear differential equations with algebraic constraints: I Convergence results for backward differentiation formulas*, Math. Comp. **46** (1986), 491–516.
11. S. P. Nørsett, *Semi explicit Runge-Kutta methods*, Report No. 6/74, Div. of Numerical Mathematics, The Norwegian Institute of Technology, Norway, 1974.
12. S. P. Nørsett and P. G. Thomsen, *Local error control in SDIRK-methods*, BIT **26** (1986), 100–113.
13. L. R. Petzold, *Order results for implicit Runge-Kutta methods applied to differential/algebraic system*, SIAM J. Numer. Anal. **23** (1986), 837–852.
14. M. Roche, *Rosenbrock methods for differential algebraic equations*, Numer. Math. **52** (1988), 45–63.
15. ———, *Implicit Runge-Kutta methods for differential algebraic equations*, Report, Dept. de Mathématiques, Université de Genève, 1987.

DIVISION OF MATHEMATICAL SCIENCES, THE NORWEGIAN INSTITUTE OF TECHNOLOGY, 7034 TRONDHEIM-NTH, NORWAY. E-mail: anne@imf.unit.no