SIAM J. NUMER. ANAL.
Vol. 26, No. 4, pp. 976–996, August 1989

THE NUMERICAL SOLUTION OF HIGHER INDEX DIFFERENTIAL/ALGEBRAIC EQUATIONS BY IMPLICIT METHODS*

KATHRYN E. BRENAN† AND LINDA R. PETZOLD‡

Abstract. This paper studies the order, stability, and convergence properties of implicit Runge–Kutta (IRK) methods applied to differential/algebraic systems with index greater than one. These methods do not in general attain the same order of accuracy for higher index differential/algebraic systems as they do for index 1 systems or for purely differential systems. Necessary and sufficient conditions on the method coefficients are derived to ensure that the local and global errors of the method attain a given order of accuracy for higher index linear constant coefficient systems. IRK methods applied to nonlinear semi-explicit index 2 systems are studied, and a sufficient set of conditions is derived which ensures that a method is accurate to a given order for these systems. Finally, some numerical experiments are presented that illustrate the theoretical results and demonstrate the effects of roundoff errors on the solution.

Key words. differential/algebraic equations, Runge–Kutta methods, higher index

AMS(MOS) subject classification. 65L05

1. Introduction.

In this paper we extend the results for order, stability, and convergence of implicit Runge–Kutta (IRK) methods derived for index 1 systems by Petzold [16] to higher index differential/algebraic systems. It is well known that these methods often do not attain the same order of accuracy for differential/algebraic systems as they do for purely differential systems. Petzold has studied their behavior on uniformly index 1 systems of the form,

$$F(y, y', t) = 0 \tag{1.1}$$

when consistent initial values $y(t_0)$ are given. We examine two classes of DAEs not considered in [16], to understand how the order of accuracy of an IRK method depends on the index of the DAE system as well as the method coefficients. First we study solvable linear constant coefficient systems

$$Ay' + By = g(t), \tag{1.2}$$

of arbitrary index ν, where A and B are square constant matrices and $g(t)$ is a smooth function. Then we develop a convergence theory for IRK methods applied to nonlinear semi-explicit index 2 systems,

$$f(x, x', y, t) = 0 \qquad g(x, y, t) = 0, \tag{1.3}$$

* Received by the editors January 5, 1987; accepted for publication (in revised form) April 14, 1988.

† The Aerospace Corporation, P. O. Box 92957, Los Angeles, California 90009. The work of this author was supported in part by National Science Foundation grant DMS-85-03294, while the author was at the University of California at Los Angeles.

‡ Lawrence Livermore National Laboratory, P. O. Box 808, Livermore, California 94550. This work was supported in part by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U. S. Department of Energy, by Lawrence Livermore National Laboratory under contract W-7405-Eng-48.

where $(\partial f/\partial x')^{-1}$ exists and is bounded in some neighborhood of the solution and the square matrix $\partial g/\partial y$ has constant rank.

We formally apply an M-stage IRK method to a DAE (1.1) to obtain the system of difference equations

$$
\begin{aligned}
&F\left(y_{n-1}+h\sum_{j=1}^{M}a_{ij}Y_j',\,Y_i',\,t_{n-1}+c_ih\right)=0 \quad i=1,2,\cdots,M\\
&y_n=y_{n-1}+h\sum_{i=1}^{M}b_iY_i'
\end{aligned}
\tag{1.4}
$$

where $h=t_n-t_{n-1}$. We will assume throughout this paper that the coefficient matrix $\mathcal{A}=(a_{ij})$ of the Runge–Kutta method is nonsingular. Note that this method reduces to a standard IRK method when applied to a system of explicit ordinary differential equations (ODEs).

In § 2 we study linear constant coefficient systems of arbitrary index ν. We derive necessary and sufficient conditions on the method coefficients to ensure that the local error of the method attains a given order of accuracy for these systems. We also investigate the error propagation properties of IRK methods applied to these systems and derive an expression for the global error.

In § 3 we develop a convergence theory for IRK methods applied to nonlinear semi-explicit index 2 systems of the form (1.3). We derive a sufficient set of conditions which ensure that a method is accurate to a given order for these systems.

In the last section we describe some numerical experiments that illustrate the order reduction effects predicted by the theory and also raise some interesting questions for future research.

It is important to note that the numerical solution of higher index systems can be quite sensitive to small perturbations such as those introduced by roundoff error. In an index ν system, such as (1.2), there are variables that depend directly on the $(\nu-1)$st derivative of the input function $g(t)$. Thus if $g(t)$ is subject to small perturbations, these variables can experience errors which are in the worst case proportional to the size of the perturbation divided by $h^{\nu-1}$. Terms of this form appear in the analysis in §§ 2 and 3, and in § 4 a comparison of numerical experiments performed in single and double precision illustrates the effects of these errors. It is our experience that on machines with long word lengths, or in double precision, and for stepsizes that one might encounter in a typical application, these errors are usually smaller than the method truncation errors.

2. Linear constant coefficient systems. In this section we derive conditions that are necessary and sufficient to ensure that the local error of an implicit Runge–Kutta method attains a given order when applied to linear constant coefficient systems of arbitrary index ν. Then we study error propagation for constant coefficient higher index systems, and derive an expression for the global error.

Consider the linear constant coefficient DAE (1.2)

$$
Ay'+By=g(t) \tag{2.1}
$$

of index ν. We assume this system is solvable, so that there exist nonsingular matrices P and Q which decouple the system [10],

$$
PAQ=\begin{pmatrix} I & 0\\ 0 & N\end{pmatrix} \qquad PBQ=\begin{pmatrix} C & 0\\ 0 & I\end{pmatrix} \tag{2.2}
$$

where I is an identity matrix and where N is a block diagonal matrix, $N = \text{diag}(N_1, N_2, \cdots, N_L)$ composed of blocks of the form

$$N_i = \begin{pmatrix} 0 & & & \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{pmatrix}. \tag{2.3}$$

Applying the IRK method to (2.1), we have

$$\begin{aligned} &AY_i' + B\left(y_{n-1} + h\sum_{j=1}^{M} a_{ij}Y_j'\right) = g(t_{n-1} + c_i h) \qquad i = 1, 2, \cdots, M \\ &y_n = y_{n-1} + h\sum_{i=1}^{M} b_i Y_i'. \end{aligned} \tag{2.4}$$

Premultiplying these difference equations by P and letting $\tilde{y}_n = Q^{-1}y_n, \tilde{Y}_i' = Q^{-1}Y_i'$, $\tilde{g}(t) = Pg(t)$, we obtain

$$\begin{aligned} &(PAQ)\tilde{Y}_i' + (PBQ)\left(\tilde{y}_{n-1} + h\sum_{j=1}^{M} a_{ij}\tilde{Y}_j'\right) = \tilde{g}(t_{n-1} + c_i h) \quad i = 1, 2, \cdots, M \\ &\tilde{y}_n = \tilde{y}_{n-1} + h\sum_{i=1}^{M} b_i \tilde{Y}_i'. \end{aligned}$$

Note that the differential and algebraic parts of the system are decoupled in this form. In addition, the algebraic subsystems are decoupled from one another. Thus it is sufficient to study the behavior of the IRK method on a canonical algebraic subsystem to understand its behavior on general linear constant coefficient systems.

Consider then a canonical algebraic subsystem of index ν

$$Ny' + y = g(t) \tag{2.5}$$

where N is a $\nu \times \nu$ matrix of the form (2.3), $g(t) = (g_1(t), g_2(t), \cdots, g_\nu(t))^T$, and $y(t) = (y_1(t), y_2(t), \cdots, y_\nu(t))^T$. The solution to (2.5) is given by

$$\begin{aligned} y_1(t) &= g_1(t) \\ y_2(t) &= g_2(t) - g_1'(t) \\ &\vdots \\ y_\nu(t) &= g_\nu(t) + \sum_{i=1}^{\nu-1} (-1)^{\nu-i} g_i^{(\nu-i)}(t). \end{aligned}$$

Note that the jth component exhibits the index j behavior of the system, in the sense that $y_j(t)$ depends on the $(j-1)$st derivative of the input function $g(t)$. Applying the IRK method to (2.5), we obtain

$$\begin{aligned} &NY_i' + \left(y_{n-1} + h\sum_{j=1}^{M} a_{ij}Y_j'\right) = g(t_{n-1} + c_i h) \qquad i = 1, 2, \cdots, M \\ &y_n = y_{n-1} + h\sum_{i=1}^{M} b_i Y_i'. \end{aligned} \tag{2.6}$$

Let $Y_i' = (Y_{i,1}', Y_{i,2}', \cdots, Y_{i,\nu}')^T$, where $Y_{i,j}'$ denotes the ith-stage derivative corresponding to the jth component of the solution vector. Because of the structure of N, the difference equations (2.6) can be solved first for the stage derivatives $Y_{1,1}', Y_{2,1}', \cdots, Y_{M,1}'$ corresponding to $y_1(t)$, second for the stage derivatives $Y_{1,2}', Y_{2,2}', \cdots, Y_{M,2}'$ corresponding to $y_2(t)$, etc. More simply, the difference equations (2.6) can be solved uniquely for the stage derivatives $Y' = (Y_1', Y_2', \cdots, Y_M')^T$ because the coefficient matrix $\mathcal{A} = (a_{ij})$ of the IRK method is nonsingular. In [16], conditions on the method coefficients were derived to ensure that the local error of an IRK method attains a given order when applied to linear constant coefficient index 1 systems. Here we extend these conditions to more general index ν systems.

In general, for each component of the solution $y_j(t), j = 1, 2, \cdots, \nu$, we solve a subset of M equations from system (2.6) for the corresponding stage derivatives:

$$
\begin{aligned}
&(Y_{1,j}', Y_{2,j}', \cdots, Y_{M,j}')^T = \\
&\quad (1/h)\mathcal{A}^{-1}\Big((g_j(t_{n-1} + c_1 h), g_j(t_{n-1} + c_2 h), \cdots, g_j(t_{n-1} + c_M h))^T \\
&\qquad -\epsilon_M y_{n-1,j} - (Y_{1,j-1}', Y_{2,j-1}', \cdots, Y_{M,j-1}')^T\Big)
\end{aligned} \tag{2.7}
$$

where $\epsilon_M = (1, 1, \cdots, 1)^T$. Note that the stage derivatives depend only on the numerical solution $y_{n-1,j}$ and on the stage derivatives of the $(j-1)$st component. Define

$$
\begin{aligned}
G_i &= \begin{pmatrix} (g_i(t_{n-1} + c_1 h) - g_i(t_{n-1}))/h \\ (g_i(t_{n-1} + c_2 h) - g_i(t_{n-1}))/h \\ \vdots \\ (g_i(t_{n-1} + c_M h) - g_i(t_{n-1}))/h \end{pmatrix} \\
c^i &= (c_1^i, c_2^i, \cdots, c_M^i)^T.
\end{aligned}
$$

Utilizing the local error assumption,

$$
y_{n-1} = y(t_{n-1}) \tag{2.8}
$$

and substituting for the $(j-1)$st-stage derivatives in (2.7), we find the following expressions for the *local* stage derivatives:

$$
\begin{aligned}
(Y_{1,1}', Y_{2,1}', \cdots, Y_{M,1}')^T &= \mathcal{A}^{-1}G_1 \\
(Y_{1,2}', Y_{2,2}', \cdots, Y_{M,2}')^T &= \mathcal{A}^{-1}G_2 - (1/h)\mathcal{A}^{-2}G_1 + (1/h)\mathcal{A}^{-1}\epsilon_M g_1'(t_{n-1}) \\
(Y_{1,3}', Y_{2,3}', \cdots, Y_{M,3}')^T &= \mathcal{A}^{-1}G_3 + (1/h)\mathcal{A}^{-1}\epsilon_M[g_2'(t_{n-1}) - g_1''(t_{n-1})] \\
&\quad - (1/h)\mathcal{A}^{-2}G_2 + (1/h^2)\mathcal{A}^{-3}G_1 \\
&\quad - (1/h^2)\mathcal{A}^{-2}\epsilon_M g_1'(t_{n-1})
\end{aligned} \tag{2.9}
$$

and similar expressions for the remaining components. We define the *local error* d_n by

$$
d_n = y(t_{n-1}) + h\sum_{i=1}^{M} b_i Y_i' - y(t_n) \tag{2.10}
$$

where $d_n = (d_{n,1}, d_{n,2}, \cdots, d_{n,\nu})^T$ and Y_i' represent the local stage derivatives given by (2.9). Expanding (2.10) in a Taylor series about t_{n-1}, and equating like powers of h, it is easy to see as in [16] that the local error $d_{n,1}$ in the first component satisfies

$$d_{n,1} = O(h^{k_{a,1}+1}), \tag{2.11}$$

when $b^T\mathcal{A}^{-1}c^i = 1$ for $i = 1, 2, \cdots, k_{a,1}$. $k_{a,1}$ is the *algebraic order* of the IRK method applied to index 1 constant coefficient systems. For the second component, the local error $d_{n,2}$ is given by

$$\begin{aligned} d_{n,2} &= y_2(t_{n-1}) - y_2(t_n) + hb^T\mathcal{A}^{-1}G_2 \\ &\quad + b^T\mathcal{A}^{-1}\epsilon_M g_1'(t_{n-1}) - b^T\mathcal{A}^{-2}G_1. \end{aligned} \tag{2.12}$$

Note that $g_2(t_{n-1}) - g_2(t_n) + hb^T\mathcal{A}^{-1}G_2 = O(h^{k_{a,1}+1})$ if we assume the IRK method has algebraic order $k_{a,1}$ on index 1 problems. Then expand the remaining terms in (2.12) in a Taylor series about t_{n-1} and equate like powers of h to obtain the following set of order conditions for index 2 constant coefficient systems,

$$\begin{aligned} &b^T\mathcal{A}^{-1}\epsilon_M = b^T\mathcal{A}^{-2}c^1 \\ &b^T\mathcal{A}^{-2}c^i = i, \quad i = 2, 3, \cdots, k_{a,2}. \end{aligned}$$

We define the *algebraic order* of the IRK method applied to index 2 constant coefficient canonical systems to be $k_{a,2}$ if these conditions are satisfied. The local error for a general index 2 constant coefficient system thus satisfies $d_{n,2} = O(h^{k_{a,2}}) + O(h^{k_{a,1}+1})$.

This analysis can be extended in a straightforward manner to the general index ν case. For completeness we list here the additional algebraic order conditions and the corresponding asymptotic behavior of the local error for both the index 3 case and the most general case. For an index 3 system,

$$d_{n,3} = O(h^{k_{a,3}-1}) + O(h^{k_{a,2}}) + O(h^{k_{a,1}+1})$$

where $k_{a,3}$ is the largest integer such that

$$\begin{aligned} &b^T\mathcal{A}^{-2}\epsilon_M = b^T\mathcal{A}^{-3}c^1 \\ &b^T\mathcal{A}^{-1}\epsilon_M = b^T\mathcal{A}^{-3}c^2/2 \\ &b^T\mathcal{A}^{-3}c^i = i(i-1), \quad i = 3, 4, \cdots, k_{a,3}. \end{aligned}$$

Finally, for a general constant coefficient index ν system, the local error satisfies

$$d_{n,\nu} = O(h^{k_{a,\nu}-\nu+2}) + O(h^{k_{a,\nu-1}-\nu+3}) + \cdots + O(h^{k_{a,1}+1}) \tag{2.13}$$

where $k_{a,\nu}$ is the largest integer such that

$$\begin{aligned} &b^T\mathcal{A}^{-i}\epsilon_M = b^T\mathcal{A}^{-\nu}c^{\nu-i}/(\nu-i)! \qquad i = 1, 2, \cdots, \nu-1 \\ &b^T\mathcal{A}^{-\nu}c^i = i(i-1)\cdots(i-\nu+1), \qquad i = \nu, \nu+1, \cdots, k_{a,\nu}. \end{aligned}$$

Clearly, the higher the index the more difficult it is to find IRK methods that are convergent in all the components. Finding an IRK method having the same rate of convergence in all of the components similarly poses severe restrictions on the coefficients.

Next we examine the propagation of errors for IRK methods applied to linear constant coefficient systems. Consider solving (2.5) by the perturbed Runge–Kutta method,

$$
\begin{aligned}
&NZ_i' + \left(z_{n-1} + h\sum_{j=1}^{M} a_{ij} Z_j' - \delta_n^{(i)} \right) = g(t_{n-1} + c_i h) \quad i = 1, 2, \cdots, M \\
&z_n = z_{n-1} + h\sum_{i=1}^{M} b_i Z_i' - \delta_n^{(M+1)},
\end{aligned} \tag{2.14}
$$

where the perturbations $\delta_n^{(i)} = (\delta_{n,1}^{(i)}, \delta_{n,2}^{(i)}, \cdots, \delta_{n,\nu}^{(i)})^T$ satisfy $\|\delta_n^{(i)}\| \leq \Delta$ for $i = 1, 2, \cdots, M+1$. The perturbations could be due to roundoff error, errors in solving the linear systems at each stage, or could be interpreted as truncation errors at each stage (see § 3). Subtracting (2.14) from (2.6), and defining $e_n = y_n - z_n$, $E_i' = Y_i' - Z_i'$, we obtain an expression for the difference between these two solutions

$$
\begin{aligned}
&NE_i' + \left(e_{n-1} + h\sum_{j=1}^{M} a_{ij} E_j' + \delta_n^{(i)} \right) = 0, \quad i = 1, 2, \cdots, M \\
&e_n = e_{n-1} + h\sum_{i=1}^{M} b_i E_i' + \delta_n^{(M+1)}.
\end{aligned} \tag{2.15}
$$

By solving the first equation in (2.15) for E_i' and substituting into the second equation, we can obtain a relation describing the error propagation of the method. For linear constant coefficient index 1 systems Petzold did this for [16] and obtained

$$
e_{n,1} = (1 - b^T \mathcal{A}^{-1} \epsilon_M) e_{n-1,1} - (b^T \mathcal{A}^{-1} \delta_{n,1} - \delta_{n,1}^{(M+1)}), \tag{2.16}
$$

where $\delta_{n,j} = (\delta_{n,j}^{(1)}, \delta_{n,j}^{(2)}, \cdots, \delta_{n,j}^{(M)})^T$. The recurrence (2.16) is unstable unless $|1 - b^T \mathcal{A}^{-1} \epsilon_M| < 1$. Hence we will require as in [16] that the IRK method satisfy the *strict stability condition*

$$
| 1 - b^T \mathcal{A}^{-1} \epsilon_M | < 1. \tag{2.17}
$$

The error propagation relation for the second component is given by

$$
\begin{aligned}
e_{n,2} \;=\; & (1 - b^T \mathcal{A}^{-1} \epsilon_M) e_{n-1,2} - (b^T \mathcal{A}^{-1} \delta_{n,2} - \delta_{n,2}^{(M+1)}) \\
& + (1/h) b^T \mathcal{A}^{-2} (\delta_{n,1} + \epsilon_M e_{n-1,1}),
\end{aligned} \tag{2.18}
$$

while for the third component it is

$$
\begin{aligned}
e_{n,3} \;=\; & (1 - b^T \mathcal{A}^{-1} \epsilon_M) e_{n-1,3} - (b^T \mathcal{A}^{-1} \delta_{n,3} - \delta_{n,3}^{(M+1)}) \\
& + (1/h) b^T \mathcal{A}^{-2} (\delta_{n,2} + \epsilon_M e_{n-1,2}) \\
& - (1/h^2) b^T \mathcal{A}^{-3} (\delta_{n,1} + \epsilon_M e_{n-1,1}).
\end{aligned} \tag{2.19}
$$

Finally, for the νth component, which would occur in an index ν system, the stability relation can be shown to be

$$
\begin{aligned}
e_{n,\nu} = \; & (1 - b^T \mathcal{A}^{-1} \epsilon_M) e_{n-1,\nu} - (b^T \mathcal{A}^{-1} \delta_{n,\nu} - \delta_{n,\nu}^{(M+1)}) \\
& - \sum_{i=1}^{\nu-1} \frac{(-1)^i}{h^i} b^T \mathcal{A}^{-i-1} (\delta_{n,\nu-i} + \epsilon_M e_{n-1,\nu-i}).
\end{aligned} \tag{2.20}
$$

Note that the strict stability condition is no longer sufficient to insure stability, in a strict mathematical sense, of the IRK method when applied to linear constant coefficient systems of index greater than one. For small stepsizes, roundoff errors can be significant for the solution components which occur in higher index systems.

These methods can be useful for the solution of higher index systems, provided that we understand the implications of the error propagation relations given above. We can see that the sensitivity to roundoff errors is confined to the later components of the system and does not propagate back into the earlier components. This observation holds also for the nonlinear semi-explicit index 2 systems that we study in the next section. Finally, using the error propagation relations above, we can extend the conclusions of Petzold [16] for global error in solving linear constant coefficient index 1 systems to higher index systems as follows.

DEFINITION 2.1. The *constant coefficient order* of an IRK method (1.4) is equal to $k_{c,\nu}$ if the method converges with global error $O(h^{k_{c,\nu}})$ for all solvable linear constant coefficient systems (1.2) of index $\leq \nu$.

THEOREM 2.1. *Suppose the* IRK *method* (1.4) *satisfies the strict stability condition. Then the constant coefficient order* $k_{c,\nu}$ *of the global error of this method is given by*

$$k_{c,\nu} = \min_{1 \leq i \leq \nu} (k_d, k_{a,i} - \nu + 2) \tag{2.21}$$

where k_d *is the order of the method for purely differential (nonstiff) systems.*

Finally, we present some results on the order of accuracy of some IRK methods from the stiff ODE literature applied to index 1 and index 2 linear constant coefficient systems. We have chosen to investigate these particular methods because our numerical experience [2] with IRK methods applied to DAEs has led us to conclude that it is very desirable for a method to be L-stable, or even better to be stiffly accurate, and also because these methods can be implemented efficiently. One reason why L-stable methods appear promising is that they perform very well when applied to index 1 and semi-explicit index 2 and index 3 systems, even when the initial values contain small errors. Recall that a method is L-stable if it is A-stable and if $\lim_{\mathrm{Re}(h\lambda) \to -\infty} |y_{n+1}/y_n| = 0$, when applied to the test problem $y' = \lambda y$. For IRK methods, this condition is equivalent to requiring that $|1 - b^T \mathcal{A}^{-1} \epsilon_M| = 0$. *Stiffly accurate* methods [17] are L-stable methods that satisfy the additional requirement that $c_M = 1$, $a_{Mj} = b_j$ for $j = 1, 2, \cdots, M$. Thus $b^T \mathcal{A}^{-1} = (0, 0, \cdots, 0, 1)^T$ for stiffly accurate methods. The L-stable methods we have chosen to investigate here are:

(1) 2-stage, "2nd-order" Singly Implicit method (SIRK) [4], with $\lambda = 1 - \sqrt{2}/2$,

(2) 5-stage, "4th-order" Diagonally Implicit method (DIRK) [1], [7],

(3) 3-stage, "3rd-order" Singly Implicit method (SIRK) [4], with $1/\lambda$ the root of the Laguerre polynomial of degree 3,

(4) 7-stage, "3rd-order" Extrapolation method based on fully implicit backward Euler and polynomial extrapolation, written as a semi-explicit Runge–Kutta method.

Methods (1) and (2) are stiffly accurate. The algebraic orders of all these methods are given in Table 2.1, where it can be seen that, as observed above, it is difficult to maintain the same rate of convergence in all of the variables for linear constant coefficient index 2 systems.

TABLE 2.1
Order of consistency for linear constant coefficient systems.

L-Stable methods	*ODE order* k_d	*Index 1 Order* $k_{a,1}$	*Index 2 Order* $k_{a,2}$
1. Two-stage SIRK	2	∞	2
2. Five-stage DIRK	4	∞	1
3. Three-stage SIRK	3	3	2
4. Seven-stage Extrp.	3	∞	3

3. Semi-explicit nonlinear index 2 systems. In this section we study nonlinear semi-explicit index 2 systems of the form

$$\begin{aligned} f(x,x',y,t) &= 0 \\ g(x,y,t) &= 0, \end{aligned} \tag{3.1}$$

where we will assume that $(\partial f/\partial x')^{-1}$ exists and is bounded in some neighborhood of the solution, $\partial g/\partial y$ has constant rank, and f and g have as many continuous partial derivatives as desired in a neighborhood of the solution. We give a set of order conditions which are sufficient to ensure that a method is accurate to a given order for these systems.

To state our results, we first recall the definitions of internal order and internal local truncation error given in [16].

DEFINITION 3.1. The ith internal local truncation error $\delta_i^{(n)}$ at t_n of an M-stage implicit Runge–Kutta method (1.4) applied to $F(y,y',t)=0$ is given by

$$\begin{aligned} \delta_i^{(n)} &= y(t_{n-1}) + h\sum_{j=1}^{M} a_{ij}y'(t_{n-1}+c_jh) - y(t_{n-1}+c_ih), \quad i=1,\cdots,M, \\ \delta_{M+1}^{(n)} &= y(t_{n-1}) + h\sum_{i=1}^{M} b_iy'(t_{n-1}+c_ih) - y(t_n). \end{aligned} \tag{3.2}$$

DEFINITION 3.2. The internal order k_I of an M-stage implicit Runge–Kutta method (1.4) is given by

$$k_I = \min(k_1,\cdots,k_M,k_{M+1})$$

where

$$\delta_i^{(n)} = O(h^{k_i+1}), \quad i=1,\cdots,(M+1).$$

As in [16], [9], it is simple to find the internal order of an implicit Runge–Kutta method in terms of its coefficients by expanding (3.2) in Taylor series around t_{n-1}, leading to the result that the internal order of an M-stage implicit Runge–Kutta method is equal to k_I if and only if the method coefficients satisfy

$$\sum_{j=1}^{M} a_{ij}c_j^{k-1} = \frac{c_i^k}{k}, \quad i=1,\cdots,M,$$

$$\sum_{j=1}^{M} b_jc_j^{k-1} = \frac{1}{k}$$

for $k = 1, \cdots, k_I$.

Then we can state the following result.

THEOREM 3.1. *Given the nonlinear, semi-explicit index 2 system* (3.1) *to be solved numerically by the M-stage* IRK *method* (1.4), *suppose*

(1) *The* IRK *method has internal order* k_I;

(2) *The* IRK *method satisfies the strict stability condition;*

(3) *The initial conditions satisfy* $\| (I - H)e_0^x \| = O(h^{k_G})$, *where* $I - H$ *is a projection operator defined below,* $e_0^x = x_0 - x(t_0)$ *and* $k_G = \min(k_d, k_I + 1)$.

Then the global errors in the numerical solution x_n *and* y_n *are* $O(h^{k_G})$ *and* $O(h^{k_I})$, *respectively.*

Proof. We will first prove this theorem for the simpler index 2 system

$$\begin{aligned} &x' + g_1(x, y, t) = 0 \\ &g_2(x, t) = 0, \end{aligned} \tag{3.3}$$

where $[(\partial g_2/\partial x)(\partial g_1/\partial y)]^{-1}$ exists and is bounded in a neighborhood of the solution, and then show that the results extend to the more general system (3.1).

Consider the M-stage IRK applied to (3.3):

$$\begin{aligned} &X_i' + g_1\left(x_{n-1} + h\sum_{j=1}^{M} a_{ij}X_j', y_{n-1} + h\sum_{j=1}^{M} a_{ij}Y_j', t_i\right) = 0 \\ &g_2\left(x_{n-1} + h\sum_{j=1}^{M} a_{ij}X_j', t_i\right) = 0 \qquad i = 1, 2, \cdots, M \\ &x_n = x_{n-1} + h\sum_{i=1}^{M} b_i X_i' \\ &y_n = y_{n-1} + h\sum_{i=1}^{M} b_i Y_i', \end{aligned} \tag{3.4}$$

where $t_i = t_{n-1} + c_i h$. It is convenient to define intermediate stage values for x and y at t_i:

$$\begin{aligned} X_i &= x_{n-1} + h\sum_{j=1}^{M} a_{ij}X_j' \\ Y_i &= y_{n-1} + h\sum_{j=1}^{M} a_{ij}Y_j'. \end{aligned}$$

The true solution satisfies

$$\begin{aligned} &x'(t_i) + g_1(x(t_i), y(t_i), t_i) = 0 \\ &g_2(x(t_i), t_i) = 0 \qquad i = 1, 2, \cdots, M \\ &x(t_n) = x(t_{n-1}) + h\sum_{i=1}^{M} b_i x'(t_{n-1} + c_i h) - \delta_{M+1}^{x(n)} \\ &y(t_n) = y(t_{n-1}) + h\sum_{i=1}^{M} b_i y'(t_{n-1} + c_i h) - \delta_{M+1}^{y(n)}, \end{aligned} \tag{3.5}$$

where

$$x(t_i) = x(t_{n-1}) + h\sum_{j=1}^{M} a_{ij}x'(t_{n-1} + c_j h) - \delta_i^{x(n)}$$
$$y(t_i) = y(t_{n-1}) + h\sum_{j=1}^{M} a_{ij}y'(t_{n-1} + c_j h) - \delta_i^{y(n)} \quad i = 1, 2, \cdots, M.$$

Let $G_{21}(t_i) = \partial g_2/\partial x$, $G_{12}(t_i) = \partial g_1/\partial y$, and $G_{11}(t_i) = \partial g_1/\partial x$, where the partial derivatives are evaluated along the true solution at t_i. Subtracting (3.5) from (3.4), we obtain

$$\begin{aligned} &E_i^{x\prime} + G_{11}(t_i)E_i^x + G_{12}(t_i)E_i^y = \eta_i^x \\ &G_{21}(t_i)E_i^x = \eta_i^y, \quad i = 1, 2, \cdots, M \\ &e_n^x = e_{n-1}^x + h\sum_{i=1}^{M} b_i E_i^{x\prime} + \delta_{M+1}^{x(n)} \\ &e_n^y = e_{n-1}^y + h\sum_{i=1}^{M} b_i E_i^{y\prime} + \delta_{M+1}^{y(n)}, \end{aligned} \tag{3.6}$$

where

$$\begin{aligned} E_i^y &= e_{n-1}^y + h\sum_{j=1}^{M} a_{ij}E_j^{y\prime} + \delta_i^{y(n)} \\ E_i^x &= e_{n-1}^x + h\sum_{j=1}^{M} a_{ij}E_j^{x\prime} + \delta_i^{x(n)}, \end{aligned} \tag{3.7}$$

and $E_i^{x\prime} = X_i' - x'(t_i)$, $E_i^{y\prime} = Y_i' - y'(t_i)$, $E_i^y = Y_i - y(t_i)$, $E_i^x = X_i - x(t_i)$, $e_n^x = x_n - x(t_n)$, and $e_n^y = y_n - y(t_n)$. The η_i terms are the sum of residuals from the Newton iteration and higher-order terms in E_i^x and E_i^y.

We can eliminate E_i^y in terms of $E_i^{x'}$ by multiplying the first equation in (3.6) by $G_{21}(t_i)$ and solving for E_i^y,

$$E_i^y = -M_i E_i^{x'} - M_i G_{11}(t_i)E_i^x + M_i\eta_i^x, \tag{3.8}$$

where $M_i = (G_{21}(t_i)G_{12}(t_i))^{-1}G_{21}(t_i)$.

Let $H_i = G_{12}(t_i)M_i$. Multiply the first equation in (3.6) by $I - H_i$ and substitute (3.8) for E_i^y. Multiply the second equation in (3.6) by $G_{12}(t_i)(G_{21}(t_i)G_{12}(t_i))^{-1}$. Equations (3.6) may now be written as

$$\begin{aligned} &(I - H_i)E_i^{x'} + N_i E_i^x = \tilde{\eta}_i^x \\ &H_i E_i^x = \tilde{\eta}_i^y \quad i = 1, 2, \cdots, M \\ &e_n^x = e_{n-1}^x + h\sum_{i=1}^{M} b_i E_i^{x'} + \delta_{M+1}^{x(n)} \\ &e_n^y = e_{n-1}^y + h\sum_{i=1}^{M} b_i E_i^{y'} + \delta_{M+1}^{y(n)}, \end{aligned} \tag{3.9}$$

where $\tilde{\eta}_i^y = G_{12}(t_i)(G_{21}(t_i)G_{12}(t_i))^{-1}\eta_i^y$, $\tilde{\eta}_i^x = (I - H_i)\eta_i^x$ and $N_i = (I - H_i)G_{11}(t_i)$.

Define

$$\begin{aligned} \tilde{E}_i^{x'} &= (I - H_i)E_i^{x'} & \tilde{\tilde{E}}_i^{x'} &= H_i E_i^{x'} \\ \tilde{\delta}_i^{x(n)} &= (I - H_i)\delta_i^{x(n)} & \tilde{\tilde{\delta}}_i^{x(n)} &= H_i \delta_i^{x(n)} \\ \tilde{e}_n^x &= (I - H_n)e_n^x & \tilde{\tilde{e}}_n^x &= H_n e_n^x. \end{aligned}$$

Rewrite the first equation in (3.9),

$$(I - H_i)E_i^{x'} + N_i\left((I - H_i)E_i^x + H_i E_i^x\right) = \tilde{\eta}_i^x.$$

Substituting the definition of E_i^x given in equation (3.7) and using the second equation in (3.9), we have an expression for $\tilde{E}_i^{x'}$,

$$\begin{aligned} \tilde{E}_i^{x'} + N_i \left((I - H_i)e_{n-1}^x + h\sum_{j=1}^{M} a_{ij}\tilde{E}_j^{x'} + h\sum_{j=1}^{M} a_{ij}(H_j - H_i)E_j^{x'} + \tilde{\delta}_i^{x(n)} \right) \\ = \tilde{\eta}_i^x - N_i\tilde{\eta}_i^y. \end{aligned} \tag{3.10}$$

To find an expression for $\tilde{\tilde{E}}_i^{x'}$, from the second equation in (3.9) we have

$$H_i e_{n-1}^x + h\sum_{j=1}^{M} a_{ij} H_i E_j^{x'} + H_i \delta_i^{x(n)} = \tilde{\eta}_i^y.$$

Thus,

$$H_i e_{n-1}^x + h\sum_{j=1}^{M} a_{ij}\tilde{\tilde{E}}_j^{x'} + h\sum_{j=1}^{M} a_{ij}(H_i - H_j)E_j^{x'} + \tilde{\tilde{\delta}}_i^{x(n)} = \tilde{\eta}_i^y. \tag{3.11}$$

Now we can rewrite (3.10) and (3.11) noting that $E_j^{x'} = \tilde{E}_j^{x'} + \tilde{\tilde{E}}_j^{x'}$ and $\delta_j^{x(n)} = \tilde{\delta}_j^{x(n)} + \tilde{\tilde{\delta}}_j^{x(n)}$, to obtain

$$\begin{aligned} &\tilde{E}_i^{x'} + h\sum_{j=1}^{M} a_{ij}N_i\tilde{E}_j^{x'} + N_i(I - H_i)\left(e_{n-1}^x + \delta_i^{x(n)}\right) \\ &\quad + h\sum_{j=1}^{M} a_{ij}N_i(H_j - H_i)(\tilde{E}_j^{x'} + \tilde{\tilde{E}}_j^{x'}) = \tilde{\eta}_i^x - N_i\tilde{\eta}_i^y \\ &h\sum_{j=1}^{M} a_{ij}\tilde{\tilde{E}}_j^{x'} + H_i\left(e_{n-1}^x + \delta_i^{x(n)}\right) \\ &\quad + h\sum_{j=1}^{M} a_{ij}(H_i - H_j)(\tilde{E}_j^{x'} + \tilde{\tilde{E}}_j^{x'}) = \tilde{\eta}_i^y, \end{aligned} \tag{3.12}$$

for $i = 1, 2, \cdots, M$. Using their Taylor series expansions about the true solution at t_{n-1}, express N_i and H_i as $N_i = N + O(h)$ and $H_i = H + O(h)$, where N and H

are evaluated along the true solution at time t_{n-1}. Then rewrite (3.12) in matrix notation to obtain:

$$\begin{pmatrix} T_1 & h^2T_2 \\ h^2T_3 & hT_4 \end{pmatrix} \begin{pmatrix} \tilde{E}^{x'} \\ \tilde{\tilde{E}}^{x'} \end{pmatrix} = - \begin{pmatrix} S_1 & 0 \\ 0 & S_4 \end{pmatrix} \begin{pmatrix} \mathbf{e}^x_{n-1} + \delta^{x(n)} \\ \mathbf{e}^x_{n-1} + \delta^{x(n)} \end{pmatrix} + \begin{pmatrix} \bar{\bar{\eta}}^x \\ \bar{\eta}^y \end{pmatrix} \tag{3.13}$$

where

$$\begin{aligned} \tilde{E}^{x'} &= (\tilde{E}^{x'}_1, \tilde{E}^{x'}_2, \cdots, \tilde{E}^{x'}_M)^T \\ \tilde{\tilde{E}}^{x'} &= (\tilde{\tilde{E}}^{x'}_1, \tilde{\tilde{E}}^{x'}_2, \cdots, \tilde{\tilde{E}}^{x'}_M)^T \\ \bar{\bar{\eta}}^x &= (\tilde{\eta}^x_1 - N_1\tilde{\eta}^y_1, \tilde{\eta}^x_2 - N_2\tilde{\eta}^y_2, \cdots, \tilde{\eta}^x_M - N_M\tilde{\eta}^y_M)^T \\ \bar{\eta}^y &= (\tilde{\eta}^y_1, \tilde{\eta}^y_2, \cdots, \tilde{\eta}^y_M)^T \\ \mathbf{e}^x_{n-1} &= (e^x_{n-1}, e^x_{n-1}, \cdots, e^x_{n-1})^T \\ \delta^{x(n)} &= (\delta^{x(n)}_1, \delta^{x(n)}_2, \cdots, \delta^{x(n)}_M)^T. \end{aligned}$$

The matrices in (3.13) are given by

$$\begin{aligned} T_1 &= \hat{T}_1 + O(h^2) \\ T_4 &= \hat{T}_4 + O(h) \\ S_1 &= \hat{S}_1 + O(h) \\ S_4 &= \hat{S}_4 + O(h), \end{aligned}$$

where $\hat{T}_1 = I_{Md} + h\mathcal{A} \otimes N$, $\hat{T}_4 = \mathcal{A} \otimes I_d$, $\hat{S}_1 = I_M \otimes (N(I - H))$, $\hat{S}_4 = I_M \otimes H$, T_2 and T_3 are $O(1)$, and d is the dimension of x in (3.3). Here, for clarity, we have denoted the dimensions of the identity matrices by subscripts. However, since the matrix I_d occurs frequently, its subscript is omitted when its dimension is obvious from the context.

Let T_n denote the left-hand matrix in (3.13). $\hat{T}_4$ is invertible because the matrix $\mathcal{A}$ of coefficients of the method is invertible. The inverse of T_n is given by

$$T_n^{-1} = \begin{pmatrix} \hat{T}_1^{-1} + O(h) & O(h) \\ O(h) & \hat{T}_4^{-1}/h + O(1) \end{pmatrix}.$$

Now we can solve for $\tilde{E}^{x'}$ and $\tilde{\tilde{E}}^{x'}$ in (3.13) to obtain

$$\begin{aligned} \begin{pmatrix} \tilde{E}^{x'} \\ \tilde{\tilde{E}}^{x'} \end{pmatrix} = & - \begin{pmatrix} \hat{T}_1^{-1}\hat{S}_1 + O(h) & O(h) \\ O(h) & \hat{T}_4^{-1}\hat{S}_4/h + O(1) \end{pmatrix} \begin{pmatrix} \mathbf{e}^x_{n-1} + \delta^{x(n)} \\ \mathbf{e}^x_{n-1} + \delta^{x(n)} \end{pmatrix} \\ & + \begin{pmatrix} \bar{\bar{\eta}}^x + O(h)\bar{\eta}^y + O(h)\bar{\bar{\eta}}^x \\ \hat{T}_4^{-1}\bar{\eta}^y/h + O(h)\bar{\bar{\eta}}^x + O(1)\bar{\eta}^y \end{pmatrix} \end{aligned} \tag{3.14}$$

Recall that $E^{x'} = \tilde{E}^{x'} + \tilde{\tilde{E}}^{x'}$, so that from (3.14),

$$\begin{aligned} E^{x'} = & -(1/h)(\hat{T}_4^{-1}\hat{S}_4)(\mathbf{e}^x_{n-1} + \delta^{x(n)}) + (1/h)\hat{T}_4^{-1}\bar{\eta}^y + \bar{\bar{\eta}}^x \\ & + O(1)(\mathbf{e}^x_{n-1} + \delta^{x(n)}) + O(1)\bar{\eta}^y + O(h)\bar{\bar{\eta}}^x. \end{aligned} \tag{3.15}$$

We can use the expression for $E^{x'}$ above to solve for e^x_n. From the third equation in (3.9), we have

$$e^x_n = e^x_{n-1} + h\mathbf{b}^T E^{x'} + \delta^{x(n)}_{M+1},$$

where $\mathbf{b}^T = (b_1 I_d, b_2 I_d, \cdots, b_M I_d) = b^T \otimes I_d$.

Substituting (3.15) into the above expression, we obtain

$$\begin{aligned} e_n^x = \; & e_{n-1}^x - (b^T \otimes I_d)(\mathcal{A}^{-1} \otimes I_d)(I_M \otimes H)(\mathbf{e}_{n-1}^x + \delta^{x(n)}) + \delta_{M+1}^{x(n)} \\ & + (b^T \otimes I_d)(\mathcal{A}^{-1} \otimes I_d)\bar{\eta}^y \\ & + O(h\delta^{x(n)}) + O(he_{n-1}^x) + O(h\bar{\eta}^y) + O(h\bar{\bar{\eta}}^x). \end{aligned} \tag{3.16}$$

Now by the *strict stability condition*, we have $|1 - b^T \mathcal{A}^{-1} \epsilon_M| < 1$, where $\epsilon_M = (1, 1, \cdots, 1)^T$. Let $b^T \mathcal{A}^{-1} \epsilon_M = 1 - \rho$. Then from (3.16) we have

$$\begin{aligned} e_n^x = \; & (I - (1-\rho)H_{n-1} + O(h))e_{n-1}^x + ((b^T \mathcal{A}^{-1}) \otimes I_d)\bar{\eta}^y \\ & - H_{n-1}((b^T \mathcal{A}^{-1}) \otimes I_d)\delta^{x(n)} + \delta_{M+1}^{x(n)} \\ & + O(h\delta^{x(n)}) + O(h\bar{\bar{\eta}}^x) + O(h\bar{\eta}^y). \end{aligned} \tag{3.17}$$

Multiplying (3.17) by H_n and using the fact that H_n is a projection, we obtain

$$\begin{aligned} \tilde{\tilde{e}}_n^x = \; & \rho(I + O(h))\tilde{\tilde{e}}_{n-1}^x - H_{n-1}\left(((b^T \mathcal{A}^{-1}) \otimes I_d)\delta^{x(n)} - \delta_{M+1}^{x(n)}\right) \\ & + O(h\delta^{x(n)}) + O(h\delta_{M+1}^{x(n)}) + O(\bar{\eta}^y) + O(h\bar{\eta}^x) \end{aligned} \tag{3.18}$$

where $\bar{\eta}^x = (\bar{\eta}_1^x, \bar{\eta}_2^x, \cdots, \bar{\eta}_M^x)^T$. Note that, by definition of the algebraic order $k_{a,1}$, we have $((b^T \mathcal{A}^{-1}) \otimes I_d)\delta^{x(n)} - \delta_{M+1}^{x(n)} = O(h^{k_{a,1}+1})$. Multiplying (3.17) by $(I - H_n)$, and noting that $(I - H_i)\tilde{\eta}_i^y = 0$ for $i = 1, 2, \cdots, M$ by definition of $\tilde{\eta}_i^y$ in (3.9), we obtain

$$\tilde{e}_n^x = (I + O(h))\tilde{e}_{n-1}^x + O(h\delta^{x(n)}) + O(\delta_{M+1}^{x(n)}) + O(h\bar{\eta}^y) + O(h\bar{\eta}^x). \tag{3.19}$$

Note that the order k_{M+1} of the last stage is always at least as large as the differential order k_d.

Now suppose that $\|\bar{\eta}^x\| \le \epsilon_1$ and $\|\bar{\eta}^y\| \le \epsilon_2$. The magnitude of ϵ_1 and ϵ_2 will be determined later. For linear problems, they are just proportional to the size of the residuals at the termination of the Newton iteration. Then rewriting (3.18) and (3.19), we have

$$\begin{aligned} \tilde{\tilde{e}}_n^x &= \rho(I + O(h))\tilde{\tilde{e}}_{n-1}^x + O(h^{k_{a,1}+1}) + O(h^{k_I+2}) + O(\epsilon_2) + O(h\epsilon_1) \\ \tilde{e}_n^x &= (I + O(h))\tilde{e}_{n-1}^x + O(h^{k_I+2}) + O(h^{k_d+1}) + O(h\epsilon_2) + O(h\epsilon_1), \end{aligned} \tag{3.20}$$

where by the strict stability condition, $-1 < \rho < 1$.

Solving the recurrence relations (3.20) and noting that $e_n^x = \tilde{\tilde{e}}_n^x + \tilde{e}_n^x$ and that $k_{a,1} \ge k_I$, we obtain

$$\|e_n^x\| = O(h^{k_G}) + O(\epsilon_2) + O(\epsilon_1) + O((I - H)e_0^x), \tag{3.21}$$

where $k_G = \min(k_d, k_I + 1)$.

Now we can bound the error in the y component. By the definition of E_i^y,

$$E^y = \mathbf{e}_{n-1}^y + h(\mathcal{A} \otimes I_d)E^{y'} + \delta^{y(n)},$$

where $E^{y'} = (E_1^{y'}, E_2^{y'}, \cdots, E_M^{y'})^T$, $E^y = (E_1^y, E_2^y, \cdots, E_M^y)^T$ and $\delta^{y(n)} = (\delta_1^{y(n)}, \delta_2^{y(n)}, \cdots, \delta_M^{y(n)})^T$. We can solve for $E^{y'}$ to obtain

$$E^{y'} = (1/h)(\mathcal{A}^{-1} \otimes I_d)(E^y - \mathbf{e}_{n-1}^y - \delta^{y(n)}).$$

In the expression for e_n^y given in (3.9), substitute for $E^{y'}$ to obtain

$$\begin{aligned} e_n^y &= (1 - b^T\mathcal{A}^{-1}\epsilon_M)e_{n-1}^y - (((b^T\mathcal{A}^{-1}) \otimes I_d)\delta^{y(n)} - \delta_{M+1}^{y(n)}) \\ &\quad + \mathbf{b}^T(\mathcal{A}^{-1} \otimes I_d)E^y. \end{aligned}$$

Note that $((b^T\mathcal{A}^{-1}) \otimes I_d)\delta^{y(n)} - \delta_{M+1}^{y(n)} = O(h^{k_{a,1}+1})$. Substitute for E^y from (3.8) and simplify to obtain

$$\begin{aligned} e_n^y = \; & \rho e_{n-1}^y - \mathbf{b}^T(\mathcal{A}^{-1} \otimes I_d)(I_M \otimes M + O(h))(E^{x'} + (I_M \otimes G_{11})E^x) \\ & + O(\eta^x) + O(h^{k_{a,1}+1}), \end{aligned} \tag{3.22}$$

where M and G_{11} are evaluated along the true solution at t_{n-1}. Now substitute for $E^{x'}$ from (3.15) and simplify, to obtain

$$\begin{aligned} e_n^y &= \rho e_{n-1}^y + (1/h)((b^T(\mathcal{A}^{-1})^2) \otimes I_d)(I_M \otimes M)(I_M \otimes H)(\mathbf{e}_{n-1}^x + \delta^{x(n)}) \\ &\quad + O(e_{n-1}^x) + O(\delta^{x(n)}) + O(\bar{\eta}^y/h) + O(\bar{\eta}^x) + O(h^{k_{a,1}+1}). \end{aligned}$$

Noting that $(I_M \otimes H)\mathbf{e}_{n-1}^x = (\tilde{\tilde{e}}_{n-1}^x, \tilde{\tilde{e}}_{n-1}^x, \cdots, \tilde{\tilde{e}}_{n-1}^x)^T$, and solving the recurrence in (3.20) for $\tilde{\tilde{e}}_{n-1}^x$ and simplifying, we have

$$e_n^y = \rho e_{n-1}^y + O(h^{k_I}) + O(\epsilon_2/h) + O(\epsilon_1).$$

Solving the above recurrence for e_n^y, we obtain

$$\|e_n^y\| = O(h^{k_I}) + O(\epsilon_2/h) + O(\epsilon_1). \tag{3.23}$$

Now for a linear problem, if we assume that the residuals from the Newton iteration satisfy $\|\eta_i^x\| = O(h^{k_G})$ and $\|\eta_i^y\| = O(h^{k_I+1})$, then we have shown the desired result for (3.3). For the nonlinear analysis, we will follow a strategy similar to that used in [13]. Recall that, by definition, η_i^x consists of terms of the form

$$\frac{\partial^2 g_1}{\partial x^2}E_i^x E_i^x, \quad \frac{\partial^2 g_1}{\partial x \partial y}E_i^x E_i^y, \quad \frac{\partial^2 g_1}{\partial y^2}E_i^y E_i^y.$$

while η_i^y is composed of terms of the form

$$\frac{\partial^2 g_2}{\partial x^2}E_i^x E_i^x.$$

Using the solution $\tilde{\tilde{e}}_{n-1}^x$ to (3.20) and equations (3.7), (3.8), (3.15), and (3.21), it can be shown that $\|E_i^y\|$ and $\|E_i^x\|$ satisfy

$$\begin{aligned} \|E_i^y\| &= O(h^{k_I}) + O(\epsilon_2/h) + O(\epsilon_1) + O((I - H)e_0^x) \\ \|E_i^x\| &= O(h^{k_G}) + O(\epsilon_2) + O(\epsilon_1) + O((I - H)e_0^x). \end{aligned} \tag{3.24}$$

If we assume that the residuals from the Newton iteration satisfy $\|\hat{\eta}^x\| = O(h^{k_G})$ and $\|\hat{\eta}^y\| = O(h^{k_I+1})$, where $\hat{\eta}$ is the contribution to η from errors in the Newton iteration, and that $(I - H)e_0^x = O(h^{k_G})$, then we have

$$\begin{aligned} \|\eta^x\| &\leq O(h^{k_G}) + (O(h^{k_I}) + O(\epsilon_1) + O(\epsilon_2/h))^2 \\ &\leq O(h^{k_G}) + O(h^{2k_I}) + O(\epsilon_1^2) + O(\epsilon_1\epsilon_2/h) + O(h^{k_I}\epsilon_1) \\ &\quad + O(h^{k_I-1}\epsilon_2) + O(\epsilon_2^2/h^2) \\ &\leq K_1(h^{k_G} + h^{k_I}\epsilon_1 + \epsilon_1^2 + \epsilon_1\epsilon_2/h + h^{k_I-1}\epsilon_2 + \epsilon_2^2/h^2). \end{aligned}$$

Similarly,

$$\begin{aligned} \|\eta^y\| &\leq O(h^{k_I+1}) + (O(h^{k_G}) + O(\epsilon_1) + O(\epsilon_2))^2 \\ &\leq K_2(h^{k_I+1} + h^{k_G}\epsilon_1 + \epsilon_1^2 + \epsilon_1\epsilon_2 + h^{k_G}\epsilon_2 + \epsilon_2^2). \end{aligned}$$

Now we let ϵ_1 and ϵ_2 be the solutions to

$$\begin{aligned} \epsilon_1 &= K_1(h^{k_G} + h^{k_I}\epsilon_1 + \epsilon_1^2 + \epsilon_1\epsilon_2/h + h^{k_I-1}\epsilon_2 + \epsilon_2^2/h^2) \\ \epsilon_2 &= K_2(h^{k_I+1} + h^{k_G}\epsilon_1 + \epsilon_1^2 + \epsilon_1\epsilon_2 + h^{k_G}\epsilon_2 + \epsilon_2^2). \end{aligned} \tag{3.25}$$

We would like to conclude that $\epsilon_1 = O(h^{k_G})$, $\epsilon_2 = O(h^{k_I+1})$. Now if we assume that $k_I \geq 2$ and solve (3.25) for ϵ_1 and ϵ_2 by functional iteration, starting with initial values that satisfy

$$\begin{aligned} \epsilon_2^{(0)} &= O(h^{k_I+1}) \\ \epsilon_1^{(0)} &= O(h^{k_G}), \end{aligned}$$

then it is easy to see that the spectral radius of the iteration matrix is less than one, and we can use the contraction mapping theorem to conclude that it converges to a solution that satisfies $\epsilon_1 = O(h^{k_G})$, $\epsilon_2 = O(h^{k_I+1})$. For $k_I = 1$, we cannot apply the theorem directly because the spectral radius is larger than one, but if we scale the variables by $\bar{\epsilon}_1 = \epsilon_1/\sqrt{h}$ and $\bar{\epsilon}_2 = \epsilon_2/h$, we can then apply the same strategy to reach the conclusion.

We have shown the result for (3.3), and it remains for us to demonstrate that we can extend the conclusions to (3.1). This is easy to do, following arguments similar to those used in [11]. For systems with index 1 constraints mixed with index 2 constraints,

$$x' + f_1(x, y, z, t) = 0, \quad f_2(x, y, t) = 0, \quad f_3(x, t) = 0, \tag{3.26}$$

where $\partial f_2/\partial y$ and $[(\partial f_3/\partial x)(\partial f_1/\partial z)]$ are both nonsingular, we can solve the difference equations corresponding to the second equation in (3.26) for Y_i at each stage i, so that the results just shown for the order of accuracy of x_n and z_n are valid. By solving for the error in y_n in terms of the error in y_{n-1} and the internal stage errors in x, it is easy to see that for strictly stable implicit Runge–Kutta methods and consistent initial conditions, the error in y_n is no worse than $O(h^{k_G})$.

Similarly, we can see that the result extends to systems of the form

$$F(x', x, y, z, t) = 0, \quad f_2(x, y, t) = 0, \quad f_3(x, t) = 0, \tag{3.27}$$

where $\partial F/\partial x'$ is nonsingular, by noting that the first equation in (3.27) can be solved for x' to obtain a system of the form (3.26).

Finally, if in (3.1) $\partial g/\partial y$ is not identically zero but is singular and has constant rank, then we can use a result of Dolezal [8] that there exist smooth nonsingular transformations which bring the system to the form (3.27), and which do not include any change of variables involving x. Thus the conclusions are valid for (3.1).

We should note that this theorem gives only a lower bound on the order of the method, and therefore does not exclude the possibility of a more accurate solution. However, numerical experiments in the next section demonstrate that some implicit Runge–Kutta methods do indeed suffer this order reduction.

4. Numerical experiments. In this section we present the results of some numerical experiments on linear and nonlinear index 2 semi-explicit systems. The experiments confirm that the order reduction effects predicted in § 3 can occur in practice, and also raise some interesting questions for future research.

The numerical experiments described in this section were restricted to the four L-stable formulae discussed in § 2. The results given here were obtained using a fixed stepsize code which implements a general M-stage IRK method, given the method coefficients. The nonlinear equations at each timestep were solved by Newton iteration. An analytic iteration matrix was provided to the code for all of the problems. The computations were performed in double and single precision on an Alliant FX/8 computer. The unit roundoff error on this machine is approximately 6×10^{-8} in single precision and 1×10^{-16} in double precision.

The first test problem was a linear problem having four differential equations and one algebraic equation [2]:

$$
\begin{aligned}
x_1' &= -e^t x_1 + x_2 + x_4 + y - e^{-t} \\
x_2' &= -x_1 + x_2 - \sin(t)x_3 + y - \cos(t) \\
x_3' &= \sin(t)x_1 + x_3 + \sin(t)x_4 - \sin^2(t) - e^{-t}\sin(t) \\
x_4' &= \cos(t)x_2 + x_3 + \sin(t)x_4 - e^{-t}(1+\sin(t)) - \cos^2(t) - e^t \\
0 &= x_1 \sin^2(t) + x_2 \cos^2(t) + (x_3 - e^t)(\sin(t) + 2\cos(t)) \\
&\quad + \sin(t)(x_4 - e^{-t})(\sin(t) + \cos(t) - 1) - \sin^3(t) - \cos^3(t)
\end{aligned}
\tag{4.1}
$$

The exact solution to this system is $x_1 = \sin(t)$, $x_2 = \cos(t)$, $x_3 = e^t$, $x_4 = e^{-t}$, and $y(t) = e^t \sin(t)$. It is easy to verify that system (4.1) is index 2 for all t. We solved this test problem for a sequence of fixed stepsizes on the interval [0,1] in double precision using the four IRK methods. Consistent initial values were specified at $t = 0$. After computing the global error at $t = 1$, an observed rate of convergence was determined by computing the ratio of global errors when successively halving the stepsize. The number of steps taken in [0,1] ranged from 2 to 16384. The observed order of the global error was two in all variables when the test problem was solved by the two-stage SIRK method, agreeing with the order predicted by the theory. However, when the five-stage DIRK was used, we found that the state variables x were computed to an accuracy of $O(h^4)$ (i.e., the nonstiff ODE order k_d), thereby exceeding the lower bound k_G^x on the order predicted in § 3. The algebraic variable y was computed to only $O(h)$ accuracy, which agrees with the lower bound value of k_G^y. The three-stage SIRK method, as expected, determined the algebraic variable to $O(h^2)$ accuracy and the state variables to $O(h^3)$ accuracy. Note that the SIRK methods, as well as the DIRK method, achieved the nonstiff ODE order of accuracy in the state variables for this linear test problem. Finally, we found that the seven-stage extrapolation formula was order 3 in all variables, thereby exceeding the order predicted by the lower bound. From these results, it might be tempting to conclude that the convergence theorem could be strengthened to predict that IRK methods will compute the state variables x to $O(h^{k_d})$ accuracy. However, this is not the case, as we can see from the next two examples. The orders k_g^x and k_g^y observed for this linear test problem are summarized with the predicted lower bounds k_G^x and k_G^y (recall that $k_G^x = \min(k_d, k_I + 1)$ and $k_G^y = k_I$) and the nonstiff ODE and internal orders, k_d and k_I respectively in Table 4.1.

To illustrate the effects of roundoff error on the solution accuracy of an index 2 system, we performed the same experiments in single precision. As to be expected from the analysis, roundoff plays an important part in the calculations, especially for

TABLE 4.1
Predicted/observed orders for linear test problem.

L-stable methods	k_d	k_I	k_G^x	k_g^x	k_G^y	k_g^y
1. Two-stage SIRK	2	2	2	2	2	2
2. Five-stage DIRK	4	1	2	4	1	1
3. Three-stage SIRK	3	2	3	3	2	2
4. Seven-stage Extrapolation	3	1	2	3	1	3

very small stepsizes. For this problem, the effects of roundoff become apparent in the y variable for most of the methods at a stepsize corresponding to about 128 timesteps in [0,1]. For smaller stepsizes, the errors in this variable can actually increase as the stepsize decreases. In double precision, these roundoff errors become dominant only for a much greater number of timesteps. In the single precision tests, we were able to compute the x components to an accuracy roughly comparable to the machine epsilon.

Next we investigated the behavior of the IRK formulae on two nonlinear problems. We chose to study the index 3 pendulum problem simply because it has been studied so frequently by DAE researchers [11],[12] and can be posed as an index 2 problem [11]. The other nonlinear problem considered arises in the context of trajectory prescribed path control problems [3]. The exact solution is not available for either problem, so we first had to generate a "true" solution that could be used for comparisons. The corresponding index 1 systems were formulated and solved by the code DASSL [15] in double precision with extremely tight error tolerances.

Consider the pendulum problem as formulated in [11]. Note that this formulation ensures that the original index 3 algebraic constraint is satisfied even though the index of the system has been reduced to two.

$$
\begin{aligned}
x_1' &= x_3 - x_1 y_2 \\
x_2' &= x_4 - x_2 y_2 \\
x_3' &= -y_1 x_1 \\
x_4' &= -y_1 x_2 - 1 \\
0 &= (1 - x_1^2 - x_2^2)/2 \\
0 &= x_1 x_3 + x_2 x_4
\end{aligned} \tag{4.2}
$$

The algebraic constraints in this problem are nonlinear, yet for a constant state the algebraic variables appear only linearly in the system. The pendulum problem was solved using the fixed stepsize IRK code on the interval [0,1] for a sequence of stepsizes with each particular IRK formula. Consistent initial conditions were specified, namely $x_1 = 1$, $x_2 = x_3 = x_4 = y_1 = y_2 = 0$. Rates of convergence for each method were estimated as in the linear problem by comparing the global errors at $t = 1$ for numerical solutions produced by successively halving the stepsize. The computations were performed in double precision, with the number of steps in [0,1] ranging again from 2 to 16384. Unlike the results for the linear problem, it does not appear that these formulae determine the state variables x to the ODE order of accuracy. In particular, the five-stage DIRK method behaved as expected from the index 2 convergence theorem, finding the state variables x to no more than $O(h^2)$ accuracy, and the algebraic variables y to $O(h)$ accuracy. Meanwhile, the three-stage SIRK method still appeared to be third-order in the state variables, while the algebraic variables were determined with close to second-order accuracy. The seven-stage extrapolation

method continued to perform admirably, yielding third order accuracy in all variables. Finally, the two-stage SIRK method remained second-order accurate for all variables. The numerical results for the pendulum problem are summarized in Table 4.2.

TABLE 4.2
Predicted/observed orders on pendulum problem.

L-stable methods	k_d	k_I	k_G^x	k_q^x	k_G^y	k_q^y
1. Two-stage SIRK	2	2	2	2	2	2
2. Five-stage DIRK	4	1	2	2	1	1
3. Three-stage SIRK	3	2	3	3	2	2
4. Seven-stage Extrapolation	3	1	2	3	1	3

We repeated these experiments in single precision and found, as in the linear test problem, that the accuracy of the y components deteriorated starting at around 128 timesteps in [0,1]. We were able to compute the x components to an accuracy comparable to the machine epsilon, even for very small stepsizes.

The trajectory problem was posed in [2] as representative of the type of trajectory prescribed path control problems of current interest. The constraints are quite nonlinear in both the state and algebraic variables, while the two algebraic constraints are designed to simply prescribe two of the state variables as functions of time. Initial values for the state variables are known exactly, but initial values for the two algebraic variables (namely, angle of attack α and bank angle β) were determined numerically from the corresponding index 1 system. Specifically, the test problem used the following initial values for the state variables: altitude $H = 100,000$ feet, longitude $\xi = 0°$, latitude $\lambda = 0°$, relative velocity $V_R = 12000$ feet/second, flight path angle $\gamma = -1°$, and azimuth $A = 45°$. Angle of attack and bank angle were initialized to $\alpha = 2.672870042°$ and $\beta = -.0522095861634°$, respectively. The "small" errors in the initial values for the algebraic variables are annihilated in one step by the IRK methods chosen, as a result of their L-stability property. This problem was solved in double precision for fixed stepsizes on the interval [0,300], and the global errors in the solution were computed at $t = 300$ using the "true" solution described earlier. The system is composed of the following six equations of motion and two prescribed path control constraints:

$$
\begin{aligned}
H' &= V_R \sin(\gamma) \\
\xi' &= V_R \cos(\gamma) \sin(A)/(r \cos(\lambda)) \\
\lambda' &= V_R \cos(\gamma) \cos(A)/r \\
V_R' &= -D/m - g \sin(\gamma) \\
&\quad - \Omega_E^2 r \cos(\lambda)(\sin(\lambda) \cos(A) \cos(\gamma) - \cos(\lambda) \sin(\gamma)) \\
\gamma' &= L \cos(\beta)/(mV_R) + \cos(\gamma)(V_R^2/r - g)/V_R + 2\Omega_E \cos(\lambda) \sin(A) \\
&\quad + \Omega_E^2 r \cos(\lambda)(\sin(\lambda) \cos(A) \sin(\gamma) + \cos(\lambda) \cos(\gamma))/V_R \\
A' &= L \sin(\beta)/(mV_R \cos(\gamma)) + V_R \cos(\gamma) \sin(A) \tan(\lambda)/r \\
&\quad - 2\Omega_E(\cos(\lambda) \cos(A) \tan(\gamma) - \sin(\lambda)) \\
&\quad + \Omega_E^2 r \cos(\lambda) \sin(\lambda) \sin(A)/(V_R \cos(\gamma)) \\
0 &= \gamma + 1 + 9(t/300)^2 \\
0 &= A - 45 - 90(t/300)^2
\end{aligned}
$$

where

$$
\begin{aligned}
r &= H + a_e \\
a_e &= 20902900 \text{ feet, the earth radius} \\
g &= \mu/r^2, \text{ the gravity force} \\
\mu &= .14076539E + 17 \\
\Omega_E &= .72921159E - 4 \\
m &= 2.890532728, \text{ the mass of the vehicle} \\
L &= .5\rho C_L S V_R^2, \text{ the aerodynamic lift force} \\
\rho &= .002378e^{-H/23800.}, \text{ the atmospheric density} \\
C_L &= .01\alpha, \text{ the aerodynamic lift coefficient} \\
S &= 1, \text{ the vehicle cross-sectional reference area} \\
D &= .5\rho C_D S V_R^2, \text{ the aerodynamic drag force} \\
C_D &= .04 + .1C_L^2, \text{ the aerodynamic drag coefficient.}
\end{aligned}
$$

Since the corrector iteration was terminated with a fairly tight tolerance, the values of the two state variables prescribed by the algebraic constraints (namely, the flight path angle γ and the azimuth A) were computed almost exactly for all the IRK methods considered. Rates of convergence for all the other variables have been estimated as we described earlier for the other test problems. The results were similar to those obtained for the pendulum problem. The numerical solution produced by the DIRK method was close to second-order accurate in the state variables, and first-order accurate in the algebraic variables. The extrapolation formula yielded close to third-order accurate solutions in all variables, while the two-stage SIRK method was clearly second-order for all variables. The three-stage SIRK method surprised us by producing a third-order accurate solution for the algebraic variables as well as for the state variables. We suspect this difference in performance for this particular IRK method, when compared to its results on the linear test problem and the pendulum problem, must be due to the specific coupling of the state and algebraic variables in this nonlinear system. The numerical results for the trajectory problem are summarized in Table 4.3.

TABLE 4.3
Predicted/observed orders on trajectory problem.

L-stable methods	k_d	k_I	k_G^x	k_g^x	k_G^y	k_g^y
1. Two-stage SIRK	2	2	2	2	2	2
2. Five-stage DIRK	4	1	2	2	1	1
3. Three-stage SIRK	3	2	3	3	2	3
4. Seven-stage Extrapolation	3	1	2	3	1	3

We repeated these experiments in single precision and found, as in the other problems, that the accuracy of the algebraic components α and β deteriorated starting at approximately 128 timesteps in [0,300]. The accuracy of the state variables, computed for small stepsizes to an accuracy comparable to the machine epsilon, was significantly higher than the accuracy attained in the algebraic variables. In addition, the accuracy of the state variables did not deteriorate appreciably even for very small stepsizes in single precision.

In conclusion, we see that the observed convergence rates of these IRK methods applied to nonlinear semi-explicit index 2 systems can sometimes be as slow as the lower bounds derived in § 3 would indicate. Some formulae, in particular the extrapolation method, achieve an order of accuracy exceeding the predicted lower bounds, suggesting that a stronger convergence theorem might be possible. On the other hand, this theorem can be applied to construct relatively high-order IRK methods for semi-explicit index 2 DAEs, simply by considering methods that have an internal order as high or nearly as high as the ODE order. In particular, consider the class of M-stage singly-implicit Runge–Kutta methods (SIRKs) whose coefficient matrix $\mathcal{A}$ is characterized by its single-fold eigenvalue. Butcher [6] has shown how these IRK formulae can be implemented very efficiently. There are two types of SIRKs, the *transformed* type [5] and the *collocation* type [14]. It is easy to show that, for index 2 problems, transformed SIRKs will be at least order $M-1$ (since $k_I \geq M-1$), while collocation SIRKs will be order M (since $k_I = M$). Note that the second-order, two-stage SIRK formula which has appeared so promising in our numerical experiments is in fact a collocation method. Note also that if an L-stable SIRK formula is desired, the eigenvalue of the $\mathcal{A}$ matrix may be selected to satisfy $L_M(\lambda^{-1}) = 0$ where L_M is the Laguerre polynomial of degree M. Methods of this type have been derived for orders up to and including six. In summary, we expect the SIRK methods to perform very well on index 2 problems. However, the development of an efficient IRK code for DAEs with index greater than one remains a challenge because of the difficulties in developing appropriate error control strategies for all the variables. In addition, the code design should include a mechanism for detecting when round off errors become significant, particularly in the algebraic components. Methods based on extrapolation appear to be promising from the point of view of addressing this concern.

Acknowledgment. The authors would like to thank Clement Ulrich for his assistance in performing the numerical experiments.

REFERENCES

[1] R. ALEXANDER, *Diagonally implicit Runge–Kutta methods for stiff ODEs*, SIAM J. Numer. Anal., 14 (1977), pp. 1006–1022.

[2] K. E. BRENAN, *Stability and convergence of difference approximations for higher index differential-algebraic systems with applications in trajectory control*, Ph.D. thesis, University of California at Los Angeles, 1983.

[3] ———, *Numerical simulation of trajectory prescribed path control problems by the backward differentiation formulas*, IEEE Journal on Automatic Control, AC-31 (1986), pp. 266–269.

[4] K. BURRAGE, *Efficiently implementable algebraically stable Runge–Kutta methods*, SIAM J. Numer. Anal., 19 (1982), pp. 245–258.

[5] ———, *A special family of Runge–Kutta methods for solving stiff differential equations*, BIT, 18 (1978), pp. 22–41.

[6] J. C. BUTCHER, *On the implementation of Runge–Kutta methods*, BIT, 16 (1976), pp. 237–240.

[7] J. R. CASH, *Diagonally implicit Runge–Kutta formulae with error estimates*, J. Inst. Math. Appl., 24 (1979), pp. 293–301.

[8] V. DOLEZAL, *The existence of a continuous basis of a certain linear subspace of E_r, which depends on a parameter*, Časopis pro pěstováni matematiky, roč. 89, Praha (1964), pp. 466–468. (In Russian.)

[9] R. FRANK, J. SCHNEID, AND C. W. UEBERHUBER, *Order results for implicit Runge–Kutta methods applied to stiff systems*, SIAM J. Numer. Anal., 22 (1985), pp. 515–534.

[10] F. R. GANTMACHER, *The Theory of Matrices*, Vol. 2, Chelsea, New York, 1964.

[11] C. W. GEAR, B. LEIMKUHLER, AND G. K. GUPTA, *Automatic integration of Euler–Lagrange equations with constraints*, J. Comput. Appl. Math., 12 and 13 (1985), pp. 77–90.

[12] B. J. LEIMKUHLER, *Error estimates for differential-algebraic equations*, Report UIUCDCD-R-86-1287, Dept. of Computer Science, University of Illinois, 1986.

[13] P. LÖTSTEDT AND L. PETZOLD, *Numerical solution of nonlinear differential equations with algebraic constraints* I: *Convergence results for backward differentiation formulas*, Math. Comput., 46 (1986), pp. 491–516.

[14] S. P. NØRSETT, *Runge–Kutta methods with a multiple real eigenvalue only*, BIT, 16 (1976), pp. 388–393.

[15] L. R. PETZOLD, *A description of DASSL: A differential/algebraic system solver*, IMACS Trans. Sci. Comput., Vol. 1, R. S. Stepleman, ed., 1982.

[16] ———, *Order results for implicit Runge–Kutta methods applied to differential/algebraic systems*, SIAM J. Numer. Anal., 23 (1986), pp. 837–852.

[17] A. PROTHERO AND A. ROBINSON, *On the stability and accuracy of one-step methods for solving stiff systems of ordinary differential equations*, Math. Comp., 28 (1974), pp. 145–162.