MATHEMATICS OF COMPUTATION
VOLUME 51, NUMBER 184
OCTOBER 1988, PAGES 659-676

# Backward Differentiation Approximations of Nonlinear Differential/Algebraic Systems

By Kathryn E. Brenan and Bjorn E. Engquist*

**Abstract.** Finite difference approximations of dynamical systems modelled by nonlinear, semiexplicit, differential/algebraic equations are analyzed. Convergence for the backward differentiation method is proved for index two and index three problems when the numerical initial values obey certain constraints. The appropriate asymptotic convergence rates and the leading error terms are determined.

**1. Introduction.** The most general systems of differential/algebraic equations (DAE's) arise in the *fully implicit* form,

$$R(t, x, x') = 0, \tag{1.1}$$

where $x = x(t)$ and $R$ are vectors of dimension $r$. The Jacobian matrix $\partial R/\partial x'$ is assumed to be singular. If it is nonsingular, system (1.1) is an implicit set of ordinary differential equations (ODE's) and can be, at least theoretically, reformulated as $x' = f(t, x)$.

The $k$-step backward differentiation formula (BDF) was introduced by Gear [8] in 1971 for the numerical integration of DAE systems. This approach has been universally accepted in industry as well as in academia as the standard numerical method for solving DAE systems. To apply this method to (1.1), replace the derivatives $x'$ by their BDF approximation with constant stepsize $h$ and evaluate the equations at $t_n$, where $t_n = t_0 + nh$. The resulting system of difference equations,

$$R\left(t_n, x_n, h^{-1}\sum_{i=0}^{k}\alpha_i x_{n-i}\right) = 0 \tag{1.2}$$

is then solved for the numerical approximation $x_n$ of $x(t_n)$.

The analytical and numerical theory for linear $R$ with constant coefficients is well understood [5], [7], [25], [27]. The variable coefficient and nonlinear cases are much harder. In contrast to the theory for the numerical approximation of ODE's, the convergence results for linear, constant coefficient systems do not extend to all linear, variable coefficient or nonlinear problems. In this paper we will restrict our interest to the class of *solvable* DAE's. A solvable DAE is one for which solutions exist, and for which solutions having the same initial values are identical [9]. Now, while the $k$-step BDF have been shown to converge on solvable, linear constant

Received April 11, 1986; revised November 9, 1987.
1980 *Mathematics Subject Classification* (1985 *Revision*). Primary 65L05.
*This work was supported in part by NSF Grant No. DMS85-03294, ARO Grant No. DAAG29-85-K-0190, DARPA Grant in the ACMP Program, ONR Grant N00014-86-K-0691, and NASA Langley Grant NAG1-270.

coefficient systems [25], they may be unstable for some fully implicit DAE's. For example, in [10] a fully implicit, linear, time-varying solvable DAE system is presented for which the one-step BDF is unstable and thus does not converge. On the other hand, numerical experiments strongly suggest that the BDF are convergent for some classes of DAE's [1]. This example demonstrates that in order to establish a convergence theory for numerical approximations to the solutions of nonlinear (including linear time-varying) DAE's, a subclass of problems of the form (1.1) has to be chosen. It is the purpose of this paper to define an appropriate subclass of nonlinear DAE's which cover many of the applications, and to develop a rigorous convergence theory for that class. The stability results derived in [1] for the linear homogeneous systems corresponding to this subclass of nonlinear systems will be summarized as well. Before the results can be presented, some background material must be given.

The behavior of DAE systems is directly related to a property called the *index* [5], also sometimes referred to as the *degree of nilpotency* [28]. For example, consider the following subclass of system (1.1) which we will refer to as *semiexplicit* DAE's:

$$y' = E(t, y, u), \tag{1.3}$$

$$0 = H(t, y, u), \tag{1.4}$$

where $y = y(t)$ and $E$ are $l$ vectors, and $u = u(t)$ and $H$ are $m$ vectors. Semiexplicit DAE systems arise in dynamic simulations of mechanical problems [17], fluid flow problems [18], [21], and optimal [4] and trajectory control [2] problems. Essentially, the index of system (1.3), (1.4) is one more than the number of times it is necessary to differentiate the algebraic equations before the algebraic variables $u$ can be explicitly determined. Hence, if the Jacobian matrix $\partial H/\partial u$ is nonsingular, then the system has index one [11], since no differentiation is required. If $\partial H/\partial u$ is singular, then the index is at least two. If the algebraic subsystem (1.4) is not present, or if the DAE system is simply a set of implicit ODE's, the index is zero. A reduction algorithm described in [9] gives a precise definition of the index.

Recall that initial values of ODE's may be specified arbitrarily. However, initial values for a DAE system of index $\nu$ must in general satisfy consistency relations which may involve up to $\nu - 1$ derivatives of some of the variables. For example, initial values for semiexplicit DAE systems must at least satisfy the algebraic equations (1.4).

Not only is the index of the system important in the characterization of the system's solutions, but it is also critical to the convergence and stability properties of numerical approximations. For example, the relation of the index to the stability and convergence properties of the BDF is very well understood for linear, constant coefficient systems,

$$Ax' + Bx = f(t), \tag{1.5}$$

where $A$ and $B$ are constant $r \times r$ matrices. If system (1.5) with index $\nu$ is solved by the $k$-step BDF ($k \leq 6$) with constant stepsize $h$, the numerical solution converges globally to $O(h^k)$ accuracy after $(\nu - 1)k + 1$ steps [25]. If the one-step BDF (i.e., Backward Euler) is used to solve an index three system, the global error is $O(1)$

after one step [11]. However, after three steps of constant length $h$, the numerical solution is $O(h)$ accurate.

If the system's index is restricted to one, then the numerical solution determined by the $k$-step BDF converges to the solution of the general implicitly formulated problem (1.1) [9]. However, as was mentioned earlier, subclasses of (1.1) must be considered when proving convergence for higher index systems. Since numerical solutions of some higher index, semiexplicit systems have been experimentally determined to $O(h^k)$ accuracy by the $k$-step BDF, it is natural to investigate the convergence properties of the BDF on this subclass of (1.1) [1], [2].

The subclasses of DAE's studied in this paper are higher index problems (i.e., index greater than one) which are sometimes labelled *ill-posed* [19], [20] or *algebraically incomplete* [23]. The classification of these higher index systems as ill-posed should be understood in the strict sense that they may not exhibit a continuous dependence in the maximum norm on the inhomogeneous terms. In fact, it will be necessary to impose additional smoothness on the systems and to consider a weaker form of stability where the solution depends continuously on the inhomogeneous terms and some of their derivatives. Solvability of these systems also requires this regularity. It is still possible for the BDF to produce a convergent numerical solution in spite of this weaker form of stability. Furthermore, it is important in practice to understand how numerical methods behave when applied to higher index systems because higher index problems arise in applications much more frequently than first thought. Until recently, many engineers solving DAE's were not cognizant of the index. Except in special cases, it is still difficult in practice to determine what the index of a given DAE is. In any case, it is useful to identify these special cases. In particular, the nonlinear systems we study here were originally motivated by the need to solve trajectory prescribed path control problems of current interest to the aerospace industry. Typically, these problems are index two or three, semiexplicit, nonlinear DAE's. While sometimes their index can be reduced (by differentiation or some other technique), in general that approach has not been very satisfactory in practice.

In this paper, convergence of the BDF is proven for semiexplicit systems of index two and three. In particular, consider the index two system

$$y' = E(t, y, u), \tag{1.6}$$

$$0 = H(t, y), \tag{1.7}$$

where the $m \times m$ matrix $(\partial H/\partial y)\ (\partial E/\partial u)$, evaluated at the solution $(y(t), u(t))$, is assumed to be nonsingular for all $t$ in some interval $I$, and $m \leq l$. Let us also consider the index three system,

$$v' = F(t, v, w, u), \tag{1.8}$$

$$w' = G(t, v, w), \tag{1.9}$$

$$0 = H(t, w), \tag{1.10}$$

where the $m \times m$ matrix $(\partial H/\partial w)(\partial G/\partial v)(\partial F/\partial u)$ is nonsingular along the solution $(v(t), w(t), u(t))$ for all $t \in I$. In system (1.8)–(1.10), $v$ and $F$ are vectors

of dimension $p$, denoted as $\dim v = \dim F = p$, while $\dim w = \dim G = q$ and $\dim u = \dim H = m$. To avoid the case of an overdetermined system, we assume $m \leq \min(p, q)$. Throughout this paper, the functions $E, F, G$, and $H$ are assumed to be sufficiently smooth functions of all their arguments as required for the convergence analysis contained herein.

Typically, the variables for which there are explicit differential equations are referred to as *state* variables, while those variables appearing only algebraically will be called the *algebraic* or control variables. In system (1.6), (1.7) the $l$ state variables are $y$, while in system (1.8)–(1.10), the state variables are $y = (v, w)$ with $\dim y = l = p+q$. In both systems, the algebraic variables are $u$. In [1] it was proven that a linear, index two system and a linear, index three system corresponding to (1.6), (1.7) and (1.8)–(1.10), respectively, are solvable. The proofs are straightforward and involve reducing the systems to sets of explicit ODE's by differentiating the algebraic equations and substituting for the state variables' derivatives from the DAE's. For *consistent* initial values, the solvability of the DAE's is then established by applying existence and uniqueness theorems for ODE's. Note that a consistent set of initial values must satisfy not only the algebraic equations given in the DAE, but also those equations arising in the reduction process. For the nonlinear DAE's (1.6), (1.7) and (1.8)–(1.10), it is technically necessary to assume there exist solutions satisfying the algebraic equations corresponding to the associated index one problems, including (1.7) and (1.10), respectively. Then, solvability of the nonlinear DAE's may be established as in the linear case—namely, differentiate the algebraic equations and apply the implicit function theorem to solve for the algebraic variables, thereby reducing the DAE's to explicit ODE systems. This last step utilizes the assumptions on the nonsingularity of the matrix products $(\frac{\partial H}{\partial y})(\frac{\partial E}{\partial u})$ and $(\frac{\partial H}{\partial w})(\frac{\partial G}{\partial v})(\frac{\partial F}{\partial u})$.

The facts that system (1.6), (1.7) and the linear system corresponding to (1.8)–(1.10) are index two and three, respectively, was first established in [1] by applying the reduction technique described in [9]. Essentially, the index is one more than the number of times it is necessary to differentiate the algebraic constraints before the algebraic variables appear explicitly. The index of the nonlinear system (1.8)–(1.10) follows in a similar way.

In the following section, convergence theorems for the BDF methods approximating the solutions to (1.6), (1.7) and (1.8)–(1.10) are stated. The rate of convergence is equivalent to the corresponding rate for linear, constant coefficient problems. The leading error term in the asymptotic expansion is given. This section also contains necessary definitions and a brief outline of the proofs. In particular, it is pointed out how this analysis differs from the corresponding theory for ODE's. The convergence proof for the index two system is given in Section 3, while the corresponding proof for the index three system is presented in the Supplement.

This paper is an extension of the convergence results derived in [1], and is a condensed version of the earlier report [3] in which the proofs are carried out in somewhat greater detail. While a convergence analysis of the $k$-step BDF was done for the nonlinear, index two system (1.6), (1.7) in [1], the proof presented in [1] was valid for only the *linear*, time-varying index three system corresponding to the nonlinear system (1.8)–(1.10) of interest here. The analysis in this paper

also leads to new results concerning the leading error terms. Extensive numerical experiments have been conducted to verify the rates of convergence for both linear and nonlinear index two and three systems of these forms [1]. Numerical results for a particularly nonlinear, index three problem arising in trajectory control have already been presented in [2].

Using a different approach than the one employed here, convergence results for the BDF have been obtained in [18] for index two systems having a form equivalent to (1.6), (1.7) and for a subclass of the index three systems (1.8)–(1.10). The proofs presented there rely on local rates of convergence established in [1]. Throughout this paper, we will point out the key differences between the results given in [1], [18], and this paper.

**2. Notation, Definition, and Statement of Results.** Before stating the convergence results obtained for systems (1.6), (1.7) and (1.8)–(1.10), it is necessary to present some definitions. A set of initial values $(y_0, u_0)$ at $t_0$ is said to be *consistent* for a DAE system if there exists at least one solution $(y(t), u(t))$ assuming those values at $t_0$. For $k \geq 2$ the $k$-step BDF requires not only initial values, but the following set of values here called *starting values*:

$$\mathbf{y}_{k-1} = [y_{k-1}^T, y_{k-2}^T, \ldots, y_0^T]^T, \tag{2.1}$$

$$\mathbf{u}_{k-1} = [u_{k-1}^T, u_{k-2}^T, \ldots, u_0^T]^T, \tag{2.2}$$

where the superscript $T$ denotes the transpose. The $l$ vector $y_j$ and the $m$ vector $u_j$ denote the numerical solution at $t_j = t_0 + jh$ for $j = 0, 1, \ldots, k-1$ and $t_j \in J = [t_0, t_0 + T] \subseteq I$. We will say these starting values are *numerically consistent to order* $k+1$ if there exists a solution to the index $\nu$ system such that

$$\|y_j - y(t_j)\| \leq K_1 h^{k+1}, \tag{2.3}$$

$$\|H(t_j, y_j)\| \leq K_2 h^{k+\nu-1} \tag{2.4}$$

for some constants $K_1$, $K_2$, and $j = 0, 1, \ldots, k-1$. In general, we could use any vector norm and corresponding consistent matrix norm during the analysis, but it is convenient to restrict the analysis to the maximum norm. We shall show that the starting values for $u$ are not critical in the convergence analysis for systems (1.6), (1.7) and (1.8)–(1.10).

Let the BDF difference approximation (1.2) of the semiexplicit problems (1.6), (1.7) and (1.8)–(1.10) have the form,

$$\sum_{i=0}^{k} \alpha_i y_{n-i} = hE(t_n, y_n, u_n), \tag{2.5}$$

$$0 = H(t_n, y_n), \tag{2.6}$$

where $E = (F, G)$ and $y = (v, w)$ for the index three system (1.8) (1.10). Given starting values satisfying (2.3), (2.4), a numerical solution $(y_n, u_n)$ of (2.5), (2.6) *converges globally with kth order accuracy* to a solution of the DAE system (1.6), (1.7) or (1.8)–(1.10) if

$$\|y_n - y(t_n)\| \leq K_3 h^k, \qquad \|u_n - u(t_n)\| \leq K_4 h^k$$

for $t_n \in [t_\theta, t_0 + T] \subseteq J$ where the constants $K_3$, $K_4$, and $\theta$ are independent of $h$.

Since instability of the BDF is a concern for fully implicit, higher index DAE's, we summarize some stability results derived in [1] for the linear, homogeneous systems corresponding to the nonlinear systems (1.6), (1.7) and (1.8)–(1.10) of interest here. For *arbitrary* starting values $\mathbf{y}_{k-1}$, the $k$-step BDF is stable (or *weakly stable* [6]) for these semiexplicit systems because there exist constants $\mathbf{K}^*$, $\mathbf{K}^{**}$, and $\mathscr{E}$ such that the following conditions hold uniformly in $n$ and $h$ for $t_{\bar{n}} \le t_n \le t_0 + T$,

$$\|\mathbf{y}_n\| \le \mathbf{K}^* e^{nh\mathscr{E}} \|\mathbf{y}_{k-1}\|, \qquad \|\mathbf{u}_n\| \le \mathbf{K}^{**} \|\mathbf{y}_{k-1}\|$$

where $\bar{n} = 3k - 1$ (i.e., after $k+1$ steps) for index two systems and $\bar{n} = 4k - 1$ (i.e., after $2k + 1$ steps) for index three systems. In fact, the state variables $\mathbf{y}$ are always computed stably in an index two system, but only after $k + 1$ steps in an index three problem. Moreover, the BDF produces a numerical solution for the algebraic variables having a *boundary layer* of instability of length $(k+1)$ steps for index two systems and of length $(2k + 1)$ steps for index three systems. These weak stability properties have since been studied for more general linear, semiexplicit, solvable, index two systems in [6]. In practice, weak stability describes how the BDF are sensitive to roundoff errors and errors in the starting values.

While it is true that the BDF methods are weakly stable for these semiexplicit systems, it is also clear that the initial steps may contain 'large' errors. In spite of that, we show in this paper that if the starting data is sufficiently accurate, the BDF does produce a numerical solution which converges to a solution of the *nonlinear* system with the expected $O(h^k)$ rate of convergence. It is precisely one point of this paper to determine how accurate the initial data must be in order to insure convergence (in particular, for index three problems). Not only is this a theoretical concern, but it is also of practical interest as evidenced by the application of the BDF methods to real problems (see numerical results given in [1] or [2]). It is a fact that the Backward Euler method fails to converge at the end of the first integration step when applied to an index three problem. Consequently, serious difficulties arise in practice when using state of the art software implementing the BDF methods (e.g., see [22]) on index three systems. It has even been observed that it is possible for the numerical solution to converge to a solution of the given DAE which is inconsistent with the given initial values for the algebraic variables [2]. This difficulty is due not only to the behavior of the Backward Euler method, but also to the nature of the nonlinear system being solved. In this particular trajectory problem, for a given set of initial state values, there is a nonunique solution for the algebraic variable. A one-step method has no memory, so the $O(1)$ error introduced into the algebraic variable during the first step may cause the numerical solution to jump to a different solution curve. If the initial values for the state variables are sufficiently accurate, this difficulty does not occur.

Now we can state the convergence results obtained for the $k$-step BDF when applied to the semiexplicit DAE systems (1.6), (1.7) or (1.8)–(1.10).

THEOREM 1. *There exists a numerical solution of the index two system* (1.6), (1.7) *by the k-step BDF with constant stepsize h for* $k < 7$ *which converges globally with kth order accuracy to a solution of* (1.6), (1.7) *if the starting values are numerically consistent to order* $k + 1$.

THEOREM 2. *There exists a numerical solution of the index three system* (1.8)–(1.10) *by the k-step BDF with constant stepsize h for* $k < 7$ *which converges globally with kth order accuracy to a solution of* (1.8)–(1.10) *after* $k+1$ *steps if the starting values are numerically consistent to order* $k+1$.

The definition of a numerically consistent set of starting values is natural in the following way. For a system of explicit ODE's, a unique solution exists corresponding to any set of *arbitrary* initial values. The numerical solution corresponding to a convergent difference approximation of order $k$ applied to an ODE will generally require starting values of accuracy $O(h^k)$. For a solvable DAE, solutions exist and are uniquely specified when a *consistent set* of initial values are given. It is therefore natural to require that the starting values for the numerical method also satisfy some consistency conditions as well. That is, we cannot expect convergence of the difference method for any arbitrary set of starting values located in a circle of radius $O(h^k)$ about a consistent set of initial values. Specifically, they must be chosen to be sufficiently close to the manifold containing the solution to the DAE.

In [1] and [18] it appears that convergence of the BDF has been proven for these systems when the starting values are accurate only to $O(h^k)$. However, if these starting values are not *numerically consistent*, then $O(h^k)$ convergence is not obtained until an additional $k$ steps are taken, where the algebraic equations are satisfied either exactly as in [1] or at least to some sufficiently small tolerance [18] (namely, to $O(h^{k+1})$ accuracy for index two systems and to $O(h^{k+2})$ for index three problems). Hence, in [1] or [18] it is proven that the $k$-step BDF, given arbitrary starting values of accuracy $O(h^k)$, converges with $O(h^k)$ accuracy to the analytic solution after $k+1$ steps for the index two system and after $2k+1$ steps for the (linear) index three system. During the initial steps, reduced rates of convergence, namely $O(h^{k-1})$, may be observed in the algebraic variables [1]. If the starting values are *numerically consistent*, then $O(h^k)$ convergence is achieved immediately for the index two system and after $k+1$ steps for the index three system.

In [18] it is assumed the algebraic equations are solved at each step to $O(h^{k+1})$ accuracy for the index two system and to $O(h^{k+2})$ accuracy for the index three system. These requirements are equivalent in our analysis to the restriction (2.4) for the starting values. In our proofs, we will assume that the algebraic equations are satisfied exactly at each step, but we could instead relax this requirement by enforcing (2.4) on each step. Specifically, the $k$-step BDF will converge with $O(h^k)$ accuracy if the algebraic equations are satisfied to $O(h^{k+1})$ accuracy for index two problems and to $O(h^{k+2})$ accuracy ($k = 1$ requires $O(h^{k+3})$) for index three systems. If the starting values do not satisfy this requirement (i.e., are not numerically consistent), then an additional $k$ steps will be required before $O(h^k)$ convergence is attained in all variables.

*Remark.* The convergence results given in Theorems 1 and 2 are also valid, without lengthening the boundary layer, when the consistency relations for the starting values are relaxed somewhat:

1. For the index two system, (2.4) follows immediately from (2.3), and hence is not a restriction.

2. In the convergence analysis for the index three system, we could relax (2.3) to

$$\|w_j - w(t_j)\| = O(h^{k+1}) \quad \text{and} \quad \|v_j - v(t_j)\| = O(h^k).$$

3. It is possible to prove Theorems 1 and 2 given starting values with error $O(h^k)$ providing they have the following special form: $(i = 0, 1, \ldots, k-1)$

$$y_i = y(t_i) + h^k c_{1,i} + O(h^{k+1}) \quad \text{(index two)},$$
$$w_i = w(t_i) + h^k c_{1,i} + O(h^{k+1}) \quad \text{(index three)},$$

where $c_{1,i}$ is a bounded vector in the null space of the Jacobian matrix $H_y$ or $H_w$, respectively.

Conditions 2 and 3 are less restrictive than the assumed consistency conditions (2.3), (2.4), but since the convergence proof follows in the same way as given here (except for straightforward technical complications arising in the initial step of the proof), we omit the details.

Before rigorously proving the theorems in Section 3 and the Supplement section, we outline some of the key elements of the analysis, common to the proofs for both the index two and index three systems. We also introduce further notation and state two corollaries concerning the form of the principal error terms.

In the convergence analysis of linear multistep methods applied to explicit ODE's, it is common to derive the following one-step evolution equation for the numerical error $h^k v_n$ [24]:

$$\omega_n = S_n \omega_{n-1} + h\tilde{f}_n$$

for $\omega_n = (v_n, v_{n-1}, \ldots, v_{n-k+1})$ and $x_n = x(t_n) + h^k v_n$. The amplification matrix $S_n$ is bounded. The vector $\tilde{f}_n$ contains the local truncation error of the BDF method and the remaining nonlinear terms of $v_n$ after linearization.

We shall also use a one-step evolution equation for the numerical error, but when such an equation is derived for the BDF methods applied to DAE's, neither the amplification matrix nor the vector $\tilde{f}_n$ is bounded independent of $h$. However, it is still possible to prove convergence for semiexplicit DAE's by utilizing the natural structure of these systems. In particular, it is necessary to bound certain matrix products such as $S_n S_{n-1} \cdots S_1$ for $n$ sufficiently large and establish cancellation between different terms in the expression for the evolution of the error. It is of particular importance in the index three case for which even smoothness of the leading error term is needed to cancel terms from different time steps. We shall therefore express the error as a truncated asymptotic error expansion

$$y_n - y(t_n) = h^k d(t_n) + h^{k+1} r_n, \tag{2.7}$$

$$u_n - u(t_n) = h^k e(t_n) + h^{k+1} s_n, \tag{2.8}$$

where for the index three system (1.8)–(1.10) we let $d(t) = (c(t), b(t))$ and $r_n = (q_n, p_n)$. In the proofs of both theorems, these expansions simplify the analysis (compare [26]). A similar approach is utilized by Henrici in [15] to determine the asymptotic behavior of the discretization error for general linear multistep

methods applied to explicit ODE's. When expressions (2.7), (2.8) are substituted into Eqs. (2.5), (2.6) and expanded by Taylor series around the analytic solution $(y(t_n), u(t_n))$, we obtain the following linear, index two DAE system for the leading error terms of the numerical solution to the index two problem (1.6), (1.7):

$$d'(t) = A(t)\,d(t) + B(t)e(t) + \frac{1}{(k+1)}y^{(k+1)}(t), \tag{2.9}$$

$$0 = C(t)\,d(t), \tag{2.10}$$

where $A(t) = E_y(*)$, $B(t) = E_u(*)$, $C(t) = H_y(*)$, and the $*$ denotes evaluation at the analytic solution $(y(t), u(t))$. Similarly, for the index three problem (1.8)–(1.10), we get the following linear, index three DAE system for the leading error terms:

$$c'(t) = A_{11}(t)c(t) + A_{12}(t)b(t) + A_{13}(t)e(t) + \frac{1}{(k+1)}v^{(k+1)}(t), \tag{2.11}$$

$$b'(t) = A_{21}(t)c(t) + A_{22}(t)b(t) + \frac{1}{(k+1)}w^{(k+1)}(t), \tag{2.12}$$

$$0 = A_{32}(t)b(t), \tag{2.13}$$

where $A_{11}(t) = F_v(*)$, $A_{12}(t) = F_w(*)$, $A_{13}(t) = F_u(*)$, $A_{21}(t) = G_v(*)$, $A_{22}(t) = G_w(*)$, and $A_{32}(t) = H_w(*)$. There exists a unique solution to the systems (2.9), (2.10) and (2.11)–(2.13) above for each set of consistent initial values [1]. The functions $(d(t), e(t))$ are smooth for smooth $E$ and $H$. Note that the asymptotic error expansion is not only useful for the convergence proof. Such expansions are the basis for automatic error controls, initialization and extrapolation techniques [14].

To prove that the numerical solution converges to the true solution as $h \to 0$, we will show that $(r_n, hs_n)$, and for technical reasons sometimes that $(r_n, s_n)$, are uniformly bounded for all $n \geq \theta$ where $\theta$ is independent of $n$ and $h$. After substituting expressions (2.7), (2.8) into the BDF difference equations (2.5), (2.6) and using the fact that $(d(t), e(t))$ is the solution to a linear DAE system, we write the remaining difference equations for $(r_n, s_n)$ in the one-step form. An induction argument for $i = k, \ldots, n-1$ to $i = n$ on the remainders $(r_i, s_i)$ is applied. Since the relations derived during the analysis are implicit functions of the remainders, it is also necessary to construct a fixed-point iteration for $(r_n, hs_n)$ or $(r_n, s_n)$. The implicit function theorem is applied. The particular restrictions on the initial conditions (i.e., (2.3), (2.4)) are essential in this context.

Throughout the analysis, certain special properties of the matrices arising must be utilized. For example, factors of order $O(1/h)$ are annihilated by projection matrices which appear naturally. Frequently, it is necessary to bound the product of time-dependent matrices having a distinctive block structure. For convenience, we now introduce a notation used to represent these block companion matrices.

*Notation.* Let $I_s$ be an $s \times s$ identity matrix, $0_s$ be an $s \times s$ zero matrix, and $Y$ be any $s \times s$ matrix. Let $\eta_\mu$, $\mu = 1, 2, \ldots, k$, be scalars. Then we define the $sk \times sk$ *block companion matrix*

$$\mathrm{CPM}(\eta_\mu Y) = \begin{pmatrix} \eta_1 Y & \eta_2 Y & \cdots & \eta_{k-1} Y & \eta_k Y \\ I_s & 0_s & \cdots & 0_s & 0_s \\ 0_s & I_s & \ddots & 0_s & 0_s \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0_s & 0_s & \cdots & 0_s & 0_s \\ 0_s & 0_s & \cdots & I_s & 0_s \end{pmatrix}.$$

We shall let $\prod_{i=1}^{n} Y_i$ denote the product of time-dependent matrices $Y_i = Y(t_i)$ given in descending order such as

$$\prod_{i=1}^{n} Y_i = Y_n Y_{n-1} \cdots Y_1.$$

To prove convergence, it is necessary only to show that $(hr_n, hs_n)$ is uniformly bounded for all $n \geq \theta$. However, we will in fact prove that $(r_n, hs_n)$ is uniformly bounded. The convergence proof directly establishes the fact that the principal leading error term in $y_n$ is smooth, but shows only that the principal leading error term in $u_n$ is $h^k(e(t_n) + hs_n)$. The smoothness of the leading error term in $u_n$ is established in the following corollaries, the proofs of which are given in Section 3 and the Supplement section, respectively:

COROLLARY 1. *If the conditions in Theorem* 1 *are valid and the algebraic equations are satisfied to* $O(h^{k+2})$ *accuracy for the starting values* $\mathbf{y}_{k-1}$ *and at the end of each integration step, then* $s_n$ *is uniformly bounded for all* $n \geq k$ *and the principal leading error term in* $u_n$ *is* $h^k e(t_n)$.

COROLLARY 2. *If the conditions in Theorem* 2 *are valid, then* $s_n$ *is uniformly bounded for all* $n \geq 3k$ *and the principal leading error term in* $u_n$ *is* $h^k e(t_n)$.

For the state variables, the leading error terms are smooth immediately ($n \geq k$) if the starting values are numerically consistent and if the algebraic equations are satisfied as in (2.4) at the end of each integration step. However, to derive the leading error terms corresponding to the algebraic variables, it is necessary in general to require the algebraic equations to be satisfied even more accurately than required for convergence. This additional regularity should not be too surprising since even for explicit ODE's, increased accuracy of the starting values is needed in the derivation of the smooth leading error terms for general linear multistep methods [15].

In the analysis given in Section 3 and the Supplement section, we assume for technical simplicity that the difference equations are solved exactly on each integration step. However, the convergence analysis could be modified to include residual terms representing the effect of terminating the iteration process. We have already explained how to relax the requirement of satisfying the algebraic equations exactly. Residuals could also be added to the difference equations corresponding to

the ODE's. A straightforward generalization of the analysis is possible by simply forcing these residuals to be sufficiently small so as to not affect the behavior of the derived error estimates (e.g., in [18] it is assumed that the residuals corresponding to the ODE part are $O(h^k)$ for the index two system). Slightly stronger conditions may have to be imposed to obtain the results derived here, in particular for those results concerning the leading error terms.

Convergence has recently been proved for variable step meshes for nonlinear, semiexplicit, index two systems in [12]. However, the rate of convergence is not established. We cannot extend the convergence analysis for the index three systems to hold for variable stepsize meshes without loss of accuracy because each time the stepsize is changed, a new boundary layer of reduced convergence rates would be initiated. In particular, the first-order BDF would fail to converge at the end of the first step, following every change in the stepsize.

**3. Convergence Proof for the Index Two System (Theorem 1).** We have already stated that there exists a unique solution $(d(t), e(t))$ to the linear, index two system (2.9), (2.10) for each set of consistent initial values, but we must show one can always find a set of consistent initial values for any set of starting values satisfying (2.3), (2.4). Clearly, from (2.10), we must select $d(t_{k-1})$ in the nullspace of $C(t_{k-1})$, denoted $\mathcal{N}(C(t_{k-1}))$. By assumption (2.3), it follows that $y_{k-1} = y(t_{k-1}) + h^{k+1}c_{1,k-1}$ for some bounded vector $c_{1,k-1}$. Therefore, since we want (2.7) to be consistent with our initial values at $t_{k-1}$, we select $d(t_{k-1}) = 0$ and $r_{k-1} = c_{1,k-1}$. We then select $e(t_{k-1})$ to satisfy the first derivative of (2.10), namely

$$e(t) = -[C(t)B(t)]^{-1}C(t)y^{(k+1)}(t)/(k+1) \tag{3.1}$$

for $t = t_{k-1}$. Since $(d(t_{k-1}), e(t_{k-1}))$ is a consistent set of initial values for (2.9), (2.10), we can define $(d(t), e(t))$ for $t \geq t_{k-1}$ to be the solution. In the more general case that $d(t_{k-1}) \in \mathcal{N}(C(t_{k-1}))$ but is not zero, as discussed in an earlier remark, one can also always find a consistent set of initial values for the DAE system (2.9), (2.10).

Since the $k$-step BDF requires a numerically consistent starting vector $\mathbf{y}_{k-1}$, we will define $d(t_i) = 0$ and $r_i = c_{1,i}$ for $i = 0, 1, \ldots, k-2$, where $c_{1,i}$ are bounded vectors such that $y_i = y(t_i) + h^{k+1}c_{1,i}$. Let $e(t_i)$ be defined as in (3.1) for $t = t_i$, $i = 0, 1, \ldots, k-2$. Then, there exists a constant $\eta_0$ such that $\|\mathbf{r}_{k-1}\| \leq \eta_0$ where $\mathbf{r}_{k-1} = [r_{k-1}^T, \ldots, r_0^T]^T$.

After expanding by Taylor series about the analytic solution $(y(t_n), u(t_n))$, and using the fact that $(d(t), e(t))$ is the solution to (2.9), (2.10) with associated initial conditions, Eqs. (2.5), (2.6) can be rewritten in the form,

$$S(t_n)\begin{bmatrix} r_n \\ hs_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{k} \gamma_i r_{n-i} \\ 0 \end{bmatrix} + \begin{bmatrix} \psi_1(t_n) \\ \psi_2(t_n) \end{bmatrix}, \tag{3.2}$$

where

$$S(t_n) = \begin{bmatrix} I_l - \frac{h}{\alpha_0}A(t_n) & -\frac{1}{\alpha_0}B(t_n) \\ C(t_n) & 0_m \end{bmatrix},$$

$$\gamma_i = -\alpha_i/\alpha_0 \quad \text{where } \alpha_i \text{ are the BDF coefficients,}$$

$$\psi_1(t) = -\frac{h}{\alpha_0}R_1(\xi) + \frac{h^k}{\alpha_0(k+1)}d^{(k+1)}(t) - \frac{h^{k+1}}{\alpha_0}R_{12}(\xi) + h^k Q_1(t) + h^{k+1}(W^{(1)}(r) + W^{(2)}(s)) + h^{k+2}g_1(r,s) + O(h^{2k}), \tag{3.3}$$

$$\psi_2(t) = -h^{k-1}Q_2(t) - h^k W^{(3)}(r) - h^{k+1}g_2(r) + O(h^{2k}), \tag{3.4}$$

$$Q_1(t) = \frac{1}{2\alpha_0}\{d(t)^T[E_{yy}\,d(t) + E_{uy}e(t)] + e(t)^T[E_{yu}\,d(t) + E_{uu}e(t)]\} + O(h^k),$$

$$Q_2(t) = \frac{1}{2}d(t)^T H_{yy}\,d(t) + O(h^k),$$

$$W^{(1)}(r) = \frac{1}{2\alpha_0}\{r^T[E_{yy}\,d(t) + E_{uy}e(t)] + [d(t)^T E_{yy} + e(t)^T E_{yu}]r\},$$

$$W^{(2)}(s) = \frac{1}{2\alpha_0}\{s^T[E_{yu}\,d(t) + E_{uu}e(t)] + [d(t)^T E_{uy} + e(t)^T E_{uu}]s\},$$

$$W^{(3)}(r) = \frac{1}{2}\{r^T H_{yy}\,d(t) + d(t)^T H_{yy}r\},$$

$$g_1(r,s) = \frac{1}{2\alpha_0}\{r^T(E_{yy}r + E_{uy}s) + s^T(E_{yu}r + E_{uu}s)\},$$

$$g_2(r) = \frac{1}{2}r^T H_{yy}r$$

for $t = t_n$, $r = r_n$, $s = s_n$ and $\xi = \xi_n$, $t_{n-k} \le \xi_n \le t_n$, and where all the partial derivatives of $E$ and $H$ are evaluated at the analytic solution $(y(t_n), u(t_n))$. The functions $R_1(\xi)$ and $R_{12}(\xi)$ are the remainders from the Taylor series expansions, and hence are bounded functions of $y^{(k+2)}(\xi)$ and $d^{(k+2)}(\xi)$, respectively.

The functions $Q_1(t)$ and $Q_2(t)$ contain all the inhomogeneous terms which are functions only of $d(t)$, $e(t)$, and partial derivatives of $E$ and $H$. We will carefully analyze the leading terms in the Taylor series part of $\psi_1$ and $\psi_2$ (through the quadratic terms in $r$ and $s$). The inclusion of higher-order terms from the Taylor series will only introduce terms of order higher than those already present. These higher-order terms have no influence on the order of the estimates derived in this analysis, and are presented by the final terms $O(h^{2k})$ in the definition of $\psi_1$ and $\psi_2$.

Since $W^{(i)}$ $(i = 1, 2, 3)$ are linear, bounded functions of their arguments, there exist bounded matrix operators $W_i$ such that

$$W^{(i)}(\alpha) = W_i\alpha, \qquad \|W_i\| \le \omega_i \quad \text{for } i = 1, 2, 3.$$

Define $\mathbf{s}_n = [s_n^T, s_{n-1}^T, \ldots, s_{n-k+1}^T]^T$. Then, if

$$S^{-1}(t_n) = \begin{pmatrix} X_n & Y_n \\ U_n & V_n \end{pmatrix}$$

we can rewrite (3.2) in the one-step form,

$$\mathbf{r}_n = F_n\mathbf{r}_{n-1} + f_{1,n}, \tag{3.5}$$

$$h\mathbf{s}_n = G_n\mathbf{r}_{n-1} + f_{2,n} + hH\mathbf{s}_{n-1}, \tag{3.6}$$

where

$$F_n = \mathrm{CPM}(\gamma_\mu X_n) \qquad (\mu = 1, 2, \ldots, k),$$
$$f_{1,n} = [(X_n\psi_1(t_n) + Y_n\psi_2(t_n))^T, 0, \ldots, 0]^T,$$
$$G_n = (I_m, 0, \ldots, 0)^T(\gamma_1 U_n, \ldots, \gamma_k U_n),$$
$$f_{2,n} = [(U_n\psi_1(t_n) + V_n\psi_2(t_n))^T, 0, \ldots, 0]^T,$$
$$H = \mathrm{CPM}(0_m).$$

Closed form expressions for the block elements of the matrix inverse of $S(t_n)$ can easily be derived, but the following approximate relations are sufficient for the analysis ($t$-dependence is suppressed):

$$X = I_l - B(CB)^{-1}C + O(h), \tag{3.7}$$

$$Y = B(CB)^{-1} + O(h), \tag{3.8}$$

$$U = -\alpha_0(CB)^{-1}C + O(h), \tag{3.9}$$

$$V = \alpha_0(CB)^{-1} + O(h). \tag{3.10}$$

Since $C(t)B(t)$ is nonsingular for all $t \in I$, it follows that $S^{-1}(t)$ is as smooth as $S(t)$, and hence the block elements of $S^{-1}(t)$ can be bounded uniformly on $I$. Therefore, there exists a constant $K$ such that

$$\|X(t)\|, \|Y(t)\|, \|U(t)\|, \|V(t)\| \leq K$$

for all $t \in I$. In general, we will let the symbol $K$ denote a generic constant, independent of $n$ and $h$.

Note that Eqs. (3.5), (3.6) are implicit in $r_n$ and $hs_n$. We shall construct a fixed point iteration for their solution $r_n$ and $hs_n$, and an induction argument concerning the past values of $\mathbf{r}_{n-1}$ and $h\mathbf{s}_{n-1}$ will be required. Specifically, let the *induction assumption* be the following:

$$\|\mathbf{r}_i\| \leq \eta_1 e^{(i-k+1)hL}, \qquad \|h\mathbf{s}_i\| \leq \eta_2 e^{(i-k+1)hL} \tag{3.11}$$

for $i = k, k+1, \ldots, n-1$. We will prove there exist such constants $\eta_1, \eta_2$ and $L$ and the equations (3.11) are satisfied for $i = n$. It follows then from the induction assumption and the definitions of $\psi_1(t)$ and $\psi_2(t)$ given in (3.3), (3.4) that for $i = k, k+1, \ldots, n-1$

$$\|\psi_1(t_i)\| \leq \sigma_1 h, \qquad \|\psi_2(t_i)\| \leq \sigma_2 h^{k-1}, \tag{3.12}$$

where to $O(h)$ accuracy $\sigma_1$ is independent of $\eta_1, \eta_2$ and $L$ if $k \geq 2$, $\sigma_1$ is dependent on $\eta_2$ and $L$ if $k = 1$ (i.e., $\sigma_1 = \tilde{\sigma}_1(\eta_2, L) + O(h)$), and $\sigma_2$ is independent of $\eta_1, \eta_2$ and $L$ for all $k$.

Define the iterates as

$$\mathbf{r}_n^{(\nu)} = [(r_n^{(\nu)})^T, r_{n-1}^T, \ldots, r_{n-k+1}^T]^T,$$
$$\mathbf{s}_n^{(\nu)} = [(s_n^{(\nu)})^T, s_{n-1}^T, \ldots, s_{n-k+1}^T]^T,$$

and the fixed-point iteration as

$$\mathbf{r}_n^{(\nu+1)} = \mathbf{r}_n^{(0)} + [Z_1(r_n^{(\nu)}, hs_n^{(\nu)})^T, 0, \ldots, 0]^T, \tag{3.13}$$

$$h\mathbf{s}_n^{(\nu+1)} = h\mathbf{s}_n^{(0)} + [Z_2(r_n^{(\nu)}, hs_n^{(\nu)})^T, 0, \ldots, 0]^T, \tag{3.14}$$

where

$$\begin{aligned} Z_1(r, hs) &= X_n[h^{k+1}(W_1 r + W_2 s) + h^{k+2} g_1(r, s)] \\ &\quad + Y_n[-h^k W_3 r - h^{k+1} g_2(r)] + O(h^{2k}), \\ Z_2(r, hs) &= U_n[h^{k+1}(W_1 r + W_2 s) + h^{k+2} g_1(r, s)] \\ &\quad + V_n[-h^k W_3 r - h^{k+1} g_2(r)] + O(h^{2k}) \end{aligned}$$

for $r = r_n^{(\nu)}$ and $s = s_n^{(\nu)}$. For starting iterates select

$$\mathbf{r}_n^{(0)} = F_n \mathbf{r}_{n-1} + \tilde{f}_{1,n}, \tag{3.15}$$

$$h\mathbf{s}_n^{(0)} = G_n \mathbf{r}_{n-1} + \tilde{f}_{2,n} + hH\mathbf{s}_{n-1}, \tag{3.16}$$

where

$$\begin{aligned} \tilde{f}_{1,n} &= [(X_n \tilde{\psi}_1(t_n) + Y_n \tilde{\psi}_2(t_n))^T, 0, \ldots, 0]^T, \\ \tilde{f}_{2,n} &= [(U_n \tilde{\psi}_1(t_n) + V_n \tilde{\psi}_2(t_n))^T, 0, \ldots, 0]^T, \\ \tilde{\psi}_1(t_n) &= -\frac{h}{\alpha_0} R_1(\xi_n) + \frac{h^k}{\alpha_0(k+1)} d^{k+1}(t_n) - \frac{h^{k+1}}{\alpha_0} R_{12}(\xi_n) + h^k Q_1(t_n), \\ \tilde{\psi}_2(t_n) &= -h^{k-1} Q_2(t_n). \end{aligned}$$

The following three conditions corresponding to the implicit function theorem are sufficient to prove the fixed-point iteration defined in (3.13), (3.14) converges to a solution:

$$\|\mathbf{r}_n^{(0)}\| \le \eta_1 e^{(n-k+1)hL}, \qquad \|h\mathbf{s}_n^{(0)}\| \le \eta_2 e^{(n-k+1)hL}; \tag{3.17}$$

$$\|Z(r_n^{(0)}, hs_n^{(0)})\| \le \delta/2, \qquad \delta > 0, \quad Z = (Z_1^T, Z_2^T)^T; \tag{3.18}$$

$$\|J\| \le \tfrac{1}{2} \text{ for any } r \text{ and } hs \text{ such that } \|r - r_n^{(0)}\| \le \delta \text{ and } \|hs - hs_n^{(0)}\| \le \delta, \tag{3.19}$$

where $J$ is the Jacobian matrix of $Z(r, hs)$ with respect to $r$ and $hs$.

If $r$ and $hs$ are bounded, conditions (3.18) and (3.19) are relatively straightforward to verify for the iteration defined in (3.13), (3.14). The expressions for $Z_1(r, hs)$ and $Z_2(r, hs)$ imply that there exists a $\delta, \delta = Kh^k$, which satisfies (3.18) providing $r = r_n^{(0)}$ and $hs = hs_n^{(0)}$ are bounded. The Jacobian matrix $J$ will be linear in $r$ and $hs$, so condition (3.19) will be satisfied for sufficiently small $h$ and for all $r$ and $hs$ such that $\|r - r_n^{(0)}\| \le \delta$ and $\|hs - hs_n^{(0)}\| \le \delta$. Hence, it only remains to prove condition (3.17).

It is quite easy to see the starting iterates for the fixed point iteration as defined by (3.15), (3.16) are bounded for $n = k$. The remaining conditions for convergence of the fixed point iteration are also satisfied, so the induction assumption is satisfied at the first step.

Now we shall assume the induction assumption (3.11) is true and prove the initial guesses $\mathbf{r}_n^{(0)}$ and $h\mathbf{s}_n^{(0)}$ are bounded as in (3.17). We will bound $h\mathbf{s}_n^{(0)}$ first because when $k = 1$ the constant $\eta_1$ will be chosen dependent on $\eta_2$ and $L$. This

dependency is a result of our decision to bound only $h\mathbf{s}_n^{(0)}$, and not $\mathbf{s}_n^{(0)}$. To bound $\mathbf{s}_n^{(0)}$ requires much more algebraic manipulation of the structure of the system [1]. This result is obtained in another way in the corollary.

To verify (3.17), we must utilize important cancellation properties in certain matrix products involving the $O(1)$ matrices $X, Y$, and $U$ defined in (3.7)–(3.9):

$$\|X(t_i)Y(t_j)\|,\ \|U(t_i)X(t_j)\| \le Kh \tag{3.20}$$

for any $|t_i - t_j| = O(h)$ and some generic constant $K$ independent of $h$. In addition, the proof requires that the matrix product $\prod_{j=i+1}^{n} F_j$ be bounded uniformly in $n$ for $i = 0, 1, \ldots, n-1$. Since $F_j = \mathrm{CPM}(\gamma_\mu X_j)$ where $X_j$ is a projection matrix to $O(h)$ accuracy, it follows from a result given by Kreiss [16] and Strang [26] that there exist constants $K^*$ and $\mathscr{E}$ such that

$$\left\| \prod_{j=i+1}^{n} F_j \right\| \le K^* e^{(n-i)h\mathscr{E}}$$

for all $n$ and $i = 0, 1, \ldots, n-1$. Moreover, using the structure of the block companion matrix $F_j$ and (3.20), it follows that

$$\left\| G_{n-l} \prod_{j=i+1}^{n-l-1} F_j \right\| \le \tilde{N}h$$

for all $n - l - i - 1 \ge k$ and some constant $\tilde{N}$. In bounding the starting iterates, we will also need the fact that there exist constants $\rho_1$ and $\rho_2$ such that

$$\|\tilde{\psi}_1(t_n)\| \le \rho_1 h, \qquad \|\tilde{\psi}_2(t_n)\| \le \rho_2 h^{k-1}$$

for any $t_n \in [t_k, t_0 + T]$. Now rewrite Eq. (3.16) for the initial guess $h\mathbf{s}_n^{(0)}$ as

$$\begin{aligned} h\mathbf{s}_n^{(0)} = {} & \sum_{l=0}^{k-1} H^l G_{n-l} \left( \prod_{i=k}^{n-l-1} F_i \right) \mathbf{r}_{k-1} + \sum_{l=1}^{k-1} H^l f_{2,n-l} + \tilde{f}_{2,n} \\ & + \sum_{l=0}^{k-1} H^l G_{n-l} \left( \sum_{i=k}^{n-l-2} \left( \prod_{j=i+1}^{n-l-1} F_j \right) f_{1,i} + f_{1,n-l-1} \right). \end{aligned}$$

It is then possible to bound $h\mathbf{s}_n^{(0)}$ using these results in the following way:

$$\begin{aligned} \|h\mathbf{s}_n^{(0)}\| \le {} & kK \max_{1 \le i \le k} |\gamma_i| K^* e^{(n-k)h\mathscr{E}} \eta_0 + K(\rho_1 h + \rho_2 h^{k-1}) \\ & + K(\sigma_1 h + \sigma_2 h^{k-1})[(n-2k)hk\tilde{N} + k^2 K K^* e^{(k-1)h\mathscr{E}} + (k-1)] \end{aligned}$$

for all $n \ge k$. This bound for $h\mathbf{s}_n^{(0)}$ is independent of $\eta_1, \eta_2$, and $L$ to $O(h)$ accuracy for all $k$. Therefore, there exists an $\eta_2$ such that for sufficiently small $h$ and $L = \mathscr{E}$, we have

$$\|h\mathbf{s}_n^{(0)}\| < \eta_2 e^{(n-k+1)h\mathscr{E}}.$$

This result is valid for $n \to \infty$, $h \to 0$ and $t_n$ such that $t_n \in [t_k, t_0 + T] \subseteq I$.

Next we bound $\mathbf{r}_n^{(0)}$ defined by (3.15), which can be rewritten as

$$\mathbf{r}_n^{(0)} = \left( \prod_{j=k}^{n} F_j \right) \mathbf{r}_{k-1} + \sum_{i=k}^{n-1} \left( \prod_{j=i+1}^{n} F_j \right) f_{1,i} + \tilde{f}_{1,n}.$$

Using relation (3.20), the structure of $F_j$, and the bounds (3.12), it is possible to show that there exists a constant $\Delta_1$ dependent on $\eta_2$ and $L$ if $k = 1$, such that $\|(\prod_{j=i+1}^{n} F_j) f_{1,i}\| \leq \Delta_1 h$ for $n - i \geq k$. Then it follows from this and earlier bounds that

$$\|\mathbf{r}_n^{(0)}\| \leq K^* e^{(n-k+1)h\mathscr{E}} \eta_0 + (n - 2k + 1)h\Delta_1 + (k-1)K(\sigma_1 h + \sigma_2 h^{k-1})K^* e^{(k-1)h\mathscr{E}} + K(\rho_1 h + \rho_2 h^{k-1})$$

for all $n \geq k$. If $k = 1$ the bound depends on $\eta_2$ and $\mathscr{E}$ (through $\Delta_1$), and hence a constant $\eta_1$ must be chosen dependent on $\eta_2$ and $\mathscr{E}$ such that

$$\|\mathbf{r}_n^{(0)}\| < \eta_1 e^{(n-k+1)h\mathscr{E}}.$$

If $k \geq 2$, $\eta_1$ may be chosen independent of $\eta_2$ and $\mathscr{E}$.

Since both $h\mathbf{s}_n^{(0)}$ and $\mathbf{r}_n^{(0)}$ are uniformly bounded as in (3.17), it follows immediately that conditions (3.18) and (3.19) of the implicit function theorem are satisfied. Therefore, there exists a solution $\mathbf{r}_n$ and $h\mathbf{s}_n$ satisfying Eqs. (3.5) and (3.6), which can be computed by straightforward iteration. Moreover, there exists a $\delta$, $\delta = O(h^k)$, such that $\|\mathbf{r}_n - \mathbf{r}_n^{(0)}\| = \|Z_1(r_n, hs_n)\| \leq \delta$ and $\|h\mathbf{s}_n - h\mathbf{s}_n^{(0)}\| = \|Z_2(r_n, hs_n)\| \leq \delta$. For sufficiently small $h$, it follows that

$$\|\mathbf{r}_n\| \leq \|\mathbf{r}_n^{(0)}\| + \delta \leq \eta_1 e^{(n-k+1)h\mathscr{E}},$$
$$\|h\mathbf{s}_n\| \leq \|h\mathbf{s}_n^{(0)}\| + \delta \leq \eta_2 e^{(n-k+1)h\mathscr{E}}.$$

This concludes the induction argument. The global rate of convergence is $O(h^k)$, since

$$\|y_n - y(t_n)\| \leq h^k \|d(t_n)\| + h^{k+1}\|r_n\|,$$
$$\|u_n - u(t_n)\| \leq h^k(\|e(t_n)\| + \|hs_n\|)$$

for all $n \geq k$.

*Proof of Corollary* 1. Since the algebraic equations are satisfied to $O(h^{k+2})$ accuracy, both at the starting values and at the end of each integration step, it follows that $C(t_i)r_i = \psi_2(t_i) + O(h)$ for $i = 0, 1, \ldots$. From the difference equations (3.2) corresponding to the ODE part, we have

$$\left[I_l - \frac{h}{\alpha_0} A(t_n)\right] r_n - \frac{h}{\alpha_0} B(t_n) s_n = \sum_{i=1}^{k} \gamma_i r_{n-i} + \psi_1(t_n).$$

Let these equations also be solved to accuracy $O(h^{k+2})$. Then the residual error may be included as a term of $O(h)$ accuracy in $\psi_1$. Multiply by $C(t_n)$, an $m \times n$ matrix of rank $m$, and invert $C(t_n)B(t_n)$ to obtain a relation for $s_n$:

$$s_n = -\frac{\alpha_0}{h}(C(t_n)B(t_n))^{-1}C(t_n)\left(\sum_{i=1}^{k} \gamma_i r_{n-i} + \psi_1(t_n) - \left[I_l - \frac{h}{\alpha_0} A(t_n)\right] r_n\right). \tag{3.21}$$

In the proof of Theorem 1, it is shown that $r_n$ and $hs_n$ are uniformly bounded for $n \geq k$. Then, $\|\psi_1(t_n)\| \leq Kh$ for some constant $K$ independent of $n$ and $h$ for all $n \geq k$. For $k \geq 2$ we have

$$\|C(t_n)r_{n-i}\| \leq \|C(t_{n-i})r_{n-i}\| + O(h) \leq \|\psi_2(t_{n-i})\| + O(h) = O(h)$$

for all $n \geq k$ and $i = 0, 1, 2, \ldots, k-1$. Thus, for $k \geq 2$ it follows that the $1/h$ factor in (3.21) is annihilated, leaving only terms of order $O(1)$. If $k = 1$ we must be a bit more careful and utilize a cancellation property. The expression (3.21) simplifies to

$$s_n = -\frac{1}{h}(C(t_n)B(t_n))^{-1}C(t_n)\left(r_{n-1} + \psi_1(t_n) - \left[I_l - \frac{h}{\alpha_0}A(t_n)\right]r_n\right).$$

Using the smoothness of $Q_2(t)$, the lowest order (i.e., $O(1)$) term in $\psi_2(t)$, it follows that

$$\begin{aligned} -C(t_n)r_{n-1} + C(t_n)r_n &= -C(t_{n-1})r_{n-1} + C(t_n)r_n + O(h) \\ &= -\psi_2(t_{n-1}) + \psi_2(t_n) + O(h) = O(h). \end{aligned}$$

As for $k \geq 2$ then, it follows that $s_n$ is bounded for all $n \geq k$, and these bounds can be chosen uniformly since they depend on the uniform bounds for $r_n$ and $hs_n$. Therefore, Eqs. (2.7), (2.8) accurately reflect the relation of the numerical solution to the true solution for all $n \geq k$. The functions $d(t)$ and $e(t)$ in the principal leading error terms are the unique solutions to the linear, index two DAE system with consistent initial conditions described earlier.

*Remark.* Note that the algebraic equations must be satisfied more closely to prove that $s_n$ is bounded, than required in the proof of convergence. If they are satisfied only to $O(h^{k+1})$ accuracy, convergence is still obtained but the error is proportional to $e(t_n) + hs_n$.

The Aerospace Corporation
2350 E. El Segundo Boulevard
El Segundo, California 90245
*E-mail:* brenan@aerospace.aero.org

Department of Mathematics
University of California
Los Angeles, California 90024

1. 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.

2. K. E. BRENAN, "Numerical simulation of trajectory prescribed path control problems by the backward differentiation formulas," *IEEE Trans. Automat. Control*, v. AC-31, 1986, pp. 266–269.

3. K. E. BRENAN & B. E. ENGQUIST, *Backward Difference Approximations of Nonlinear Differential-Algebraic Equations*, ATR-85(9990)-5, The Aerospace Corporation, El Segundo, Ca., 1985 or Dept. of Computer Science, Uppsala University Report #101, Uppsala, Sweden.

4. A. E. BRYSON & Y. C. HO, *Applied Optimal Control*, Hemisphere Publishing Corp., Washington, D.C., 1975.

5. S. L. CAMPBELL, *Singular Systems of Differential Equations* I–II, Pitman, Marshfield, Mass., 1980, 1982.

6. K. D. CLARK, "The numerical solution of some higher index time varying semistate systems by difference methods," *Circuits Systems Signal Process.* v. 6, 1987, pp. 61–75.

7. F. R. GANTMACHER, *The Theory of Matrices*, Vol. I–II, Chelsea, New York, 1959.

8. C. W. GEAR, "The simultaneous numerical solution of differential-algebraic equations," *IEEE Trans. Circuit Theory*, v. CT-18, 1971, pp. 89–95.

9. C. W. GEAR & L. R. PETZOLD, "ODE methods for the solution of differential/algebraic systems," *SIAM J. Numer. Anal.*, v. 21, 1984, pp. 367–384.

10. C. W. GEAR & L. R. PETZOLD, "Differential/algebraic systems and matrix pencils," in *Matrix Pencils* (B. Kågström and A. Ruhe, eds.), Lecture Notes in Math., vol. 973, 1983, pp. 75–89.

11. C. W. GEAR, H. H. HSU & L. PETZOLD, *Differential-Algebraic Equations Revisited*, Proc. Conference on Matrix Pencils, Piteå, Sweden, 1982.

12. C. W. GEAR, B. LEIMKUHLER & G. K. GUPTA, "Automatic integration of Euler-Lagrange equations with constraints," *J. Comput. Appl. Math.*, v. 12 and 13, 1985, pp. 77–90.

13. H. GINGOLD, "A method of global blockdiagonalization for matrix-valued functions," *SIAM J. Numer. Anal.*, v. 9, 1978, pp. 1076–1082.

14. P. HENRICI, *Discrete Variable Methods in Ordinary Differential Equations*, Wiley, New York, 1962.

15. P. HENRICI, *Error Propagation for Difference Methods*, Wiley, New York, 1963.

16. H. O. KREISS, "Über implizite Differenzmethoden für partielle Differentialgleichungen," *Numer. Math.*, v. 5, 1963, pp. 24–47.

17. P. LÖTSTEDT, "Mechanical systems of rigid bodies subject to unilateral constraints," *SIAM J. Appl. Math.*, v. 42, 1982, pp. 281–296.

18. P. LÖTSTEDT & L. R. PETZOLD, "Numerical solution of nonlinear differential equations with algebraic constraints I: Convergence results for backward differentiation formulas," *Math. Comp.*, v. 46, 1986, pp. 491–516.

19. R. MÄRZ, *Multistep Methods for Initial Value Problems in Implicit Differential-Algebraic Equations*, Preprint 22, Humboldt-Univ., Berlin, Sektion Mathematik, 1981.

20. R. MÄRZ, "On initial value problems in differential-algebraic equations and their numerical treatment," *Computing*, v. 35, 1985, pp. 13–37.

21. J. F. PAINTER, *Solving the Navier-Stokes Equations with LSODI and the Method of Lines*, Report UCID-19262, Lawrence Livermore National Laboratory, Livermore, Ca., 1981.

22. L. R. PETZOLD, "Differential/algebraic equations are not ODE's," *SIAM J. Sci. Statist. Comput.*, v. 3, 1982, pp. 367–384.

23. W. C. RHEINBOLDT, "Differential-algebraic systems as differential equations on manifolds," *Math. Comp.*, v. 43, 1984, pp. 473–482.

24. R. D. RICHTMYER & R. W. MORTON, *Differential Methods for Initial-Value Problems*, Interscience, New York, 1967.

25. R. F. SINCOVEC, A. M. ERISMAN, E. L. YIP & M. A. EPTON, "Analysis of descriptor systems using numerical algorithms," *IEEE Trans. Automat. Control*, v. AC-26, 1981, pp. 139–147.

26. G. STRANG, "Accurate partial difference methods II–Nonlinear problems," *Numer. Math.*, v. 6, 1964, pp. 37–46.

27. J. H. WILKINSON, "Linear differential equations and Kronecker's canonical form" in *Recent Applications in Numerical Analysis* (C. deBoor and G. H. Golub, eds.), Academic Press, New York, 1978, pp. 231–241.

28. E. L. YIP & R. F. SINCOVEC, "Solvability, controllability, and observability of continuous descriptor systems," *IEEE Trans. Automat. Control*, v. AC-26, 1981, pp. 702–707.