MATHEMATICS OF COMPUTATION
VOLUME 46, NUMBER 174
APRIL 1986, PAGES 491-516

# Numerical Solution of Nonlinear Differential Equations with Algebraic Constraints I: Convergence Results for Backward Differentiation Formulas

By Per Lötstedt* and Linda Petzold**

**Abstract.** In this paper we investigate the behavior of numerical ODE methods for the solution of systems of differential equations coupled with algebraic constraints. Systems of this form arise frequently in the modelling of problems from physics and engineering; we study some particular examples from electrical networks, fluid dynamics and constrained mechanical systems. We show that backward differentiation formulas converge with the expected order of accuracy for these systems.

**1. Introduction.** In this paper we investigate the behavior of numerical ODE methods for the solution of systems of differential/algebraic equations (DAE) of the form

$$0 = F_1(x, x', y, t), \qquad 0 = F_2(x, y, t), \tag{1.1}$$

where the initial values of at least $x$ are given at $t = 0$ and $\partial F_1/\partial x'$ is nonsingular. Systems of this form arise frequently in the modelling of problems in physics and engineering. In general, there are many DAE systems, including simple linear systems, which cannot be solved by numerical ODE methods because of numerical instability [14]. However, the DAE systems arising in several areas of application (for example, the simulation of electrical networks and mechanical systems, and the solution of the equations of fluid dynamics) share certain properties that make them amenable to solution by ODE methods, even though they are complicated nonlinear systems. Our objective here is to develop an understanding of the behavior of numerical methods for solving these special DAE systems. Strategies for solving them reliably and efficiently are suggested in [22, 26].

The basic idea of using a numerical ODE method for solving DAE systems was introduced by Gear [13], and consists of replacing $x'$ in (1.1) by a difference approximation, and then solving the resulting equations for approximations to $x$ and $y$. The simplest example of a numerical ODE method for (1.1) is the backward Euler method. Using this approach, the derivative $x'(t_{n+1})$ at time $t_{n+1}$ is approximated

Received July 18, 1984; revised May 28, 1985.

1980 *Mathematics Subject Classification*. Primary 65L05, 65L07.

*Work supported in part by the National Swedish Board for Technical Development when the author was at the Department of Numerical Analysis and Computing Science, Royal Institute of Technology, S-10044 Stockholm, Sweden.

**Work supported by the U. S. Department of Energy Office of Basic Energy Sciences.

by a backward difference of $x(t)$, and the resulting system of nonlinear equations is solved for $x_{n+1}$ and $y_{n+1}$,

$$(1.2)\quad 0 = F_1\left(x_{n+1}, \frac{x_{n+1} - x_n}{t_{n+1} - t_n}, y_{n+1}, t_{n+1}\right), \qquad 0 = F_2(x_{n+1}, y_{n+1}, t_{n+1}).$$

In this way the solution is advanced from time $t_n$ to $t_{n+1}$. In this paper we consider several different higher-order numerical ODE methods for the solution of (1.1).

Not all systems of the form (1.1) can be solved using numerical ODE methods, even though the solutions to these systems are well defined. An important characteristic for understanding both the properties of solutions to DAE systems and the behavior of numerical methods for solving these systems is the *index* of the system, for which a precise definition is given later in this section. The system (1.1) has index zero when the second equation in (1.1) is missing, and index one when $\partial F_2/\partial y$ is nonsingular. Numerical ODE methods can be used to solve linear and nonlinear problems of index no greater than one with no great difficulty. The situation for problems whose index exceeds one is considerably more complicated. The nonlinear problems that interest us here have an index of either two or three, hence we can expect some difficulties in trying to solve them. In order to better understand the index and its role in the structure and solution of DAE systems, we give a brief review of the properties of general linear differential/algebraic systems in the next few paragraphs. For further details see [25].

In some sense the simplest DAE systems are linear constant-coefficient systems

$$(1.3)\qquad Ax'(t) + Bx(t) = g(t).$$

The equation (1.3) is easily understood by transforming the system to Kronecker canonical form. The main idea is that there exists a nonsingular row scaling $P$ and a nonsingular change of variables $Q$ that transform the system to a canonical form. Now, if the system is *solvable*, that is, if solutions to (1.3) exist for all sufficiently smooth input functions $g(t)$, and solutions are uniquely specified by their values at any time in the interval of interest [14], then we can find $P$ and $Q$ that decouple the system into a "differential"part and a "singular" part

$$(1.4)\qquad x_1'(t) + Cx_1(t) = g_1(t), \qquad Ex_2'(t) + x_2(t) = g_2(t),$$

where

$$Q^{-1}x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}, \qquad Pg(t) = \begin{bmatrix} g_1(t) \\ g_2(t) \end{bmatrix}$$

and $E$ has the property that either there exists an integer $m$ such that $E^m = 0$, $E^{m-1} \neq 0$, or $E$ is the "empty" matrix. The value of $m$ is defined to be the index of the system. The matrix $E$ is composed of Jordan blocks of the form

$$\begin{pmatrix} 0 & & & \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{pmatrix}$$

and $m$ is the size of the largest of these blocks. If $E$ is empty, then $m = 0$, and the system is just a standard ODE system.

Systems of index greater than one have several properties that are not shared by the lower index systems. To explain these properties, we study the simplest index three problem, written in the form (1.1),

$$x_1' - x_2 = 0, \quad x_2' - y = 0, \quad x_1 - g(t) = 0. \tag{1.5}$$

The solution to this problem is $x_1 = g(t)$, $x_2 = g'(t)$, $y = g''(t)$. When initial values are given for $x$ and $y$ which are not compatible with the solution given above, then the solution will initially exhibit a jump discontinuity or possibly even an impulse. Similarly, if the driving term $g(t)$ is not twice differentiable everywhere, the solution will not exist everywhere.

The surprising fact about linear constant coefficient DAE systems is that even though they are so unlike ODE systems, there is a large class of numerical ODE methods that work for solving them [25]. For example, when the backward Euler method with constant stepsize $h$ is applied to (1.5), we find that the solution is accurate to $O(h)$ on all steps after the third step. However, the solution can be very inaccurate on the first two steps, and for several steps after a change of stepsize. This situation causes problems for the types of error estimators that are commonly used in variable-stepsize codes. Therefore, to use numerical ODE methods effectively for the solution of high-index constant-coefficient DAE systems, we must be clever about how we construct the algorithms and the error estimates. We discuss these issues further in [26].

Now that we have defined the index for linear constant-coefficient systems, we can think about generalizing the concept for more complicated systems. For general linear problems (where the matrices $A$ and $B$ in (1.3) are time-dependent), there are several possible ways to define the index of a system [14]. The *local index* is the index of the local constant-coefficient system (where $A(t)$ and $B(t)$ are held constant at some fixed time $t$, and we study the structure of the resulting constant-coefficient problem). The *global index* is defined in terms of a reduction of the DAE to a semicanonical form. The idea behind this reduction, which is analogous to the transformation to canonical form in the constant-coefficient case, is to decouple the system via nonsingular time-dependent transformations into a "differential" part and a "singular" part. For many practical problems, we can find a nonsingular (time-dependent) change of variables and a nonsingular (time-dependent) scaling that transform the system to the semicanonical form

$$x_1'(t) + C(t)x_1(t) = g_1(t), \qquad Ex_2'(t) + x_2(t) = g_2(t),$$

where the matrices are all constant except $C(t)$ [14]. If the index of $E$ is $m$, then we say that the system has a global index of $m$.

The definitions for the local and global index are useful from the point of view that they provide a means for studying the underlying structure of complicated DAE systems, but they fail to provide us with any kind of a simple procedure for finding the index of a given DAE system. The following algorithm, described in [14], is useful for these purposes. We describe the technique below for linear nonconstant coefficient systems, but it applies directly to nonlinear problems (1.1) when $F_1$ is linear in $x'$, and we will make use of it several times later in this paper.

ALGORITHM 1.1

(1) If $A$ in (1.3) is nonsingular, then we are done.

(2) Otherwise, premultiply (1.3) by a nonsingular matrix $P(t)$ to "zero out" a maximal number of rows of $A$ and permute the zero rows to the bottom to obtain:

$$\begin{bmatrix} A_{11} \\ 0 \end{bmatrix} x' + \begin{bmatrix} B_{11} \\ B_{12} \end{bmatrix} x = g(t).$$

(3) Differentiate the bottom half of the system to obtain the new system

$$\begin{bmatrix} A_{11} \\ B_{12} \end{bmatrix} x' + \begin{bmatrix} B_{11} \\ B'_{12} \end{bmatrix} x = \hat{g}(t).$$

Now apply this process again to the new system.

The idea behind Algorithm 1.1 is that by differentiating the "algebraic" constraints of the system, we can reduce the system's index without changing its solution. For solvable systems with no turning points (this includes all of the systems that we study here), this algorithm terminates in $m$ iterations if and only if the global index is $m$ [14]. Thus we can use this procedure for determining the global index of a system. The algorithm can also be used to find the local index of a system by considering the matrices $A$ and $B$ at some time to be constant, and then applying the algorithm to the resulting system. In this case, the algorithm terminates in $m$ iterations if and only if the local index is $m$. For the systems that we study in this paper, the local index will always be equal to the global index. However, for more general DAE systems, this is not always the case [14].

Now we are able to extend our definitions of local and global index to nonlinear systems as follows. The local index is the index of the linear constant-coefficient system that results from linearizing a nonlinear system at a given fixed time. The global index is the number of iterations before Algorithm 1.1 terminates when it is applied directly to the nonlinear system (without linearization). The application of the algorithm to particular nonlinear systems, where $F_1$ is not necessarily linear in $x'$, is exemplified at the end of Section 2.

Unfortunately, numerical ODE methods that work for linear constant-coefficient systems break down when the matrices are time-dependent and the (global) index is greater than one [14]. In fact, we are not aware of any numerical ODE methods for solving general linear DAE systems, let alone general nonlinear systems. (There is a method which has recently been suggested by Campbell [6], [7] for solving linear DAE systems, but it is not a numerical ODE method in the sense that we have described here because it involves differentiating the coefficient matrices.) Thus, it comes as quite a surprise that there is a large class of numerical ODE methods that work for solving the special high-index nonlinear DAE systems that we study in this paper. In many ways the behavior of these complicated nonlinear systems is suggestive of the behavior of the much simpler constant-coefficient systems, and numerical methods that work for constant-coefficient systems appear also to be useful for solving the nonlinear systems that we study here.

Now that we have finally dispensed with the background information that is necessary for understanding our results, the outline of the rest of this paper is as follows. In Section 2 we study the error propagation properties of the backward differentiation formulas (BDF) applied to systems of the form (1.1). We show that when the functions $F_1$ and $F_2$ satisfy a set of assumptions that are physically reasonable for the problems studied in Section 3, then a zero-stable $k$-step BDF converges with a global error of $O(h^k)$. The assumptions in Section 2 guarantee that the global index of the systems analyzed will be either one or two, that the local index will be equal to the global index, and that there will be no turning points in the system (turning points are points at which the system structure changes [9]).

In Section 3 we discuss several classes of problems arising in three application areas. First, we describe the DAEs arising in the modelling of electrical networks. These problems usually have an index of either one or two, and we can apply the results of Section 2 to show that the BDF methods converge with the expected order of accuracy (assuming sufficiently accurate starting values). Secondly, we discuss DAEs that describe the flow of fluids. For example, when the incompressible Navier-Stokes equations are discretized via the numerical method of lines, they yield a DAE system that satisfies the assumptions of Section 2. Lastly, we investigate DAEs that arise in the simulation of mechanical systems of rigid bodies. These systems may be written in several different, but equivalent (in the sense that they have the same analytical solutions) forms. In one form, which is probably the most natural, it turns out that the index (both local and global) of the system is three. Hence, the theory in Section 2 does not apply to these problems. We study the error propagation properties of the BDF applied to these special index three systems, and show that the BDF methods again converge with the expected order of accuracy. Similar results on the convergence of BDF methods for the index two and index three nonlinear DAE systems have recently been obtained independently by Brenan [4].

**2. Global Error Analysis for Index One and Two Nonlinear Systems.** In this section we analyze the errors introduced in the solution of the differential/algebraic system (1.1) when it is discretized by a constant stepsize BDF method. We find that, provided the functions $F_1$ and $F_2$ are such that the system is solvable and has a global index of either one or two (for index two, one additional assumption must be satisfied), then the global error of a zero-stable $k$-step constant stepsize BDF method is $O(h^k)$. A BDF method is zero-stable if $1 \leqslant k \leqslant 6$.

2.1. *Error Analysis.* We begin by analyzing the error propagation properties for functions satisfying the assumptions that we require to prove convergence. At the end of the section, we demonstrate that these assumptions are equivalent to assuming that the system has global (and local) index of either one or two, plus the additional assumption, if the index is two, that the nonzero rows of $\partial F_2/\partial y$ are linearly independent. Now, to begin the analysis, consider the system (1.1),

$$0 = F_1(x, x', y, t), \qquad 0 = F_2(x, y, t). \tag{2.1}$$

Let $F = (F_1^T, F_2^T)^T$ and assume that an analytical solution to (2.1) exists. To solve (2.1) numerically at $t_n$, Gear [13] replaces $x'(t_n)$ in (2.1) by $\rho x_n/h$, where $\rho$ is the difference operator defined by

$$\rho x_n = \sum_{i=0}^{k} \alpha_i x_{n-i}, \tag{2.2}$$

$h = t_n - t_{n-1}$ and $\alpha_i$ are the BDF coefficients, to obtain the following system of nonlinear equations:

$$F\left(x_n, \frac{\rho x_n}{h}, y_n, t_n\right) = 0. \tag{2.3}$$

The solution $(x_n, y_n)$ of (2.3) is calculated by Newton's method. According to the implicit function theorem, [12, Theorem 10.2.2] (2.3) has a unique solution if the inverse of the Jacobian

$$J = \begin{pmatrix} \dfrac{\partial F_1}{\partial x} + \dfrac{\alpha_0}{h}\dfrac{\partial F_1}{\partial x'} & \dfrac{\partial F_1}{\partial y} \\ \dfrac{\partial F_2}{\partial x} & \dfrac{\partial F_2}{\partial y} \end{pmatrix}$$

exists.

Denote the global truncation error in $x_n$ and $y_n$ by $e_n^x$ and $e_n^y$, respectively, and the discretization error when (2.2) is substituted for $hx'(t_n)$ by $\tau_n$. The truncation error of the BDF-$k$ method has the asymptotic behavior $\|\tau_n\| = O(h^k)$. The residual of (2.3), when the Newton iterations are interrupted, is denoted by $\eta = (\eta_1^T, \eta_2^T)^T$. The computed solution solves

$$\begin{aligned} \eta &= F\left(x_n, \frac{\rho x_n}{h}, y_n, t_n\right) \\ &= F\left(x(t_n) + e_n^x, \sum_{i=0}^{k} \alpha_i \frac{(x(t_{n-i}) + e_{n-i}^x)}{h}, y(t_n) + e_n^y, t_n\right) \\ &= F\left(x(t_n) + e_n^x, x'(t_n) + \tau_n + \frac{\rho e_n^x}{h}, y(t_n) + e_n^y, t_n\right) \\ &= F(x(t_n), x'(t_n), y(t_n), t_n) + \frac{\partial F}{\partial x} e_n^x + \frac{\partial F}{\partial x'}\left(\frac{\rho e_n^x}{h} + \tau_n\right) \\ &\quad + \frac{\partial F}{\partial y} e_n^y - \{\text{terms of higher order in } e_n^x \text{ and } e_n^y \text{ and } \tau_n\}. \end{aligned}$$

Let $\delta = (\delta_1^T, \delta_2^T)^T$ be the sum of $\eta$ and the high-order terms in $e_n^x$ and $e_n^y$. Then the errors $e_n^x$ and $e_n^y$ satisfy

$$\begin{pmatrix} \alpha_0 \dfrac{\partial F_1}{\partial x'} + h\dfrac{\partial F_1}{\partial x} & h\dfrac{\partial F_1}{\partial y} \\ h\dfrac{\partial F_2}{\partial x} & h\dfrac{\partial F_2}{\partial y} \end{pmatrix} \begin{pmatrix} e_n^x \\ e_n^y \end{pmatrix} = \begin{pmatrix} -\dfrac{\partial F_1}{\partial x'}(c_n + h\tau_n) + h\delta_1 \\ h\delta_2 \end{pmatrix}, \tag{2.4}$$

where $c_n = \sum_{i=1}^{k} \alpha_i e_{n-i}^x$ and the derivatives of $F_1$ and $F_2$ are evaluated at $x(t_n)$, $x'(t_n)$, $y(t_n)$ and $t_n$. For notational convenience, we introduce $F_{1x} = \partial F_1/\partial x$, $F'_{1x} = \partial F_1/\partial x'$, $A_1 = \alpha_0 F'_{1x} + hF_{1x}$, $A_2 = \partial F_1/\partial y$, $A_3 = \partial F_2/\partial x$, and $A_4 = \partial F_2/\partial y$ in (2.4) to obtain

$$\begin{matrix} (2.5a) \\ (2.5b) \end{matrix} \qquad \begin{pmatrix} A_1 & hA_2 \\ hA_3 & hA_4 \end{pmatrix} \begin{pmatrix} e_n^x \\ e_n^y \end{pmatrix} = hJ_n \begin{pmatrix} e_n^x \\ e_n^y \end{pmatrix} = \begin{pmatrix} -F'_{1x}(c_n + h\tau_n) + h\delta_1 \\ h\delta_2 \end{pmatrix}.$$

To continue the analysis, we make the following assumption.

*Assumption* 2.1. For all $t_n$, $F'_{1x}$, $F_{1x}$, $A_2$, $A_3$ and $A_4$ are bounded, $F'_{1x}$ is a square matrix and the inverses of $F'_{1x}$ and the Schur complement $A_4 - hA_3A_1^{-1}A_2$ [10] exist for $0 < h \leqslant h_0$.

It follows from the definition of $A_1$ that $A_1^{-1}$ exists if $h$ is sufficiently small. $A_1$ and $A_4$ are both square matrices. The inverse of $hJ_n$ is computed by Gaussian elimination

$$(2.6)\quad (hJ_n)^{-1} = \begin{pmatrix} A_1^{-1} + hA_1^{-1}A_2\left(A_4 - hA_3A_1^{-1}A_2\right)^{-1}A_3A_1^{-1} & -A_1^{-1}A_2\left(A_4 - hA_3A_1^{-1}A_2\right)^{-1} \\ -\left(A_4 - hA_3A_1^{-1}A_2\right)^{-1}A_3A_1^{-1} & \left(A_4 - hA_3A_1^{-1}A_2\right)^{-1}/h \end{pmatrix}.$$

Thus, if Assumption 2.1 is satisfied, then a solution to (2.1) exists. Using (2.6), we find that the explicit expressions for the errors are

$$(2.7a)\quad \begin{aligned} e_n^x = &-\left(A_1^{-1} + hA_1^{-1}A_2\left(A_4 - hA_3A_1^{-1}A_2\right)^{-1}A_3A_1^{-1}\right)\left(F'_{1x}(c_n + h\tau_n) - h\delta_1\right) \\ &-hA_1^{-1}A_2\left(A_4 - hA_3A_1^{-1}A_2\right)^{-1}\delta_2, \end{aligned}$$

$$(2.7b)\quad \begin{aligned} e_n^y = &\left(A_4 - hA_3A_1^{-1}A_2\right)^{-1}A_3A_1^{-1}\left(F'_{1x}(c_n + h\tau_n) - h\delta_1\right) \\ &+\left(A_4 - hA_3A_1^{-1}A_2\right)^{-1}\delta_2. \end{aligned}$$

Note that the errors in the "algebraic"variable $y$ on previous time steps $e_i^y$, $i < n$ do not influence the errors in any of the variables at the current time $t_n$.

Consider two different possibilities: I. $A_4$ is nonsingular, II. $A_4$ is identically equal to zero. (Later, we will show that it is sufficient to consider these two cases.) An error analysis for Case I is carried through in Liniger [20]. In this case, it follows from (2.5b) that

$$(2.8)\quad e_n^y = A_4^{-1}(\delta_2 - A_3e_n^x).$$

For Case II, by the assumption that $A_4 - hA_3A_1^{-1}A_2$ is nonsingular, $A_2$ has full column rank when $A_4 = 0$. Hence, $e_n^y$ is uniquely determined by (2.5a). Multiply (2.5a) from the left by $A_2^T$ and solve the resulting equation for $e_n^y$ in Case II:

$$(2.9)\quad e_n^y = \left(A_2^TA_2\right)^{-1}A_2^T\left(\delta_1 - \left(F'_{1x}\frac{\rho e_n^x}{h} + F_{1x}e_n^x + F'_{1x}\tau_n\right)\right).$$

The asymptotic expansion in power of $h$ for $e_n^y$ in (2.8) and (2.9) are directly dependent on the corresponding expansions for $e_n^x$. These expansions will now be investigated.

The nonsingular $F'_{1x}$ has an LU-factorization $F'_{1x} = B_1B_2$. Let $\gamma$ denote $h/\alpha_0$ and $B = (F'_{1x})^{-1} = B_2^{-1}B_1^{-1}$. It follows from the definition of $A_1$ that

$$A_1 = \alpha_0B_1\left(I + \gamma B_1^{-1}F_{1x}B_2^{-1}\right)B_2,$$

and if $h$ is sufficiently small, $h\|B_1^{-1}F_{1x}B_2^{-1}\|/\alpha_0 < 1$, then

$$(2.10)\quad A_1^{-1} = B_2^{-1}\left(I + \gamma A'_{11}\right)B_1^{-1}\alpha_0^{-1} = \left(B + \gamma A_{11}\right)\alpha_0^{-1},$$

where

$$\|A_{11}\| = O(h^0) = O(1).$$

Let

$$A'_2 = \alpha_0A_1^{-1}A_2 = BA_2 + \gamma A_{11}A_2$$

and substitute (2.10) into the Schur complement

$$D = A_4 - \gamma A_3 A_2' = A_4 - \gamma A_3 B A_2 - \gamma^2 A_3 A_{11} A_2. \tag{2.11}$$

Substitute (2.10) and (2.11) into (2.7a) and simplify to obtain an expression for $e_n^x$ which is valid for both Case I and Case II:

$$\alpha_0 e_n^x = -\left(I + \gamma A_2' D^{-1} A_3\right)\left(I + \gamma A_{11} B^{-1}\right)\left(c_n + h\tau_n - hB\delta_1\right) - hA_2' D^{-1}\delta_2. \tag{2.12}$$

Now we will proceed to bound the size of $e_n^x$ and $e_n^y$ for both cases. For Case I, $A_4$ is nonsingular, so that for $h$ sufficiently small, the inverse of $D$ in (2.11) can be written

$$D^{-1} = A_4^{-1} + \gamma D_1, \qquad \|D_1\| = O(1).$$

It follows from (2.12) that $e_n^x$ satisfies the difference equation

$$\begin{aligned} \rho e_n^x = -h\tau_n - \gamma\left(A_{11}B^{-1} + A_2' D^{-1} A_3\right)&\left(\sum_{i=1}^{k} \alpha_i e_{n-i}^x + h\tau_n\right) \\ &+ hB\delta_1 - hA_2' D^{-1}\delta_2 + \gamma^2 b + h\gamma\delta_3, \end{aligned} \tag{2.13}$$

where $\|b\| \leqslant k_b \|c_n + h\tau_n\|$ and $\|\delta_3\| \leqslant k_\delta \|\delta\|$. Suppose that $h$ is sufficiently small and that the initial global errors are bounded, $\|e_i^x\| \leqslant \xi$, $i = 0, 1, \ldots, k-1$, and that $\|\delta\| \leqslant \varepsilon$. Then, it follows from Lemma 3.2 in Henrici [18] that there are positive constants $K_1$, $K_2$, $K_3$, and $K_4$ such that

$$\|e_n^x\| \leqslant \left(K_1\xi + K_2 t_n h^k + K_3 t_n \varepsilon\right)\exp(K_4 t_n).$$

Hence, the asymptotic behavior as $\xi \to 0$, $h \to 0$, and $\varepsilon \to 0$ of the global error for a given $t_n$ is

$$\|e_n^x\| = O(\xi) + O(h^k) + O(\varepsilon). \tag{2.14}$$

The global error in a purely differential system (1.1) without algebraic equations would exhibit the same dependence on $\xi, h$, and $\varepsilon$ as $e_n^x$ in (2.14). The conclusion from (2.8) is that

$$\|e_n^y\| = O(\xi) + O(h^k) + O(\varepsilon). \tag{2.15}$$

For nonlinear systems, $\delta_1$ and $\delta_2$ in (2.9) and (2.13) depend on $e_n^x$ and $e_n^y$ because they are higher-order terms that were neglected in forming (2.4). Apart from the residuals from the Newton iteration, $\delta_1$ and $\delta_2$ consist of terms of the form

$$\frac{\partial^2 F}{\partial x^2} e_n^x e_n^x, \quad \frac{\partial^2 F}{\partial x \partial y} e_n^x e_n^y, \quad \frac{\partial^2 F}{\partial y^2} e_n^y e_n^y,$$

$$\frac{\partial^2 F}{(\partial x')^2}\left(\frac{\rho e_n^x}{h}\right)\left(\frac{\rho e_n^x}{h}\right), \quad \frac{\partial^2 F}{\partial x \partial x'} e_n^x\left(\frac{\rho e_n^x}{h}\right), \quad \frac{\partial^2 F}{\partial y \partial x'} e_n^y\left(\frac{\rho e_n^x}{h}\right),$$

where the derivatives are evaluated somewhere between the true solution and the numerical solution.

It follows from the definition of $\delta$ that

$$\|\delta\| \leqslant \|\eta\| + C_0\left(\|e_n^x\| + \|e_n^y\| + \|\rho e_n^x / h\|\right)^2.$$

Furthermore, by (2.13), $\|\rho e_n^x / h\| \leqslant C'(h^k + \|e_n^x\| + \|\delta\|)$. Thus, if $\xi = O(h^k)$ and $\|\eta\| = O(h^k)$ in Case I, then we infer from (2.13), (2.14) and (2.15) that there is an $\varepsilon$ of $O(h^k)$ and the magnitude of the errors $\|e_n^x\|$ and $\|e_n^y\|$ is of $O(h^k)$.

Now in Case II, we have $A_4 = 0$, and if we assume additionally that in (2.11) $A_3BA_2$ is nonsingular, then the inverse of $D$ for sufficiently small values of $h$ is

$$D^{-1} = -\left((A_3BA_2)^{-1} + \gamma D_2\right)/\gamma, \qquad \|D_2\| = O(1).$$

Recall that $\gamma = h/\alpha_0$. The equation in this case that corresponds to (2.12) is

$$\begin{aligned}\alpha_0 e_n^x = &-\left(I - BA_2(A_3BA_2)^{-1}A_3 + \gamma A_5\right)\left(I + \gamma A_{11}B^{-1}\right)(c_n + h\tau_n - hB\delta_1)\\ &+\alpha_0\left(BA_2(A_3BA_2)^{-1} + \gamma A_6\right)\delta_2.\end{aligned} \tag{2.16}$$

The terms proportional to $\gamma^i$, $i \geqslant 1$, in the first matrices multiplying the two error terms in (2.16) are collected in $A_5$, $\|A_5\| = O(1)$, and $A_6$, $\|A_6\| = O(1)$. The assumption guarantees that the index of the system will be two. An example where this assumption does not hold is analyzed in Section 3.

Let $H = BA_2(A_3BA_2)^{-1}A_3$. Then the matrix $H$ is a projector with the properties [1]:

(1) $H^2 = H$,

(2) An eigenvalue of $H$ is either 0 or 1,

(3) If $z \in N(A_3) = N(H)$ then $Hz = 0$, and if $z \in R(BA_2) = R(H)$, then $Hz = z$,

where $N(H)$ and $R(H)$ are the null space and the range of $H$, respectively. It follows from (2.5b) that the component of $e_n^x$ in $R(H)$ fulfills

$$He_n^x = BA_2(A_3BA_2)^{-1}A_3e_n^x = BA_2(A_3BA_2)^{-1}\delta_2. \tag{2.17}$$

The errors $e_i^x$, $i = n-k,\ n-k+1, \ldots, n-1$, also satisfy relations similar to (2.17),

$$H_ie_i^x = B_iA_{2i}(A_{3i}B_iA_{2i})^{-1}A_{3i}e_i^x = B_iA_{2i}(A_{3i}B_iA_{2i})^{-1}\delta_{2i} = C_i\delta_{2i}, \tag{2.18}$$

where the subscript $i$ indicates that an array is evaluated at $x(t_i)$, $x'(t_i)$, $y(t_i)$, $t_i$, and $H = H_n$. By virtue of (2.17), the difference equation (2.16) can be written

$$\begin{aligned}\rho e_n^x = &-\gamma\left(A_5 + (I - H + \gamma A_5)A_{11}B^{-1}\right)c_n\\ &+H\rho e_n^x - h(I - H + \gamma A_5)\left(I + \gamma A_{11}B^{-1}\right)(\tau_n - B\delta_1) + \alpha_0\gamma A_6\delta_2.\end{aligned}$$

The following identity is derived using (2.18):

$$\begin{aligned}H\rho e_n^x &= \sum_{i=0}^{k}\alpha_iH_{n-i}e_{n-i}^x + \sum_{i=1}^{k}\alpha_i(H - H_{n-i})e_{n-i}^x\\ &= \sum_{i=n-k}^{n}\alpha_{n-i}C_i\delta_{2i} + \sum_{i=1}^{k}\alpha_i(H - H_{n-i})e_{n-i}^x.\end{aligned}$$

Thus,

$$\begin{aligned}\rho(e_n^x - C_n\delta_{2n}) &= h\sum_{i=1}^{k}\alpha_iW_{n-i}e_{n-i}^x - hZ(\tau_n - B\delta_1) + hA_6\delta_2\\ &= h\sum_{i=1}^{k}\alpha_iW_{n-i}\left(e_{n-i}^x - C_{n-i}\delta_{2,n-i}\right)\\ &\quad+h\sum_{i=1}^{k}\alpha_iW_{n-i}C_{n-i}\delta_{2,n-i} - hZ(\tau_n - B\delta_1) + hA_6\delta_2,\end{aligned} \tag{2.19}$$

where

$$W_{n-i} = \frac{(H - H_{n-i})}{h} - \alpha_0^{-1}\left(A_5 + (I - H + \gamma A_5)A_{11}B^{-1}\right),$$
$$Z = (I - H + \gamma A_5)(I + \gamma A_{11}B^{-1}).$$

Since $A_{3i}B_iA_{2i}$, $i = 0, 1, \ldots, n$ are nonsingular, and the matrices defining $C_i$ and $H_i$ are smooth, there exist constants $\mathscr{H}_0$, $\mathscr{H}_1$ and $\mathscr{H}_2$ independent of $h$ with the properties

$$\|C_i\| \leqslant \mathscr{H}_0, \quad \|H - H_{n-i}\| \leqslant ih\mathscr{H}_1, \quad \|W_i\| \leqslant \mathscr{H}_2. \tag{2.20}$$

Note that (2.18) is valid only for $i \geqslant k$ because $i < k$ are the initial values. Thus, (2.19) and the identity preceding (2.19) are valid only for $n \geqslant 2k$, because they involve $k$ past values of $e_n^x$. It is possible to show [4] that $e_n^x = O(h^k)$, $e_n^y = O(h^{k-1})$ for $n = k, \ldots, 2k - 1$. In the remainder of the analysis, we assume that we are starting with these approximations as the 'initial values'.

Suppose that $\|e_i^x\| \leqslant \xi$, $i = 0, 1, \ldots, k$, $\|\delta_1\| \leqslant \varepsilon_1$ and $\|\delta_{2i}\| \leqslant \varepsilon_2$. Then, $\|e_i^x - C_i\delta_{2i}\| \leqslant \xi + \mathscr{H}_0\varepsilon_2$ for $i = 0, 1, \ldots, k$. Then, Lemma 3.2 of Henrici [18] asserts that there is an upper bound on the solution to (2.19),

$$\|e_n^x - C_n\delta_2\| \leqslant \left(k_1\xi + k_2t_nh^k + k_3t_n\varepsilon_1 + k_4t_n\varepsilon_2\right)\exp(k_5t_n),$$

where $k_i$, $i = 1, \ldots, 5$ are positive constants. The bound on $\|e_n^x\|$ is

$$\begin{aligned}\|e_n^x\| &\leqslant \|C_n\delta_2\| + \|e_n^x - C_n\delta_2\| \\ &\leqslant \mathscr{H}_0\varepsilon_2 + \left(k_1\xi + k_2t_nh^k + k_3t_n\varepsilon_1 + k_4t_n\varepsilon_2\right)\exp(k_5t_n).\end{aligned}$$

The asymptotic behavior of the global error in $x$ as $\xi \to 0$, $h \to 0$, $\varepsilon_1 \to 0$ and $\varepsilon_2 \to 0$ is

$$\|e_n^x\| = O(\xi) + O(h^k) + O(\varepsilon_1) + O(\varepsilon_2). \tag{2.21}$$

Introduce the expression for $\rho e_n^x$ from (2.19) into (2.9) and utilize (2.21). The asymptotic behavior of $e_n^y$ is

$$\|e_n^y\| = O(\xi) + O(h^k) + O(\varepsilon_1) + O(\varepsilon_2/h). \tag{2.22}$$

Since $A_4 \equiv 0$ in Case II, $\delta_2$ is a function only of $e_n^x$ and not of $e_n^y$. Suppose that $\xi = O(h^k)$, $\eta_1 = O(h^k)$ and $\eta_2 = O(h^{k+1})$. It follows from (2.19) that

$$\|\rho e_n^x/h\| \leqslant k'\left(\|e_n^x\| + h^k + \|\delta_1\| + \|\delta_2\| + \|\rho C_n\delta_{2n}\|/h\right) \leqslant k''\left(h^k + \varepsilon_1 + \varepsilon_2/h\right).$$

According to the definition and (2.21), (2.22) and the above estimate, $\delta_1$ and $\delta_2$ are bounded by

$$\begin{aligned}\|\delta_1\| &\leqslant \|\eta_1\| + k_0\left(\|e_n^x\| + \|e_n^y\| + \|\rho e_n^x/h\|\right)^2 \\ &\leqslant k_1\left(h^k + h^{2k} + \varepsilon_1^2 + \frac{\varepsilon_2^2}{h^2} + h^k\varepsilon_1 + h^{k-1}\varepsilon_2 + \frac{\varepsilon_1\varepsilon_2}{h}\right), \\ \|\delta_2\| &\leqslant \|\eta_2\| + k_2\|e_n^x\|^2 \\ &\leqslant k_3\left(h^{k+1} + h^{2k} + \varepsilon_1^2 + \varepsilon_2^2 + h^k\varepsilon_1 + h^k\varepsilon_2 + \varepsilon_1\varepsilon_2\right).\end{aligned} \tag{2.23}$$

Determine $\varepsilon_1$ and $\varepsilon_2$ as solutions of

$$\begin{aligned}\varepsilon_1 &= k_1\left(h^k + h^{2k} + \varepsilon_1^2 + \frac{\varepsilon_2^2}{h^2} + h^k\varepsilon_1 + h^{k-1}\varepsilon_2 + \frac{\varepsilon_1\varepsilon_2}{h}\right), \\ \varepsilon_2 &= k_3\left(h^{k+1} + h^{2k} + \varepsilon_1^2 + \varepsilon_2^2 + h^k\varepsilon_1 + h^k\varepsilon_2 + \varepsilon_1\varepsilon_2\right),\end{aligned} \tag{2.24}$$

and solve (2.24) by functional iteration $\underline{\varepsilon} = \underline{G}(\underline{\varepsilon})$ with the initial value $\underline{\varepsilon}^{(0)}$ satisfying

$$\varepsilon_1^{(0)} = k_1 h^k, \qquad \varepsilon_2^{(0)} = k_3 h^{k+1}.$$

Then $\underline{\varepsilon}^{(1)} = \underline{G}(\underline{\varepsilon}^{(0)})$ has the property

$$\varepsilon_1^{(1)} = O(h^k), \qquad \varepsilon_2^{(1)} = O(h^{k+1}),$$

as before and $\|\partial\underline{G}/\partial\underline{\varepsilon}\| = O(h^{k-1})$, where the derivative is evaluated at $\underline{\varepsilon}^{(0)}$. For $k > 1$, we can use the contraction mapping theorem to conclude that $\varepsilon_1 = O(h^k)$ and $\varepsilon_2 = O(h^{k+1})$. For $k = 1$, we cannot apply the theorem directly because $\|\partial\underline{G}/\partial\underline{\varepsilon}\| = O(1)$. But, if we scale the variables by $\bar{\varepsilon}_1 = \varepsilon_1/\sqrt{h}$, $\bar{\varepsilon}_2 = \varepsilon_2/h$, we can then apply the same strategy as above to reach the conclusion.

A complication to the above analysis is that the matrices $\partial^2F/\partial x^2$, $\partial^2F/\partial x\,\partial y$, $\partial^2F/\partial y^2$, etc., depend on $e_n^x$ and $e_n^y$, and we have not taken this into account in (2.23). Modify $\underline{G}$ by multiplying by $C'(1 + h + \|e_n^x\| + \|e_n^y\|)$ where the extra terms arise because the matrices depend on $t$, $e_n^x$ and $e_n^y$. Then solve the new system by functional iteration with the initial value $\underline{\varepsilon}^{(0)}$ taken to be the converged solution of the previous iterative procedure. The same analysis applies also in this case. Hence, there exist $\varepsilon_1 = O(h^k)$ and $\varepsilon_2 = O(h^{k+1})$ for $k \geqslant 1$ such that the inequalities $\|\delta_1\| \leqslant \varepsilon_1$ and $\|\delta_2\| \leqslant \varepsilon_2$ are satisfied. Moreover, it follows from (2.21) and (2.22) that the error in Case II behaves asymptotically as we would expect with a BDF-$k$ method,

$$\|e_n^x\| = O(h^k), \qquad \|e_n^y\| = O(h^k).$$

Examples of systems of physical and technical importance with the above structure are discussed in the next section.

The case when $A_4$ is singular but $A_4 \neq 0$ can easily be brought to the form of Case II by a linear transformation if we assume:

*Assumption* 2.2. If $A_4$ is singular, then the rows in $A_4$ different from zero are linearly independent.

Permute the rows of $A_3$, $A_4$ and $\delta_2$ such that they have the structure

$$A_3 = \begin{pmatrix} A_{31} \\ A_{32} \end{pmatrix}, \quad A_4 = \begin{pmatrix} A_{41} \\ 0 \end{pmatrix}, \quad \delta_2 = \begin{pmatrix} \delta_{21} \\ \delta_{22} \end{pmatrix},$$

where the rows of $A_{41}$ are not identically equal to zero. If Assumption 2.2 holds, then there is a permutation of the columns of $A_2$ and $A_{41}$ and of $e_n^y$,

$$A_{41} = (A_{42}, A_{43}), \qquad (e_n^y)^T = \left((e_{n2}^y)^T, (e_{n3}^y)^T\right),$$

where $A_{42}$ is nonsingular. The components $e_{n2}^y$ of $e_n^y$ corresponding to $A_{42}$ can be expressed as

$$e_{n2}^y = A_{42}^{-1}(\delta_{21} - A_{31}e_n^x - A_{43}e_{n3}^y), \tag{2.25}$$

see (2.5b). Insert (2.25) into (2.5a) and remove the rows of $A_{31}$, $A_{41}$ and $\delta_{21}$. The original system has been reduced to a Case II system with the global errors $e_n^x$ and $e_{n3}^y$ where $\delta_{22}$ depends only on $e_n^x$. The Case II analysis is directly applicable. The asymptotic behavior of $\|e_{n2}^y\|$ is obtained by (2.25).

2.2. *Index of ODEs Coupled with Constraints.* The error analysis of Section 2 is now complete. The structure of the DAE systems in the above analysis is either

$$F_1(x, x', y, t) = 0, \tag{2.26a}$$

$$F_2(x, y, t) = 0, \tag{2.26b}$$

where $\partial F_2/\partial y$ is nonsingular, or

$$F_1(x, x', y, t) = 0, \tag{2.27a}$$

$$F_2(x) = 0. \tag{2.27b}$$

The number of equations in $F_1$ ($F_2$) is the same as the number of components of $x$ ($y$).

In both (2.26) and (2.27), $F'_{1x} = \partial F_1/\partial x'$ is nonsingular and in (2.26), also $A_4 = \partial F_2/\partial y$ is nonsingular. Apply Algorithm 1.1 to (2.26) and take the time derivative of (2.26b) to obtain

$$A_3(x, y, t)x' + A_4(x, y, t)y' + \frac{\partial F_2}{\partial t} = 0. \tag{2.28}$$

The matrix $A$ in the algorithm is

$$A = \begin{pmatrix} F'_{1x} & 0 \\ A_3 & A_4 \end{pmatrix}.$$

$A$ is invertible and by the implicit function theorem [12, Theorem 10.2.2] we can solve (2.26a) and (2.28) for $x'$ and $y'$. The result is an ODE system. The algorithm has terminated and the global index of (2.26) is one. Since $F'_{1x}$ and $A_4$ are invertible, Assumption 2.1 is fulfilled by (2.26). On the other hand, if Assumption 2.1 is valid for (2.26) and $A_4$ is nonsingular, then the global index is one.

In (2.27) compute the time derivative of (2.27b),

$$A_3(x)x' = 0. \tag{2.29}$$

Solve (2.27a) for $x'$ as a function of $x$, $y$, and $t$:

$$x' = \psi(x, y, t). \tag{2.30}$$

This is possible since $F'_{1x}$ is nonsingular. Insert (2.30) into (2.29) and compute a new time derivative of $A_3\psi$. According to [12, Theorem 10.2.2],

$$(A_3\psi)' = \frac{\partial A_3}{\partial x}\psi x' - A_3BF_{1x}x' - A_3BA_2y' + A_3\frac{\partial \psi}{\partial t} = 0,$$

where $B = (F'_{1x})^{-1}$. Our matrix $A$ in the algorithm is

$$A = \begin{pmatrix} F'_{1x} & 0 \\ \dfrac{\partial A_3}{\partial x}\psi - A_3BF_{1x} & -A_3BA_2 \end{pmatrix}.$$

If $A_3BA_2$ is nonsingular, then $A$ shares this property. We can solve the equations for $x'$ and $y'$ to obtain an ODE system. The algorithm has terminated. The global index of (2.27) is two. If $F'_{1x}$ and $A_3BA_2$ are invertible and $A_4 = 0$, then Assumption 2.1 is valid for (2.27). Conversely, if Assumption 2.1 is satisfied for a system with the structure (2.27) and the inverse of the Schur complement is of $O(1/h)$ when $h \to 0$ then the inverse of $A_3BA_2$ exists and the global index is two. The asymptotic behavior of the global error follows from the analysis of Case II.

In order to determine the local index of the general system (2.1), we linearize the system and freeze the coefficients. The following system has the same local index as (2.1),

$$\begin{pmatrix} F'_{1x} & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} x' \\ y' \end{pmatrix} + \begin{pmatrix} F_{1x} & A_2 \\ A_3 & A_4 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = 0. \tag{2.31}$$

Compute the time derivative of the last row of (2.31). The local index is one if and only if $A_4$ is nonsingular, and then we are done. Now if $A_4$ is singular, let the singular value decomposition (SVD) of $A_4$ be

$$A_4 = U\Sigma V^T = (U_1, U_2)\begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix} = U_1\Sigma_1 V_1^T. \tag{2.32}$$

Then $N(A_4)$ is spanned by $V_2$ and $R(A_4)$ is spanned by $U_1$. The diagonal matrix $\Sigma_1$ has positive diagonal elements [1]. Split $y$ into two parts,

$$y = V_1 y_1 + V_2 y_2. \tag{2.33}$$

Differentiate (2.33) and the last row of (2.31). The result is

$$A_3 x' + A_4 V_1 y_1' = 0. \tag{2.34}$$

Premultiply (2.34) by $U^T$ and eliminate $x'$ by means of (2.31). Let $C = A_3 B A_2$. Then

$$\begin{pmatrix} F_{1x}' & 0 & 0 \\ 0 & \Sigma & 0 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} x' \\ y_1' \\ y_2' \end{pmatrix} + \begin{pmatrix} F_{1x} & A_2V_1 & A_2V_2 \\ -U_1^TA_3BF_{1x} & -U_1^TCV_1 & -U_1^TCV_2 \\ -U_2^TA_3BF_{1x} & -U_2^TCV_1 & -U_2^TCV_2 \end{pmatrix}\begin{pmatrix} x \\ y_1 \\ y_2 \end{pmatrix} = 0. \tag{2.35}$$

Take the time derivative of the last row of (2.35). Our new matrix $A$ is nonsingular if and only if $U_2^TA_3BA_2V_2$ is nonsingular. According to Algorithm 1.1 this is the necessary and sufficient condition for the local index of (2.1) to be two. If $A_4 \equiv 0$ (Case II), then we can take $U_2 = V_2 = I$ and $\Sigma = 0$ in (2.32). We have found that the global index is equal to the local index for index one systems of the form (2.26) and if the global (local) index of system (2.27) is two, so is the local (global) index.

**3. Systems of Importance in Physics and Engineering.** Three areas of application are discussed in this section: the simulation of electrical networks, the solution of certain equations in fluid dynamics, and the simulation of mechanical systems. Other areas of application, which we do not discuss here, are control theory [8], power systems [27], and heat flow [3]. The last reference discusses common properties in models of physics with a network structure. It is often natural and convenient to pose and solve these problems as DAE systems. In the examples presented below, the index (both global and local) never exceeds three. All of these problems can be solved by numerical ODE methods, although as we point out in [22], [26], some special care must be taken in implementing these methods so that they are reliable and efficient for solving these complicated nonlinear problems.

3.1. *Electrical Networks.* The computer-aided design of electrical networks has been the subject of many papers. Here we are interested only in design problems requiring the numerical solution of a DAE system of the form (1.1). The formulation of these problems is treated in more detail in [2], [5], [17], [19], [27].

The electrical networks under consideration consist of branches and nodes. The variables $x$ and $y$ in (1.1) correspond to currents and voltages in the network. The constitutive relationship between the current and the voltage in a branch with an inductor or a capacitor is an ODE. In a resistor branch the relationship is an algebraic equation. Furthermore, the topology of the network introduces constraints on the variables. These constraints are linear algebraic equations representing

Kirchhoff's current law and Kirchhoff's voltage law. These relations together form a DAE system (1.1).

Sincovec et al. [27] state the equations characterizing a network containing voltage sources and linear capacitors and resistors:

$$(3.1)\qquad \begin{pmatrix} C\dot{V}_C \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 0 & -I & 0 & 0 \\ 0 & I & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 & -R & 0 \\ 0 & -I & 0 & 0 & 0 & 0 & A_E \\ -I & 0 & 0 & 0 & 0 & 0 & A_C \\ 0 & 0 & -I & 0 & 0 & 0 & A_R \\ 0 & 0 & 0 & A_E^T & A_C^T & A_R^T & 0 \end{pmatrix} \begin{pmatrix} V_C \\ V_E \\ V_R \\ I_E \\ I_C \\ I_R \\ V_N \end{pmatrix} = \begin{pmatrix} 0 \\ E \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

The variables $V_i$ and $I_i$ represent the voltages and currents in branches with voltage sources, capacitors or resistors ($i = E, C$ or $R$). The node voltage vector is $V_N$. The diagonal matrices $C$ and $R$ have positive diagonal elements and may be time-dependent. The matrices $A_i$, $i = E$, $C$ or $R$, reflect the structure of the network, and the equations in which they appear are derived from Kirchhoff's laws. The driving term $E = E(t)$ on the right-hand side of (3.1) characterizes the voltage sources. The Jacobian of a nonlinear problem has a similar structure [27].

Let $A_*$ and $A'$ be defined by

$$A_* = \begin{pmatrix} A_E \\ A_C \\ A_R \end{pmatrix}, \qquad A' = \begin{pmatrix} A_E \\ A_C \end{pmatrix}.$$

It is shown in [22] and [27] that if the columns of $A_*$ and the rows of $A'$ are linearly independent, then the system (3.1) has index one.

In Sincovec et al. [27] the conditions on $A_*$ and $A'$ are given physical interpretations. The columns of $A_*$ are linearly independent if and only if there is a path from every node through branches to the ground node, i.e., the network is "connected". A loop in the network is a path from one node via branches in the loop back to the original node. The rows of $A'$ are linearly dependent if and only if there is a loop in the network with branches containing only voltage sources and capacitors.

The size of the DAE system containing the differential equations can be reduced by eliminating variables. There are several ways of performing this elimination systematically. The objective of Kuh and Rohrer [19] is to obtain a system of ODEs satisfied by the state variables. This is always possible by Algorithm 1.1 when the system is linear and the matrix corresponding to $A_4$ in the Jacobian of (3.1) is nonsingular. Then the rest of the variables are expressed as linear functions of the state variables. These reductions are also carried out in Hachtel et al. [17] and Sincovec et al. [27].

In the case when the rows of $A'$ are linearly dependent, Sincovec et al. [27] operate on the original DAE system with linear transformations, yielding a system of the form

$$(3.2)\qquad \begin{pmatrix} A_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} y' + \begin{pmatrix} 0 & A_2 & 0 \\ A_2^T & 0 & 0 \\ B_1 & B_2 & I \end{pmatrix} y = g(t),$$

where $A_1$ is nonsingular. This system has a global and local index of two and satisfies Assumptions 2.1 and 2.2, and so the error analysis of Section 2 applies. We can see that (3.2) has a global index of two by applying one iteration of Algorithm 1.1 to (3.2). To do this, differentiate the second and third rows and substitute for the upper part of $y'$ from the first equation to obtain

$$(3.3)\qquad \begin{pmatrix} A_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & B_2 & I \end{pmatrix} y' + \begin{pmatrix} 0 & A_2 & 0 \\ A_2^{T'} & -A_2^T A_1^{-1} A_2 & 0 \\ B_1' & B_2' - B_1 A_1^{-1} A_2 & 0 \end{pmatrix} y = \hat{g}(t).$$

By interchanging column and row two and column and row three, we can see that the index of (3.3) is one if the inverse of $A_2^T A_1^{-1} A_2$ exists, i.e., if and only if $A_2$ has linearly independent columns. Since (3.3) was obtained by applying one iteration of Algorithm 1.1 to (3.2), and each iteration of that algorithm reduces the index of a system by one [14], it follows that the global index of (3.2) is two. Using a virtually identical argument on the local linearization of (3.2), we can show that the local index of (3.2) is two.

The index of the subsystem in (3.2) defined by

$$(3.4)\qquad \begin{pmatrix} A_1 & 0 \\ 0 & 0 \end{pmatrix} y' + \begin{pmatrix} 0 & A_2 \\ A_2^T & 0 \end{pmatrix} y = g(t)$$

is also two, by a similar argument. The reduction of this subsystem (3.4) is continued in Sincovec et al. [27] by differentiating the algebraic constraints and using the differential equations to eliminate the variables appearing in the equations without derivatives. The result is a subsystem of only ODEs.

3.2. *Fluid Dynamics*. The flow of an incompressible, viscous fluid is described by the Navier-Stokes equations

$$(3.5)\qquad \frac{\partial u}{\partial t} + (u \cdot \nabla) u = -\nabla p + \gamma \nabla^2 u,$$

$$(3.6)\qquad \nabla \cdot u = 0,$$

where $u$ is the velocity in two or three space dimensions, $p$ is the pressure and $\gamma$ is the kinematic viscosity. Equation (3.5) is the momentum equation and (3.6) is the incompressibility condition. In addition to (3.5), (3.6), the flow satisfies problem dependent boundary conditions. After spatial discretization of the equations with a finite difference or finite element method, the vectors $U$ and $P$ approximating $u$ and $p$ satisfy

$$(3.7)\qquad M\dot{U} + (K + N(U))U + CP = f(U, P),$$

$$(3.8)\qquad C^T U = 0,$$

according to Gresho et al. [15]. The system of partial differential equations (3.5) and algebraic equations (3.6) has been transformed by the "method of lines" into the DAE system (3.7), (3.8). The mass matrix $M$ is the identity matrix (finite differences) or a symmetric positive-definite matrix (finite elements). The discretization of the operator $\nabla$ is $C$ and the forcing function $f$ emanates from the boundary conditions.

The system (3.7), (3.8) has the same structure as (2.27) with $A_2 = C - \partial f/\partial P$, $A_3 = C^T$ and $B = M^{-1}$. Suppose that either $\partial f/\partial P = 0$ and $C$ has linearly independent columns, or $C^T M^{-1} A_2$ is nonsingular. One of these conditions is usually satisfied. Then, under these conditions, the system has local and global index of two. Thus the error analysis of Section 2 applies to these systems. These results about the local index were also obtained by Painter [23] by explicitly computing the transformation matrices $P$ and $Q$ in (1.4). Note the similarity between (3.4) and (3.7), (3.8). The algebraic constraint is defined by $A_2^T$ and $C^T$, and the variable without derivative in the system is multiplied by $A_2$ and $C$.

An alternative formulation to (3.7) and (3.8) is used by Gresho et al. [16] to solve the Navier-Stokes equations in three dimensions. The DAE system to be solved is written as

$$M\dot{U} + (K + N(U))U + CP = f(U,P), \tag{3.9}$$

$$AP = C^T M^{-1}(f - (K + N(U))U), \tag{3.10}$$

where $A = C^T M^{-1} C$ is a discretized approximation to the Laplacian operator into which the velocity boundary conditions have been (automatically) incorporated. Equation (3.10) is derived from (3.7) and (3.8) by first differentiating (3.8) with respect to time, and then multiplying (3.7) by $C^T M^{-1}$. The fact that $C^T \dot{U} = 0$ is then used to eliminate $\dot{U}$ from (3.7) and the resulting equation is solved for $P$. Note that this is exactly one iteration of Algorithm 1.1, hence it follows that the index of the reformulated system (3.9) is one. This formulation has the advantages that first, since the index is one, there are more algorithms and better theory available for solving the problem; and second, (3.9) can now be solved with an explicit method, while (3.10) is essentially a Poisson equation and can be solved by an implicit method. A possible disadvantage is that for finite elements, the matrix $A$ may not be very sparse (although it is sparse for finite differences). Because the constraint (3.8) was linear with constant coefficients, it turns out that using the formulation (3.10) the original constraint (3.8) is satisfied exactly at every time step [16] (as opposed to "drifting off" the original constraint due to truncation errors in the solution, a phenomenon that we discuss in [26]) even though (3.9), (3.10) is implicitly using only the differential constraint $C^T \dot{U} = 0$.

The equations satisfied by a compressible, inviscid, isentropic medium are the Euler equations [11]:

$$\frac{\partial \rho}{\partial t} + (u \cdot \nabla)\rho + \rho\nabla \cdot u = 0, \tag{3.11a}$$

$$\frac{\partial u}{\partial t} + (u \cdot \nabla)u + \frac{\nabla p}{\rho} = 0, \tag{3.11b}$$

$$p - f(\rho) = 0, \tag{3.11c}$$

where the scalar $\rho$ is the density. The first equation (3.11a) represents the conservation of mass, (3.11b) is the momentum equation, and (3.11c) is the equation of state. If we, to simplify the discussion, ignore the boundary conditions, discretize (3.11) in space, and substitute the spatial derivatives by finite difference approximations, then the density vector $D$ and the velocity and pressure vectors $U$ and $P$ satisfy a DAE system:

$$\dot{D} + C_1(U)D + C_2(D)U = 0, \tag{3.12a}$$

$$\dot{U} + C_3(U)U + C_4(D)P = 0, \tag{3.12b}$$

$$P - F(D) = 0. \tag{3.12c}$$

This system has the same structure as (2.26). We find immediately that the matrix in (3.12) corresponding to $A_4$ in (2.5) is $I$ and consequently, the index of the system is one. To create a subsystem of (3.12) with index zero, simply replace $P$ in (3.12b) by $F(D)$. Then (3.12a) and (3.12b) constitute an ODE system. The addition of viscous terms dependent on $U$ on the right-hand sides of (3.11a) and (3.11b), or an extra entropy equation, does not change the index of the discretized problem.

### 3.3. *Systems of Rigid Bodies.*

3.3.1. *Problem Description*. The mechanical systems considered here consist of rigid bodies interconnected directly by joints or via other components such as springs and dampers. The vector $q$ of coordinates of the bodies satisfies the following equations [28],

$$M(q)q'' = f(q,q',t) + G(q)\lambda, \tag{3.13a}$$

$$\Phi(q) = 0. \tag{3.13b}$$

The mass matrix $M$ is nonsingular almost everywhere, $\lambda$ is the Lagrange multiplier vector and $\partial\Phi/\partial q = G^T$. The algebraic equation (3.13b) often represents geometrical constraints on the system.

A simple example of a system such as (3.13) is the physical pendulum. Let $L$ denote the length of the bar, $\lambda$ the force in the bar, and $x$ and $y$ the Cartesian coordinates of an infinitesimal ball of mass one in one end of the bar. Then $x$, $y$, and $\lambda$ solve the DAE system

$$x'' = 2x\lambda, \quad y'' = 2y\lambda - g, \quad x^2 + y^2 - L^2 = 0. \tag{3.14}$$

Here, $G^T = (2x, 2y)$ and $g$ is the gravity constant.

If the initial condition $q_0 = q(0)$ is consistent with (3.13b), $\Phi(q_0) = 0$, then the algebraic constraint (3.13b) can be replaced by its differentiated form

$$G^T(q)q' = 0. \tag{3.15}$$

Moreover, if $\Phi(q_0) = 0$, $q_0' = q'(0)$ and $G^T(q_0)q_0' = 0$ then the condition (3.15) is equivalent to

$$d(G^Tq')/dt = G^Tq'' + G'^Tq' = 0. \tag{3.16}$$

We obtain a system of linear equations satisfied by $\lambda$ by introducing $q''$ from (3.13a) into (3.16),

$$G^TM^{-1}G\lambda + G^TM^{-1}f + G'^Tq' = 0. \tag{3.17}$$

It is shown in Lötstedt [21] that even if the columns of $G$ are linearly dependent in (3.17) and (3.13a) and $\lambda$ is not unique, the contribution $G\lambda$ to the equations of motion (3.13a) is unique. Furthermore, there is a matrix $G_*$ consisting of linearly independent columns of $G$ which can replace $G$ in (3.13a) and (3.17) such that the solutions to (3.13) with $G$ and $G_*$ are identical. Henceforth, we assume that the columns of $G$ are linearly independent.

We eliminate $\lambda$ in (3.13a), using the solution of (3.17), and arrive at an ODE system

$$Mq'' + G(G^TM^{-1}G)^{-1}G'^Tq' = \left(I - G(G^TM^{-1}G)^{-1}G^TM^{-1}\right)f \tag{3.18}$$

to solve for the coordinates. Note that each time the constraints are differentiated in (3.15), (3.16), this is equivalent to one iteration of Algorithm 1.1. Hence, the global index of the system with the constraint (3.15) is one less than the index of the system with the constraint (3.13b), and the index of the system with the constraint (3.16) is one less than the index of the system with the constraint (3.15). Now, since (3.13a) coupled with (3.16) is equivalent to (3.17) coupled with (3.18) by simple substitutions which do not change the index, and the index of (3.17) coupled with (3.18) is obviously one (as long as $G^TM^{-1}G$ is nonsingular), it follows that the index of (3.13a) coupled with (3.15) is two, and the index of (3.13) is three. By a similar argument based on the local linearization of (3.13), we can show that the local index of (3.13) is three, and that the local index is reduced by one whenever the constraint is differentiated. The systems (3.13) and ((3.13a), (3.15)) have the form (2.27). The matrices corresponding to $A_3BA_2$ in the error analysis are identically zero for (3.13), and $G^TM^{-1}G$ for ((3.13a), (3.15)). The DAE system ((3.13a), (3.15)) has the same symmetry property as (3.4) and ((3.7), (3.8)).

There are other methods based on classical analytical mechanics aimed at reducing the number of algebraic equations in (3.13). A particular method is developed in Wittenburg [28] and several other techniques are reviewed in Paul [24].

For the numerical treatment of (3.13), the differential equations are rewritten in first-order form:

$$u' = v, \tag{3.19a}$$

$$M(u)v' = f(u,v,t) + G(u)\lambda. \tag{3.19b}$$

The scaled Jacobian of (3.19) and (3.13b) is

$$hJ_n = \begin{pmatrix} \alpha_0 I & -hI & 0 \\ hX & \alpha_0 M + hY & -hG \\ hG^T & 0 & 0 \end{pmatrix}, \tag{3.20}$$

where

$$X = \frac{\partial M}{\partial u}v' - \frac{\partial f}{\partial u} - \frac{\partial G}{\partial u}\lambda, \qquad Y = -\frac{\partial f}{\partial v}.$$

If instead of solving (3.13), we choose to solve the (analytically) equivalent system (3.13a) coupled with (3.15), then the scaled Jacobian of this new system is

$$hJ_n = \begin{pmatrix} \alpha_0 I & -hI & 0 \\ hX & \alpha_0 M + hY & -hG \\ hZ & hG^T & 0 \end{pmatrix}, \tag{3.21}$$

where

$$Z = \frac{\partial G^T}{\partial u}v.$$

Since the index of this system is two, the error analysis in Section 2 applies here. However, it does not apply to the original system (3.13). We shall now investigate the propagation of errors in this case where the index is three.

3.3.2. *Error Analysis for Index Three Case.* The global error in $u = q$ at $t_n$ is denoted by $e_n^u$ and the error in $v = q'$ by $e_n^v$. Let $\delta_1 = ((\delta_1^u)^T, (\delta_1^v)^T)^T$ in (2.7). For nonlinear systems $\delta_1$ and $\delta_2$ depend on $e_n^u$, $e_n^v$ and $e_n^\lambda$ (the error in $\lambda$). Apart from the residuals from the Newton iteration $\eta_1$, the leading terms of $\delta_1^v$ are of the form

$$\frac{\partial^2 f}{\partial u^2} e_n^u e_n^u, \quad \frac{\partial^2 f}{\partial u \partial v} e_n^u e_n^v, \quad \frac{\partial^2 f}{\partial v^2} e_n^v e_n^v$$

(i.e., terms which are of order two in $e_n^u$ and $e_n^v$), as well as

$$\frac{\partial G}{\partial u} e_n^u e_n^\lambda \quad \text{and} \quad \frac{\partial M}{\partial u} e_n^u \frac{\rho e_n^v}{h}.$$

In the analysis to follow, we will define $\delta_1^v$ to consist of the terms which are of order two or higher in $e_n^u$ and $e_n^v$. The other neglected term, $\delta_2$, is of the form

$$\delta_2 = \frac{1}{2} \frac{\partial^2 \Phi}{\partial u^2} e_n^u e_n^u + \{\text{higher-order terms in } e_n^u\} + \eta_2,$$

where $\eta_2$ consists of the residuals from the Newton iteration and round-off errors. The equation for the errors corresponding to (2.4) is then given by

$$(3.22) \qquad \begin{pmatrix} \alpha_0 I & -hI & 0 \\ hX & \alpha_0 M + hY & -hG \\ hG^T & 0 & 0 \end{pmatrix} \begin{pmatrix} e_n^u \\ e_n^v \\ e_n^\lambda \end{pmatrix} = \begin{pmatrix} -h\tau_n^u - c_n^u + h\delta_1^u \\ -hM\tau_n^v - Mc_n^v + h\delta_1^v \\ h\delta_2 \end{pmatrix}.$$

We can solve for $e_n^\lambda$ as in (2.9) to obtain

$$(3.23) \quad e_n^\lambda = (G^T M^{-1} G)^{-1} G^T \left( \tau_n^v + \frac{\rho e_n^v}{h} + M^{-1} Y e_n^v + M^{-1} X e_n^u - M^{-1} \delta_1^v \right).$$

Insert the matrices from (3.22) corresponding to $A_1$, $A_2$, $A_3$ and $A_4$ in (2.7a), and form the matrix $D$ in (2.11),

$$(3.24) \qquad \begin{aligned} D &= h(G^T\ 0) \begin{pmatrix} \alpha_0 I & -hI \\ hX & \alpha_0 M + hY \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ G \end{pmatrix} \\ &= \gamma (G^T\ 0) \begin{pmatrix} I - \gamma^2 S^{-1} X & \gamma S^{-1} \\ -\gamma S^{-1} X & S^{-1} \end{pmatrix} \begin{pmatrix} 0 \\ G \end{pmatrix} \\ &= \gamma^2 G^T S^{-1} G, \end{aligned}$$

where $S = M + \gamma Y + \gamma^2 X$ and $\gamma = h/\alpha_0$. There is a matrix $S_1$, for $\gamma$ sufficiently small, such that

$$(3.25) \qquad S^{-1} = M^{-1} + \gamma S_1, \qquad \|S_1\| = O(1).$$

Now, utilizing (3.24) and (2.12), the errors $e_n^u$ and $e_n^v$ satisfy

$$(3.26) \qquad \begin{aligned} \alpha_0 \begin{pmatrix} e_n^u \\ e_n^v \end{pmatrix} = &- \begin{pmatrix} I - S^{-1} G (G^T S^{-1} G)^{-1} G^T & 0 \\ -S^{-1} G (G^T S^{-1} G)^{-1} G^T / \gamma & I \end{pmatrix} \\ &\cdot (I + \gamma A_{11} B^{-1})(c_n + h\tau_n - hB\delta_1) \\ &+ \alpha_0 \begin{pmatrix} S^{-1} G (G^T S^{-1} G)^{-1} \\ S^{-1} G (G^T S^{-1} G)^{-1} / \gamma \end{pmatrix} \delta_2. \end{aligned}$$

Let

$$\begin{gathered} K_j = S_j^{-1}G_j\left(G_j^TS_j^{-1}G_j\right)^{-1}, \\ H_j = K_jG_j^T, \quad H = H_n, \quad H' = \begin{pmatrix} H & 0 \\ 0 & H \end{pmatrix}. \end{gathered} \tag{3.27}$$

Premultiply (3.26) by the projector $(I - H')$ to obtain

$$\alpha_0(I - H')e_n^z = -(I - H')\left(I + \gamma A_{11}B^{-1}\right)(c_n + h\tau_n - hB\delta_1), \tag{3.28}$$

where $(e_n^z)^T = ((e_n^u)^T, (e_n^v)^T)$. We have from the definition of $B$ that

$$B\delta_1 = \begin{pmatrix} \delta_1^u \\ M^{-1}\delta_1^v \end{pmatrix}.$$

In the absence of round-off errors, $\delta_1^u = 0$ by virtue of the linearity of (3.19a). Since $G_i^TS_i^{-1}G_i$, $i = 0, 1, \ldots, n$, are nonsingular, and $G_i$ and $S_i$ are smooth, then we have that there is an $\mathscr{K}_3$ such that

$$\|H' - H'_{n-i}\| \leqslant ih\mathscr{K}_3, \tag{3.29}$$

cf. (2.20).

It follows from (3.26) that

$$H_je_j^u = K_j\delta_{2,j}, \tag{3.30}$$

where $\delta_{2,j}$ is evaluated at $t_j$. Let the error in the differentiated constraint at $t_j$ be denoted by $\delta_{3,j}$,

$$\delta_{3,j} = G_j^Te_j^v = G_j^Tv_j.$$

Let $L_{j-i} = (G_j - G_{j-1})^T/h$. Since $G$ is smooth, $L_{j-i}$ is bounded independently of $h$. From the discretization of the equation $u' = v$ and (3.22) we have that

$$\delta_{3,j} = G_j^Te_j^v = \frac{\rho(\delta_{2,j})}{h} + G_j^T\tau_j^u + \sum_{i=0}^{k} \alpha_iL_{j-i}e_{j-i}^u, \tag{3.31a}$$

and hence that

$$H_je_j^v = K_j\delta_{3,j} = K_j\frac{\rho(\delta_{2,j})}{h} + H_j\tau_j^u + K_j\sum_{i=0}^{k} \alpha_iL_{j-i}e_{j-i}^u. \tag{3.31b}$$

Rewrite the right-hand side of (3.28),

$$\begin{aligned} \alpha_0(I - H')e_n^z = &-\sum_{i=1}^{k} \alpha_i\left(I - H'_{n-i}\right)e_{n-i}^z \\ &- h\sum_{i=1}^{k} \alpha_i\left[\frac{\left(H'_{n-i} - H'\right)}{h} + \alpha_0^{-1}(I - H')A_{11}B^{-1}\right]e_{n-i}^z \\ &+ h(I - H')\left(I + \gamma A_{11}B^{-1}\right)(B\delta_1 - \tau_n). \end{aligned} \tag{3.32}$$

Split $e_j^z$ into two parts:

$$\begin{aligned} e_j^z &= \left(I - H_j'\right)e_j^z + H_j'e_j^z \\ &= \left(I - H_j'\right)e_j^z + \begin{pmatrix} K_j & 0 \\ 0 & K_j \end{pmatrix}\begin{pmatrix} \delta_{2,j} \\ \delta_{3,j} \end{pmatrix} \qquad (3.33) \\ &= \left(I - H_j'\right)e_j^z + K_j'\delta_j^z. \end{aligned}$$

Insert (3.33) into (3.32) and rearrange the terms. The difference equation satisfied by $(I - H_j')e_j^z$ is

$$\begin{aligned} &\rho\left[\left(I - H_n'\right)e_n^z\right] \\ &\quad = -h\sum_{i=1}^{k}\alpha_i W_{n-i}\left(I - H_{n-i}'\right)e_{n-i}^z + hZ(B\delta_1 - \tau_n) \\ &\qquad - h\sum_{i=1}^{k}\alpha_i W_{n-i}K_{n-i}'\begin{pmatrix} \delta_{2,n-i} \\ \rho(\delta_{2,n-i}/h) + G_{n-i}^T\tau_{n-i}^u \end{pmatrix} \\ (3.34) &\qquad - h\sum_{i=1}^{k}\alpha_i W_{n-i}K_{n-i}'\begin{pmatrix} 0 \\ \sum_{m=0}^{k}\alpha_m L_{n-i-m}(I - H_{n-i-m})e_{n-i-m}^u \end{pmatrix} \\ &\qquad - h\sum_{i=1}^{k}\alpha_i W_{n-i}K_{n-i}'\begin{pmatrix} 0 \\ \sum_{m=0}^{k}\alpha_m L_{n-i-m}K_{n-i-m}\delta_{2,n-i-m} \end{pmatrix}, \end{aligned}$$

where

$$W_{n-i} = \frac{\left(H_{n-i}' - H'\right)}{h} + \alpha_0^{-1}(I + H')A_{11}B^{-1},$$
$$Z = (I - H')\left(I + \gamma A_{11}B^{-1}\right),$$

and we have used (3.30) and (3.31b) to find $H_je_j^u$ and $H_je_j^v$, and $e_{j-i}^u$ in (3.31b) has been split into two parts.

Note that, as in Section 2, (3.30) is not satisfied for the initial values. Consequently, (3.34), which involves $2k$ past values of $\delta_{2,j}$, is not valid for $n < 3k$. In the analysis to follow we assume that we are starting with $n \geqslant 3k$ with the previously computed approximations as the 'initial values'. A detailed analysis of the errors in this initial region is given in [4]. The errors after these few initial steps are of $O(h^k)$, as we require below.

By (3.29), $\|W_{n-i}\|$ is bounded independently of $h$. $\|Z\|$ has the same property. Change the order of summation in

$$\begin{aligned} (3.35)\quad h\sum_{i=1}^{k}\alpha_i W_{n-i}K_{n-i}'\begin{pmatrix} 0 \\ \rho(\delta_{2,n-i})/h \end{pmatrix} &= \sum_{i=1}^{k}\alpha_i W_{n-i}K_{n-i}'\sum_{j=0}^{k}\alpha_j\begin{pmatrix} 0 \\ \delta_{2,n-i-j} \end{pmatrix} \\ &= \sum_{j=0}^{k}\alpha_j d_{n-j} = \rho d_n, \end{aligned}$$

where

$$d_{n-j} = \sum_{i=1}^{k}\alpha_i W_{n-i}K_{n-i}'\begin{pmatrix} 0 \\ \delta_{2,n-i-j} \end{pmatrix}.$$

The equation (3.34) can be rewritten using (3.35),

$$
\begin{aligned}
&\rho\left[\left(I-H_n'\right)e_n^z+d_n\right]\\
&\quad=-h\sum_{i=1}^{k}\alpha_iW_{n-i}\left[\left(I-H_{n-i}'\right)e_{n-i}^z+d_{n-i}\right]\\
&\qquad+h\sum_{i=1}^{k}\alpha_iW_{n-i}d_{n-i}+hZ(B\delta_1-\tau_n)-h\sum_{i=1}^{k}\alpha_iW_{n-i}K_{n-i}'\begin{pmatrix}\delta_{2,n-i}\\ G_{n-i}^T\tau_{n-i}^u\end{pmatrix}\\
&\qquad-h\sum_{i=1}^{k}\alpha_iW_{n-i}K_{n-i}'\sum_{m=0}^{k}\alpha_mL_{n-i-m}'\left[\left(I-H_{n-i-m}'\right)e_{n-i-m}^z+d_{n-i-m}\right]\\
&\qquad-h\sum_{i=1}^{k}\alpha_iW_{n-i}K_{n-i}'\sum_{m=0}^{k}\alpha_mL_{n-i-m}'\left[K_{n-i-m}'\begin{pmatrix}0\\ \delta_{2,n-i-m}\end{pmatrix}-d_{n-i-m}\right].
\end{aligned}
\tag{3.36}
$$

The definition of $L_{n-i-m}'$ is

$$L_{n-i-m}'=\begin{pmatrix}0&0\\0&L_{n-i-m}\end{pmatrix}.$$

Assume that $\|\delta_{1,i}\|\leqslant\varepsilon_1$, $\|\delta_{2,i}\|\leqslant\varepsilon_2$ and $\|e_j^z\|\leqslant\xi$ for all $i$ and $j=0,1,\dots,2k$. It follows from the definition of $d_i$ that $\|d_i\|\leqslant C_0\varepsilon_2$ for some $C_0>0$. Apply Henrici's lemma to (3.36). The upper bound on $\|(I-H_n')e_n^z+d_n\|$ in terms of $h$, $\varepsilon_1$, $\varepsilon_2$ and $\xi$ is

$$\left\|\left(I-H_n'\right)e_n^z+d_n\right\|\leqslant\left(k_1\xi+k_2t_nh^k+k_3t_n\varepsilon_1+k_4t_n\varepsilon_2\right)\exp(k_5t_n),\tag{3.37}$$

where $k_i$, $i=1,2,\dots,5$ are positive constants. If $\xi=O(h^k)$, then by (3.30), (3.31) and (3.37),

$$
\begin{aligned}
&\left\|\begin{pmatrix}e_n^u\\ e_n^v-K_n\left(\rho\delta_{2,n}/h+\sum_{i=0}^{k}\alpha_iL_{n-i}e_{n-i}^u\right)\end{pmatrix}\right\|\\
&\qquad\leqslant\left\|\left(I-H_n'\right)e_n^z+d_n\right\|+\|d_n\|+\left\|\begin{pmatrix}K_n\delta_2\\ H_n\tau_n^u\end{pmatrix}\right\|\\
&\qquad\leqslant C_1\left(h^k+\varepsilon_1+\varepsilon_2\right).
\end{aligned}
\tag{3.38}
$$

Hence,

$$\|e_n^u\|\leqslant C_1\left(h^k+\varepsilon_1+\varepsilon_2\right)\tag{3.39}$$

and if the Newton residual parts $\eta_1$ and $\eta_2$ of $\delta_1$ and $\delta_2$ satisfy $\|\eta_1\|=O(h^{k+1})$ and $\|\eta_2\|=O(h^{k+2})$, then by the definition of $\delta_2$ and (3.39),

$$\|\delta_2\|\leqslant C_2\left(h^{k+2}+h^k(\varepsilon_1+\varepsilon_2)+(\varepsilon_1+\varepsilon_2)^2\right).\tag{3.40}$$

An estimate of $e_n^v$ is derived from (3.38) and (3.39):

$$\|e_n^v\|\leqslant C_3\left(h^k+\varepsilon_1+\varepsilon_2+\varepsilon_3\right),\tag{3.41}$$

where $\|\rho\delta_{2n}/h\|\leqslant\varepsilon_3$.

According to the definition, an upper bound on the leading terms of $\delta_1$ is

$$\|\delta_1\|\leqslant C_4\left(\left(\|\eta_1\|+\|e_n^u\|+\|e_n^v\|\right)^2+\|e_n^u\|\left(\|e_n^\lambda\|+\|\rho e_n^v/h\|\right)\right).\tag{3.42}$$

A bound on $\|e_n^\lambda\|$ is obtained from (3.23),

$$\|e_n^\lambda\| \leqslant C_5\left(\|\tau_n^v\| + \|e_n^u\| + \|e_n^v\| + \|\delta_1\| + \|\rho e_n^v/h\|\right). \tag{3.43}$$

In order to determine the asymptotic behavior of $\|e_n^\lambda\|$ and $\|\delta_1\|$, we need an estimate of $\rho e_n^v/h$ as $h \to 0$. By (3.31), (3.33) and (3.36), $\|\rho e_n^v/h\|$ is bounded by

$$\begin{aligned}\left\|\frac{\rho e_n^v}{h}\right\| &\leqslant \left\|\rho\left[\frac{(I - H_n)e_n^v}{h} + d_n^v\right]\right\| + \left\|\frac{\rho(K_n\delta_{3n})}{h}\right\| + \left\|\frac{\rho d_n^v}{h}\right\| \\ &\leqslant C_6\left(\|e_n^u\| + \|e_n^v\| + \|\delta_2\| + \|\delta_1\| + \|\tau_n\|\right) \\ &\quad + \left\|\frac{\rho(K_n\rho\delta_{2n})}{h^2}\right\| + \left\|\frac{\rho(H_n\tau_n^u)}{h}\right\| + \left\|\frac{\rho(K_n\rho L_n e_n^u)}{h}\right\| + \left\|\frac{\rho d_n}{h}\right\|,\end{aligned} \tag{3.44}$$

where $d_n^v$ is the part of $d_n$ corresponding to $e_n^v$.

The bound

$$\left\|\frac{\rho(H_n\tau_n^u)}{h}\right\| \leqslant C_7h^k \tag{3.45}$$

follows from the fact that both $H_n$ and $\tau_n^u = -h^k u^{(k+1)}/(k+1)$ are smooth functions.

In the fourth term on the right-hand side of (3.44), $\rho L_n e_n^u$ can be written as

$$\rho L_n e_n^u = \sum_{i=0}^{k} \alpha_i(L_{n-i} - L_n)e_{n-i}^u + L_n\rho e_n^u. \tag{3.46}$$

Since $L_n$ is smooth, and $\|e_n^u\|$ is bounded by (3.39), the summation in (3.46) is of $O(h^{k+1} + h\varepsilon_1 + h\varepsilon_2)$. The last term in (3.46) is rewritten by means of (3.22). Then

$$\|L_n\rho e_n^u\| \leqslant h\|L_n\|\left(\|e_n^v\| + \|\tau_n^u\| + \|\delta_1^u\|\right).$$

Hence, from (3.41),

$$\|\rho L_n e_n^u\| \leqslant hC_8\left(h^k + \varepsilon_1 + \varepsilon_2 + \varepsilon_3\right)$$

and consequently

$$\left\|\frac{\rho(K_n\rho L_n e_n^u)}{h}\right\| \leqslant C_9\left(h^k + \varepsilon_1 + \varepsilon_2 + \varepsilon_3\right). \tag{3.47}$$

The fifth term in (3.44) is defined by (3.35) and is bounded by

$$\|\rho d_n/h\| \leqslant C_{10}\varepsilon_3. \tag{3.48}$$

The second term in (3.44) calls for a closer examination. Let $Q_n$ denote $0.5\partial^2\phi/\partial u^2$ at $t_n$, and $r_n$ contain the terms of higher order in $e_n^u$ in $\delta_{2n}$ such that

$$\delta_{2n} = Q_n e_n^u e_n^u + r_n + \eta_{2n}.$$

We have for $k = 1$ that $e_n^u - e_{n-1}^u = he_n^v$ and

$$\begin{aligned}\rho\delta_{2n} &= Q_n e_n^u e_n^u - Q_{n-1}e_{n-1}^u e_{n-1}^u + \eta_{2n} - \eta_{2,n-1} + s_n \\ &= hQ_n\left(e_n^u e_n^v + e_n^v e_{n-1}^u\right) + \left(Q_n - Q_{n-1}\right)e_{n-1}^u e_{n-1}^u + \eta_{2n} - \eta_{2,n-1} + s_n,\end{aligned} \tag{3.49}$$

where $s_n = r_n - r_{n-1}$. It follows from the fact that $\|\eta_2\| = O(h^3)$ (3.49), (3.39) and (3.41) that

$$\begin{aligned}\left\|\frac{\rho\delta_{2n}}{h}\right\| &\leqslant C_{11}\left(h^2 + \varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_1\varepsilon_2 + h\varepsilon_1 + h\varepsilon_2 + h\varepsilon_3 + \varepsilon_1\varepsilon_3 + \varepsilon_2\varepsilon_3 + \|s_n\|/h\right) \\ &= C(1).\end{aligned} \tag{3.50}$$

When $k \geqslant 2$, we accept the bound

$$\left\| \frac{\rho \delta_{2n}}{h} \right\| \leqslant C_{12} \frac{\varepsilon_2}{h} \leqslant \varepsilon_3 = C(k), \qquad k = 2,3,\ldots. \tag{3.51}$$

Therefore, there is a $C_{13}$ such that

$$\left\| \frac{\rho(K_n \rho \delta_{2n})}{h^2} \right\| \leqslant \frac{C_{13} C(k)}{h}, \qquad k = 1,2,3,\ldots. \tag{3.52}$$

The upper bound on $\|\rho e_n^v/h\|$ is now derived from (3.44), (3.39), (3.40), (3.41), (3.45), (3.47), (3.48), and (3.52),

$$\left\| \frac{\rho e_n^v}{h} \right\| \leqslant C_{14}\left( h^k + \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \frac{C(k)}{h} \right). \tag{3.53}$$

The same bound is obtained for $\|e_n^\lambda\|$ in (3.43) except for the constant $C_{14}$. Insert the estimates (3.39), (3.41) and (3.53) into (3.42). Then

$$\begin{aligned} \|\delta_1\| \leqslant C_{15}\Big( & h^{k+1} + \varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2 + h^k \varepsilon_1 + h^k \varepsilon_2 + h^k \varepsilon_3 \\ & + \varepsilon_1 \varepsilon_2 + \varepsilon_2 \varepsilon_3 + \varepsilon_1 \varepsilon_3 + h^{k-1} C(K) + \frac{\varepsilon_1 C(k)}{h} + \frac{\varepsilon_2 C(k)}{h} \Big). \end{aligned} \tag{3.54}$$

Suppose that $\varepsilon_1 = k_1 h^i$, $\varepsilon_2 = k_2 h^j$ and $\varepsilon_3 = k_3 h^l$, and choose $i$, $j$, and $l$ as large as possible such that $\|\delta_1\| \leqslant \varepsilon_1$, $\|\delta_2\| \leqslant \varepsilon_2$ and $\|\rho \delta_{2n}/h\| \leqslant \varepsilon_3$.

Take $i = k + 1$, $j = k + 2$ and $l = k + 1$. Then, for $k \geqslant 2$, the conclusion from (3.40) is that $k_2 = C_2(1 + \tilde{C}_2)$ and from (3.51), that $k_3 = C_{12} k_2$, and from (3.54), that $k_1 = C_{15}(1 + \tilde{C}_{15})$, where $\tilde{C}_2$ and $\tilde{C}_{15}$ are bounded since $h \leqslant h_0$. When $k = 1$, it can be shown by the techniques used in (3.49) that the cubic terms in $s_n$ satisfy $\|s_n\| = O(h^4)$. Then the leading term of $C(1)$ in (3.50) is $C_{11} h^2$. Therefore, $k_3 = C_{11}(1 + \tilde{C}_{11})$, $k_1$ derived from (3.54) is $C_{15}(1 + C_{11} + \tilde{C}_{15})$, and $k_2$ in the bound $\varepsilon_2$ of $\|\delta_2\|$ in (3.40) is $C_2(1 + C_{15}(1 + C_{11}) + \tilde{C}_2)$, where $\tilde{C}_{11}$, $\tilde{C}_{15}$ and $\tilde{C}_2$ are bounded. Thus, there are constants $k_1$, $k_2$, and $k_3$ for $k \geqslant 1$ such that

$$\begin{gathered} \|\delta_1\| \leqslant \varepsilon_1 = k_1 h^{k+1}, \\ \|\delta_2\| \leqslant \varepsilon_2 = k_2 h^{k+2}, \\ \left\| \frac{\rho \delta_{2n}}{h} \right\| \leqslant \varepsilon_3 = k_3 h^{k+1}. \end{gathered} \tag{3.55}$$

The asymptotic behavior of $e_n^u$, $e_n^v$ and $e_n^\lambda$ is by (3.39), (3.41) and (3.43):

$$\|e_n^u\| = O(h^k), \quad \|e_n^v\| = O(h^k), \quad \|e_n^\lambda\| = O(h^k). \tag{3.56}$$

Finally, we observe that $\delta_{3n}$, the error in the differentiated constraint defined in (3.31), also is of $O(h^k)$.

In the above analysis, only the leading terms of the Taylor series part of $\delta_1$ and $\delta_2$ have been taken into account. However, the inclusion of the higher-order terms in $e_n^u$ and $e_n^v$ will only lead to terms of higher order in $h$ than $h^{k+1}$ in $\delta_1$ and $h^{k+2}$ in $\delta_2$. These higher-order terms have no influence on the order of the estimates in (3.55) and (3.56).

3.4. *Concluding Remarks.* In all of the examples considered in this section, the number of equations is equal to the number of variables. The matrix $F'_{1x}$ is nonsingular almost everywhere and in (3.1), (3.7) and (3.9) $F'_{1x}$ is symmetric and positive definite. If $F'_{1x}$ is nonsingular, then $A_1$ is nonsingular if $h$ is sufficiently small. A standard Newton solver of nonlinear equations (2.3) requires $hJ_n$ to be nonsingular. Then, by the identity [10]

$$\det(hJ_n) = \det(A_1)\det\left(A_4 - hA_3A_1^{-1}A_2\right)$$

the Schur complement $A_4 - hA_3A_1^{-1}A_2$ is nonsingular and Assumption 2.1 is fulfilled. The assumptions that we have made to complete the error analysis in Section 2 and Subsection 3.3 seem to be reasonable, and are usually fulfilled by certain equations of fluid dynamics and by important subclasses of electrical networks and mechanical systems.

Aerospace Division
SAAB-Scania
S-58188 Linköping, Sweden

Computing and Mathematics Research Division
Lawrence Livermore National Laboratory
Livermore, California 94550

1. A. Ben-Israel and T. N. E. Greville, *Generalized Inverses: Theory and Applications*, Wiley, New York, 1974.

2. F. H. Branin, Jr., "Computer methods of network analysis," *Proc. IEE-E*, v. 55, 1967, pp. 1787–1801.

3. F. H. Branin, Jr., "The network concept as a unifying principle in engineering and the physical sciences," in *Problem Analysis in Science and Engineering* (F. H. Branin, Jr. and K. Huseyin, eds.), Academic Press, New York, 1977.

4. K. 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.

5. D. A. Calahan, *Computer-Aided Network Design*, rev. ed., McGraw-Hill, New York, 1972.

6. S. L. Campbell, "The numerical solution of higher index linear time varying singular systems of differential equations," *SIAM J. Sci. Statist. Comput.*, v. 6, 1985, pp. 334–348.

7. S. L. Campbell, *Explicit Methods for Solving Singular Differential Equation Systems*, North Carolina State University, Raleigh, North Carolina, 1984. (Preprint.)

8. S. L. Campbell, *Singular Systems of Differential Equations*, Pitman, San Francisco, 1979.

9. S. L. Campbell & L. R. Petzold, "Canonical forms and solvable singular systems of differential equations," *SIAM J. Algebraic Discrete Methods*, v. 4, 1983, pp. 517–521.

10. R. W. Cottle, "Manifestations of the Schur complement," *Linear Algebra Appl.*, v. 8, 1974, pp. 189–211.

11. R. Courant & K. O. Friedrichs, *Supersonic Flow and Shock Waves*, Springer-Verlag, Berlin and New York, 1948.

12. J. Dieudonné, *Foundations of Modern Analysis*, Academic Press, New York, 1969.

13. C. W. Gear, "Simultaneous numerical solution of differential/algebraic equations," *IEEE Trans. Circuit Theory*, CT-18, 1971, pp. 89–95.

14. C. W. Gear & L. R. Petzold, "ODE methods for the solution of differential/algebraic systems," *SIAM J. Numer. Anal.*, v. 21, no. 4, 1984, pp. 367–384.

15. P. M. Gresho, R. L. Lee & R. L. Sani, "On the time-dependent solution of the incompressible Navier-Stokes equations in two and three dimensions," *Recent Advances in Numerical Methods in Fluids*, vol. 1, Pineridge, Swansea, 1980.

16. P. M. Gresho & C. D. Upson, *Current Progress in Solving the Time-Dependent, Incompressible Navier-Stokes Equations in Three-Dimensions by (Almost) the FEM*, UCRL-87445, Lawrence Livermore National Laboratory, Livermore, California, 1982.

17. G. D. Hachtel, R. K. Brayton & F. G. Gustavson, "The sparse tableau approach to network analysis and deisgn," *IEEE Trans. Circuit Theory*, CT-18, 1971, pp. 101–113.

18. P. Henrici, *Error Propagation for Difference Methods*, Wiley, New York, 1963.

19. E. S. Kuh & R. A. Rohrer, "The state-variable approach to network analysis," *Proc. IEE-E*, v. 53, 1965, pp. 672–686.

20. W. LINIGER, "Multistep and one-leg methods for implicit mixed differential algebraic systems," *IEEE Trans. Circuits and Systems*, CAS-26, 1979, pp. 755–762.

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

22. P. LÖTSTEDT & L. R. PETZOLD, *Numerical Solution of Nonlinear Differential Equations with Algebraic Constraints*, SAND 83-8877, Sandia National Laboratories, Livermore, California, 1983.

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

24. B. PAUL, "Analytical dynamics of mechanisms—A computer oriented overview," *Mech. Mach. Theory*, v. 10, 1975, pp. 481–507.

25. L. PETZOLD, "Differential/algebraic equations are not ODEs," *SIAM J. Sci. Statist. Comput.*, v. 3, no. 3, 1982, pp. 367–384.

26. L. R. PETZOLD & P. LÖTSTEDT, "Numerical solution of nonlinear differential equations with algebraic constraints II: Practical implications," *SIAM J. Sci. Statist. Comput.* (To appear.)

27. R. F. SINCOVEC, B. DEMBART, M. A. EPTON, A. M. ERISMAN, J. W. MANKE & E. L.YIP, *Solvability of Large-Scale Descriptor Systems*, Report, Boeing Computer Services Company, Seattle, Washington, 1979.

28. J. WITTENBURG, *Dynamics of Systems of Rigid Bodies*, Teubner, Stuttgart, 1977.