APPROXIMATION METHODS FOR THE CONSISTENT INTIALIZATION OF DIFFERENTIAL-ALGEBRAIC EQUATIONS*

B. LEIMKUHLER†, L. R. PETZOLD,‡ AND C. W. GEAR§

Abstract. The algebraic constraints in a system of differential-algebraic equations (DAEs) impose a consistency requirement on the initial values that can be difficult to satisfy. In this paper the consistency requirement is characterized by a system of equations. An approximation method is introduced for these equations, and the numerical solution of the resulting system is analyzed for certain important classes of DAEs. Finally, a numerical experiment is described.

Key words. differential-algebraic equations, consistant initial values, numerical methods

AMS(MOS) subject classification. 65L05

1. Introduction.

In this paper, we present a computational method for the initialization of the differential-algebraic equation (DAE)

$$F(x, x', t) = 0, \tag{1}$$

where $x : [t_0, T] \to \mathbf{R}^n$ and $F : \mathbf{R}^{2n+1} \to \mathbf{R}^n$ is C'.

When $|\partial F/\partial x'| = 0$, there is an underlying algebraic subsystem of (1) that imposes a consistency requirement on the solution. Thus there is not a solution through every given initial value. When an ordinary differential equation (ODE) method such as BDF is used to solve a DAE, small inconsistencies in the initial values may cause the method to fail or to become extremely inefficient.

Consistent initial values can sometimes be determined from physical considerations, but more often their calculation requires a greater understanding of the underlying relationships between variables than is needed for the actual problem statement. The very nature of a DAE implies that certain algebraic constraints are present which restrict the choice of initial values. Numerical methods for DAEs place a difficult and unnecessary burden on the user by requiring initial values for all components. Computing consistent derivatives of the solution (which are typically needed by numerical ODE methods) adds yet another level of complication, since, for a DAE, the initial derivatives cannot be determined by a simple function evaluation (as is the case for an explicit ODE) but may only be found through one or more differentiations of the constraint equations and the solution of a nonlinear system of equations.

The subject of this paper is the *consistent initialization problem:* given specified information about the initial state of the problem that is sufficient to specify a unique solution to a DAE, determine the complete initial vector $(x(t_0), x'(t_0))$ corresponding to this unique solution.[1]

We begin in §2 by discussing elementary concepts that are useful for the analysis of the consistency requirement and some prior work on the consistent initialization

* Received by the editors December 21, 1987; accepted for publication (in revised form) November 23, 1989.

† Department of Mathematics, University of Kansas, Lawrence, KS 66045. The work of this author was supported in part by the U.S. Department of Energy grant DEFG02-87-ER25026.

‡ Computing and Mathematics Research Division, Lawrence Livermore National Laboratory, L-316, Livermore, California 94550. The work of this author was supported in part by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U.S. Department of Energy, Lawrence Livermore National Laboratory contract W-7405-Eng-48.

§ Department of Computer Science, University of Illinois, Urbana, Illinois 61801.

[1] We will use the terms "initial state" and "initial vector" interchangeably to refer to the entire vector $(x(t_0), x'(t_0))$.

problem, before describing our approach. We develop a system of equations involving derivatives of (1) that an initial value $x(t_0)$ and derivative $x'(t_0)$ must satisfy in order that they lie on a smooth solution of the DAE. Since the derivatives of the DAE are often difficult to obtain analytically, in §3 we construct a family of one-sided finite difference approximations, and show how these differences can be used to form an accurate approximation of the original equations. The solution of the resulting approximate equations is considered in §4, and convergence results demonstrating the applicability of the method to some important classes of problems are given. In particular, it is shown that reductions of the consistency equations are often possible, leading to well-behaved nonlinear systems. (Even when such a regularization is not easily performed, the approximate consistency equations may still be numerically tractable.) Finally, in §5 we describe an experiment in which an implementation, **DAIS**, of the initialization procedure is applied to a practical problem.

2. Fundamentals and problem statement. In this section we define a few basic concepts such as "solvability" and "index" and then use them to make a clearer statement of the consistent initialization problem. We first describe some important classes of DAEs which will be referred to frequently.

2.1. Common classes of DAEs. In a fundamental book by Campbell [4], and in papers by various authors, the following structure is considered:

$$\begin{aligned} Ex' + Ax &= f, \\ x(t_0) &= x_0, \end{aligned} \tag{2}$$

where E and A are $n \times n$, x_0 is an n-vector, and $f : [t_0, T] \to \mathbf{R}^n$. We call this the *linear, constant-coefficient problem.*

The coefficient matrices E and A in (2) may also be allowed to vary with time, yielding a class of problems with more complex properties due to time-dependent coupling between solution components.

In considering nonlinear DAEs, it is frequently assumed that the problem is in the *semi-explicit* form:

$$\begin{aligned} u' &= f(u, v, t), \\ 0 &= g(u, v, t), \end{aligned} \tag{3}$$

where the vectors u and v are in $\mathbf{R}^{m_1}$ and $\mathbf{R}^{m_2}$, respectively, and f and g are functions defined on the domain $\mathbf{R}^{m_1} \times \mathbf{R}^{m_2} \times \mathbf{R}$ which map into, respectively, $\mathbf{R}^{m_1}$ and $\mathbf{R}^{m_2}$.

When the problem can be written

$$u_2' = f_1(u_1, u_2, u_3, \cdots, u_k, t), \tag{4}$$

$$u_3' = f_2(u_2, u_3, \cdots, u_k, t), \tag{5}$$

$$\vdots$$

$$u_k' = f_{k-1}(u_{k-1}, u_k, t), \tag{6}$$

$$0 = f_k(u_k, t) \tag{7}$$

for some partition $(u_1, \cdots, u_k)$ of x in (1), we have the so-called *Hessenberg form* of order k. Note that a DAE in the semi-explicit form (3) is also in Hessenberg form (of order two) provided $\partial g/\partial v = 0$, and that any DAE which is in the $k+1$st order Hessenberg form can be expressed in the Hessenberg form of order k, for $k \geq 2$. Many writers refer to the Hessenberg form as the *triangular* form, especially when referring to the common cases $k = 2$ and $k = 3$.

2.2. Solvability and index. The structure of the general DAE (1) is usually identified in the literature with a quantity called its *index*. Many problems arising in practice are in the Hessenberg form, for which the index is apparent [3]. In this paper we assume that the index is known a priori. In this subsection we define the notions of solvability and index so as to restrict the problem classes we study and guide the construction of our initialization method.

DEFINITION 2.1 (Solvability). The DAE (1) is *solvable* on a subset Ω of $\mathbf{R}^{n+1}$ if, for some $r \geq 1$, there is an r-parameter family of solutions $x(t, c)$ (defined for $t \in \mathbf{R}, c \in \mathbf{R}^r$) such that

1. $(t, x(t, c)) \in \Omega$,
2. If $\hat{x}(t)$ is any solution whose graph lies in Ω, then $\hat{x}(t) = x(t, \hat{c})$ for some $\hat{c}$,
3. The graph of $x(t, c)$ is an $(r + 1)$-dimensional manifold.

We generally take $\Omega = [t_0, T] \times \mathbf{R}^n$. This definition ensures that solvable problems have solutions, that the solutions are uniquely defined by their value at $t = t_0$, and that there is no bifurcation of solutions [7].

We use an alternative, more restrictive definition of solvability in this paper. We first define the pseudoderivative to simplify the notation and allow us to treat the derivatives of the solution as independent variables. If in Definition 2.2 below, ξ_i is identified with the ith derivative of $x(t)$, then the pseudoderivative is just the usual total derivative. We introduce the notion of the pseudoderivative to stress the fact that we restrict our attention to the single point $t = t_0$, and are interested only in how the various derivatives of the solution at this point are related by the DAE.

DEFINITION 2.2 (Pseudoderivative). Given a function $G : \mathbf{R}^{(p+1)n+1} \to \mathbf{R}^n$, $G = G(\xi_0, \xi_1, \cdots, \xi_p, t)$, the *pseudoderivative*, DG of G is a new function on $\mathbf{R}^{(p+2)n+1}$ satisfying

$$DG(\xi_0, \xi_1, \cdots, \xi_p, \xi_{p+1}, t) = \frac{\partial G}{\partial \xi_0}\xi_1 + \frac{\partial G}{\partial \xi_1}\xi_2 + \cdots + \frac{\partial G}{\partial \xi_p}\xi_{p+1} + \frac{\partial G}{\partial t},$$

where the partials are evaluated at $(\xi_0, \xi_1, \cdots, \xi_p, t)$.

The powers of the operator D are defined in the usual manner: $D^m G = D(D^{m-1}G)$. (If $D^{m-1}G$ is a function on $\mathbf{R}^{sn+1}$, then $D^m G$ will be a function defined on $\mathbf{R}^{(s+1)n+1}$.)

DEFINITION 2.3 (Smooth solvability). The DAE (1) is *smoothly solvable* on $[t_0, T]$ if there exists a nonnegative integer m such that F has m continuous derivatives and the nonlinear system of "derivative equations"

$$\begin{aligned} F(\xi_0, \xi_1, t) &= 0, \\ DF(\xi_0, \xi_1, \xi_2, t) &= 0, \\ &\vdots \\ D^m F(\xi_0, \xi_1, \xi_2, \cdots, \xi_{m+1}, t) &= 0 \end{aligned}$$

when viewed as relating the independent symbols $\xi_0, \xi_1, \cdots, \xi_{m+1}, t$, can be solved on $[t_0, T]$ for ξ_1 in terms of ξ_0 and t: $\xi_1 = \phi(\xi_0, t)$, where the mapping ϕ so defined is Lipschitz continuous.

An alternative formulation that assumes constant rank for the system of derivative equations can be found in Campbell [6]. Problems that are smoothly solvable have smooth solutions that are solutions of the Lipschitz continuous ODE, $x' = \phi(x, t)$. Since such solutions are smooth and well behaved, smoothly solvable DAEs are also solvable.

DEFINITION 2.4 (Index). The *global index* of a smoothly solvable DAE is the smallest nonnegative integer m such that the DAE and the first m derivative equations given in Definition 2.3 can be solved for $\xi_1 = \phi(\xi_0, t)$.

Note that the ξ_i, $i = 2, \cdots, m+1$, are not uniquely determined by the system of derivatives. A definition in the same spirit was originally given in Rheinboldt [21].

The equation $x' = \phi(x,t)$ is defined at every point in $[t_0, T]$, and so forms an ODE (which we refer to as the *underlying* ODE). In our case, we will assume that this equation (and hence (1)) has a very smooth (C^{m+1}) solution in a neighborhood of t_0. This is a strong requirement—one which may not be satisfied in practice.

Definition 2.4 suggests that by differentiating a DAE one or more times we could construct an ODE with the same solution, to which we could presumably apply an ODE method. However, this is not usually a practical approach because the differentiation may be expensive, it is necessary to determine constants of integration accurately, and the numerical solution can drift from the constraints and out of the physically meaningful solution space. The reader is referred to Gear [8] for a discussion. On the other hand, a numerical method based on differentiation of the problem may be feasible, as we shall see below, for determining a consistent initial condition for the problem since accumulation of error is not a concern.

As a simple example, a DAE in the semi-explicit form (3) has index one if $\partial g/\partial v$ is nonsingular. In, among other places, [9] it is shown that the index of a DAE in the semi-explicit triangular form ((3) with $\partial g/\partial v = 0$) is two provided $(\partial g/\partial u)(\partial f/\partial v)$ is nonsingular, and that the third-order Hessenberg form ((4)–(7) with $k = 3$) has index three provided $(\partial f_3/\partial u_3)(\partial f_2/\partial u_2)(\partial f_1/\partial u_1)$ is nonsingular.

2.3. The consistent initialization problem.

We are now prepared to begin in detail our study of the initialization problem. We present some examples and survey the literature before laying out a system of consistency equations which will form the basis of our method. Later sections will be concerned with obtaining approximations to the consistency equations and with the numerical solution of those equations.

2.3.1. Incompletely specified initial conditions.

In practice, it is often difficult to determine the complete initial vector (x_0, x_0'). Certain variables may have readily apparent physical significance, leading to obvious initial values. Others may have an obscure dependence on the initial state. The following examples illustrate this fact.

Example 2.5. This DAE has index one:

$$\begin{aligned} y_1' + y_2' + y_1 &= g_1(t), \\ y_2 &= g_2(t). \end{aligned}$$

If either $y_1(t_0)$ or $y_1'(t_0)$ is known, then a unique solution to the DAE can be found, but in either case it is necessary to differentiate the constraint in order to solve for the unknown components of the initial vector.

Example 2.6. Consider a plane pendulum consisting of a unit point mass swinging on a massless inelastic rod of length one, anchored at the origin. We can write the equations as a DAE:

$$\begin{aligned} x'' &= -\lambda x, \\ y'' &= -\lambda y - g, \\ x^2 + y^2 &= 1. \end{aligned}$$

Here g is the gravitational constant, (x, y) represents the position of the pendulum mass, and λ is a Lagrange multiplier. This problem has index three. The initial position can be any point on the unit circle, but the initial velocity must be consistent with the differentiated constraint. The Lagrange multiplier (representing the tension in the rod) must be derived from physical considerations or by another differentiation of the constraint. To determine the derivatives of the Lagrange multipliers requires yet a third differentiation of the constraint. In practice, most numerical methods do not require the initial values for λ. On the other hand, if the problem is given in transformed variables, or in partially determined nonphysical generalized coordinates, initial values for *all* components may need to be determined.

Other examples may be found in [15].

In these examples of the consistent initialization problem, the striking feature is that in order to apply any of the standard methods, we must have information about the state of the system which is hard to find, even though that information is contained in the system itself.

2.3.2. The effect of poor initial values and derivatives. When consistent initial values are difficult to determine exactly, the problem-solver may resort to crude and even nonphysical assumptions in the hope that any gross initial errors will be rapidly damped as the solution settles into a steady state. (See [12] for a particularly difficult case.)

Present numerical software for solving DAEs often fails or becomes extremely inefficient when the initial values are inconsistent with the DAE. The reasons for this behavior were first discussed in Petzold [20], where a modification to the usual error estimate can also be found . This filters and scales the usual predictor-corrector difference according to the structure of the DAE and effectively eliminates the difficulty for some problems. (See also [14].) However, even when such a filtered estimate is used, poor initial values and approximations for the derivatives of the solution variables at the initial point can lead to gross inefficiency. Let us consider what happens when the initial derivative approximations are very inaccurate (often zero values are used).

In most modern ODE and DAE solvers, the number of reevaluations and refactorizations of the Jacobian is kept to a minimum because these operations are so costly; instead, many steps will use the same approximate Jacobian. If the initial derivatives are not known, the initial predictor will be very poor, and an exceptionally small stepsize will be necessary to make the predictor-corrector difference small. Poor appoximations to the initial derivatives may lead to a sequence of costly failures as the step is reduced (each stepsize change forces a reevaluation and refactorization of the Jacobian) and then, subsequently, to even more Jacobian evaluations and factorizations as the stepsize is increased to the natural level over the first few integration steps.

In fact, if the initial values are sufficiently inaccurate, or the problem highly nonlinear, the Newton iteration at the first step either may not converge, may converge very slowly, or may converge to an undesired solution. (Just such a "bad" problem is discussed in Brenan [2, p. 285].)

See [11] for a detailed discussion of the effects of inconsistent initial conditions in the context of Runge–Kutta methods.

2.3.3. Previous work on the initialization problem. For determining the initial derivatives when the initial solution values are known, Petzold [19], in her code **DASSL**, takes a small backward-Euler step and relies on a damped Newton iteration to determine the correct values. The method used in the code **SPRINT** [1] can be viewed as a precursor of the general scheme we develop here.

Campbell [5] suggests an approach based on Taylor's series expansions for determining consistent initial conditions for the linear time-varying problem

$$A(t)x'(t) + B(t)x(t) = f(t).$$

This method is limited because it requires a knowledge of the derivatives of the matrices. Campbell [6] has recently extended his method to nonlinear problems and developed a general time-stepping method based on those ideas, but it is unlikely that the method will be competitive for complex scientific problems in which the partial Jacobian matrices are unknown or difficult to compute. There are many parallels between Campbell's work and our own, but the emphasis is quite different.

Pantelides [18] uses a graph-theoretic algorithm to determine the minimal set of equations to be differentiated in order to solve for consistent initial values. The differentiations are then to be carried out exactly, as in Campbell's algorithm.

Mrziglod [17] bases his algorithm for determining consistent initial values on a thorough decomposition of the system structure. Unfortunately, this approach depends on the isolation of algebraic constraints by index reduction and requires knowledge of analytical expressions for various derivatives of the problem.

Some other, related techniques are described in [15].

2.3.4. The consistency equations. In this paper, we adopt the following formal definition.

DEFINITION 2.7 (Consistent initial vector). A *consistent initial vector* (x_0, x'_0) for the DAE (1) at $t = t_0$ is such that the DAE has a C^{m+1} solution through that vector.

Throughout the following, we assume that (1) has index m. In general, the user may possess information about the initial state in the form of some algebraic equations:

$$B(x_0, x'_0) = 0. \tag{8}$$

In addition to these user-controlled specifications, any valid initial point must satisfy the problem itself:

$$F(x_0, x'_0, t_0) = 0. \tag{9}$$

Also, depending on the index, (x_0, x'_0) must satisfy the m derivative equations (here stated in terms of the pseudodifferentiation operator defined in the last section):

$$\begin{aligned} DF(x_0, x'_0, x''_0, t_0) &= \frac{\partial F}{\partial x}x'_0 + \frac{\partial F}{\partial x'}x''_0 + \frac{\partial F}{\partial t} = 0, && (10)\\ &\vdots \\ D^m(x_0, x'_0, x''_0, \cdots, x_0^{(m+1)}, t_0) &= 0 && (11) \end{aligned}$$

for some choice of $(x''_0, x'''_0, \cdots, x_0^{(m+1)})$. Here all the partial derivatives of F are evaluated at (x_0, x'_0) and $t = t_0$.

We call the $m + 2$ equations (8)–(11) the *consistency equations.* Theorem 2.8 states a condition under which the consistency equations uniquely determine x_0 and x'_0.

THEOREM 2.8. *Suppose the problem*

$$(12)\qquad \begin{aligned} F(x(t),x'(t),t) &= 0, \\ B(x(t_0),x'(t_0)) &= 0, \end{aligned}$$

has precisely one continuous solution $\bar{x}(t)$ *and it is smooth:* $\bar{x}(t) \in C^{m+1}[t_0,T]$. *Then the consistency equations* (8)–(11) *posess a solution* $(x_0, x_0', \cdots, x_0^{(m+1)})$, *and the first two components are uniquely determined:* $x_0 = \bar{x}(t_0), x_0' = \bar{x}'(t_0)$.

Proof (existence). Differentiate $\omega(t) = F(\bar{x}(t), \bar{x}'(t), t) = 0$ m times. Since

$$\omega(t_0) = 0; \omega'(t_0) = 0; \cdots ; \omega^{(m)}(t_0) = 0$$

coupled with (8) are precisely the consistency equations satisfied at $(\bar{x}(t_0), \cdots, \bar{x}^{(m+1)}(t_0))$, this vector is a solution.

(Uniqueness). Suppose $(\tilde{x}_0, \tilde{x}_0', \cdots, \tilde{x}_0^{(m+1)})$ is a solution to the consistency equations. Let $x' = \phi(x,t)$ be the underlying ODE from the definition of index (Definition 2.4), and consider the solution $\tilde{x}(t)$ to the IVP $x' = \phi(x,t)$, $t \in [t_0,T]$; $x(t_0) = \tilde{x}_0$. Obviously $\tilde{x}(t)$ is a solution to the DAE (1) which satisfies the initial condition (8). However there is only one such solution; we must have $\tilde{x}(t) = \bar{x}(t)$. □

We now have a framework for initializing DAEs, or at least smooth ones with a known index. Assuming that enough information is known about the initial state, we can, in principle, differentiate the constraints a few times and manipulate the resulting equations to find the initial values. However, this strategy is inadequate as it stands because the system of consistency equations, as described, requires the formation of pseudoderivatives of the problem which are expensive to compute, and because the equations (8)–(11) are hard to solve in general. In the next section we demonstrate that the pseudoderivatives can be approximated by relatively easy-to-compute finite differences. Some techniques for the efficient solution of the approximated equations are given in the following section.

3. Computable approximations to the consistency equations. In this section, we show how the pseudoderivatives can be approximated by finite differences.

3.1. First-order derivatives. To motivate the general approach, we consider a particular approximation to the first derivative in the index-1 case. We simplify the notation by defining $z(t) = (x(t), x'(t), t)$, $z_0 = (x_0, x_0', t_0)$, and $z_0' = (x_0', x_0'', 1)$. We seek an approximation to the derivative

$$\frac{dF(z(t))}{dt}\ |_{t=t_0} = J_F(z(t_0))z'(t_0),$$

where J_F is the Jacobian. An obvious choice is the simple one-sided difference

$$\frac{F(z_0 + hz_0') - F(z_0)}{h}.$$

When $z_0 = z(t_0)$ and $z_0' = z'(t_0)$, $z_0 + hz_0'$ is a truncated Taylor's series expansion for $z(t_0 + h)$ and we assume F is C^2, so the truncation error in the derivative approximation is $O(h)$.

Berzins, Dew, and Furzeland [1] use this approach in **SPRINT**.

We would like to determine a family of higher-order methods for approximating these and higher pseudoderivatives of the DAE. There are several points worth noting:

(1) because F is only defined for $t \geq t_0$, we are generally limited to one-sided differences; (2) the displacement h need not be related to the eventual starting stepsize of the integration procedure; (3) numerical differentiation is ill-conditioned, so that the roundoff errors incurred in the computation of F will be magnified by $O(\frac{1}{h})$ or $O(\frac{1}{h^k})$ when approximating the kth derivative; see below.

The idea is this: some points $z_1, \cdots, z_s$ are chosen as truncated Taylor's series expansions of $z(t_0 + c_1 h), \cdots, z(t_0 + c_s h)$. Then values of F at those points are combined with coefficients $\alpha_1, \cdots, \alpha_s$ chosen to annihilate certain of the high order terms in the Taylor expansion of the approximation. This idea leads us to the class of approximation methods described by

$$D_h^k F = \frac{k!}{h^k}\left[\sum_{i=1}^{s} \alpha_i F\left(z_0 + c_i h z_0' + \frac{(c_i h)^2}{2} z_0'' + \cdots + \frac{(c_i h)^r}{r!} z_0^{(r)}\right) - \left(\sum_{i=1}^{s} \alpha_i\right) F(z_0)\right] \tag{13}$$

for $r \geq k$. (Allowing $r > k$ allows for the efficient approximation of several derivatives simultaneously; see [15].)

Theorem 3.1 states conditions under which these methods approximate the derivatives, and Theorem 3.3 establishes the attainable order of the approximations.

THEOREM 3.1. *Let $F : \mathbf{R}^{n_1} \to \mathbf{R}^{n_2}$, $z : \mathbf{R} \to \mathbf{R}^{n_1}$. Suppose $r \geq k$. If z is $r+1$ times differentiable in a neighborhood of $t = t_0$ and F is C^{p+1} in a neighborhood of $z(t_0)$ for some $p \geq r$, then the truncation error of* (13), *as an approximation to the kth derivative, satisfies*

$$L_h^k F \equiv D_h^k F - \frac{d^k F(z(t))}{dt^k} \Big|_{t=t_0} = O(h^l)$$

provided that $p \geq k + l - 1$, and

$$\sum_{i=1}^{s} \alpha_i c_i^j = \begin{cases} 1, j = k, \\ 0, j \neq k, \end{cases} \quad j = 1, \cdots, k + l - 1.$$

Proof. Let $\omega_*(\tau) = F(z(t_0 + \tau))$; clearly $\omega_*^{(k)}(0) = (d^k F(z(t))/dt^k) \big|_{t=t_0}$. Substituting the true solution $z(t)$ in (13) we have

$$D_h^k F = \frac{k!}{h^k} \sum_{i=1}^{s} \alpha_i \left[F(\bar{z}(c_i h)) - F(\bar{z}(0))\right],$$

where

$$\bar{z}(c_i h) = z(t_0) + c_i h z'(t_0) + \cdots + \frac{(c_i h)^r}{r!} z^{(r)}(t_0).$$

Define $\omega(c_i h) = F(\bar{z}(c_i h))$. $\bar{z}$ is a polynomial in $c_i h$, so it is C^∞ and the Taylor expansion of ω is only limited by the continuity of F:

$$\omega(c_i h) = \omega(0) + c_i h \omega'(0) + \cdots + \frac{(c_i h)^p}{p!} \omega^{(p)}(0) + O(h^{p+1}).$$

Note that $\omega(0) = F(z(t_0))$. We now have an expression for the truncation error in powers of h:

$$L_h^k F = D_h^k F - \frac{d^k F(z(t))}{dt^k} \Big|_{t=t_0}$$

$$
\begin{aligned}
&= \frac{k!}{h^k}\sum_{i=1}^{s}\alpha_i(\omega(c_ih)-\omega(0)) - \omega_*^{(k)}(0) \\
(14)\qquad &= \frac{k!}{h^k}\sum_{i=1}^{s}\alpha_i\left[\sum_{\substack{j=1\\ j\neq k}}^{p}\frac{(c_ih)^j}{j!}\omega^{(j)}(0)+\frac{(c_ih)^k}{k!}\omega^{(k)}(0)\right] - \omega_*^{(k)}(0)+O(h^{p+1-k}) \\
&= \sum_{\substack{j=1\\ j\neq k}}^{p} h^{j-k}\left(\sum_{i=1}^{s}\alpha_i c_i^j\right)\frac{k!}{j!}\omega^{(j)}(0)+\left(\sum_{i=1}^{s}\alpha_i c_i^k\right)\omega^{(k)}(0)-\omega_*^{(k)}(0)+O(h^{p+1-k}).
\end{aligned}
$$

If $\omega^{(k)}(0)=\omega_*^{(k)}(0)$, then this last expression yields the stated order conditions,

$$
\begin{aligned}
\omega^{(k)}(0)-\omega_*^{(k)}(0) &= \frac{d^k}{d\tau^k}[\omega(\tau)-\omega_*(\tau)]\ \Big|_{\tau=0} \\
&= \frac{d^k}{d\tau^k}[F(\bar{z}(\tau))-F(z(t_0+\tau))]\ \Big|_{\tau=0} \\
&= \frac{d^k}{d\tau^k}[F(\bar{z}(\tau))-F(\bar{z}(\tau)+R(\tau))]\ \Big|_{\tau=0},
\end{aligned}
$$

where $R(\tau)$ is an $O(\tau^{r+1})$ remainder term. We see that this is the multiplier of $\tau^k/k!$ in the Maclaurin series expansion of

$$
\phi(\tau)=F(\bar{z}(\tau))-F(\bar{z}(\tau)+R(\tau)).
$$

But F is smooth, so $\phi(\tau)$ is $O(\tau^{r+1})$. Hence the terms in the expansion of ϕ through τ^r must vanish for τ sufficiently small. Since $r\geq k$ we have established the fact that $\omega^{(k)}(0)=\omega_*^{(k)}(0)$.

Replacing $\omega_*^{(k)}(0)$ by $\omega^{(k)}(0)$ in (14) and setting the coefficients of powers of h (through h^{l-1}) to zero, we get the stated order conditions. □

LEMMA 3.2. *Let $c_1, c_2, \cdots, c_s$ be s distinct, positive real numbers. For $\nu = k, k+1, \cdots$, let $\mathbf{e}_k^\nu$ be the kth standard basis vector in $\mathbf{R}^\nu$ and define $\alpha=(\alpha_1,\cdots,\alpha_s)^T$ and*

$$
\mathbf{M}=\begin{bmatrix}
c_1 & c_2 & \cdot\ \cdot & c_s \\
c_1^2 & c_2^2 & \cdot\ \cdot & c_s^2 \\
\cdot & \cdot & & \cdot \\
\cdot & \cdot & & \cdot \\
c_1^\nu & c_2^\nu & \cdot\ \cdot & c_s^\nu
\end{bmatrix}.
$$

Then the system

$$
\mathbf{M}\alpha=\mathbf{e}_k^\nu
$$

has a solution if and only if $\nu\leq s$.

Proof. Let $\mathbf{V}_s$ be the Vandermonde matrix of order s, $\mathbf{C}=\mathrm{diag}(c_1,\cdots,c_s)$.

If $\nu=s$ then $\mathbf{M}$ is the column-scaled Vandermonde matrix $\mathbf{V}_s\mathbf{C}$, so it is nonsingular for positive, distinct $c_1,\cdots,c_s$. Also, the first $j\leq s$ rows are independent and span $\mathbf{R}^j$, thus there is a solution provided $\nu\leq s$.

Now let $\nu>s$ and suppose (by way of contradiction) that the system has a solution. If $k>s$, then we must have at least that $\mathbf{V}_s\mathbf{C}\alpha=0$ (the first s equations), but then $\alpha = 0$, so we cannot satisfy the kth equation: $\sum_{i=1}^{s} \alpha_i c_i^k = 1$. Hence the only possibility is that $1 \le k \le s$. We must (at least) satisfy the first $s+1$ equations. Let $\tilde{\alpha} = \mathbf{C}\alpha$, then we must have

$$\mathbf{V}_s \tilde{\alpha} = \mathbf{e}_k^s$$

and

$$\sum_{i=1}^{s} \tilde{\alpha}_i c_i^s = 0.$$

Let $\mathbf{c}^q = (c_1^q, \cdots, c_s^q)$. The above equations imply that the matrix L consisting of the first $k-1$ and the last $s-k$ rows of $\mathbf{V}_s$ together with the row $\mathbf{c}^s$ have a nontrivial null space. If $k=1$, L is precisely the scaled Vandermonde matrix, M, which we know is nonsingular.

This leaves only the possibility $2 \le k \le s$. L is singular (and hence has a nontrivial null space) only if $\mathbf{c}^0, \mathbf{c}^1, \cdots, \mathbf{c}^{k-2}, \mathbf{c}^k, \cdots, \mathbf{c}^s$ are linearly dependent; in this case, there exist s coefficients γ_i, $i = 0, \cdots, k-2, k, \cdots, s$ not all zero so that each c_i is a root of the polynomial

$$P(x) = \sum_{\substack{i=0 \\ i \ne k-1}}^{s} \gamma_i x^i.$$

Note that this means, in particular, that P has degree s. Applying Rolle's theorem on each of the $s-1$ subintervals bounded by zeros of P, all of the zeros of $P'(x)$ are distinct and lie on the positive real axis (in between the zeros of P), and so for $P''(x)$, etc. The important point is that all the zeros of any of the first $s-1$ derivatives of P are positive. However, we can easily compute the $(k-1)$st derivative:

$$P^{(k-1)}(x) = \sum_{i=1}^{s-k+1} \gamma_{i+k-1}(i+k-1)(i+k-2)\cdots(i+1)x^i,$$

which has a root at zero. This is the contradiction we have been seeking. □

THEOREM 3.3. *Let z, F be as in Theorem* 3.1, *then the maximum attainable order of the kth derivative approximation* (13) *is* $\min(p-k+1, s-k+1)$.

Proof. The proof follows directly from Lemma 3.2 since the coefficients can only be determined if $\nu = k+l-1$ satisfies $\nu \le s$ and $\nu \le p$. □

4. The numerical solution of the consistency equations.

In this section we consider what happens when we introduce approximations for the derivatives in the system of consistency equations. The problem is that of understanding the convergence behavior of a sequence of approximate minimizers to approximate problems.

In §2 we showed how the consistency requirements on the initial values and solution derivatives lead to a system of nonlinear equations (called the consistency equations) involving derivatives of the problem. In §3, we constructed approximations to these equations based on finite differences. To prevent confusion, we refer to the equations developed in §2 as the "exact" consistency system and the equations resulting from the substitution of difference approximations as the "approximate" system.

Here we first show how certain classes of problems can be greatly simplified by reductions that yield problems with full column rank Jacobians. (Such reductions apply to both the exact and approximate consistency equations.) Next we consider other classes of DAEs which lead to more difficult, rank-deficient problems, and show that the special structure of the consistency equations allows the possibility of solving these problems.

The classes of problems to which we restrict our attention here are precisely those classes of DAEs whose solution by ODE methods has been considered in the literature. (See, for example, the recent treatise [11] on Runge–Kutta methods for DAEs.)

4.1. The approximate consistency equations. If we substitute derivative approximations into the consistency equations, the resulting system will probably not have a solution. The difficulty occurs because, in general, the human problem solver does not know which of the variables are determined by the constraints (or the constraint derivatives) and which are arbitrary. If information redundant with the analytical DAE or its derivatives is specified, the approximation of the problem can lead to an inconsistency. (See [16, p. 23] for an example.)

For the class of problems linear in x' that they consider, Berzins, Dew, and Furzeland [1] avoid many difficulties by taking only a single step of Newton for the one-stage, forward difference approximation. This is cheap, but probably not accurate. There is no guarantee that, in solving the linearized derivative equation, the approximation will not be forced out of the manifold described by the DAE and the user's known initial conditions.

To achieve better accuracy, the natural approach is to solve the consistency system in the least-squares sense (i.e., to minimize the squared 2-norm of the residual). The application of this method is complicated by the known rank deficiency of the Jacobian. In certain cases, however, the rank-deficient part of the consistency equations can be removed by taking advantage of the particular structure of those equations.

In the next subsection, we consider reductions to full column rank for a variety of common problems. We demonstrate in § 4.3 that solutions of approximations to the resulting regular least-squares problems are convergent under mild conditions. We then turn our attention to the approximate consistency equations arising from linear, constant-coefficient problems (§ 4.4) and index-1 DAEs which are linear in x' (§ 4.5). Finally in § 4.6 we summarize our method in an algorithm for the initialization of DAEs.

4.2. Reductions to full column rank. What we mean by a "reduction to full column rank" is a manipulation of the exact consistency equations that applies as well to the approximate equations and results in systems of nonlinear equations for x and x' having full column rank Jacobians. Here we consider the manipulations principally as they relate to the exact consistency equations; in § 4.3 we show that for the difference approximations we are using, the Jacobians of the reduced approximate consistency equations will typically also have full column rank.

Consider an index-1 linear, constant-coefficient problem

$$Ex' + Ax = f,$$

where sufficient user conditions of the form $Px + Qx' = \xi$ are prescribed to specify a unique smooth solution to the DAE. The consistency equations take the form

$$\begin{bmatrix} A & E & 0 \\ 0 & A & E \\ P & Q & 0 \end{bmatrix} \begin{bmatrix} x \\ x' \\ x'' \end{bmatrix} = \begin{bmatrix} f(t_0) \\ f'(t_0) \\ \xi \end{bmatrix}. \tag{15}$$

Now suppose the matrix E has a known column basis given by, say, the first r columns ($E = (E_1, E_2)$, where E_1 is $n \times r$ and $\mathcal{R}(E_1) = \mathcal{R}(E)$). Then we can solve the reduced consistency system given by:

$$\begin{bmatrix} A & E & 0 \\ 0 & A & E_1 \\ P & Q & 0 \end{bmatrix} \begin{bmatrix} x \\ x' \\ x_1'' \end{bmatrix} = \begin{bmatrix} f(t_0) \\ f'(t_0) \\ \xi \end{bmatrix},$$

where x_1'' is an r-vector. By using some properties of linear consistency systems that are developed below, it is a simple matter to show that the matrix multiplying (x, x', x_1'') in the reduced equations has full rank and that the reduced system has a unique solution which corresponds to a solution of the (unreduced) consistency system with $x'' = (x_1'', 0)$. The approximate consistency equations corresponding to (15) are unchanged except that f' is approximated by a difference approximation $\hat{D}_h f$. This idea is explored below for problems that are linear in x' in order to demonstrate the existence of a convergent solution sequence to those problems.

Another reduction can be performed on DAEs in the Hessenberg form (4)–(7). The consistency equations for an index-2 problem in the triangular form ((3) with $g = g(u,t)$) are given by:

$$\begin{aligned} u_0' &= f(u_0, v_0), \\ 0 &= g(u_0), \\ u_0'' &= \frac{\partial f}{\partial u} u_0' + \frac{\partial f}{\partial v} v_0', \\ 0 &= \frac{\partial g}{\partial u} u_0', \\ u_0''' &= \psi(u_0, v_0, u_0', v_0', u_0'', v_0''), \\ 0 &= \frac{\partial^2 g}{\partial u^2}[u_0', u_0'] + \frac{\partial g}{\partial u} u_0'', \end{aligned}$$

where ψ is some function involving first and second derivatives of f, and all partials are evaluated at $u = u_0$ and $v = v_0$.

This system can easily be reduced, since we are only interested in determining (u_0, v_0, u_0', v_0'):

$$\begin{aligned} &(16) & u_0' &= f(u_0, v_0), \\ &(17) & 0 &= g(u_0), \\ &(18) & 0 &= \frac{\partial g}{\partial u} f(u_0, v_0), \\ &(19) & 0 &= \frac{\partial^2 g}{\partial u^2}[f(u_0, v_0), f(u_0, v_0)] + \frac{\partial g}{\partial u}\left[\frac{\partial f}{\partial u} f(u_0, v_0) + \frac{\partial f}{\partial v} v_0'\right]. \end{aligned}$$

The reduction is performed using only knowledge of the separation according to the triangular form. A similar reduction can be applied to the approximate consistency equations; see [15].

Details of a reduction along the same lines for the index-3 triangular form and an efficient technique for initializing index-2 and index-3 triangular forms in the steady state can be also be found in [15].

4.3. Accuracy of solutions for full column rank problems. We have seen that, in many cases, it is possible to reduce the complexity of the initialization problem significantly by some simple analysis of the equations. Under certain conditions, when a full column rank approximate system, such as that derived by manipulations of the previous section, is solved in the least-squares sense, the error in the solution will be proportional to that present in the equations. The proof, contained in [15], is a straightforward application of the implicit function and local approximation theorems.

THEOREM 4.1. *Let a mapping $G(u,h)$ satisfy $G(u_0,0) = 0$, and suppose that $\|G(u,h) - G(u,0)\| = O(h^p)$ as $h \to 0$, uniformly for u in a neighborhood N of u_0. Suppose G is C'' in u and C' in h and that $\partial G/\partial u$ has full column rank in the same neighborhood (and for h sufficiently small). Then for any fixed h sufficiently small, the local minimum (in u) of the functional $\|G(u,h)\|^2$ is unique and with respect to the parameter h, the local minima so defined form a C' function, $u = u(h)$, and furthermore, $\|u(h) - u_0\| = O(h^p)$ as $h \to 0$.*

Note that an important assumption here is that the Jacobian of the approximate function G has full rank. Since we know that the Jacobian of the exact analogue has full column rank (for reduced problems), we need only establish that the Jacobian of the approximate system converges with h. In Theorem 4.2, the convergence is shown for the autonomous case, which is easily seen to be adequate.

THEOREM 4.2. *For a C' function $g : \mathbf{R}^n \times \mathbf{R}^n \mapsto \mathbf{R}^n$, $g = g(x,y)$, consider the operator*

$$\hat{D}_h^k g = \frac{k!}{h^k}\sum_{i=0}^{s}\alpha_i g\left(\sum_{j=0}^{r}\frac{(hc_i)^j}{j!}x^{(j)}, \sum_{j=0}^{r}\frac{(hc_i)^j}{j!}x^{(j+1)}\right)$$

as an approximation to the kth pseudoderivative of g defined, recursively, by

$$D^k g = \frac{\partial D^{k-1}g}{\partial x}x' + \frac{\partial D^{k-1}g}{\partial x'}x'' + \cdots + \frac{\partial D^{k-1}g}{\partial x^{(k)}}x^{(k+1)},$$

for $k \geq 1$, with

$$D^0 g = g(x,x').$$

We assume here that $r \geq k$ and we identify x and x^0. Thus $\hat{D}_h^k g$ is a function of $(x,x',\cdots,x^{(r+1)})$ and $D^k g$ is a function of $(x,x',\cdots,x^{(k+1)})$. Let $\hat{D}_h^k g - D^k g = O(h^p)$, then, for $l = 0,1,\cdots,k+1$,

$$\frac{\partial \hat{D}_h^k g}{\partial x^{(l)}} - \frac{\partial D^k g}{\partial x^{(l)}} = O(h^p)$$

and if $r > k$, then for $l = k+2,\cdots,r+1$,

$$\frac{\partial \hat{D}_h^k g}{\partial x^{(l)}} = O(h^p).$$

The body of the proof is found in the following two lemmas which describe the exact and approximate Jacobians.

LEMMA 4.3. *For $l = 1,\cdots,k$,*

$$\frac{\partial D^k g}{\partial x^{(l)}} = \binom{k}{l} D^{k-l} g_x + \binom{k}{l-1} D^{k-l+1} g_{x'}.$$

Proof. Using the recursive definition of the pseudoderivative and the equality of mixed partials, we may write (for $l = 1, \cdots, k$):

$$\begin{aligned}
\frac{\partial D^k g}{\partial x^{(l)}} &= \frac{\partial}{\partial x^{(0)}}\frac{\partial D^{k-1}g}{\partial x^{(l)}}x^{(1)} + \cdots + \frac{\partial}{\partial x^{(l-2)}}\frac{\partial D^{k-1}g}{\partial x^{(l)}}x^{(l-1)} \\
&\quad + \frac{\partial D^{k-1}g}{\partial x^{(l-1)}} + \frac{\partial}{\partial x^{(l-1)}}\frac{\partial D^{k-1}g}{\partial x^{(l)}}x^{(l)} \\
&\quad + \frac{\partial}{\partial x^{(l)}}\frac{\partial D^{k-1}g}{\partial x^{(l)}}x^{(l+1)} + \cdots + \frac{\partial}{\partial x^{(k)}}\frac{\partial D^{k-1}g}{\partial x^{(l)}}x^{(k+1)} \\
&= D\frac{\partial D^{k-1}g}{\partial x^{(l)}} + \frac{\partial D^{k-1}g}{\partial x^{(l-1)}}.
\end{aligned}$$

The other two cases are given by

$$\frac{\partial D^k g}{\partial x^{(0)}} = D\frac{\partial D^{k-1}g}{\partial x^{(0)}} = \cdots = D^k g_x$$

and

$$\frac{\partial D^k g}{\partial x^{(k+1)}} = \frac{\partial D^{k-1}g}{\partial x^{(k)}} = \cdots = g_{x'}.$$

The proof of the lemma now follows by a simple induction; see [15] for details. □

Because there is nothing unique about $O(h^p)$ approximations to D^k there is no analogue of Lemma 4.3 for the approximations. On the other hand, that would be more than we really require. Here we establish a result that serves our purposes.

LEMMA 4.4. *For an $O(h^p)$ difference approximation $\hat{D}_h^k$, and for $l = 1, \cdots, k$,*

$$\frac{\partial \hat{D}_h^k g}{\partial x^{(l)}} = \begin{pmatrix} k \\ l \end{pmatrix} D^{k-l} g_x + \begin{pmatrix} k \\ l-1 \end{pmatrix} D^{k-l+1} g_{x'} + O(h^p).$$

Proof. To simplify the notation, we define

$$\bar{z}(hc_i) := \left(\sum_{j=0}^{r} \frac{(hc_i)^j}{j!}x^{(j)}, \sum_{j=0}^{r} \frac{(hc_i)^j}{j!}x^{(j+1)}\right).$$

For $l = 1, \cdots, k$,

$$\begin{aligned}
\frac{\partial \hat{D}_h^k g}{\partial x^{(l)}} &= \frac{k!}{h^k}\sum_{i=0}^{s} \alpha_i \frac{(hc_i)^l}{l!} g_x(\bar{z}(hc_i)) + \frac{k!}{h^k}\sum_{i=0}^{s} \alpha_i \frac{(hc_i)^{l-1}}{(l-1)!} g_{x'}(\bar{z}(hc_i)) \\
&= \begin{pmatrix} k \\ l \end{pmatrix} \frac{(k-l)!}{h^{k-l}} \sum_{i=0}^{s} \alpha_i c_i^l g_x(\bar{z}(hc_i)) \\
&\quad + \begin{pmatrix} k \\ l-1 \end{pmatrix} \frac{(k-l+1)!}{h^{k-l+1}} \sum_{i=0}^{s} \alpha_i c_i^{l-1} g_{x'}(\bar{z}(hc_i)).
\end{aligned}$$

Now, assuming that $\hat{D}_h^k$ satisfies the order conditions of Theorem 3.1 which ensure that it is an $O(h^p)$ approximation to the kth pseudoderivative, we know that

$$\sum_{i=0}^{s} \alpha_i c_i^j = \begin{cases} 1, j = k, \\ 0, j \neq k, \end{cases} \quad j = 0, \cdots, k+p-1.$$

Rewriting these conditions, for $l = 1, \cdots, k$, we can say

$$\sum_{i=0}^{s} (\alpha_i c_i^l) c_i^j = \begin{cases} 1, j = k-l, \\ 0, j \neq k-l, \end{cases} \quad j = 0, \cdots, k+p-1-l.$$

Now we simply note that this ensures that the coefficients $\alpha_i c_i^l$ are those of $O(h^p)$ approximation to D^{k-l}, and a similar result holds for $\alpha_i c_i^{l-1}$. □

The proof of Theorem 4.2 follows trivially by comparison of Lemmas 4.3 and 4.4 for $l = 1, \cdots, k$. The cases $l = 0$ and $l = k+1$ are easy to verify individually, using propositions developed in the proofs of the lemmas.

When $r > k$, an examination of the order conditions along the lines of the above discussion shows that the partial derivatives with respect to $x^{(l)}$, $l = k+2, \cdots, r+1$ are all $O(h^p)$.

4.4. Solving the linear consistency equations. In [15], a detailed analysis of the linear case is given based on Lemma 4.5 below, and standard results on the stability of linear least-squares problems.

Consider the consistency system for a linear DAE:

$$M \begin{pmatrix} u \\ v \end{pmatrix} = b, \tag{20}$$

which we assume has a solution, the u-part of which is unique. (Here u represents the initial value and derivative, $u = \binom{x_0}{x_0'}$, and v consists of the higher derivatives which do not particularly interest us.)

LEMMA 4.5. *Let M be a $q \times (p_1 + p_2)$ real matrix. Partition M as (U, V), where U is $q \times p_1$ and V is $q \times p_2$. Then* (20) *has a solution $(u, v) \in \mathbf{R}^{p_1} \times \mathbf{R}^{p_2}$ and u is uniquely determined if and only if the following three propositions hold:* (i) $b \in \mathcal{R}(M)$, (ii) *U has full column rank, and* (iii) $\mathcal{R}(U) \cap \mathcal{R}(V) = \{0\}$.

Proof. Suppose (20) has a solution $(u^*, v^*)^T$ and that u^* is uniquely determined. Obviously (i) holds. If U does not have full column rank, then U has a nontrivial right null space: $U\hat{u} = 0$ for some $\hat{u} \neq 0$. Hence $(u^* + \hat{u}, v^*)^T$ is another solution to (20). This contradicts the assumed uniqueness of u^*, so (ii) holds. Finally let $z \in \mathcal{R}(U) \cap \mathcal{R}(V)$ with $z = Uz_1 = Vz_2$. Therefore $U(z_1 + u^*) + V(v^* - z_2) = b$, so that $(z_1 + u^*, v^* - z_2)^T$ also solves the system. By uniqueness, we must have $z_1 = 0$ which implies that $z = Uz_1 = 0$, so (iii) is established.

We now assume that all three propositions hold. Clearly (i) implies the existence of solutions. Suppose $(u_0, v_0)^T$ and $(u_1, v_1)^T$ are two solutions, then

$$U(u_0 - u_1) = V(v_1 - v_0).$$

However, the intersection of range spaces is $\{0\}$ by (iii), so, in particular, $U(u_0 - u_1) = 0$. Since U has a trivial null space by (ii), $u_0 = u_1$. □

4.5. Nonlinear consistency equations. In this section we extend our analysis to include the consistency equations for the subset of index-1 DAEs which are linear in x'.

We consider a system of DAEs of the form

$$E(x,t)x' + f(x,t) = 0, \tag{21}$$

where $x, x' \in \mathbf{R}^n$, E is a $n \times n$ matrix-valued function assumed to be C' in its arguments, and $f : \mathbf{R}^n \times \mathbf{R} \rightarrow \mathbf{R}^n$ is also C'. We assume that the user has supplied p consistent initial conditions $B(x, x') = 0$, which are sufficient in the sense of Theorem 2.8 to specify a unique solution to the DAE. We further assume that $E(x, t)$ has constant rank in a neighborhood of the solution.

This is an important class of problems from a practical point of view, including most of the index-1 DAEs arising in engineering and scientific problems. Incidentally, this is the class of problems which the popular codes **SPRINT** [1] and **LSODI** [13] are designed to handle. The assumption that $E(x, t)$ has constant rank is necessary to insure solvability.

Consider an approximation to the consistency function (the left-hand side of the consistency equations):

$$\begin{bmatrix} E(x,t_0)x' + f(x,t_0) \\ \frac{1}{h}\sum_{i=0}^{s}\alpha_i\{E(x+hc_ix',t_0+hc_i)x' + f(x+hc_ix',t_0+hc_i)\} \\ +\left[\sum_{i=0}^{s}\alpha_ic_iE(x+hc_ix',t_0+hc_i)\right]x'' \\ B(x,x') \end{bmatrix}.$$

Note that if the derivative approximation used is $O(h)$, then the term $\sum_{i=0}^{s}\alpha_i c_i \cdot E(x+hc_ix', t_0+hc_i)$ which multiplies x'' in the derivative equation is itself an approximation to $E(x, t_0)$. Since the latter expression is readily computable, and in the exact equations, x'' occurs premultiplied by $E(x, t_0)$, we choose to work instead with the somewhat simpler function

$$M(x,x',x'';h) \\ := \begin{bmatrix} E(x,t_0)x' + f(x,t_0) \\ \frac{1}{h}\sum_{i=0}^{s}\alpha_i\{E(x+hc_ix',t_0+hc_i)x' + f(x+hc_ix',t+hc_i)\} + E(x,t)x'' \\ B(x,x') \end{bmatrix}.$$

Note that $\lim_{h\rightarrow 0} M(x,x',x'';h) = \bar{M}(x,x',x'')$, where $\bar{M}(x,x',x'')$ represents the exact consistency function.

This slight modification facilitates the analysis; indeed it is the assumption that we always work with these slightly modified equations which rules out the possibility of a catastrophic rank-change in the Jacobian of the consistency function. In the sequel, we make the following additional assumption.

Assumption 4.6. The Jacobian $\partial\bar{M}(x,x',x'')/\partial(x,x',x'')$ of the exact consistency equations satisfies the following properties in the neighborhood of the solution: (i) $\partial\bar{M}(x,x',x'')/\partial(x,x')$ has full column rank and (ii) $\mathcal{R}\left(\partial\bar{M}(x,x',x'')/\partial(x,x')\right) \cap \mathcal{R}\left(\partial\bar{M}(x,x',x'')/\partial x''\right) = \{0\}$.

This assumption essentially says that the linearization of the consistency equations has the same structure as a system of consistency equations for a linear DAE; it eliminates the possibility of a degenerate case in which, for example, the partial Jacobian $\partial\bar{M}(x,x',x'')/\partial(x,x')$ does not have full column rank at the initial point, even though the solution is uniquely determined at that point and in a neighborhood.

Let $\bar{z} = (\bar{x}, \bar{x}')$ be the (unique) consistent initial vector for the DAE (21). We are interested in the the local minima (relative to (x, x', x'')) of the functional

$$\phi(x, x', x''; h) := \|M(x, x', x''; h)\|^2$$

for various choices of h. Because the Jacobian of M is generally rank deficient, we cannot assume that ϕ is strictly convex, so for any fixed value of h, ϕ will not have a unique local minima. On the other hand, Theorem 4.7 establishes that the local minima for sufficiently small h can be chosen as a continuous function of h whose (x, x')-part converges to $\bar{z}$ as $h \to 0$. (From this point, $z(h)$ represents the (x, x')-part of a local minimum of the consistency functional approximated with displacement h.)

4.5.1. Existence of a solution to the approximation.

THEOREM 4.7. *Let M, ϕ, and $\bar{z}$ be defined as in the previous paragraphs. Suppose that the Jacobian obeys the conditions of Assumption* 4.6. *Then there exists an $H > 0$ and continuous functions $z(h)$ and $x''(h)$ (defined for $0 < h < H$) such that (for h in $(0, H)$) $(z(h), x''(h))$ is a local minimum of $\phi(z, x''; h)$ and $\lim_{h\to 0} z(h) = \bar{z}$* [2].

Before proving this result, we give a couple of simple lemmas which allow us to reduce the structure of the rank-deficient problem to that of a full rank system.

The first lemma implies that we can always select a column basis for $E(x, t)$ in a neighborhood of the initial point.

LEMMA 4.8. *Let $A(u)$ be a C' matrix-valued function. Suppose that A has constant rank r in a neighborhood of a point u_0. Then if some set of r columns of $A(u)$ is linearly independent at $u = u_0$, the set remains linearly independent for u in a neighborhood of this point.*

Proof. Let $i_1, i_2, \cdots, i_r$ be the indices of r linearly independent columns of $A(u_0)$. Define a new matrix function $W(u)$ with just those r columns. Since small perturbations of u lead to small perturbations of W (A is C'), the result follows immediately by the fact that sufficiently small perturbations cannot lower the rank of a matrix. □

The next result shows how we can reduce the rank-deficient problem to a full rank problem.

LEMMA 4.9. *Let G be defined on $\mathbf{R}^p \times \mathbf{R}^q \times \mathbf{R}^r \to \mathbf{R}^s$ have the form $G(u, v_1, v_2) = G_1(u) + A_1(u)v_1 + A_2(u)v_2$, where G_1 is some differentiable function in u and A_1 and A_2 are differentiable matrix-valued functions. Suppose further that $A_2(u) = A_1(u)B(u)$, where B is differentiable. Then $\|G\|^2$ has a local minimum at (u, v_1, v_2) if and only if $\|\tilde{G}\|^2$ has a local minimum at $(u, v_1 + B(u)v_2)$, where $\tilde{G}$ is the restriction defined by $\tilde{G}(x, y) = G(x, y, 0)$.*

Proof. The condition for the squared norm of a smooth mapping $f(w)$ to have a local minimum at a point w_0 is that $(\partial f/\partial w)|_{w=w_0}^T f(w_0) = 0$. Therefore, for $\|G\|^2$ to have a local minimum at (u, v_1, v_2) we must have that

$$\left(\frac{\partial G_1}{\partial u} + \frac{\partial A_1}{\partial u}v_1 + \frac{\partial A_2}{\partial u}v_2\right)^T G(u, v_1, v_2) = 0, \tag{22}$$

$$A_1(u)^T G(u, v_1, v_2) = 0, \tag{23}$$

$$A_2(u)^T G(u, v_1, v_2) = 0, \tag{24}$$

[2] To simplify the details of the proof, we have here neglected showing the existence of an $O(h^p)$ solution sequence when the approximation has this order. In fact, the sequence constructed in the proof, below, has this property.

where all the derivatives are evaluated at (u, v_1, v_2).

It is readily apparent that (24) is a direct consequence of (23), since $A_2(u) = A_1(u)B(u)$, so we work with just the first two equations.

The equations satisfied by local minima of $\|\tilde{G}\|^2$ are

$$\left(\frac{\partial G_1}{\partial u} + \frac{\partial A_1}{\partial u}(v_1 + B(u)v_2)\right)^T \tilde{G}(u, v_1 + B(u)v_2) = 0, \tag{25}$$

$$A_1(u)^T \tilde{G}(u, v_1 + B(u)v_2) = 0. \tag{26}$$

Note that $A_1(u)(v_1 + B(u)v_2) = A_1(u)v_1 + A_2(u)v_2$, so $\tilde{G}(u, v_1 + B(u)v_2) = G(u, v_1, v_2)$, and upon comparing the two sets of equations with this in mind we find that (23) and (26) are the same equation. Subtracting (25) from (22), we arrive at

$$\left[\frac{\partial A_1}{\partial u}B(u) - \frac{\partial A_2}{\partial u}\right]^T G(u, v_1, v_2) = 0. \tag{27}$$

If this holds given either set of equations, the equivalence is established. Differentiate $A_2(u) = A_1(u)B(u)$ to get

$$\frac{\partial A_2}{\partial u} = A_1(u)\frac{\partial B}{\partial u} + \frac{\partial A_1}{\partial u}B(u).$$

Upon substitution in (27) we need only that $(A_1(u)(\partial B/\partial u)v_2)^T G(u, v_1, v_2) = 0$, which follows directly from (23) or (26). □

With these lemmas we are ready to prove Theorem 4.7.

Proof. From the constant rank assumption and Lemma 4.8, we know that for x in some neighborhood of $\bar{x}$, there is a fixed choice of r columns of $E(x, t_0)$ which form a basis for $\mathcal{R}(E(x, t_0))$, where r is the rank of $E(\bar{x}, t_0)$. Without loss of generality, we may assume that the first r columns, $E_1(x, t)$, form such a basis. Let $E_2(x, t)$ represent the remaining $n - r$ columns. $E_2(x, t_0) = E_1(x, t_0)V(x)$, for some V which has the same smoothness as E. Partition x'' as (x_1'', x_2''), according to the above partition of $E(x, t_0)$.

We next apply Lemma 4.9 to see that $(\hat{x}, \hat{x}', \hat{x}_1'', \hat{x}_2'')$ is a local minimum of ϕ if and only if $(\hat{x}, \hat{x}', \hat{x}_1'' + B(\hat{x})\hat{x}_2'')$ is a local minimum of the restriction $\tilde{\phi}$ of $\phi(x, x', x_1'', 0)$ to (x, x', x_1'')-space. This means that we can arbitrarily select $x_2'' = 0$.

Thus the reduced consistency function may be written

$$\begin{aligned}&\tilde{M}(x, x', x_1''; h)\\ &:= \begin{bmatrix} E(x, t_0)x' + f(x, t_0) \\ \frac{1}{h}\sum_{i=0}^{s} \alpha_i\{E(x + hc_i x', t_0 + hc_i)x' + f(x + hc_i x', t + hc_i)\} + E_1(x, t)x_1'' \\ B(x, x') \end{bmatrix}.\end{aligned}$$

To establish the existence of a convergent sequence of solutions, we will demonstrate that this reduced function has a full rank Jacobian, so that Theorem 4.1 insures the existence of a locally unique minimum for h sufficiently small.

First note that Assumption 4.6 means that the reduced function which is the limit of $\tilde{M}$ as $h \to 0$ has a full column rank Jacobian.

The difference approximations we consider here, being linear combinations of values of F at various points, possess the same continuity as F itself. Lemma 4.2

establishes that the Jacobian of the approximation is an approximation to the Jacobian of the exact system. Therefore, in particular, $J_{\tilde{M}} = \partial\tilde{M}(x, x', x_1'')/\partial(x, x', x_1'')$ has locally full column rank and, since rank cannot decrease in a sufficiently small neighborhood of its matrix argument, the continuity of the approximation means that $J_{\tilde{M}}$ has full column rank in some neighborhood of the initial point.

Thus we have reduced our problem to a full column rank problem, whose solvability is guaranteed by Theorem 4.1. □

If we knew which columns of E form the basis E_1, we could construct a full rank approximation with the same solution as the orignal rank-deficient system of consistency equations. But even when such a basis cannot be explicitly determined, its existence alone evidently serves to guarantee the existence of a convergent solution sequence.

Unfortunately, this result does not tell us that the (x, x')-part of solutions we actually obtain in solving the approximate consistency equations in the least squares sense will necessarily converge. The problem comes from the fact that some components of the vector of derivatives (components which are not being sought) may tend to infinity as $h \to 0$. In Lemma 5.4.5 of [15], it is shown that as long as a solution sequence remains bounded as $h \to 0$, the (x, x')-part must converge; this result follows from the Bolzano–Weierstrass theorem.

4.6. Algorithm. We are now ready to state a general algorithm for the initialization of DAEs.

ALGORITHM 4.1. DAE Initialization.

```
1. Set coefficient arrays for difference approximations,
   unit round-off error, initial displacement $h$, tolerance $\tau$
2. Repeat
   2.1. Solve $\|\tilde{M}_h(x, x', \cdots, x^{(m+1)})\|^2 = \min$
        (or the reduction) for $(x_0, x_0')$ (and higher derivatives)
   2.2. Estimate error-norm $\epsilon_h$ in $(x_0, x_0')$
   2.3. Estimate improved displacement, $h$
   2.4. Until $\epsilon_h \leq \tau$.
```

5. Numerical experiment: trajectory control. In [15], a number of important implementation questions are considered. In particular, it is shown there how to exploit bandedness of the partial Jacobian matrices of the DAE during the formation and solution of the consistency equations and how to choose a differencing parameter for the approximation of the Jacobian of the consistency function; also described there is a particular implementation of the **DAIS** method.

Here we briefly describe one numerical experiment. The following system, called a "trajectory prescribed path control problem," is a DAE for which initialization is a nontrivial problem. The problem consists of six differential relations:

$$\begin{aligned}
H' &= V_R \sin(\gamma), \\
\xi' &= \frac{V_R \cos(\gamma) \sin(A)}{r \cos(\lambda)}, \\
\lambda' &= \frac{V_R}{r} \cos(\gamma) \cos(A), \\
V_R' &= \frac{-D}{m} - g \sin(\gamma),
\end{aligned}$$

TABLE 1
Constants and derived variables.

r	$H + a_e$
g	μ/r^2
L	$.5\rho C_L S V_R^2$
$C_L(\alpha)$	$.01\alpha\pi/180$
$C_D(\alpha)$	$.04 + .1C_L^2$
D	$.5\rho C_D S V_R^2$
$\rho(H)$	$.002378e^{-H/23800}$
a_e	.20902900e+9
μ	.1407653916e+17
Ω_E	.72921159e−4
m	.2890532728e+1
S	1.0

$$\begin{aligned}
& & & -\Omega_E^2 r\cos(\lambda)\Big(\sin(\lambda)\cos(A)\cos(\gamma) - \cos(\lambda)\sin(\gamma)\Big),\\
\gamma' & = & & \frac{L\cos(\beta)}{mV_R} + v\frac{\cos(\gamma)}{V_R}\left(\frac{V_R^2}{r} - g\right) + 2\Omega_E\cos(\lambda)\sin(A),\\
& & & +\frac{\Omega_E^2 r\cos(\lambda)}{V_R}\Big(\sin(\lambda)\cos(A)\sin(\gamma) + \cos(\lambda)\cos(\gamma)\Big),\\
A' & = & & \frac{L\sin(\beta)}{mV_R\cos(\gamma)} + \frac{V_R}{r}\cos(\gamma)\sin(A)\tan(\lambda),\\
& & & -2\Omega_E\Big(\cos(\lambda)\cos(A)\tan(\gamma) - \sin(\lambda)\Big),\\
& & & +\frac{\Omega_E^2 r\cos(\lambda)\sin(\lambda)\sin(A)}{V_R\cos(\gamma)},
\end{aligned}$$

coupled with two constraints (the desired path):

$$\begin{aligned}
\gamma + 1 + 9(t/300)^2 &= 0,\\
A - 45 - 90(t/300)^2 &= 0,
\end{aligned}$$

where the differential variables are $(H, \xi, \lambda, V_R, \gamma, A)$, the algebraic variables are (α, β) and constants and "derived" variables are given in Table 1. The (provided) initial values for the differential variables are (to numerical precision)

$$(H_0, \xi_0, \lambda_0, V_R, 0, \gamma_0, A_0) = (12000, -\pi/180, 100000, 45\pi/180, 0, 0).$$

For units and physical interpretations, see [2]; for our present purpose, we ignore the physical significance of the problem to treat it solely as a test case for the initialization method.

In this example, the initial conditions are given by initial values for the differential variables. In [2] it is demonstrated that this problem has index two. The problem has interesting features with respect to initialization: the equations are nonlinear and, what is relatively rare in the study of high index problems, even the algebraic variables (here α and β) occur nonlinearly. The problem is in the Hessenberg form and results in consistency equations of the form (16)–(19).

TABLE 2
Trajectory control problem.

Order	h_1	h_2	error(u)	error(u')	error(v)	error(v')
1	10.0e−06	5.85e−04	1.77e−17	2.62e−09	9.77e−05	1.35e−06
2	36.8e−05	3.16e−03	1.77e−17	3.18e−10	3.04e−07	7.39e−06
3	24.0e−04	9.22e−03	1.77e−17	1.05e−10	2.51e−08	3.69e−07

DAIS solves the reduced equations (reduction of §4.2) by replacing the first derivatives and second derivatives by independent finite difference approximations constructed from coefficients computed according to Lemma 3.2, using $c_i = i$.

In [15], the same displacement was used for each of the derivative approximations, and this was provided a priori to **DAIS**. In the current version of the code, the parameters h_1 and h_2 used in the difference approximations to the first and second derivatives respectively are computed automatically according to the formula

$$h_i = \left(\frac{i(|\Gamma| + 1)\epsilon_M}{p|\Gamma|} \right)^{1/(p+i)}, \tag{28}$$

where ϵ_M is machine epsilon, p is the order of the approximation being used, and Γ is some measure of smallest scale present in the problem. The estimate is the approximate optimum choice of displacement (under certain simplifying assumptions on the behavior of the rounding error) if the pseudoderivatives of the ODE

$$y' = \Gamma y$$

are computed using the finite-difference approximations of §3. When the smallest scale present in the problem is used as the basis for the displacement selection, it will typically lead to an overestimate of the optimum choice for the larger modes. No claim is made here that (28) is the best displacement selection strategy, neither do we suggest a method for estimating the smallest scale of the problem, Γ.

Heuristics for the selection of the displacements for the difference approximations to the Jacobian can be found in [15].

In Table 2 we give the relative error in Euclidean norm of the computed initial solution and derivatives for the trajectory problem. Here $u = (H, \xi, \lambda, V_R, \gamma, A)$ is a vector of differential variables, and $v = (\alpha, \beta)$ is the vector of algebraic variables, in which we would expect to find the larger errors. In this example, we chose $\Gamma = 10^{-5}$.

The problem was solved on a Digital Equipment Corporation μVAX in double precision ($\epsilon_M \approx 10^{-17}$). In this environment, the most accuracy we would expect to be able to achieve in the derivatives and algebraic variables would be around eight digits. The results given in Table 2 are therefore as expected. It is interesting that the higher-order approximations seem to perform best. Similar behavior was observed in some other experiments.

6. Conclusion. In this paper we described a practical method for initializing a differential-algebraic equation and its numerical implementation. We showed how the consistency conditions for an index-m DAE lead to a system of equations involving derivatives of the problem. We then showed how these equations can be approximated by finite differences and discussed the solution of the resulting system. Finally we described a numerical experiment in which the authors' program, **DAIS**, was used to initialize a highly nonlinear DAE arising in trajectory control simulation.

Many important questions have not been addressed here, including, in particular, the selection and analysis of a robust nonlinear solver for the rank-deficient case and the application of the method to the important class of index-2 DAEs in the semi-explicit form but not in triangular form (e.g., those arising typically from the method of lines solution of PDEs).

Acknowledgments. The authors thank K. E. Brenan and S. L. Campbell for their generous assistance and the referees for their detailed and constructive criticism.

REFERENCES

[1] M. Berzins, P. M. Dew, and R. M. Furzeland, *Developing software for time dependent problems using the method of lines and differential-algebraic integrators,* Appl. Numer. Math., (1988).

[2] K. E. Brenan, *Stability and convergence of difference approximations for higher index differential-algebraic systems with applications in trajectory control,* Ph.D. thesis, Department of Mathematics, UCLA, Los Angeles, CA, 1983.

[3] K. E. Brenan, S. L. Campbell, and L. R. Petzold, *The Numerical Solution of Initial Value Problems in Ordinary Differential-Algebraic Equations,* North-Holland, Amsterdam, 1989.

[4] S. L. Campbell, *Singular Systems of Differential Equations,* Res. Notes in Math. 40, Pitman Advanced Publishing Program, Boston, 1980.

[5] S. L. Campbell, *Consistent initial conditions for linear time varying singular systems,* in Frequency Domain and State Space Methods for Linear Systems, Elsevier Science (North-Holland), Amsterdam, 1986.

[6] ———, *A Computational Method for General Higher-Index Nonlinear Singular Systems of Differential Equations,* Proceedings of 1988 IMACS Conference,

[7] ———, *private communication,* 1987.

[8] C. W. Gear, *Differential-algebraic equation index transformations,* SIAM J. Sci. Statist. Comput., 9 (1988), pp. 39–48.

[9] C. W. Gear, G. K. Gupta, and B. Leimkuhler, *Automatic integration of Euler-Lagrange equations with constraints,* J. Comput. Appl. Math., 12-13 (1985), pp. 77–90.

[10] G. K. Gupta, C. W. Gear, and B. Leimkuhler, *Implementing linear multistep formulas for solving rm DAEs,* Report No. UIUCDCS-R-85-1205, Dept. of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, 1985.

[11] E. Hairer, C. H. Lubich, and M. Roche, *The numerical solution of differential-algebraic equations by Runge–Kutta methods,* Report, Departement de Mathematiques, Universitè de Genève, Geneva, 1988.

[12] A. C. Hindmarsh and H. J. Johnson, *Dynamic simulation of reversible solid-fluid reactions in nonisothermal porous spheres with Stefan-Maxwell diffusion,* Chem. Eng. Sci., 43 (1988), pp. 3235–3258.

[13] A. C. Hindmarsh, LSODE *and* LSODI, *two new initial value ordinary differential equation solvers,* ACM-SIGNUM Newsletters, 15 (1980), pp. 10–11.

[14] B. Leimkuhler, *Error estimates for differential-algebraic equations,* Report No. UIUCDCS-R-86-1287, Dept. of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, 1986.

[15] ———, *Approximation methods for the consistent initialization of differential-algebraic equations,* Ph.D. thesis (Report No. UIUCDCS-R-88-1450), Dept. of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, 1988.

[16] B. Leimkuhler, L. R. Petzold, and C. W. Gear, *On obtaining a consistent set of initial values for a system of differential-algebraic equations,* Report No. UIUCDCS-R-87-1344, Dept. of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, 1987.

[17] T. Mrziglod, *Zur Theorie und numerischen Realisierung von Lösungsmethoden bei Differentialgleichungen mit angekoppelten algebraischen Gleichungen,* Ph.D. thesis, Universität zu Köln, Cologne, FRG, 1987.

[18] C. C. Pantelides, *The consistent initialization of differential-algebraic systems,* SIAM J. Sci. Statist. Comp., 9 (1988), pp. 213–232.

[19] L. R. Petzold, *A description of* DASSL: *a differential/algebraic system solver,* Proceedings of IMACS World Congress, Montreal, Canada, 1982.

[20] ———, *Differential/algebraic equations are not* ODEs, SIAM J. Sci. Statist. Comp., 3 (1982), pp. 367–384.

[21] W. C. Rheinboldt, *Differential-algebraic equations as differential equations on manifolds,* Math. Comput., 43 (1984), pp. 473–482.