

NUMERICAL SOLUTION OF BOUNDARY VALUE PROBLEMS IN DIFFERENTIAL-ALGEBRAIC SYSTEMS*

KENNETH D. CLARK[†] AND LINDA R. PETZOLD[‡]

Abstract. This paper extends the theory of shooting and finite-difference methods for linear boundary value problems (BVPs) in ordinary differential equations (ODEs) to BVPs in differential-algebraic equations (DAEs) of the form

$$\begin{aligned} E(t)y'(t) + F(t)y(t) &= f(t), \quad t \in [a, b], \\ B_a y(a) + B_b y(b) &= \beta, \end{aligned}$$

where $E(\cdot)$, $F(\cdot)$, and $f(\cdot)$ are sufficiently smooth and the DAE initial value problem (IVP) is solvable. $E(t)$ may be singular on $[a, b]$ with variable rank, and the DAE may have an index that is larger than one. When $E(t)$ is nonsingular, the singular theory reduces to the standard theory for ODEs. The convergence results for backward differentiation formulas and Runge-Kutta methods for several classes of DAE IVPs are applied to obtain convergence of the corresponding shooting and finite-difference methods for these DAE boundary value problems. These methods can be implemented directly without having to (1) regularize the system to a lower index DAE or ODE or (2) convert the system to a particular canonical structure. Finally, some numerical experiments that illustrate these results are presented.

Key words. differential-algebraic systems, boundary value problems, higher index, shooting methods, finite-difference methods, backward differentiation formulas, implicit Runge-Kutta methods

AMS(MOS) subject classifications. 65L05, 34A08

1. Introduction. In this paper we extend the theory of shooting and finite-difference methods for linear boundary value problems (BVPs) in ordinary differential equations (ODEs) to BVPs in *differential-algebraic systems* (DAEs) of the form

$$(1.1a) \quad \mathcal{L}y(t) \equiv E(t)y'(t) + F(t)y(t) = f(t), \quad t \in [a, b],$$

$$(1.1b) \quad \mathcal{B}y(t) \equiv B_a y(a) + B_b y(b) = \beta,$$

where $E(\cdot)$, $F(\cdot)$, and $f(\cdot)$ are sufficiently smooth and the DAE initial value problem (IVP) is solvable. We allow $E(t)$ to be singular on $[a, b]$ with variable rank, and the DAE (1.1a) may have an index that is larger than one. See [14] or [33] for a detailed discussion of the index of a DAE. Intuitively, ODEs have index 0, while the solutions to higher index DAE systems (index > 1) involve derivatives of the coefficients E , F , and the input f . Index one systems contain algebraic variables that are uniquely determined by the state variables (not including derivatives).

IVPs in DAEs have been extensively studied in recent years from both a theoretical and a numerical perspective. These problems arise frequently in applications,

* Received by the editors April 25, 1988; accepted for publication (in revised form) January 27, 1989.

[†] Mathematical Sciences Division, U.S. Army Research Office, Research Triangle Park, North Carolina 17709. The research of this author was performed while he was a visiting faculty member at Lawrence Livermore National Laboratory, Livermore, California.

[‡] Computing & Mathematics Research Division, L-316, Lawrence Livermore National Laboratory, P. O. Box 808, Livermore, California 94550. The work of this author was partially supported by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U.S. Department of Energy, by Lawrence Livermore National Laboratory under contract W-7405-Eng-48.

including circuit and control theory [6], [17], [36]; chemical kinetics [25]; fluid dynamics [33], [38]; and robotics [35]. In some cases, the models lead to the nonlinear semi-explicit formulation

$$(1.2a) \quad y' = f(y, z, t),$$

$$(1.2b) \quad 0 = g(y, z, t)$$

and with it the interpretation of (1.1a), (1.2) as constrained ODEs or differential equations on manifolds [42]. However, in many applications the fully-implicit formulation (cf. (1.3)) is more appropriate.

As an extension to the initial value theory, it is natural to consider BVPs in DAEs. DAE BVPs arise in the modeling of semiconductor devices [1]; control theory [4], [17]; detonation modeling [28]; the design of heat exchangers [37]; and in parameter-estimation problems for multibody systems [5]. We believe that as information regarding DAEs and software for these problems becomes more widely disseminated in the scientific and engineering community, the number and variety of applications will increase.

In recent years, several researchers have studied various approaches to the general solution of DAE BVPs. The work of März and Griepentrog [34], [23] focuses on difference and shooting methods for BVPs for nonlinear fully-implicit systems:

$$(1.3a) \quad F(y', y, t) = 0,$$

$$(1.3b) \quad G(y(a), y(b)) = 0$$

under a *transferability* hypothesis that guarantees that (1.3a) is a regular, index one system and the nullspace of $F_{y'}$ is independent of y', y and has constant dimension. All linear solvable index one systems (1.1a) are transferable, as are semi-explicit systems (1.2) where g_z is bounded and invertible. The numerical approach in [23] requires knowledge of some projector onto $\ker(F_{y'})$ and its derivative at each meshpoint. There are some theoretical results for the subclass of *tractable* index two systems [22], but it is implied that a successful numerical approach involves regularizing the DAE to a nonsingular or index one system and then numerically solving the regularization (cf. [31], [32]).

Ascher [1] gives a convergence result for collocation schemes applied to semi-explicit index one DAEs, where the collocation methods are applied in such a way that the algebraic components of the system are approximated in a piecewise discontinuous space. In Ascher [1], a convergence result is outlined and order conditions are given for Gaussian collocation methods applied directly to fully-implicit index one systems. Hanke [24] describes a least-squares collocation method for linear differential-algebraic equations that is applicable to higher index systems.

Bock, Eich, and Schlöder [5] describe numerical methods based on multiple shooting and collocation for equality—and inequality—constrained DAE BVPs arising from parameter-identification problems for multibody systems. Their approach is restricted primarily to semi-explicit index one systems, and the methods distinguish the algebraic from the differential components in their numerical treatment. This distinction in the method between the algebraic and differential components—a distinction that is inherent in methods proposed for semi-explicit systems by März and Griepentrog; Ascher; and Bock, Eich, and Schlöder—is natural and highly appropriate in the semi-explicit index one case, but for the fully-implicit case it is unclear how to accomplish the distinction in general without the expensive computation of projectors at each meshpoint.

This paper serves several purposes. First, we show that under an appropriate formulation, the theory of shooting and finite-difference methods for linear systems (e.g., simple and parallel shooting with partially or completely separated boundary conditions, one-step difference schemes with extrapolation) for ODE BVPs can be formally extended to DAE BVPs, using the characterization of the solution manifold given in [11]. We note that we have not addressed here the issue of conditioning for the DAE BVP and for the numerical methods. For a discussion of these issues for the DAE BVP, see Lentini and März [29], [30]; and for ODE BVPs and numerical methods for ODE BVPs, see Ascher, Mattheij, and Russell [3]. Ascher [1] addresses these issues for a limited class of numerical methods for DAE BVPs. We make no assumptions on the index of the DAE (1.1a), except what is required for convergence of the corresponding methods applied to related DAE IVPs. Thus for many of the methods, the results apply to the solution of higher index systems. The results and details of the theory are straightforward extensions of the results in [26], [27] and reduce to the ODE case when $E(t)$ is nonsingular; thus it is possible to treat ODE and DAE BVPs within the same theoretical framework. For the purposes of clarity and consistency, we will adopt the notation and presentation in these papers to the greatest possible extent. As in the ODE case, the shooting theory provides a necessary theoretical basis for the development of more direct techniques, such as finite differences. We discuss shooting methods in § 3 and treat the finite-difference case in § 4. In § 5 we present the results of some numerical experiments that reinforce the theory of the previous sections.

A consequence of this approach is that the initial value methods that exhibit the restricted convergence and stability properties for certain subclasses of numerically solvable DAEs (1.1a), e.g., *backward differentiation formulas* (BDF) ([6], [7], [16]–[21], [33], [41], [43]); *implicit Runge–Kutta methods* (IRK) ([8], [40]); or the *i th order j th block* (i.e., (i, j) -) series methods ([11], [12]) can in principle be used to construct convergent approximations to the BVP (1.1) under similar restrictions. Furthermore, the DAE can be solved directly by these methods without having to convert the system to a canonical structure. In particular, it is unnecessary to transform the DAE BVP to an ODE BVP on a lower-dimensional space or to regularize the DAE. Knowledge of the solution manifold (which may require derivative information) is required only at the initial time point $t_0 = a$, or in the case of parallel shooting at each parallel node τ_j , and not at every numerical meshpoint t_n .

2. Background and terminology. We assume that $E(\cdot)$, $F(\cdot)$, and $f(\cdot)$ are real matrix- and vector-valued functions of $t \in I = [a, b]$, with dimensions $m \times m$ and $m \times 1$, respectively. The space of s -times continuously differentiable functions on I is denoted by $C^s(I)$, or more conveniently C^s , with the range (e.g., matrix- or vector-valued) understood from the context. Throughout this paper we assume that E, F are at least C^{2m} , while f is at least C^m , although in many cases it suffices to have $E, F, f \in C^\sigma$, where σ is the global index (see the discussion below). As in [11], we adopt the following definition of solvability for (1.1a).

DEFINITION 2.1. The system (1.1a) is *solvable* on I if and only if

- (1) for all f there exists a C^1 solution y ;
- (2) all solutions corresponding to f are defined and at least C^1 on the entire interval I and are uniquely determined by their values $y(t)$ for each $t \in I$;
- (3) all solutions of the homogeneous system $\mathcal{L}y = 0$ are at least C^{2m+1} ; and
- (4) if $f \in C^s$ for $m \leq s \leq 2m$, then any corresponding solution y is C^{s-m+1} .

Conditions (1) and (2) of the definition constitute the standard definition of solvability (cf. [14], [16]) and imply that solutions are pointwise linearly independent

form a basis for $\ker(\mathbf{E}_j(\tau)^T)$. Then $Q = W(W^T W)^{-1}W^T$ is an orthogonal projector onto $\ker(\mathbf{E}_j^T)$ and $I - Q$ is an orthogonal projector onto $\text{im}(\mathbf{E}_j) = \ker(\mathbf{E}_j^T)^\perp$. Clearly (2.2) implies

$$(2.4) \quad W(\tau)^T \mathbf{F}_j(\tau) y_0 = W(\tau)^T \mathbf{f}_j(\tau).$$

But (2.4) implies $f_j - \mathbf{F}_j y_0 \in \ker(Q) = \text{im}(I - Q) = \text{im}(\mathbf{E}_j)$, hence (2.2) and (2.4) are equivalent. In § 4, we write the system (2.4) as

$$(2.5) \quad M_\tau y_0 = g(\tau),$$

where M_τ is $\eta \times m$ with full row rank, $g(\tau) = 0$ if $f \equiv 0$ and $r = \dim \ker(M_\tau) = m - \eta$. It follows that $M_0 = \ker(M_a)$, where $M_a = M_\tau$ at $\tau = a$. Although numerically impracticable or expensive in many cases, this tells us that if r is not known a priori, in principle it can be determined from the rank of \mathbf{E}_j .

Frequently, it is possible to take $j < m + 1$ in (2.2). If $\sigma + 1$ is the smallest integer such that $\mathbf{E}_{\sigma+1}$ is 1-full and constant rank, then σ is the *global index* of (1.1a) and it suffices to take $j = \sigma + 1$. It is relatively straightforward to show that this definition agrees with the definition of global index given in [21]. (See also [17] for an equivalent definition of global index in a slightly different context.) If for each $\tau \in I$ there exists a scalar $\lambda_\tau \in \mathfrak{R}$ such that $(\lambda_\tau E(\tau) + F(\tau))^{-1}$ exists, then (1.1a) is a regular system and the *local index* of (1.1a) at $t = \tau$ is the index of nilpotency of the matrix $E_\lambda(\tau) = (\lambda_\tau E(\tau) + F(\tau))^{-1} E(\tau)$, denoted $\text{ind}(E(\tau), F(\tau))$. It is well known that the local and global indices for higher index (index > 1) systems may differ when E, F are time varying, although for index one systems they are the same.

The main results of this paper depend only on solvability and (2.4) and therefore are independent of the index. However, since the index one systems are well understood and arise most frequently in applications, we briefly review several facts for this case. In the following proposition let $(\cdot)^\dagger$ denote the *Moore-Penrose inverse*, while $(\cdot)^D$ is the *Drazin inverse* [15].

PROPOSITION 2.1. *If (1.1a) is solvable, then $\text{im}([I - EE^\dagger]F) = \text{im}([I - EE^\dagger])$ for all $t \in I$. Consequently, if $\text{rank}(E(t)) = r$, then $\text{rank}([I - E(t)E(t)^\dagger]F(t)) = m - r$ and $\dim \ker([I - E(t)E(t)^\dagger]F(t)) = r$.*

Proof. Solvability of (1.1a) implies for every solution $y(t)$,

$$(2.6) \quad [I - EE^\dagger]Fy = [I - EE^\dagger]f \quad \text{for all } t \in [a, b],$$

since $[I - EE^\dagger]$ is the orthogonal projector onto $\ker(E^T) = \text{im}(E)^\perp$. Clearly, $\text{im}([I - EE^\dagger]) \supseteq \text{im}([I - EE^\dagger]F)$ but if the inequality is strict, there exists smooth $f(t)$ such that (2.6) is not satisfied for some $t^* \in [a, b]$. This contradicts solvability, hence $\text{im}([I - EE^\dagger]) \equiv \text{im}([I - EE^\dagger]F)$. \square

Proposition 2.1 follows directly from the proof of Theorem 2.2 in [16] and is independent of the index of (1.1a). In the special case $\text{ind}(E, F) = 1$, the system (1.1a) is solvable if and only if $\text{rank}(E(t))$ is constant on $[a, b]$. Equivalently, there exist invertible $P(t), Q(t)$ as smooth as $E(t)$ and $F(t)$ such that

$$(2.7) \quad P(t)E(t)Q(t) = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}, \quad P(t)F(t)Q(t) = \begin{pmatrix} C(t) & 0 \\ 0 & I \end{pmatrix},$$

where the identity block in PEQ has size $\text{rank}(E(t))$. Thus the dimension r of the solution manifold for solvable index one systems is $\text{rank}(E(t))$, while if the index is

greater than one, $r < \text{rank}(E(t))$. In particular, from Proposition 2.1 we have the following.

COROLLARY 2.1. *If (1.1a) is solvable index 1, the $m - r$ linearly independent relations in (2.6), evaluated at $t = \tau$, completely specify the set of consistent initial conditions for (1.1a) starting at $t = \tau$.*

Of course, the additional consistency requirements for higher index systems are obtained from differentiations of (1.1a), as in (2.4). Note that if only one consistent initial condition y_p^0 is needed and $F(\tau)v = f(\tau)$ is a consistent linear system, then we may choose $y_p^0 = v$. Also, compare (2.6) with the equivalent characterization (2.4) with $j = 2$, or the system

$$(2.8) \quad (I - E_\lambda^D E_\lambda)y = (I - E_\lambda^D E_\lambda)F_\lambda f_\lambda, \text{ evaluated at } t = \tau,$$

where $q_\lambda = (\lambda E + F)^{-1}q$ for $q = E, F, f$, which is derived in [14]. Using the approach in [23], one can also show that $M_f(\tau)$ is the set of all x such that

$$(2.9) \quad x = (P - Q[E + FQ]^{-1}FP)x_0 + Q[E + FQ]^{-1}f, \quad x_0 \text{ arbitrary}$$

evaluated at $t = \tau$, where Q is an arbitrary projector onto $\text{ker}(E)$ and $P = I - Q$. Which characterization is appropriate depends on the circumstances, but for our purposes it will be convenient to use (2.5), since it is independent of the index.

3. Shooting theory for linear DAEs. Assume that $r = m - \text{dim}(\text{ker}(E_j))$ has been determined by a careful rank determination of E_j using, for example, a singular value decomposition [22] or Gauss elimination with pivoting and scaling if the linearly independent rows of E_j are not nearly linearly dependent in the numerical sense. In many cases of interest (e.g., constant coefficients or the structural forms discussed in [13], [16], and [17]), $r = \text{core-rank}(E_\lambda) = \text{rank}(E_\lambda^D E_\lambda)$. The core-rank of a matrix A is the size of the block corresponding to the nonzero eigenvalues in the Jordan form for A . In order for the boundary conditions (1.1b) to uniquely determine solutions for all β , it is necessary that $B_a, B_b \in \mathbb{R}^{r \times m}$ with $\text{rank}[B_a, B_b] = r$. Thus a correct formulation for the BVP is

$$(3.1a) \quad \mathcal{L}y(t) \equiv E(t)y'(t) + F(t)y(t) = f(t), \quad t \in I,$$

$$(3.1b) \quad \mathcal{B}y(t) \equiv B_a y(a) + B_b y(b) = \beta, \quad B_a, B_b \in \mathbb{R}^{r \times m}, \quad \beta \in \mathbb{R}^r.$$

DEFINITION 3.1. The BVP (3.1) is *solvable* if and only if (3.1a) is a solvable DAE and for every $\beta \in \mathbb{R}^r$ there exists a unique solution y to (3.1).

Let $y_p^0 \in M_f(a)$ and assume $(\phi_i^0)_1^r$ is any basis for $M_0 = \text{ker}(M_a)$, where M_a is given in (2.5). Correspondingly, let $y_p(t)$ and the fundamental matrix $Y(t) = [\phi_1(t), \dots, \phi_r(t)]$ be the solutions to the $r + 1$ IVPs

$$(3.2a) \quad \mathcal{L}y_p(t) = f(t), \quad y_p(a) = y_p^0 \in M_f(a),$$

$$(3.2b) \quad \mathcal{L}Y(t) = 0, \quad Y(a) = Y_0 = [\phi_1^0, \dots, \phi_r^0].$$

Note that $Y(t)$ has full column rank for all $t \in I$, since (3.1a) is solvable. Using the representation (2.1a) and imposing the boundary conditions (3.1b), we find that y is a solution of (3.1) if and only if the vector $\xi = (\xi_1, \dots, \xi_r)^T$ satisfies

$$(3.3) \quad [B_a Y_0 + B_b Y(b)]\xi = \beta - (B_a y_p^0 + B_b y_p(b)).$$

As in the ODE case, the $r \times r$ matrix

$$(3.4) \quad S = B_a Y_0 + B_b Y(b)$$

is the *shooting matrix* for (3.1) and is unique up to a change of basis for M_0 , i.e., if \tilde{S} is any shooting matrix for (3.1), there exists a constant invertible $r \times r$ matrix Q such that $\tilde{S} = SQ$. Thus we have Theorem 3.1.

THEOREM 3.1. *The BVP (3.1) is solvable if and only if S is invertible. The desired solution is given by (2.1a), where y_p^0, Y satisfy (3.2) and ξ is the solution to (3.3).*

Therefore, *simple shooting* consists of solving the $r + 1$ IVPs in (3.2) over $[a, b]$, forming the shooting matrix (3.3), solving the linear system (3.4) for ξ , and finally solving the IVP

$$(3.5) \quad \mathcal{L}y(t) = f(t), \quad y(a) = y_p^0 + Y_0 \xi$$

for the unique solution y to (3.1). At the end of this section we discuss numerical implementations of shooting procedures. Unfortunately, if the differential part of (3.1a) exhibits stiffness, the shooting matrix may be ill conditioned. One way to inhibit the effects of exponential growth of solutions on the conditioning of the shooting equation is to consider separate shooting problems on smaller subintervals of $[a, b]$ and join the resulting solutions by imposing continuity. This technique of *parallel* or *multiple* shooting is discussed momentarily. As we have already noted, the details are essentially the same as for ODEs, with the exception that we must incorporate information about the solution manifold. To this end the following intuitive result is very useful.

PROPOSITION 3.1. *Suppose (3.1) is solvable with solution manifold determined by (2.5). Let the $m \times 2m$ matrix X and $m \times (r + m)$ matrix Z be defined by*

$$X = \begin{pmatrix} B_a & B_b \\ M_a & 0 \end{pmatrix}, \quad Z = \begin{pmatrix} S & B_a \\ 0 & M_a \end{pmatrix}.$$

Then $\text{rank}(X) = \text{rank}(Z) = m$.

Proof. Clearly $\text{rank}(Z) = m$, since S is nonsingular and M_a has full row rank $m - r$. It suffices to show that $\text{im}(X) \supseteq \text{im}(Z)$. Suppose $z = (z_1^T, z_2^T)^T \in \text{im}(Z)$. Then there exists $v = (v_1^T, v_2^T)^T$ such that

$$\begin{aligned} z_1 &= S v_1 + B_a v_2 = [B_a Y_0 + B_b Y(b)] v_1 + B_a v_2 \\ &= B_a (Y_0 v_1 + v_2) + B_b Y(b) v_1, \\ z_2 &= M_a v_2. \end{aligned}$$

Let $u_1 = Y_0 v_1 + v_2, u_2 = Y(b) v_1, u = (u_1^T, u_2^T)^T$. Then $z = Xu$, since $M_a Y_0 = 0$. Therefore $\text{im}(X) \supseteq \text{im}(Z)$ and more specifically $\text{im}(X) = \text{im}(Z)$, implying $\text{rank}(X) = m$. \square

Partially separated boundary conditions. If either B_a or B_b are rank deficient, the number of IVPs to be solved in (3.2) can be reduced to $q + 1$, where $q = \min(\text{rank}(B_a), \text{rank}(B_b))$. In this case the boundary conditions are partially separated. Without loss of generality, assume $\text{rank}(B_b) = q < r$. There exists a nonsingular matrix R such that premultiplying (3.1b) by R yields

$$(3.6a) \quad C_a y(a) = \beta_a,$$

$$(3.6b) \quad C_{ba} y(a) + C_b y(b) = \beta_b,$$

where $\beta_a \in \mathbb{R}^{r-q}$, $\beta_b \in \mathbb{R}^q$, C_a is $(r-q) \times m$ with full row rank $(r-q)$ and C_b is $q \times m$ with full row rank q . If $C_{ba} = 0$, the boundary conditions are completely separated.

From Proposition 3.1, the matrix $[C_a^T, M_a^T]^T$ has full row rank $m-q$. Let D_a be a $q \times m$ matrix such that $U = [C_a^T, M_a^T, D_a^T]^T$ is invertible. We impose the left-hand boundary conditions (3.6a) on the particular solution y_p . Thus, suppose y_p satisfies

$$(3.7) \quad \mathcal{L}y_p(t) = f(t), \quad Uy_p(a) = [\beta_a^T, g(a)^T, \gamma^T]^T$$

for an arbitrary vector $\gamma \in \mathbb{R}^q$ (e.g., take $\gamma = 0$). Partition U^{-1} as

$$U^{-1} = [U_1, U_2, U_a], \quad U_a \in \mathbb{R}^{m \times q},$$

and let $V(t)$ satisfy

$$(3.8) \quad \mathcal{L}V(t) = 0, \quad V(a) = U_a.$$

Note that a unique solution to (3.8) exists, since (3.1a) is solvable and $M_a U_a = 0$ by definition of U_a . Now let

$$(3.9) \quad y(t) = y_p(t) + V(t)\mu, \quad \mu \in \mathbb{R}^q,$$

and impose the q remaining boundary conditions (3.6b) to get a linear system for the parameters μ

$$(3.10) \quad [C_{ba}U_a + C_bV(b)]\mu = \beta_b - (C_{ba}y_p(a) + C_b y_p(b)).$$

To see that $[C_{ba}U_a + C_bV(b)]$ is invertible, note that RS is invertible, where

$$RS = \begin{pmatrix} C_a y_0 \\ C_{ba} Y_0 + C_b Y(b) \end{pmatrix}$$

where Y_0 is any basis for $\ker(M_a)$ and $Y(t)$ is the corresponding homogeneous solution. In particular, we can let Y_0 have the form $[Z(a), U_a]$, hence $Y(b) = [Z(b), V(b)]$. By definition of U_a , $C_a U_a = 0$, which implies

$$RS = \begin{pmatrix} C_a Z(a) & 0 \\ C_{ba} Z(a) + C_b Z(b) & C_{ba} U_a + C_b V(b) \end{pmatrix}.$$

Thus $(C_{ba}U_a + C_bV(b))$ is invertible.

Parallel Shooting. Suppose (3.1) is solvable, and let the nodes $(\tau_j)_0^J$ define a partition of $[a, b]$

$$\tau_0 = a < \tau_1 < \dots < \tau_J = b.$$

On each subinterval $[\tau_{j-1}, \tau_j]$, $j = 1, \dots, J$, the BVP solution $y(t)$ can be represented as

$$(3.11) \quad y(t) = y_j(t) = v_j(t) + V_j(t)\xi_j, \quad t \in [\tau_{j-1}, \tau_j],$$

where v_j is any particular solution to

$$(3.12) \quad \mathcal{L}v_j(t) = f(t), \quad v_j(\tau_{j-1}) = v_j^0 \in M_f(\tau_{j-1}),$$

least one important variation, the *stabilized march technique* [27], is legitimate in that the selections for v_j^0, V_j^0 automatically satisfy the constraint equations. We briefly sketch the details.

Suppose the boundary conditions are partially separated as in (3.6b). Represent $y_j(t)$ as in (3.9). Let $V_1^0 = U_a(m \times q), v_1^0 = y_p(a)$, as in (3.7), (3.8). Inductively define V_{j+1}^0 , for $j = 1, \dots, J$ to be the columns of $V_j(\tau_j)$ orthogonalized, i.e.,

$$(3.18) \quad V_{j+1}^0 = V_j(\tau_j)P_j,$$

where P_j is a $q \times q$ nonsingular upper triangular matrix. But $\text{im}(V_j(\tau_j)) = \text{ker}(M_{\tau_j})$ so that

$$M_{\tau_j}V_{j+1}^0 = 0,$$

that is, V_{j+1}^0 is consistent for the homogeneous problem at $t_0 = \tau_j$. Defining v_{j+1}^0 as the projected component of $v_j(\tau_j)$ in $[\text{im}(V_{j+1}^0)]^\perp = \text{im}(M_{\tau_j}^T)$,

$$(3.19) \quad v_{j+1}^0 = [I - V_{j+1}^0(V_{j+1}^0)^T]v_j(\tau_j),$$

we have $M_{\tau_j}v_{j+1}^0 = M_{\tau_j}v_j(\tau_j) = g(\tau_j)$ so that v_{j+1}^0 is also consistent. Therefore algebraic manipulation of (3.14) leads to the system

$$(3.20) \quad \begin{pmatrix} -I & P_1 & & & & \\ & -I & P_2 & & & \\ & & \ddots & \ddots & & \\ & & & -I & P_J & \\ C_{ba}U_a & & & & C_bV_{J+1}^0 & \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_J \\ \mu_{J+1} \end{pmatrix} = \begin{pmatrix} P_1(V_2^0)^T v_1(\tau_1) \\ P_2(V_3^0)^T v_2(\tau_2) \\ \vdots \\ P_J(V_{J+1}^0)^T v_J(\tau_J) \\ \beta_b - [C_{ba}v_1^0 + C_bv_{J+1}^0] \end{pmatrix},$$

which is square and nonsingular of size $(J + 1)q$. Note that if $q = r$ (i.e., the boundary conditions are completely mixed), then the stabilized march version of (3.14) will be square and invertible also so that it will not be necessary to solve (3.14) as a linear least-squares problem.

Numerical methods. Let $G = (t_n)_0^N$ define a grid on $[a, b]$,

$$t_0 = a, \quad t_n = t_{n-1} + h_n \quad (1 \leq n \leq N), \quad t_N = b,$$

where the stepsizes satisfy the boundedness criterion

$$(3.21) \quad h = \max(h_n) \leq \theta \cdot \min(h_n)$$

for some $\theta > 0$, independent of n, h . Here we briefly discuss the use of numerical initial value methods to approximate the BVP solution $y(t)$ on the grid G by employing the shooting strategy previously described.

Suppose P is a globally $O(h^S)$ -convergent method for the DAE IVP (3.1a), given $O(h^S)$ accurate starting values. Let $(u_{v,n})_{n=0}^N$ ($v = p, 1 \leq v \leq r$) denote the numerical approximations for the $r + 1$ IVPs in (3.2) generated by P . Then

$$(3.22) \quad \begin{aligned} \|u_{p,n} - y_p(t_n)\| &\leq Lh^S, & 0 \leq n \leq N, \\ \|u_{v,n} - \phi_v(t_n)\| &\leq Lh^S, & 0 \leq n \leq N, \quad 1 \leq v \leq r \end{aligned}$$

for some constant $L > 0$ (independent of n , $h \leq h_0$). Define the matrix $U_n = [u_{1,n}, u_{2,n}, \dots, u_{r,n}]$ and let S_h be defined by

$$(3.23) \quad S_h = B_a U_0 + B_b U_N = S + O(h^S),$$

since $U_0 = Y(a)$ and $U_N = Y(b) + O(h^S)$. For $h > 0$ sufficiently small, $S_h^{-1} = S^{-1} + O(h^S)$, hence if ξ_h is the solution to the system

$$(3.24) \quad S_h \xi_h = \beta - [B_a y_p^0 + B_b u_{p,n}] = \beta - [B_a u_{p,0} + B_b u_{p,n}],$$

then $\xi_h = \xi + O(h^S)$, where ξ is the exact solution to the shooting equation (3.3). Note that the initial condition $y(a) = u_{p,0} + U_0 \xi_h$ is consistent if $u_{p,0}$ and U_0 are exact. Therefore, if (u_n) is the numerical approximation to the IVP

$$(3.25) \quad \mathcal{L}y(t) = f(t), \quad y(a) = u_{p,0} + U_0 \xi_h$$

and P is stable under $O(h^S)$ perturbations in the initial starting values, then

$$(3.26) \quad \|u_n - y(t_n)\| = O(h^S), \quad 0 \leq n \leq N,$$

where $y(t)$ is the unique solution to the BVP (3.1).

Clearly, any stable method that converges for the DAE IVP can be used to solve the BVP by shooting. But the class of methods that can be used to solve DAE IVPs is limited. Explicit methods may lead to systems of equations that cannot be solved uniquely for solution vectors at each timestep, and otherwise may not be applicable unless the constraint manifold is explicitly known at each timestep. Symmetric schemes have the problem that for fully-implicit index one and related higher index systems, there is a potential instability. This instability can often be corrected for BVPs by locating some of the consistency conditions at the correct boundary [2]. Even for IVPs, methods for DAEs must be very carefully chosen and implemented.

The traditional ODE methods that have been successful, namely BDF and some implicit Runge-Kutta (IRK) methods, converge and are stable for index one systems [21], [40] but do not converge for all higher index systems except in some cases where the system has a special structure ([20], [13], [17]). Even when these methods do work, they usually exhibit numerical behavior that is not characteristic of the same methods applied to nonstiff ODEs, although there are similarities to stiff ODEs. For example, constant stepsize BDF methods exhibit numerical boundary layers of instability and reduced order (or non-) convergence due to inconsistencies in the starting values. That is, the numerical solutions evolve for some fixed number of steps before achieving the order of convergence expected for ODEs (differential order). The instability is either local or transient in nature and is not a serious problem, unless the stepsize is extremely small. Furthermore, even bounded stepsize variation (3.21) will initiate new boundary layers if the index is greater than 2 [20], [39]. For parallel shooting, these considerations imply the existence of boundary layers on each subinterval $[\tau_{j-1}, \tau_j]$. Thus it is important to take h sufficiently small that the entries in the shooting equations (3.14), (3.20) are accurate to the desired order and not taken from the boundary layer, and large enough so that rounding errors do not dominate.

IRK methods are prone to global order reduction unless the method coefficients satisfy order conditions in addition to the differential order conditions [8], [40], [1], [9]. On the other hand, there exist IRKs that do not exhibit the boundary layer. For example, one can construct extrapolation methods based on the implicit Euler

method by taking enough steps at each stepsize so that the boundary layer has already passed. But here the boundary layer is hidden inside the stages. We do not know whether there exists an IRK method where the intermediate stages do not exhibit boundary layers.

More recently, the (i, j) methods based on (2.2) have been developed and analyzed for linear systems in [11], [12] and extended to nonlinear systems in [10]. The (i, j) methods are based on solving (2.2) for y'_n (and possibly higher order derivatives) in terms of y_n, t_n and then integrating for y_{n+1} using any consistent one-step method that is stable for ODEs. In principle, these methods can be used to solve any singular system that is solvable according to Definition 2.1. However, the (i, j) methods are computation intensive, as they require solving a $(mj) \times (mj)$ singular linear system involving derivatives of the coefficients and input at each step t_n . In practical applications, it may not be easy or even possible to obtain the necessary derivatives, especially if the functions are nonlinear.

4. Finite-difference methods. In this section we show that the more direct approach of finite-difference methods can be used to solve the DAE BVP if the IVP can be numerically solved, and the constraint manifold (2.5) is given at $t = a$. We consider difference approximations to (3.1) of the form

$$(4.1a) \quad \mathcal{L}_h u_j = \sum_{k=0}^N C_{jk}(h) u_k = F_j(h, f) \quad j = 1, 2, \dots, N,$$

$$(4.1b) \quad \mathcal{B}_h \mathbf{u}_h = B_a u_0 + B_b u_N = \beta,$$

$$(4.1c) \quad M_a u_0 = g(a),$$

where $\mathbf{u}_h = \{u_n\}_0^N$ is the approximation to the solution $\{y(t_n)\}_0^N$ and $h = \max(h_n)$ as before. We also assume that (4.1a) satisfies the property that $f \equiv 0$ implies $F_j(h, f) = 0$ for every j, h . In matrix form (4.1) can be written as

$$(4.2) \quad A_h \mathbf{u}_h = \mathbf{F}(h, f),$$

where

$$(4.3) \quad A_h = \begin{pmatrix} \begin{pmatrix} B_a \\ M_a \end{pmatrix} & 0 & \dots & 0 & \begin{pmatrix} B_b \\ 0 \end{pmatrix} \\ C_{1,0} & C_{1,1} & \dots & C_{1,N-1} & C_{1,N} \\ \vdots & \vdots & & \vdots & \vdots \\ C_{N,0} & C_{N,1} & \dots & C_{N,N-1} & C_{N,N} \end{pmatrix},$$

$$\begin{aligned} \mathbf{u}_h &= (u_0^T, u_1^T, \dots, u_N^T)^T, \\ \mathbf{F}(h, f) &= ([\beta^T, g(a)^T]^T, F_1(h, f)^T, \dots, F_N(h, f)^T)^T. \end{aligned}$$

Suppose the local truncation error associated with (4.1a,b) is $O(h^S)$. If (3.1a) is an ODE, then (4.1) is stable if and only if the family of matrices $\{A_h^{-1}\}$ is uniformly bounded as $h \rightarrow 0^+$. Convergence of the method to $O(h^S)$ accuracy then follows from stability and consistency. Furthermore, convergence is independent of the forcing function $f(t)$ and the boundary value β , since A_h is independent of these parameters. Unfortunately, when $E(t)$ is singular A_h^{-1} will in general contain terms

that are unbounded, as $h \rightarrow 0^+$. For example, if (4.1) is a constant stepsize implicit Euler method, then $\|A_h^{-1}\| = O(h^{-\nu})$, where $\nu = \text{ind}(E, F)$. On the other hand, it is interesting to note that with a subtle variation the arguments used in the ODE case to relate convergence of (4.1) to the convergence for IVPs can be used when $E(t)$ is singular.

Consider two solvable DAE BVPs $BV^{(v)}$, $v = 0, 1$, with IVP (3.1a); boundary conditions

$$(4.4) \quad \mathcal{B}^{(v)}y \equiv B_a^{(v)}y(a) + B_b^{(v)}y(b) = \beta;$$

and difference matrices $A_h^{(v)}$, respectively. Before proving the main result in this section, we will need the following simple lemma.

LEMMA 4.1. *Let $y^{(v)}$, $v = 0, 1$, denote the solutions to $BV^{(v)}$ with boundary value β . Furthermore, let $Y^{(0)}$ be the fundamental solution matrix ($n \times r$) to the BVP*

$$(4.5) \quad \begin{aligned} \mathcal{L}Y^{(0)} &= 0, \\ \mathcal{B}^{(0)}Y^{(0)} &= I, \end{aligned}$$

where I is the identity matrix of size $r \times r$. Then $y^{(1)}$ is the unique solution to $BV^{(0)}$, with boundary value

$$(4.6) \quad \tilde{\beta} = (\mathcal{B}^{(1)}Y^{(0)})^{-1}(\beta - \mathcal{B}^{(1)}y^{(0)} + (\mathcal{B}^{(1)}Y^{(0)})\beta).$$

Proof. Express $y^{(1)}$ as

$$y^{(1)}(t) = y^{(0)}(t) + Y^{(0)}(t)\xi,$$

where ξ solves

$$(4.7) \quad (\mathcal{B}^{(1)}Y^{(0)})\xi = \beta - \mathcal{B}^{(1)}y^{(0)}.$$

Now apply the boundary operator $\mathcal{B}^{(0)}$ to $y^{(1)}$ to get

$$\begin{aligned} \mathcal{B}^{(0)}y^{(1)} &= \mathcal{B}^{(0)}y^{(0)} + (\mathcal{B}^{(0)}Y^{(0)})\xi \\ &= \beta + I \cdot (\mathcal{B}^{(1)}Y^{(0)})^{-1}(\beta - \mathcal{B}^{(1)}y^{(0)}) \\ &= (\mathcal{B}^{(1)}Y^{(0)})^{-1}(\beta - \mathcal{B}^{(1)}y^{(0)} + (\mathcal{B}^{(1)}Y^{(0)})\beta). \end{aligned} \quad \square$$

In the remainder of this section we prove the following theorem.

THEOREM 4.1. *Let $BV^{(v)}$, $v = 0, 1$, be solvable BVPs of the form (3.1a), (4.4) with right-hand side f_v and boundary value β_v , respectively. Consider the proposition $P^{(v)}$: $A_h^{(v)}$ is invertible and (4.1) converges to $O(h^S)$ accuracy to the solution of $BV^{(v)}$, independent of f_v and β_v . Then $P^{(0)}$ if and only if $P^{(1)}$.*

Proof. Assume $P^{(0)}$. Applying (4.1) to $BV^{(1)}$ with arbitrary but fixed $f_1 = f$ and $\beta_1 = \beta$ yields the system

$$(4.8) \quad A_h^{(1)}u_h^{(1)} = F(h, f)$$

or equivalently

$$(4.9) \quad [I + D_h(A_h^{(0)})^{-1}]A_h^{(0)}u_h^{(1)} = F(h, f),$$

where $D_h = A_h^{(1)} - A_h^{(0)}$:

$$D_h = \begin{pmatrix} \begin{pmatrix} B_a^{(1)} - B_a^{(0)} & \\ 0 & \\ 0 & \\ \vdots & \\ 0 & \end{pmatrix} & 0 & \cdots & 0 & \begin{pmatrix} B_b^{(1)} - B_b^{(0)} \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}.$$

Thus, to show $A_h^{(1)}$ is invertible it suffices to show the invertibility of $[I + D_h(A_h^{(0)})^{-1}]$.

Partition $(A_h^{(0)})^{-1}$ as $(Z_{jk}^{(0)})$, $0 \leq j, k \leq N$, where each block $Z_{jk}^{(0)}$ is $m \times m$. Then

$$(4.10) \quad I + D_h(A_h^{(0)})^{(-1)} = \begin{pmatrix} Q_{h0} & Q_{h1} & \cdots & Q_{hN} \\ 0 & I & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & I \end{pmatrix},$$

where

$$(4.11) \quad \begin{aligned} Q_{h0} &= I + \begin{pmatrix} B_a^{(1)} - B_a^{(0)} \\ 0 \end{pmatrix} Z_{0,0}^{(0)} + \begin{pmatrix} B_b^{(1)} - B_b^{(0)} \\ 0 \end{pmatrix} Z_{N,0}^{(0)} \\ &= \begin{pmatrix} B_a^{(1)} Z_{0,0}^{(0)} + B_b^{(1)} Z_{N,0}^{(0)} \\ 0 \quad I \end{pmatrix}, \end{aligned}$$

since $B_a^{(0)} Z_{0,0}^{(0)} + B_b^{(0)} Z_{N,0}^{(0)} = (I \ 0)$ by definition of the $Z_{jk}^{(0)}$. Also,

$$(4.12) \quad \begin{aligned} Q_{hj} &= \begin{pmatrix} B_a^{(1)} - B_a^{(0)} \\ 0 \end{pmatrix} Z_{0,j}^{(0)} + \begin{pmatrix} B_b^{(1)} - B_b^{(0)} \\ 0 \end{pmatrix} Z_{N,j}^{(0)} \\ &= \begin{pmatrix} B_a^{(1)} Z_{0,j}^{(0)} + B_b^{(1)} Z_{N,j}^{(0)} \\ 0 \end{pmatrix}, \quad j = 1, \dots, N. \end{aligned}$$

Partition $Z_{j,0}^{(0)} = (Z_{j,0,1}^{(0)} \mid Z_{j,0,2}^{(0)})$, where $Z_{j,0,1}^{(0)}$ is $m \times r$. Then the column of blocks $\{Z_{j,0,1}^{(0)}\}_0^N$ is the solution to

$$(4.13) \quad A_h^{(0)} \mathbf{u}_h = ((I \ 0)^T, 0, \dots, 0)^T,$$

where I is $r \times r$, $(I \ 0)^T$ is $m \times r$, and the remaining zeros are $m \times r$. That is, $\{Z_{j,0,1}^{(0)}\}_0^N$ is the difference approximation to the homogeneous BVP (4.5). Note that solvability of (4.5) follows from the solvability of $BV^{(0)}$. Furthermore, since $BV^{(1)}$ is solvable, the shooting matrix

$$B^{(1)} Y^{(0)} = B_a^{(1)} Y^{(0)}(a) + B_b^{(1)} Y^{(0)}(b)$$

is invertible. Now

$$(4.14) \quad \begin{aligned} B_a^{(1)} Z_{0,0}^{(0)} + B_b^{(1)} Z_{N,0}^{(0)} &= \left(B_a^{(1)} Z_{0,0,1}^{(0)} + B_b^{(1)} Z_{N,0,1}^{(0)} \mid B_a^{(1)} Z_{0,0,2}^{(0)} + B_b^{(1)} Z_{N,0,2}^{(0)} \right) \\ &= (\tilde{Q}_{h,0,1} \mid \tilde{Q}_{h,0,2}). \end{aligned}$$

If (4.1) globally converges to $O(h^S)$ accuracy for $BV^{(0)}$ independent of f and β , then it does so for (4.5), implying $Z_{0,0,1}^{(0)} = Y^{(0)}(a) + O(h^S)$ and $Z_{N,0,1}^{(0)} = Y^{(0)}(b) + O(h^S)$.

Hence $\tilde{Q}_{h,0,1} = \mathcal{B}^{(1)}Y^{(0)} + O(h^S)$ is invertible, for h sufficiently small. It follows that $A_h^{(1)}$ is invertible, since the (1,1) block $Q_{h,0}$ in (4.10) is given by

$$(4.15) \quad Q_{h,0} = \begin{pmatrix} \tilde{Q}_{h,0,1} & \tilde{Q}_{h,0,2} \\ 0 & I \end{pmatrix}.$$

Now we show that (4.1) converges for $BV^{(1)}$. From (4.9) we have that $u_h^{(1)}$ is the unique solution to

$$(4.16) \quad A_h^{(0)}u_h^{(1)} = [I + D_h(A_h^{(0)})^{-1}]^{-1}F = \tilde{F}.$$

From the structure of $[I + D_h(A_h^{(0)})^{-1}]$ given by (4.10), (4.11), and (4.15), we have

$$(4.17) \quad \tilde{F} = ([\tilde{\beta}_h^T, g(a)^T]^T, F_1^T, \dots, F_N^T)^T,$$

where

$$(4.18) \quad \tilde{\beta}_h = \tilde{Q}_{h,0,1}^{-1} \left(\beta - \tilde{Q}_{h,0,2}g(a) - \sum_{j=1}^N (B_a^{(1)}Z_{0,j}^{(0)} + B_b^{(1)}Z_{N,j}^{(0)})F_j \right).$$

That is, $u_h^{(1)}$ is the difference approximation to $BV^{(0)}$, with input f and boundary value $\tilde{\beta}_h$. Simplifying (4.18), we obtain

$$\begin{aligned} \tilde{\beta}_h &= \tilde{Q}_{h,0,1}^{-1} \left(\beta - \left(B_a^{(1)} \left(Z_{0,0,2}^{(0)}g(a) + \sum_{j=1}^N Z_{0,j}^{(0)}F_j \right) \right. \right. \\ &\quad \left. \left. + B_b^{(1)} \left(Z_{N,0,2}^{(0)}g(a) + \sum_{j=1}^N Z_{N,j}^{(0)}F_j \right) \right) \right) \\ &= \tilde{Q}_{h,0,1}^{-1} (\beta - (aB_a^{(1)}(u_{h,0}^{(0)} - Z_{0,0,1}^{(0)}\beta) + B_b^{(1)}(u_{h,N}^{(0)} - Z_{N,0,1}^{(0)}\beta))). \end{aligned}$$

Thus

$$(4.19) \quad \tilde{\beta}_h = \tilde{Q}_{h,0,1}^{-1} (\beta - \mathcal{B}^{(1)}u_h^{(0)} + \tilde{Q}_{h,0,1}\beta),$$

where $u_h^{(0)}$ solves $A_h^{(0)}u_h^{(0)} = F$, i.e., $u_h^{(0)}$ is the difference approximation to $BV^{(0)}$ with input f and boundary value β . Since (4.1) converges for $BV^{(0)}$ independent of f, β by hypothesis, $\tilde{\beta}_h$ is bounded independent of h for small $h > 0$.

Let $y^{(v)}, v = 0, 1$, denote the exact solution to $BV^{(v)}$ with input f and boundary value β . From (4.19), we have

$$(4.20) \quad \tilde{\beta}_h = \tilde{\beta} + \Delta,$$

where

$$(4.21) \quad \tilde{\beta} = (\mathcal{B}^{(1)}Y^{(0)})^{-1}(\beta - (\mathcal{B}^{(1)}y^{(0)} + \mathcal{B}^{(1)}Y^{(0)})\beta)$$

and $\|\Delta\| = O(h^S)$. From Lemma 4.1, $y^{(1)}$ is the solution to $BV^{(0)}$, with boundary value $\tilde{\beta}$. We wish to estimate $\|u_{h,j}^{(1)} - y^{(1)}(t_j)\|$. By the triangle inequality,

$$(4.22) \quad \|u_{h,j}^{(1)} - y^{(1)}(t_j)\| \leq \|u_{h,j}^{(1)} - \tilde{y}_h(t_j)\| + \|\tilde{y}_h(t_j) - y^{(1)}(t_j)\|,$$

where \tilde{y}_h is the exact solution to $BV^{(0)}$ with input f and boundary value $\tilde{\beta}_h$. By the argument immediately following (4.18) and hypothesis $P^{(0)}$, we have that $\|u_{h,j}^{(1)} - \tilde{y}_h(t_j)\| = O(h^S)$. Furthermore, $\|\tilde{y}_h(t_j) - y^{(1)}(t_j)\| = O(h^S)$ follows by variation of parameters using (4.20) and the uniform boundedness of $\|Y^{(0)}(t)x\|$ on $[a, b]$, i.e.,

$$\begin{aligned} \|\tilde{y}_h(t_j) - y^{(1)}(t_j)\| &= \|Y^{(0)}(t_j)(\tilde{\beta}_h - \tilde{\beta})\| \\ &\leq \|Y^{(0)}(t_j)\| \|\Delta\| \\ &= O(h^S). \end{aligned} \quad \square$$

We note that (4.1) need not be globally convergent in order for Theorem 4.1 to remain a valid result in the context of DAEs, where boundary layers may exist. We only require $\|Z_{0,0,1}^{(0)} - Y^{(0)}(a)\| = O(h^S)$ and N to be sufficiently large such that $\|Z_{N,0,1}^{(0)} - Y^{(0)}(b)\| = O(h^S)$. Thus, by associating $BV^{(0)}$ with the IVP

$$(4.23) \quad \mathcal{L}y = f, \quad y_0 \in M_f(a),$$

and taking $B_a^{(0)}$ so that $[B_a^T, M_a^T]^T$ is invertible, we get Corollary 4.1.

COROLLARY 4.1. *Suppose (4.1) converges with $O(h^S)$ accuracy to a solution of (4.23) for $n \geq J$ when initial values are consistent to $O(h^S)$. Then (4.1) converges to a solution of the boundary value problem $BV^{(1)}$ to $O(h^S)$ accuracy for $n \geq J$.*

5. Numerical experiments. In this section we present the results of some numerical experiments on linear and nonlinear index one and index two DAE BVPs. The experiments confirm the results of the theory and also raise some interesting questions.

The numerical experiments described in this section were all performed using the finite-difference methods formulated as described in § 4. The nonlinear equations at each timestep were solved by Newton iteration. The iteration was terminated when the ℓ_2 norm of the difference between two successive iterates was less than a specified tolerance. An analytic iteration matrix was provided to the code for all of the problems. All of the computations were performed in double precision on an Alliant FX/8 computer.

The first test problem was a linear variable-coefficient index one DAE BVP on $[0,1]$ given by

$$(5.1) \quad \begin{pmatrix} 1 & -t & t^2 \\ 0 & 1 & -t \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} y'_1 \\ y'_2 \\ y'_3 \end{pmatrix} + \begin{pmatrix} 1 & -(t+1) & (t^2+2t) \\ 0 & -1 & (t-1) \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \sin(t) \end{pmatrix},$$

with boundary conditions

$$\begin{aligned} y_1(0) &= 1, \\ y_2(1) - y_3(1) &= e. \end{aligned}$$

This problem has true solution

$$\begin{aligned} y_1 &= e^{-t} + te^t, \\ y_2 &= e^t + t \sin(t), \\ y_3 &= \sin(t). \end{aligned}$$

This system is related to the linear constant coefficient DAE system

$$\begin{aligned} y_1' &= -y_1, \\ y_2' &= y_2, \\ y_3 &= \sin(t), \end{aligned}$$

with boundary conditions $y_1(0) = 1, y_2(1) = e$ by the nonsingular spatially dependent change of variables $y = Q\tilde{y}$, where

$$Q = \begin{pmatrix} 1 & t & 0 \\ 0 & 1 & t \\ 0 & 0 & 1 \end{pmatrix}.$$

This problem was constructed to show the effects, if any, of spatially dependent coupling between the differential and algebraic parts of the system on solution accuracy.

We solved this linear problem using the implicit midpoint method, with the number of steps in $[0,1]$ ranging from 2 to 64. We computed the approximate order by comparing the errors of the solution with a given mesh spacing to the errors obtained by halving the mesh spacing. In our tests, the error behaved consistently as $O(h^2)$. It is to be expected from Corollary 4.2 (together with the results in [1], [9]) that the implicit midpoint method is globally $O(h^2)$ when applied directly to index one DAE IVPs.

Next we solved a simplified model of a steady-state semiconductor device described in [1] with the same method. The system has the form

$$\begin{aligned} 0 &= n - p - C(t), \\ J_n' &= 0, \\ J_p' &= 0, \\ J_n &= (n' - n\psi'), \\ J_p &= -(p' + p\psi'), \end{aligned} \tag{5.2}$$

and the boundary conditions are given by

$$\begin{aligned} n &= 1/2(C + \sqrt{C^2 + 4\delta^4}), & t = +1, -1, \\ p &= 1/2(-C + \sqrt{C^2 + 4\delta^4}), & t = +1, -1, \\ \psi &= \psi_{bi}(t) + \frac{1}{2}(t+1)V, & t = +1, -1, \end{aligned}$$

where $\psi_{bi}(t) = \ln n(t) - \ln \delta^2$. For the free parameters and functions, we have chosen $\delta = 10^{-4}$, $V = 1$, and $C(t) = 1/2 + (\tan^{-1}(\lambda t))/\pi$, where λ is a parameter that determines the steepness of $C(t)$. $C(t)$ is constructed to be an approximation to a square wave for sufficiently large λ .

We rewrote the problem in a form that was easier to understand by performing a nonsingular constant change of variables $J_+ = J_n + J_p, J_- = J_n - J_p, N_+ = n + p, N_- = n - p$ to arrive at the system

$$\begin{aligned} 0 &= N_- - C(t), \\ J_+' &= 0, \\ J_-' &= 0, \\ J_+ &= (N_+' - N_+\psi'), \\ J_- &= (N_+' - N_-\psi'), \end{aligned} \tag{5.3}$$

with boundary conditions

$$\begin{aligned} N_+ &= \sqrt{C^2 + 4\delta^4}, & t = +1, -1, \\ \psi &= \psi_{bi} + \frac{1}{2}(t+1)V, & t = +1, -1. \end{aligned}$$

This is the proper number of boundary conditions for this problem, which is index one. Note that in the original formulation of (5.2), two of the boundary conditions are redundant. The numerical solution of (5.3) will be the same as that for (5.2), apart from errors due to roundoff, because the problems are related by a constant nonsingular change of variables.

We can substitute $N_- = C(t)$ into the system (5.3), to obtain the related ODE system

$$(5.4) \quad \begin{aligned} J'_+ &= 0, \\ J'_- &= 0, \\ \psi' &= (C'(t) - J_+)/N_+, \\ N'_+ &= J_- + C(t)(C'(t) - J_+)/N_+, \end{aligned}$$

which we will refer to later.

We solved (5.3) for $\lambda = 20$ with a fixed stepsize and the number of steps ranging from 2 to 2056. We found that for large stepsizes (fewer than approximately 64 steps) and for an odd number of nodes in the interval, there is a problem with oscillations in the solution for $t > 0$. After the stepsize is decreased sufficiently so that the solution is resolved, there is no longer a problem with oscillations, and the error behaves as $O(h^2)$. For an even number of nodes, there is no problem with oscillations, and the error behaves as $O(h^2)$. We observed similar behavior for larger values of λ , where the oscillation disappears when there are a sufficient number of meshpoints to accurately resolve the solution.

There is an explanation for this oscillating behavior, which occurs only for an odd number of nodes. In this test problem, the most rapid change in the solution occurs at $t = 0$. For an even number of meshpoints, $t = 0$ is in the center of a mesh interval, whereas for an odd number of meshpoints there is a meshpoint at $t = 0$. It is easy to see that for the midpoint method applied to an algebraic equation whose solution is a step function, the solution is smooth if the step occurs at the center of a mesh interval, and it oscillates otherwise. Since $C(t)$ approximates a step function, the behavior we have observed with respect to odd and even numbers of meshpoints is what we would expect.

Of course, in this test problem if we did not discretize the algebraic equation with the implicit midpoint method but instead evaluated N_- always at the meshpoints, then the oscillation would disappear. This would give the same results as solving the related ODE system (5.4) with the implicit midpoint method. However, it is not always so easy to isolate the algebraic variable in applications, so we are interested in seeing the effects of not treating that variable specially.

In comparison with the midpoint method applied to the related ODE, the ODE formulation does not exhibit any oscillations. On the other hand, the DAE formulation is apparently much less sensitive to perturbations in the initial guess for N_+ . In our experiments, the ODE solution only converged for initial guesses for N_+ that were

very near the exact solution for that variable, whereas the DAE formulation converged for a much wider range of initial values.

The third problem that we tested was a linear index two system on $[0,1]$ given by

$$\begin{aligned}
 (5.5) \quad y_1' - t &= -y_1 + (1 - t)y_2, \\
 0 &= \beta y_1 + (-1 - \beta t)y_2 + \sin(t), \\
 z_1 &= y_1', \\
 z_2 &= y_2',
 \end{aligned}$$

with boundary conditions

$$\begin{aligned}
 y_1(0) &= 1, \\
 \beta y_1(1) + (-1 - \beta)y_2(1) &= -\sin(1).
 \end{aligned}$$

This problem is an index two extension of an index one problem proposed by Ascher [2]. The boundary conditions are posed such that the midpoint method is stable for the index one part of the problem. We solved the problem for $\beta = 10$ and for $\beta = 100$. The errors in y_1 and y_2 behaved consistently as $O(h^2)$. However, the midpoint method is not convergent for the index two variables z_1 and z_2 , even for IVPs, and for this problem the solutions for these variables exhibited oscillations whose amplitudes did not decrease as $h \rightarrow 0$. By modifying the formula slightly and discretizing the subsystem $z = y'$ by

$$z_n = \frac{y_{n+1} - y_{n-1}}{2h},$$

the formula becomes $O(h^2)$ accurate for z , IVPs and BVPs. In our experiments the modified formula was $O(h^2)$ accurate for y and z , and the numerical results exhibited no oscillations. These experiments confirm the results of § 4.

One interesting point on this index two example is that the condition number of the matrix generated by straightforward application of the finite-difference technique is $O(h^{-4})$. We experimented with several different scalings of the DAE system that reduced the condition number to $O(h^{-3})$, but the scaling had almost no effect on the errors that were obtained.

It is possible to transform a semi-explicit index two system

$$\begin{aligned}
 y' &= f(y, z), \\
 0 &= g(y, z)
 \end{aligned}$$

to an index one system

$$\begin{aligned}
 y' &= f(y, w'), \\
 0 &= g(y, w')
 \end{aligned}$$

coupled with $w' = z$ [19]. Using this transformation and then scaling the resulting matrix reduces the condition number to $O(h^{-1})$. However, in our experiments we found that the errors in z did not change appreciably from the other formulations.

The final problem that we tested was a linear index two system. This problem has the property that the matrix $E(t)$ is not of constant rank. The problem is given

on $[0, 1]$ by

$$\begin{pmatrix} (2+t) & 1 & -t \\ -2 & -1 & 0 \\ -t(t+1) & 0 & t(t+1) \end{pmatrix} y' + \begin{pmatrix} (1-t^2) & 2 & (t^2-1) \\ -3 & -1 & 1 \\ (2+t) & -(1+t) & t \end{pmatrix} y = \begin{pmatrix} \sin(t) + e^{-t}t^2 - e^t \\ e^{-t} - \sin(t) \\ e^t(t+1) - e^{-t} \end{pmatrix},$$

with boundary conditions

$$\begin{pmatrix} 3 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 1 \end{pmatrix} y(0) + \begin{pmatrix} -2 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} y(1) = \begin{pmatrix} -(e + \frac{3}{2}) \\ 1 \\ 1 \end{pmatrix}.$$

This problem can be obtained via a nonsingular constant change of variables and a nonsingular time-dependent scaling of the system from a simpler index two system for which the matrix $E(t)$ is also not of constant rank. The results of Clark [18] imply that the implicit Euler method converges with order $O(h)$ for IVPs of this type. Thus the results in § 4 imply that the implicit Euler method, formulated as in § 4, should yield $O(h)$ accuracy for this problem. Our numerical experiments confirm these conclusions. It should be noted that it is possible to use a one-sided difference scheme for this problem because it is not stiff, and also that there are index two problems that are not in semi-explicit form for which the implicit Euler method, as well as more general Runge–Kutta and multistep methods, is not stable [21]. The results of this paper imply that a method is convergent for the BVP if and only if it is convergent for the related IVP. They do not make any statements about which methods are convergent for the IVP.

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

REFERENCES

- [1] U. ASCHER, *On numerical differential algebraic problems with application to semiconductor device simulation*, SIAM J. Numer. Anal., 26(1989), pp. 517-538.
- [2] ———, *On symmetric schemes and differential-algebraic equations*, SIAM J. Sci. Statist. Comput., this issue, pp. 937-949.
- [3] U. ASCHER, R. M. MATTHEIJ, AND R. RUSSELL, *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, Englewood Cliffs, NJ, Prentice Hall, 1988.
- [4] J. BETTS, T. BAUER, W. HUFFMAN, AND K. ZONDERVAN, *Solving the optimal control problem using a nonlinear programming technique, Part I: General formulation*, AIAA-84-2037, Proc. 1984 AIAA/ AAS Astrodynamics Conference, Seattle, WA, 1984.
- [5] H. G. BOCK, E. EICH, AND J. P. SCHLÖDER, *Numerical Solution of Constrained Least Squares Boundary Value Problems in Differential-Algebraic Equations*, Universität Heidelberg, Preprint No. 440, December 1987.
- [6] K. BRENNAN, *Numerical simulation of trajectory prescribed path control problems*, IEEE Trans. Automat. Control, AC-31, (1986), pp. 266-269.
- [7] K. BRENNAN AND B. ENGQUIST, *Backward differentiation approximations of nonlinear differential-algebraic systems*, Aerospace Corp. Report ATR-85(9990)-5, 1985.
- [8] K. E. BRENNAN AND L. R. PETZOLD, *The numerical solution of higher index differential-algebraic equations by implicit Runge–Kutta methods*, UCRL-95905, Lawrence Livermore National Laboratory, 1986; SIAM J. Numer. Anal., 26(1989), pp. 981-1001.

- [9] K. BURRAGE AND L. PETZOLD, *On order reduction for Runge-Kutta methods applied to differential-algebraic systems and to stiff systems of ODEs*, UCRL-98046, Lawrence Livermore National Laboratory, Livermore, CA, 1988.
- [10] S. L. CAMPBELL, *A computational method for general higher index nonlinear singular systems of differential equations*, CSRC Tech. Report 063087-01, Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC, 1987.
- [11] ———, *A general form for solvable linear time varying singular systems of differential equations*, SIAM J. Math. Anal., 18 (1987), pp. 1101-1115.
- [12] ———, *The numerical solution of higher index, linear time varying singular systems of differential equations*, SIAM J. Sci. Statist. Comput., 6(1985), pp. 334-348.
- [13] ———, *One canonical form for higher index, linear time varying singular systems*, Circuits Systems Signal Process., 2(1983), pp. 311-326.
- [14] ———, *Singular Systems of Differential Equations II*, Pitman, Marshfield, MA, 1982.
- [15] S. L. CAMPBELL AND C. D. MEYER, JR., *Generalized Inverses of Linear Transformations*, Pitman, Marshfield, MA, 1979.
- [16] K. CLARK, *The numerical solution of some higher index time varying semistate systems by difference methods*, Circuits Systems and Signal Process., 6(1987), pp. 261-275.
- [17] ———, *A structural form for higher index semistate equations I: Theory and applications to circuit and control theory*, Linear Algebra Appl., 98 (1988), pp. 169-197.
- [18] ———, *On structure and the numerical solution of singular systems*, in Recent Advances in Singular Systems, Proc. International Conference on Singular Systems, Atlanta, GA, 1987, pp. 41-45.
- [19] C. W. GEAR, *Differential-algebraic equation index transformations*, SIAM J. Sci. Statist. Comput., 9(1988), pp. 39-47.
- [20] C. W. GEAR, G. K. GUPTA, AND B. LEIMKUEHLER, *Automatic integration of Euler-Lagrange equations with constraints*, J. Comput. Appl. Math., 12&13(1985), pp. 77-90.
- [21] C. W. GEAR AND L. R. PETZOLD, *ODE methods for the solution of differential/algebraic systems*, SIAM J. Numer. Anal., 21(1984), pp. 716-728.
- [22] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The John Hopkins University Press, Baltimore, MD, 1983.
- [23] E. GRIEPENTROG AND R. MÄRZ, *Differential-Algebraic Equations and Their Numerical Treatment*, Teubner-Texte zur Mathematik, Band 88, Leipzig, GDR, 1986.
- [24] M. HANKE, *On a least-squares collocation method for linear differential-algebraic equations*, Numer. Math., 54 (1988), pp. 79-90.
- [25] R. J. KEE, L. R. PETZOLD, M. D. SMOOKE, AND J. F. GRGAR, *Implicit methods in combustion and chemical kinetics modeling*, in Multiple Time Scales, J. Brackbill and B. Cohen, eds., Academic Press, New York, 1985, pp. 113-144.
- [26] H. B. KELLER, *Approximation methods for nonlinear problems with application to two-point boundary value problems*, Math. Comp., 29(1975), pp. 464-474.
- [27] ———, *Numerical Solution of Two-Point Boundary Value Problems*, Regional Conference Series in Applied Mathematics, Vol. 24, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1976.
- [28] I. J. KIRBY AND G. A. LEIPER, *A small divergent detonation theory for intermolecular explosives*, Proc. 8th International Symposium on Detonation, Albuquerque, NM, (1985), pp. 760-768.
- [29] M. LENTINI AND R. MÄRZ, *The condition of boundary value problems in transferable differential-algebraic equations*, Humboldt-Universität, Sect. Math., Berlin, GDR, 1987.
- [30] ———, *Conditioning and dichotomy in differential algebraic equations*, Humboldt-Universität, Sect. Math., Berlin, GDR, 1988.
- [31] P. LÖTSTEDT, *Discretization of singular perturbation problems by BDF methods*, Report No. 99, Dept. Computer Science, Uppsala University, Uppsala, Sweden, 1985.
- [32] ———, *On the relationship between singular perturbation problems and differential-algebraic equations*, Report No. 100, Dept. Computer Science, Uppsala University, Uppsala, Sweden 1985.
- [33] P. LÖTSTEDT AND L. R. PETZOLD, *Numerical solution of nonlinear differential equations with algebraic constraints: Convergence results for the backward differentiation formulas*, Math. Comp., 46(1986), pp. 491-516.
- [34] R. MÄRZ, *On difference and shooting methods for boundary value problems in differential/algebraic equations*, Preprint No. 24, Humboldt-Universität Sect. Math. Berlin, GDR, 1982.
- [35] N. H. MCCLAMROCH, *Singular systems of differential equations as dynamic models for constrained robot systems*, Tech. Report RSD-TR-2-86, Center for Research on Integrated Manufacturing, College of Engineering, Univ. Michigan, Ann Arbor, MI, 1986.

- [36] R. W. NEWCOMB, *The semistate description of nonlinear time variable circuits*, IEEE Trans. Circ. Sys., CAS-28(1981), pp. 62-71.
- [37] B. OWREN, personal communication, 1987.
- [38] J. F. PAINTER, *Solving the Navier-Stokes equations using LSODI and the method of lines*, Report UCID-19262, Lawrence Livermore National Laboratory, Livermore, CA, 1981.
- [39] L. R. PETZOLD, *Differential-algebraic equations are not ODEs*, SIAM J. Sci. Statist. Comput., 3(1982), pp. 367-384.
- [40] ———, *Order results for implicit Runge-Kutta methods applied to differential/algebraic systems*, SIAM J. Numer. Anal., 23(1986), pp. 837-852.
- [41] L. R. PETZOLD AND P. LÖTSTEDT, *Numerical solution of nonlinear differential equations with algebraic constraints II: Practical implications*, SIAM J. Sci. Statist. Comput., 7(1986), pp. 720-733.
- [42] W. RHEINBOLDT, *Differential-algebraic systems as differential equations on manifolds*, Math. Comp., 4(1984), pp. 473-482.
- [43] R. SINCOVEC, A. M. ERISMAN, E. L. YIP, AND M. A. EPTON, *Solvability of large scale descriptor systems*, Final Report, Boeing Computer Services Co., Tukwila, WA, 1979.