ON NUMERICAL DIFFERENTIAL ALGEBRAIC PROBLEMS WITH APPLICATION TO SEMICONDUCTOR DEVICE SIMULATION

Uri Ascher

Technical Report 87-27

July 1987

Abstract

This paper considers questions of conditioning of and numerical methods for certain differential algebraic equations subject to initial and boundary conditions. The approach taken is that of separating "differential" and "algebraic" solution components, at least theoretically.

This yields conditioning results for differential algebraic boundary value problems in terms of "pure" differential problems, for which existing theory is well-developed. We carry the process out for problems with (global) index 1 or 2.

For semi-explicit boundary value problems of index 1 (where solution components are separated) we give a convergence theorem for a special class of collocation methods. For general index 1 problems we discuss advantages and disadvantages of certain symmetric difference schemes. For initial value problems with index 2 we discuss the use of BDF schemes, summarizing conditions for their successful and stable utilization.

Finally, the present considerations and analysis are applied to two problems involving differential algebraic equations which arise in semiconductor device simulation.

Subject classification: AMS(MOS): 65L10.

Keywords: Differential algebraic equations, initial value problems, boundary value problems, conditioning, index, collocation, BDF, semiconductor device simulation.

1. INTRODUCTION

Often in applications, a differential problem is naturally cast in the form

$$\phi(t,\mathbf{x},\mathbf{x}') = 0 \qquad a<t<b, \tag{1}$$

subject to some boundary conditions at one or more points, i.e. $\mathbf{x}' \equiv \frac{d\mathbf{x}}{dt}$ appears implicitly, rather than explicitly as in the more usual form required by most standard software. When considering classes of problems of this form, the matrix

$$E(t) := \frac{\partial \phi(t,\mathbf{x}(t),\mathbf{x}'(t))}{\partial \mathbf{x}'} \tag{2}$$

plays a crucial role. If $E(t)$ is nonsingular, $a<t<b$, then (1) can be converted, at least in principle, to an explicit ODE form, and standard numerical ODE schemes are expected to work as usual. (In practice, however, a conversion of the nonlinear ODE (1) to an explicit form may not always be simple or even possible. Still, while standard software may be inapplicable, we know at least how to choose and implement suitable methods.)

Here we consider cases where $E(t)$ is singular, which means that in (1) there is a mixture of *differential and algebraic equations* (DAEs). Such equations have been the topic of an intense analytic and numerical investigation recently, especially for initial value problems (IVPs); see, for instance, [Ge], [GePe], [Ca], [Pe], [Ma1], [Ma2] and the monograph [GrMa].

The simplest instance of DAEs is when the differential and the algebraic equations are separated, so we have

$$0 = \mathbf{f}(t,\mathbf{y},\mathbf{z}) \tag{3a}$$

$$\mathbf{z}' = \mathbf{g}(t,\mathbf{y},\mathbf{z}) \tag{3b}$$

where $\mathbf{y}(t)\in\mathbf{R}^{n_y}$, $\mathbf{z}(t)\in\mathbf{R}^{n_z}$, $n=n_z+n_y$ and we assume that $U^{11}(t)\equiv\frac{\partial \mathbf{f}(t,\mathbf{y}(t),\mathbf{z}(t))}{\partial \mathbf{y}}$ is non-

singular for all $a<t<b$. Hence $E(t)=\begin{pmatrix} 0 & 0 \\ 0 & I_{n_z} \end{pmatrix}$, where I_l denotes the $l\times l$ identity matrix. The system (3) is subject to n_z boundary conditions (BC) involving $\mathbf{y}$ and $\mathbf{z}$. Such a system arises, for example, when attempting to find the reduced solution of a singularly perturbed system with separated (decoupled) fast and slow components, i.e. when we set $\epsilon=0$ in

$$\epsilon \mathbf{y}' = \mathbf{f}(t,\mathbf{y},\mathbf{z}) \tag{4}$$

with (3b) holding.

Despite the efforts reported in the literature, there are still many questions regarding the numerical solution of DAEs which presently are not answered satisfactorily, especially for boundary value problems (BVPs). The class of equations (1), (2) with $E(t)$ singular contains problems with a very wide variety in difficulty, from essentially trivial to very difficult to solve. An attempt must therefore be made to define subclasses of easy and of manageable DAEs. This is usually done in terms of the (global) index [GePe], [GrMa], [LoPe]. DAEs with index 1 (or transferable DAEs, of which (3) is a special case) are generally well-understood theoretically, and corresponding initial value problems are numerically handled, using BDF schemes, with ease comparable (almost) to that of handling stiff ODEs. A general-purpose code is also available [Pe]. In contrast, there are no comparably good methods developed for boundary value DAEs with index 1 (see [GrMa]).

DAEs of a higher index are less well-understood and are more difficult to solve numerically. They are noted to be ill-posed (see, e.g. [GrMa]), so corresponding numerical processess are ill-conditioned, and some efforts to regularize them have been reported (e.g. [Ma], [Ha], [Ca]). However, these efforts do not appear at the time of this writing to have yielded effective general numerical methods for practical problems of this type. Indeed, the existence of practical problems representing stable physical processes which are modeled as such "ill-posed" mathematical

equations (e.g. [PeLo]) may require some reflection. Moreover, such initial value problems are often solved successfully in practice by a BDF-based method [LoPe], [PeLo]. A careful consideration, such as given e.g. by [GrMa] and the other papers cited, is required regarding the nature of this difficulty, which goes beyond the mere increase in roundoff error accumulation.

Let us explain this last point further: To be sure, increased roundoff error amplification is expected as the global index increases. Generally, if a discretization with a mesh

$$\pi : a = t_1 < t_2 < \cdots < t_N < t_{N+1} = b \tag{5a}$$

is used, with

$$h_i := t_{i+1} - t_i, \quad h := \max_{1 \le i \le N} h_i, \tag{5b}$$

then for a problem with index m, an $O(\sum_{i=1}^{N} h_i^{-m+1})$ roundoff error amplification is expected [GrMa]. (This amplification is, incidentally, unbounded as $h \to 0$ also when $m=1$, but it does not depend on $\min_i h_i$ then.) But this alone is not necessarily a practical reason to bring in the heavy guns of regularization, with the inherent loss of information they entail. A similar roundoff error amplification is obtained when discretizing directly a scalar ODE of order m. A reformulation of the latter into a first order system, which brings the roundoff error accumulation down to $O(N)$, is usually found practically unnecessary. (A particularly poignant case in point is the code COLSYS [ACR] compared to its newer version [BaAs] which realizes such an improvement of roundoff error accumulation with the *same* discretization. This improvement is usually not crucial when working with ~15 decimal digits, but see examples and discussion in [APR], [AsBa].) Finding the numerical solution of a DAE problem with index 3, say, often presents much more severe obstacles than solving directly for an ODE of order 3.

In this paper we take the point of view that a reasonable class of DAEs to consider is such that, upon separation of differential and algebraic equations and solution components, yields a well-conditioned boundary value problem for the differential part. This excludes many pathological examples which have appeared in the literature for higher index problems. We develop conditions and conditioning constants (bounds) resulting from this criterion for DAEs with index 1 and 2 (§§ 3,4). (The principle applies to higher index problems as well.)

Consider a linear form of (1)

$$E(t)\mathbf{x}' = A(t)\mathbf{x} + \mathbf{q}(t). \tag{6}$$

We further emphasize, as in [GrMa], the smoothness of the solution components $\mathbf{x}(t)$ corresponding to the smoothness of the inhomogeneity components $\mathbf{q}(t)$. This we follow also when considering numerical methods, by considering mostly collocation methods in piecewise polynomial spaces (see, e.g., [dBSw], [As2]). Such collocation methods enjoy also the interpretation of implicit Runge-Kutta schemes, and include basic schemes like backward Euler, the trapezoidal and the midpoint schemes. Results like order reduction for certain Runge-Kutta schemes are naturally understood in the restricted collocation context (see [AsWe1], [AsWe2], [AsWe3]).

In §2 we consider the simple semi-explicit DAE case and apply the old idea of collocating for different solution components in different piecewise polynomial spaces. This yields the convergence theorem 12, which to our knowledge has not appeared before. In §3.1 we consider briefly use of marginally stable symmetric difference schemes for DAEs of index 1, and in §4.1 we look at what is needed for BDF schemes to work for initial value DAEs of index 2. In both cases a decisive consideration is whether or not the system can be brought into a semi-explicit form (where the differential and algebraic solution components are decoupled) by a constant transformation. For, none of the usual discretization schemes models a time-dependent

decoupling transformation (specifically, the term $T^{-1}T'$ in (15a) below) very well, and the effects of this get worse as the index increases.

In §5 we describe two applications of these considerations to the semiconductor device equations (see, e.g., [Mar]). First we look at a boundary value DAE of index 1 for the one-dimensional steady state problem. Then we look at the time dependent problem (with a few space variables) which, after applying the method of lines, is an initial value DAE of index "almost 2", and briefly evaluate some numerical methods which have been used for its solution. It should be remarked that numerical experience suggests that the time dependent problem is "easier" in some sense to solve than its steady state counterpart, and the latter has been shown under certain restrictions to be well-conditioned [AMSSW].

2. Simple, semi-explicit DAEs

We begin with a simple case to illustrate some ideas. Consider the DAE system (3) with $\frac{\partial \mathbf{f}(t,\mathbf{y}(t),\mathbf{z}(t))}{\partial \mathbf{y}}$ nonsingular for all $a<t<b$. This DAE does not cause much difficulty, theoretically or numerically. In fact, the numerical solution of (3) can be much easier than that of (4),(3b). It is often possible and desirable to use (3a) to eliminate $\mathbf{y}$ in terms of $\mathbf{z}$ and substitute into (3b) to obtain a system of order n_z in standard form. Even if such an explicit elimination is not possible an implicit one is, upon noticing that in (3) $\mathbf{z}(t)$ is generally one derivative smoother than $\mathbf{y}(t)$. Note that (3) is subject to n_z (not n) BC. These correspond to the integration constants in (3b), so for an IVP, $\mathbf{z}(a)$ alone (and not $\mathbf{y}(a)$) should be prescribed. For the linear BVP

$$0 = U^{11}(t)\mathbf{y} + U^{12}(t)\mathbf{z} + \mathbf{g}^1(t) \tag{7a}$$

$$\mathbf{z}' = U^{21}(t)\mathbf{y} + U^{22}(t)\mathbf{z} + \mathbf{g}^2(t) \tag{7b}$$

$$B_a\begin{pmatrix}\mathbf{y}(a)\\ \mathbf{z}(a)\end{pmatrix} + B_b\begin{pmatrix}\mathbf{y}(b)\\ \mathbf{z}(b)\end{pmatrix} = \beta, \qquad B_a, B_b \in \mathbf{R}^{n_z \times n}, \tag{7c}$$

with $U^{11}(t)$ nonsingular, we have from (7a)

$$\mathbf{y}(t) = -[U^{11}(t)]^{-1}[U^{12}(t)\mathbf{z}(t) + \mathbf{g}^1(t)], \tag{7d}$$

and this can be substituted into (7b) and (7c) to obtain a regular BVP of order n_z for $\mathbf{z}(t)$. If this latter BVP has a unique solution then the BC (7c) are said to be *consistent*. We will *assume* not only that the BC are consistent, but that the BVP for $\mathbf{z}$ is well-conditioned. Thus, if we write the BVP for $\mathbf{z}$ as

$$\mathbf{z}' = \hat{A}(t)\mathbf{z} + \hat{\mathbf{q}}(t), \tag{8a}$$

$$\hat{B}_a\mathbf{z}(a) + \hat{B}_b\mathbf{z}(b) = \hat{\beta} \tag{8b}$$

(matrices appearing in (8) are all $n_z \times n_z$), then there exists a constant κ of moderate size such that

$$||\mathbf{z}||_\infty \leq \kappa(|\hat{\beta}| + ||\hat{\mathbf{q}}||_1)$$

(see, e.g., [AMR, §3.2]). Translating this back to the DAE problem (7), and assuming that there are constants K_j such that

$$||U^{21}(\cdot)[U^{11}(\cdot)]^{-1}||_\infty \leq K_1, \quad ||[U^{11}(\cdot)]^{-1}||_\infty \leq K_2, \quad ||[U^{11}(\cdot)]^{-1}U^{12}(\cdot)||_\infty \leq K_3,$$

and that the boundary matrices are scaled to 1, we obtain the stability bounds

$$||\mathbf{z}||_\infty \leq \kappa(||\mathbf{g}^2||_1 + K_1||\mathbf{g}^1||_1 + |\beta| + K_2(|\mathbf{g}^1(a)| + |\mathbf{g}^1(b)|)), \tag{9a}$$

$$||\mathbf{y}||_\infty \leq K_3||\mathbf{z}||_\infty + K_2||\mathbf{g}^1||_\infty. \tag{9b}$$

Above we have assumed that the coefficient matrices U^{jl} are sufficiently smooth and that $\mathbf{g} \equiv \begin{pmatrix}\mathbf{g}^1\\ \mathbf{g}^2\end{pmatrix} \in L_\infty(a,b)$ (more carefully, $\mathbf{g}^1 \in L_\infty$ and $\mathbf{g}^2 \in L_1$). From (9) and (7) it is then

clear that $\mathbf{y}$ is as smooth as $\mathbf{g}^1$ is and $\mathbf{z}'$ is as smooth as $\mathbf{g}^2$ is.

Generally, $\mathbf{z}$ is one derivative smoother than $\mathbf{y}$. It is natural to require that this be reflected in the numerical approximation, as is suggested when collocation is considered. We now recall this method for a scalar ODE of order m,

$$\frac{d^m u}{dt^m} \equiv u^{(m)} = f(t,u,u',\ldots,u^{(m-1)}), \quad a<t<b, \tag{10}$$

subject to m boundary conditions. We say that a function v is in $\mathbf{P}_{k+m}$ if $v(t)$ is a polynomial of order $k+m$ (degree $< k+m$) on an appropriate interval, and that v is in $\mathbf{P}_{k+m,\pi}$ if $v(t)$ is a piecewise polynomial which is in $\mathbf{P}_{k+m}$ on each subinterval of the mesh π of (5). A k-stage collocation method under consideration is determined by a mesh π and a set of k points

$$0 \le \rho_1 < \rho_2 < \cdots < \rho_k \le 1. \tag{11a}$$

An approximate solution $u_\pi(t)$ defined on $[a,b]$ is determined such that $u_\pi \in \mathbf{P}_{k+m,\pi} \cap C^{m-1}[a,b]$, u_π satisfies the BC, and u_π satisfies the differential equation (10) at the collocation points

$$t_{ij} := t_i + h_i\rho_j, \qquad 1 \le j \le k, \;\; 1 \le i \le N. \tag{11b}$$

Thus, for various ODE orders the approximation space is determined so that the highest derivatives appearing be in $\mathbf{P}_{k,\pi}$, independent of m. In (3a) the "order" is 0, so we require the collocation approximation $\mathbf{y}_\pi(t)$ of $\mathbf{y}(t)$ to be in the piecewise *discontinuous* space $\mathbf{P}_{k,\pi}$. Following the recipe in [As2], we can represent the solution polynomials in each mesh subinterval using say a Runge-Kutta basis, which amounts to using Lagrange interpolation at collocation points for the highest derivative appearing in each component, and then locally eliminate the k coefficients of the representation of each component of $\mathbf{y}_\pi(t)$ in each subinterval of the mesh, after linearization. The nonsingularity of $U^{[1]}(t)$ guarantees that this is possible to do. This

idea is natural in the context of mixed order ODE systems, and extensions of COLSYS using it have been implemented [Ho], [Ba].

It is important to realize that while (3) can be viewed as a limit of a stiff BVP, it is not necessarily "stiff" in itself, provided that the lower smoothness of $\mathbf{y}(t)$ (hence of its approximant $\mathbf{y}_\pi(t)$) is recognized. If one follows the recipe of eliminating the local representation of $\mathbf{y}_\pi(t)$ then a collocation approximation for the BVP (8) in $\mathbf{z}(t)$ alone is obtained. (If the BC (7c) involve $\mathbf{y}(a)$ or $\mathbf{y}(b)$ then (7d) is applied to obtain (8b).) The usual collocation theory then applies. We have

Theorem 12.

Assume that there are integers $p \geq k \geq 1$ such that

(a) the linear BVP (7) is well-conditioned, in the sense that κ, K_1, K_2 and K_3 are of a moderate size; has coefficients in $C^p[a,b]$; and has a solution $\mathbf{y}(t) \in C^p[a,b]$, $\mathbf{z}(t) \in C^{p+1}[a,b]$; and

(b) the k canonical collocation points $\rho_1, \ldots, \rho_k$ of (11a) satisfy the orthogonality conditions

$$\int_0^1 \phi(t) \prod_{l=1}^{k} (t-\rho_l)\, dt = 0 \qquad \phi \in \mathbf{P}_{p-k} \tag{12a}$$

(this implies that $p \leq 2k$).

Then for h small enough the following hold:

(a) The collocation method described above has a unique solution $\mathbf{y}_\pi(t)$, $\mathbf{z}_\pi(t)$.

(b) There exists an implementation such that the solution scheme is stable, with a condition number $c\kappa N$, c a constant of moderate size.

(c) The following error estimates hold at mesh points:

$$|\mathbf{z}(t_i) - \mathbf{z}_\pi(t_i)| = O(h^p) \qquad 1 \leq i \leq N+1. \tag{12b}$$

(When $p > k+1$, this is called *superconvergence*.)

(d) At any point in [a,b], the error in $\mathbf{z}$ satisfies

$$\mathbf{z}^{(j)}(t)-\mathbf{z}_\pi^{(j)}(t) = h_i^{k+1-j}\mathbf{z}^{(k+1)}(t_i)P^{(j)}\left(\frac{t-t_i}{h_i}\right)+O(h_i^{k+2-j})+O(h^p) \tag{12c}$$

$$t_i \le t \le t_{i+1},\ 1 \le i \le N,\ 0 \le j \le k,$$

where

$$P(\xi) = \frac{1}{k!(m-1)!}\int_0^{\xi}(x-\xi)^{m-1}\prod_{l=1}^{k}(x-\rho_l)\,dx. \tag{12d}$$

(e) At collocation points the error in $\mathbf{y}_\pi$ satisfies

$$|\mathbf{y}(t_{ij}) - \mathbf{y}_\pi(t_{ij})| \le K_3|\mathbf{z}(t_{ij})-\mathbf{z}_\pi(t_{ij})| \qquad 1 \le j \le k,\ 1 \le i \le N. \tag{12e}$$

In particular, if $\rho_k=1$ then

$$|\mathbf{y}(t_i) - \mathbf{y}_\pi(t_i)| = O(h^p) \qquad 1 \le i \le N+1.$$

(f) At any point in $[a,b]$, the error in $\mathbf{y}_\pi$ satisfies only

$$|\mathbf{y}^{(j)}(t) - \mathbf{y}_\pi^{(j)}(t)| = O(h_i^{k-j}) + O(h^p h_i^{-j}), \qquad t_i \le t \le t_{i+1},\ 0 \le j \le k. \tag{12f}$$

Thus, if $\rho_k<1$ then the superconvergence order at mesh points is lost.

(g) At any point t in $[a,b]$, the error in $\mathbf{y}$ can be made comparable to that in $\mathbf{z}$ by redefining the approximation to $\mathbf{y}(t)$ using (7d) and $\mathbf{z}_\pi(t)$. In particular, defining

$$\mathbf{y}_i := -[U^{11}(t_i)]^{-1}[U^{12}(t_i)\mathbf{z}_\pi(t_i)+\mathbf{g}^1(t_i)], \tag{12g}$$

yields

$$|\mathbf{y}(t_i) - \mathbf{y}_i| = O(h^p) \qquad 1 \le i \le N+1, \tag{12h}$$

even if $\rho_k<1$.

Proof: Parts (a)-(d) are proved in [As2, Thm 11]. Part (e) follows trivially from the fact that

the algebraic equations (7a) are exactly satisfied by the approximate solution at collocation points, noting (9b). Part (f) follows from (12e) and the fact that $\mathbf{y}_\pi(t)$ is a polynomial of order k on each mesh subinterval. (Note that the superconvergence order for $\mathbf{z}(t)$ follows from orthogonality in the integration, and for $\mathbf{y}(t)$ there is no integration.) Finally, Part (g) is obvious.

□

A similar theorem holds for nonlinear problems (3) as well (using Newton's method), extending [As2, Thm 13].

Example 1

Taking k=1, ρ_1=$\frac{1}{2}$, yields the midpoint scheme for $\mathbf{z}(t)$. For (3) we obtain

$$0 = \mathbf{f}(t_{i+1/2}, \mathbf{y}_{i+1/2}, \frac{\mathbf{z}_i + \mathbf{z}_{i+1}}{2}) \qquad 1 \le i \le N$$

$$\frac{\mathbf{z}_{i+1} - \mathbf{z}_i}{h_i} = \mathbf{g}(t_{i+1/2}, \mathbf{y}_{i+1/2}, \frac{\mathbf{z}_i + \mathbf{z}_{i+1}}{2})$$

where $t_{i+1/2} = t_i + \frac{1}{2}h_i$, and $\mathbf{y}_\pi(t)$ is piecewise constant with $\mathbf{y}_\pi(t) \equiv \mathbf{y}_{i+1/2}$, $t_i \le t < t_{i+1}$. Improved values for $\mathbf{y}$ at mesh points may be obtained by post-processing, solving

$$0 = \mathbf{f}(t_i, \mathbf{y}_i, \mathbf{z}_i)$$

for $\mathbf{y}_i$, $1 \le i \le N+1$. (One Newton iteration per mesh point, for a system of size n_y, starting from $\mathbf{y}_\pi(t_i)$, should ordinarily suffice.)

□

Remark

We wish to emphasize again that the numerical methods considered in this section for (3) are *simpler* than the corresponding ones for (4), (3b). This is in contrast to the cases for more general DAEs, where numerical methods generally perform *at best* as well as for the corresponding stiff ODEs (and often worse). Indeed, regularization methods have been proposed which imbed the DAE in a stiff ODE. But if the origin of the DAE under consideration is a "simplification", e.g. a reduced problem, of a well-conditioned ODE problem, then one should keep in mind the option of not solving the DAE problem at all.

□

3. Boundary value DAEs with index 1

The situation for DAEs gets more complicated when the differential and algebraic solution components are mixed together. Thus, consider a linear DAE (6) subject to boundary conditions

$$\tilde{B}_a \mathbf{x}(a) + \tilde{B}_b \mathbf{x}(b) = \beta. \tag{13}$$

We asume that there are nonsingular matrix functions $S(t), T(t)$, T differentiable, such that

$$E(t) = S(t) \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} T^{-1}(t). \tag{14}$$

It is not difficult to see that with the notation

$$\begin{pmatrix} U^{11}(t) & U^{12}(t) \\ U^{21}(t) & U^{22}(t) \end{pmatrix} \equiv U(t) := S^{-1}(t) A(t) T(t) - \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} T^{-1}(t) T'(t), \tag{15a}$$

$$\begin{pmatrix} \mathbf{y}(t) \\ \mathbf{z}(t) \end{pmatrix} \equiv \hat{\mathbf{x}}(t) := T^{-1}(t) \mathbf{x}(t), \tag{15b}$$

$$\mathbf{g}(t) := S^{-1}(t) \mathbf{q}(t), \tag{15c}$$

we obtain the semi-explicit DAE (7a,b). Also, (7c) holds with

$$B_a = \tilde{B}_a T(a), \quad B_b = \tilde{B}_b T(b).$$

The condition of solvability for $\mathbf{y}$ in terms of $\mathbf{z}$ is again that the upper left $n_y \times n_y$ block of U, $U^{11}(t) = (S^{-1}AT)^{11}(t)$, be nonsingular for all $a \le t \le b$, and this condition defines index 1 for the DAE (or, ***transferability*** of the DAE in [GrMa]). If this condition holds then (7d) holds and $\mathbf{y}(t)$, the "algebraic part" of the solution, can in principle be eliminated from (7b,c), yielding a standard ODE problem (8) for the "differential part". Proceeding as when deriving (9) we obtain

Theorem 16

Let the DAE (6) have index 1 and assume that, with the transformation (15), (7d), the BVP (8) is well-conditioned, with a conditioning constant κ. Further, assume that there are constants K_j such that

$$||T|| \le K_1, \quad ||S^{-1}|| \le K_2, \quad ||A|| \le K_3, \tag{16a}$$

$$||T^{-1}T'|| \le K_4, \tag{16b}$$

$$||[(S^{-1}AT)^{11}]^{-1}|| \le K_5, \tag{16c}$$

$$K_6 := K_1K_2K_3, \quad K_7 := K_6 + K_4, \quad K_8 := 1 + K_6K_5. \tag{16d}$$

Then the unique solution $\mathbf{x}(t)$ of the boundary value DAE (6), (13) satisfies

$$||\mathbf{x}|| \le K_1[K_5K_2 + \kappa K_8(1 + K_5K_7)]||\mathbf{q}|| + K_1\kappa K_8|\beta|. \tag{16e}$$

In (16a-e), the maximum norm in t over $[a,b]$, with some consistent local (pointwise) vector and matrix norms, is taken.

□

Remarks

(a) The constant K_5 measures "how transferable" the DAE is: If K_5 is large then the DAE is "closer" to one of a higher index.

(b) The constant K_4 measures the rate of change with t of the null space of $E(t)$. The dimension of this null space is of course constant and equals n_y.

(c) the size of K_6 depends on the scaling of (6). Assuming that $A(t)$ has been well-scaled, so $K_3 \sim 1$, the size of $K_1 K_2$ depends on the choice of $T(t)$. Often, a pointwise SVD, or an RQ-decomposition with column pivoting, works well (i.e., K_4 is not large), yielding an orthogonal $T(t)$. Then,

$$||T(t)||_2 = ||T(t)^{-1}||_2 = 1,\ ||S(t)||_2 = ||E(t)||_2,\ ||S(t)^{-1}||_2 = ||E(t)^{+}||_2,$$

where E^{+} is the Penrose pseudo-inverse. Thus, in a sense K_6 reflects essentially the size of the problem coefficients.

(d) Obviously, the bound in (16e) is not always finely tuned. But it shows that if the problem is safely transferable to a well-conditioned ODE problem then the original DAE problem is well-conditioned.

(e) In [GrMa], [Ha], [LeMa], the decoupling of the differential and the algebraic solution components is considered via a projector $P(t)$, which may be related to our $T(t)$ by

$$P = T \begin{pmatrix} 0 & 0 \\ 0 & I_{n_z} \end{pmatrix} T^{-1}.$$

We choose (14) because for us it yields a more transparent presentation. Moreover, an ODE problem of a reduced size n_z is obtained and the usual BVP theory applies, giving the conditioning result (16e). This theory involves an $n_z \times n_z$ Green's function with non-singular fundamental solutions. In [LeMa] an $n \times n$ Green's function is evolved instead, and this is more complicated because singular fundamental solutions and shooting matrices

are encountered. (Note that [LeMa] have other uses for their Green's function as well.)

□

For the numerical solution of such problems as (6), (13) and its nonlinear counterpart, there are two basic approaches. One is to decouple the differential and the algebraic parts of the solution explicitly, as we have just done theoretically. Then a piecewise discontinuous approximation space for $\mathbf{y}(t)$ may be used as described in §2, and the problem becomes simple. However, the decoupling (14), (15) may be expensive (and tricky for nonlinear problems).

More popular in the literature hitherto has been the approach to not attempt an explicit decoupling, but to proceed with a method for stiff ODEs. Such a method necessarily considers $\mathbf{x}(t)$ to be piecewise continuous, hence approximates $\mathbf{y}(t)$ (say by $\mathbf{y}_\pi(t)$) in a "wrong" space, in the sense that $\mathbf{y}_\pi(t)$ is as smooth as the approximation to $\mathbf{z}(t)$. But if the method damps out the local error contributions arising from this excess continuity in $\mathbf{y}_\pi(t)$ then it will still work, because an implicit, approximate decoupling of algebraic and differential components results. The popular BDF schemes ([Ge], [GePe], [Pe], [LoPe]) derive their phenomenal success for IVPs from the fact that they essentially collocate the DAE at each mesh subinterval's right end point only. Thus, the algebraic relations (7a) are reproduced exactly at mesh points (cf. §2). This is the correct discretization limit when (3a) is considered as a limit of the stiff equation (4), which is independent of the sign of the eigenvalues of the Jacobian $\frac{\partial \mathbf{f}}{\partial \mathbf{y}}$. Using for instance the backward Euler scheme for the test equation

$$y' = \lambda y$$

with a step size h such that $|\lambda|h \to \infty$, we have the damping factor $|\frac{1}{1-h\lambda}| \to 0$ *regardless* of the sign of $\mathrm{Re}(\lambda)$. (Care should be taken in designing a local error control, though, because only controlling the error in $\mathbf{z}(t)$ makes sense.)

For BVPs, the BDF schemes can be used to advantage as well in a shooting setting, provided that there is no excess stiffness (in both directions) in the ODE for $\mathbf{z}(t)$. However, even this procedure, which is relatively simple but has well-known stability difficulties [AMR, §4.2], is not entirely straightforward, because only consistent initial and boundary conditions and only the differential solution components play a part, strictly speaking, in the shooting matching, and the shooting matrix is generally singular. Moreover, while for IVPs a subset of consistent initial conditions can be derived from a given set in an automatic manner, for BVPs this task can be complicated to perform numerically (cf. [FlOm]). In general, it appears that some explicit knowledge of the decoupling transformation $T(t)$ is necessary for a reasonable numerical method for BVPs.

In view of these considerations, an explicit decoupling of the differential and algebraic solution components may become a favoured option for certain BVPs [Ba].

3.1 On symmetric difference schemes

The possibility that the ODE part in a DAE would have a BVP stiffness (i.e., both fast decreasing and fast increasing solution modes, thus preventing an efficient use of a typical IVP stiff solver based on, e.g., BDF [AMR, Ch. 10]) has received little attention in the literature hitherto. A symmetric difference scheme is often inappropriate for a coarse application to (6), (13), since the growth factor of any symmetric scheme has modulus 1 (and not <1) in the limit, so any local error (e.g. BC inconsistency) will not be damped out. Unfavourable marginal stability bounds then result [AsWe2, Lemma 5.1], [AsWe3], [Ma2]. Still, if the boundary conditions are consistent then accurate approximations may usually be obtained, provided that the method successfully decouples the algebraic equations from the differential ones. Thus, collocation at Lobatto points is rejected in this context, but collocation at Gaussian points is not [As1]. To

explain this further, we consider the most transparent cases which are the midpoint ("box") and the trapezoidal schemes.

The midpoint scheme for (7a,b) reads, in case that $\mathbf{y}_\pi$ is from the same approximation space as $\mathbf{z}_\pi$,

$$0 = U^{11}(t_{i+1/2})\frac{\mathbf{y}_i+\mathbf{y}_{i+1}}{2} + U^{12}(t_{i+1/2})\frac{\mathbf{z}_i+\mathbf{z}_{i+1}}{2} + \mathbf{g}^1(t_{i+1/2}) \tag{17a}$$

$$\frac{\mathbf{z}_{i+1}-\mathbf{z}_i}{h_i} = U^{21}(t_{i+1/2})\frac{\mathbf{y}_i+\mathbf{y}_{i+1}}{2} + U^{22}(t_{i+1/2})\frac{\mathbf{z}_i+\mathbf{z}_{i+1}}{2} + \mathbf{g}^2(t_{i+1/2}) \tag{17b}$$

(cf. Example 1). An analysis for (17) was carried out in [We], [AsWe2], so we only mention essential points here. Thus, we can eliminate $\frac{\mathbf{y}_i+\mathbf{y}_{i+1}}{2}$ from the first equation and substitute into the second, obtaining an ordinary midpoint scheme for $\mathbf{z}$. Stability and second order convergence for $\mathbf{z}_i$ then follow (assuming that the BC have been prepared to be in $\mathbf{z}$ alone), as usual for an ODE. To obtain results for $\mathbf{y}_i$ as well, discussion reduces to the initial value problem

$$\mathbf{y}_{i+1} = -\mathbf{y}_i + 2\mathbf{f}(t_{i+1/2}), \quad \mathbf{y}_1 \text{ given}, \tag{18a}$$

as an approximation to the problem

$$\mathbf{y} = \mathbf{f}(t). \tag{18b}$$

The solution of the recursion is

$$\mathbf{y}_{i+1} = (-1)^i\mathbf{y}_1 + 2\sum_{j=1}^{i}(-1)^{i-j}\mathbf{f}(t_{j+1/2}). \tag{18c}$$

In this we see the unfortunate properties of the scheme, namely, that no error is damped. Hence the error is not localized. There is also a linear growth of roundoff error which usually is only of theoretical concern. Still, if $\mathbf{y}_1=\mathbf{f}(a)$ (corresponding to setting $\mathbf{y}(a)$ explicitly via (7d) in terms

of $\mathbf{z}(a)$) and $\mathbf{f}$ is smooth then the error is $O(h)$, and it is $O(h^2)$ if the mesh is locally almost uniform, i.e. if $h_{i+1} = h_i(1+O(h_i))$ for all i odd *or* for all i even.

For (6) the midpoint scheme reads

$$E(t_{i+1/2})\frac{\mathbf{x}_{i+1}-\mathbf{x}_i}{h_i} = A(t_{i+1/2})\frac{\mathbf{x}_i+\mathbf{x}_{i+1}}{2} + \mathbf{q}(t_{i+1/2}).$$

Using (14) at $t_{i+1/2}$ and multiplying through by $S^{-1}(t_{i+1/2})$, we obtain for

$$\begin{pmatrix} \mathbf{y}_i \\ \mathbf{z}_i \end{pmatrix} \equiv \hat{\mathbf{x}}_i := T^{-1}(t_i)\mathbf{x}_i$$

the system

$$0 = U^{11}(t_{i+1/2})\frac{\mathbf{y}_i+\mathbf{y}_{i+1}}{2} + U^{12}(t_{i+1/2})\frac{\mathbf{z}_i+\mathbf{z}_{i+1}}{2} + \mathbf{g}^1(t_{i+1/2}) +$$
$$+ \frac{h_i}{4}P_y(UT^{-1}T\prime)(t_{i+1/2})(\hat{\mathbf{x}}_{i+1}-\hat{\mathbf{x}}_i) + O(h_i^2)\hat{\mathbf{x}}_i + O(h_i^2)\hat{\mathbf{x}}_{i+1}\ ,$$

$$\frac{\mathbf{z}_{i+1}-\mathbf{z}_i}{h_i} = U^{21}(t_{i+1/2})\frac{\mathbf{y}_i+\mathbf{y}_{i+1}}{2} + U^{22}(t_{i+1/2})\frac{\mathbf{z}_i+\mathbf{z}_{i+1}}{2} + \mathbf{g}^2(t_{i+1/2}) +$$
$$+ O(h_i)\hat{\mathbf{x}}_i + O(h_i)\hat{\mathbf{x}}_{i+1}\ ,$$

where

$$P_y := \begin{pmatrix} I_{n_y} & 0 \end{pmatrix} \in \mathbf{R}^{n_y \times n}. \tag{19}$$

Comparing this to (17) and considering the marginal stability implied from (18c), attention reduces to the IVP for

$$0 = \frac{\mathbf{y}_i + \mathbf{y}_{i+1}}{2} + \frac{h_i}{4}M(t_{i+1/2})(\mathbf{y}_{i+1} - \mathbf{y}_i) + \mathbf{f}_{i+1/2},\ \ i\geq 1 \tag{20a}$$

where

$$M := [U^{11}]^{-1}P_yUT^{-1}T\prime P_y^T \in \mathbf{R}^{n_y \times n_y}, \tag{20b}$$

and $\mathbf{f}_{i+1/2}$ is an inhomogeneity. For

$$\hat{\mathbf{y}}_i := (-1)^i \mathbf{y}_i \tag{21a}$$

(cf. [As1] and references therein), (20a) gives

$$0 = \frac{\hat{\mathbf{y}}_{i+1} - \hat{\mathbf{y}}_i}{h_i} + M(t_{i+1/2}) \frac{\hat{\mathbf{y}}_{i+1} + \hat{\mathbf{y}}_i}{2} + (-1)^{i+1} 2 h_i^{-1} \mathbf{f}_{i+1/2}, \quad i \geq 1. \tag{21b}$$

The homogeneous part of (21b) is just a midpoint discretization for the ODE

$$\hat{\mathbf{y}}' = -M(t)\hat{\mathbf{y}}, \tag{21c}$$

and the usual zero-stability theory of one step schemes for initial value ODEs yields for h small enough,

$$|\mathbf{y}_{i+1}| \leq K \, e^{\|M\|(t_{i+1}-a)} \, (|\mathbf{y}_1| + 2|\sum_{j=1}^{i} (-1)^{i-j} \mathbf{f}_{j+1/2}| \,), \tag{22}$$

K a constant. Subsequently, the results for (17) may be retrieved here too: In $\mathbf{f}_{j+1/2}$ we have $O(h_i)$ terms in $\mathbf{z}$ which sum up to a bounded quantity, terms like $F(t_{j+1/2}) \frac{\mathbf{z}_i + \mathbf{z}_{i+1}}{2}$ which also sum up to a bounded quantity because of the sign alternation and the smoothness of $F(t)$, and $O(h_i^2)$ terms in $\mathbf{y}$ which are handled by a contraction argument. The details are sufficiently close to those of Lemma 5.1 in [AsWe2] so as to eliminate need of repeating them here. For a k-stage Gauss collocation (the 1-stage scheme is just the midpoint one), we obtain convergence with order reduction,

$$\max_i |\mathbf{y}_i - \mathbf{y}(t_i)| \leq const(|\mathbf{y}_1 - \mathbf{y}(t_1)| + h^{k+q}) \tag{23}$$

where q=1 if k is odd and the mesh is locally almost uniform, q=0 otherwise.

Remark

which is of the semi-explicit form (3). However, what corresponds to U^{11} there is $-\lambda^2\Delta\psi$. Away from junctions we therefore have $K_5 \sim \lambda^{-2}$ in (16c) and the DAE is close to one of index 2. (The index is not more than 2 even if λ=0, as before.)

Still, since there are no time-dependent transformations required to separate the different solution components, the IVP for (46) can often be satisfactorily discretized using a BDF scheme, according to Theorem 37 (or [LoPe]). This observation is borne out in practice.

In [BCFRS], a scheme based on a combination of a trapezoidal step followed by a BDF step of order 2 was proposed, working with the ψ,u,v variables. But our analysis does not lead us to recommend this scheme (cf. [PHSM]).

Note that, while in the space variable $\mathbf{x}$ the dependent variable set ψ,u,v may appear to offer an advantage, in time the more natural variable set is ψ,n,p, with ψ the algebraic and n,p the differential solution components. (The other variable set is not so decoupled.) Only initial values in n and p need be provided, and it may be argued that a reasonable discretization scheme should not require initial values for ψ either. But the closeness of this DAE to one of a higher index implies that the situation may be more complex than that covered in §2. If we perturb this DAE slightly by setting λ=0 then only initial values for n (or for p, but not both) are required, the other variable's initial values being determined by

$$p(\mathbf{x},0) = n(\mathbf{x},0) - C(\mathbf{x}). \tag{47}$$

Thus, if for $0<\lambda\ll 1$ arbitrary initial values are prescribed both for n and for p then this is almost an inconsistency and a layer adjustment with a rapid time change in all three dependent variables is needed to satisfy (38a) with $\Delta\psi$ not large (cf. [Ri]). No such initial layer in time is needed if, given e.g. $n(\mathbf{x},0)$, we prescribe $p(\mathbf{x},0)$ by (47).

Acknowledgement

I wish to thank Dr. Georg Bader and Prof. Roswitha März for many fruitful discussions.

References

[As1] U. Ascher, "On some difference schemes for singular singularly perturbed boundary value problems", Numer. Math. 46 (1986), 1-30.

[As2] U. Ascher, "Collocation for two-point boundary value problems revisited", SIAM J. Numer. Anal. 23 (1986), 596-609.

[AsBa] U. Ascher and G. Bader, "A note on conditioning, stability and collocation matrices", Tech. Rep. CMA-R16-86, Australian National University, Canbera 1986. J. Appl. Math. Comput., to appear.

[ACR] U. Ascher, J. Christiansen and R.D. Russell, "Collocation software for boundary value ODEs", Trans. Math. Software 7 (1981), 209-222.

[AsJa] U. Ascher and S. Jacobs, "On collocation implementation for singularly perturbed two-point problems", Tech. Rep. 86-19, Dept. Computer Science, UBC, 1986. Submitted.

[AMSSW] U. Ascher, P. Markowich, C. Schmeiser, H., Steinrück and R. Weiss, "Conditioning of the steady state semiconductor device problem", Tech. Rep. 86-18, Dept. Computer Science, UBC, 1986. Submitted.

[AMR] U. Ascher, R. Mattheij and R.D. Russell, *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, Prentice-Hall (estimated publication date fall 1987)

[APR] U. Ascher, S. Pruess and R.D. Russell, "On spline basis selection for solving differential equations", SIAM J. Numer. Anal. 20 (1983), 121-142.

[AsWe1] U. Ascher and R. Weiss, "Collocation for singular perturbation problem I: First order systems with constant coefficients", SIAM J. Num. Anal. 20 (1983), 537-557.

[AsWe2] U. Ascher and R. Weiss, "Collocation for singular perturbation problems II: Linear first order systems without turning points", Math. Comp. 43 (1984), 157-187.

[AsWe3] U. Ascher and R. Weiss, "Collocation for singular perturbation problems III: Nonlinear problems without turning points", SIAM J. Scient. Stat. Comp. 5 (1984), 811-829.

[Ba] G. Bader, in preparation.

[BaAs] G. Bader and U. Ascher, "A new basis implementation for a mixed order boundary value ODE solver", SIAM J. Scient. Stat. Comput. 8 (1987), 483-500.

[BCRFS] R.E. Bank, W.M. Coughran, W. Fichtner, D.J. Rose and R.K. Smith, "Computational aspects of semiconductor device simulation", Numerical Analysis Ms 85-3, Bell Labs., 1985.

[dBSw] C. de Boor and B. Swartz, "Collocation at Gaussian points", SIAM J. Numer. Anal. 10 (1973), 582-606.

[Ca] S.L. Campbell, "Regularizations of linear time varying singular systems", Automatica 20 (1984), 365-370.

[FlOm] J.E. Flaherty and R.E. O'Malley, Jr., "On the numerical integration of two-point boundary value problems for stiff systems of ordinary differential equations", in Proc. BAIL I Conf., J.J.H. Miller (Ed.), Boole Press, Dublin, 1980, 93-102.

[Ge] C.W. Gear, "The simultaneous numerical solution of differential-algebraic equations", IEEE Trans. Circuit Theory 18 (1971), 89-95.

[GePe] C.W. Gear and L.R. Petzold, "ODE methods for the solution of differential/algebraic systems", SIAM J. Numer. Anal. 21 (1984), 716-728.

[GrMa] E. Griepentrog and R. März, *Differential-Algebraic Equations and their Numerical Treatment*, Teubner-Texte zur Mathematik Band 88, Leipzig 1986.

[Ha] M. Hanke, "On the regularization of index 2 differential-algebraic equations", Report Nr. 137, Humboldt Universität zu Berlin, 1987.

[Ho] M. Ho, "A collocation solver for systems of boundary-value differential/algebraic equations", Computers and Chem. Eng. 7 (1983), 735-737.

[LeMa] M. Lentini and R. März, "The condition of boundary value problems in transferable differential-algebraic equations", Report Nr. 136, Humboldt Universität zu Berlin, 1987.

[LoPe] P. Lötstedt and L.R. Petzold, "Numerical solution of nonlinear differential equations with algebraic constraints I: Convergence results for backward differentiation formulas", Math. Comp. 46 (1986), 491-516.

[Mar] P.A. Markowich, *The Stationary Semiconductor Device Equations*, Springer, Wien New York, 1986.

[Ma1] R. März, "Multistep methods for initial value problems in implicit differential algebraic equations", Beitrage zur Num. Math. 12 (1984), 107-123.

[Ma2] R. März, "On difference and shooting methods for boundary value problems in differential-algebraic equations", ZAMM 64 (1984), 463-473.

[OMFl] R.E. O'Malley and J.E. Flaherty, "Analytical and numerical methods for nonlinear singular singularly-perturbed initial value problems", SIAM J. Appl. Math. 38 (1980), 225-248.

[Pe] L.R. Petzold, "A description of DASSL: A differential/algebraic system solver", Sandia Report SAND 82-8637, 1982.

[PeLo] L.R. Petzold and P. Lötstedt, "Numerical solution of nonlinear differential equations with algebraic constraints II: Practical implications", SIAM J. Scient. Stat. Comput. 7 (1986), 720-733.

[PHSM] S.J. Polak, C. Den Heijer, H.A. Schilders and P. Markowich, "Semiconductor device modelling from the numerical point of view", Int. J. Numer. Methods in Eng. 24 (1987), 763-838.

[Ri] C. Ringhofer, "Numerical methods for transient semiconductor device modelling", Manuscript, MRC, Madison, Wisconsin, 1983.

[Se] S. Selberherr, *Analysis and Simulation of Semiconductor Devices*, Springer, Wien New York, 1984.

[We] R. Weiss, "An analysis of the box and trapezoidal schemes for linear singularly perturbed boundary value problems", Math. Comput. 42 (1984), 41-68.