Pergamon

Appl. Math. Lett. Vol. 7, No. 2, pp. 29–34, 1994

0893-9659/94 $6.00 + 0.00

0893-9659(94)E0007-X

Finite Difference Methods for Time Dependent, Linear Differential Algebraic Equations*

P. J. RABIER AND W. C. RHEINBOLDT
Department of Mathematics and Statistics
University of Pittsburgh, Pittsburgh, PA 15260, U.S.A.

(Received October 1993; accepted November 1993)

Abstract—Recently the authors developed a global reduction procedure for linear, time-dependent DAE that transforms their solutions into solutions of smaller systems of ODE's. Here it is shown that this reduction allows for the construction of simple, convergent finite difference schemes for such equations.

Keywords—Linear DAE, Time dependent, Reduction, Finite differences, Extrapolation.

1. INTRODUCTION

Time-dependent, linear differential algebraic equations (DAEs),

$$A(t)\dot{x} + B(t)x = b(t), \qquad A(t), B(t) \in \mathcal{L}(\mathbb{R}^n),\ b(t) \in \mathbb{R}^n,\ t \in \mathbb{R} \tag{1.1}$$

arise in many circuit and control problems (see e.g., [1–3] for some references). In general, standard, ODE-type, numerical methods for (1.1) are known to fail or perform poorly when the index exceeds one, and certain Taylor-type methods developed in [2], that apply to a larger class of problems, require the solution of large augmented systems of equations.

Recently we developed a new, global reduction theory [4] which leads to existence and uniqueness results for classical as well as generalized solutions of (1.1) under rather general conditions. The aim here is to show that this reduction process allows for the construction of simple, convergent finite difference approximations for the numerical solution of (1.1) which appears to provide a new tool for the solution of general systems of the form (1.1).

2. SUMMARY OF THE REDUCTION PROCESS

In its basic form the reduction process of [4] assumes that the coefficient functions A and B are analytic. Although generalizations to the non-analytic case are also discussed in [4], we shall retain here, for simplicity, the analyticity assumption. Throughout this note analytic mappings will be referred to as "mappings of class C^ω."

A main tool for the proof of the globality of the reduction procedure is the concept of "transformation functions" introduced by Kato [5]. With it and with another result of Kato [6] the following basic result was proved:

*The work was supported in part by ONR-Grant N-00014-90-J-1025, and NSF-Grant CCR-9203488.

THEOREM 1. *Let $\mathcal{J} \subset \mathbb{R}$ be an open interval, $m \geq 1$ any integer, and $M \in C^\omega(\mathcal{J}; \mathcal{L}(\mathbb{R}^m))$ with $r = \max_{t \in \mathcal{J}} \text{rank } M(t)$. Then, there is a subset $\mathcal{S} \subset \mathcal{J}$ of isolated points such that rank $M(t) = r$ if and only if $t \in \mathcal{J} \setminus \mathcal{S}$. Furthermore, the orthogonal projections onto rge $M(t)$ and $\ker M(t)$, $t \in \mathcal{J} \setminus \mathcal{S}$ are analytic on $t \in \mathcal{J} \setminus \mathcal{S}$ and can be extended as analytic functions over the entire interval $\mathcal{J}$.*

With the notation of this theorem the orthogonal projection $P(t)$ onto rge $M(t)$, for $t \in \mathcal{J} \setminus \mathcal{S}$, is at any point of S an orthogonal projection onto a subspace containing rge $M(t)$ and this subspace, called the extended range of $M(t)$ and denoted by ext rge $M(t)$, is independent of the specific choice of P.

For ease of notation, we call a pair (A, B) of coefficent functions *admissible* if $A, B \in C^\omega(\mathcal{J}; \mathcal{L}(\mathbb{R}^n))$ on a fixed open interval $\mathcal{J} \subset \mathbb{R}$. An admissible pair is said to be *regular* if rank $[A(t)A(t)^T + B(t)B(t)^T] = n$, $\forall t \in \mathcal{J}$. Then Theorem 1 ensures that rank $A(t) = r$, $\forall t \in \mathcal{J} \setminus \mathcal{S}$ where $S \subset \mathcal{J}$ consists of isolated points, and that there is a family of projections $P \in C^\omega(\mathcal{J}; \mathcal{L}(\mathbb{R}^n))$ onto ext rge $A(t)$, $\forall t \in \mathcal{J}$. Now $Q = I - P$ can be shown to satisfy $\dim \ker Q(t)B(t) = r$, $\forall t \in \mathcal{J}$. Moreover, there exist mappings $C \in C^\omega(\mathcal{J}; \mathcal{L}(\mathbb{R}^r; \mathbb{R}^n))$ and $D \in C^\omega(\mathcal{J}; \mathcal{L}(\mathbb{R}^n, \mathbb{R}^r))$ such that

$$C(t) \in GL(\mathbb{R}^r, \ker Q(t)B(t)), \qquad \forall t \in \mathcal{J}. \tag{2.1}$$

$$D(t) \in GL(\text{ext rge } A(t), \mathbb{R}^r), \qquad \forall t \in \mathcal{J}, \tag{2.2}$$

respectively. For any such choice of mappings Q, C, D the pair of admissible (A_1, B_1) defined by

$$A_1, B_1 \in C^\omega(\mathcal{J}; \mathcal{L}(\mathbb{R}^r)), \quad A_1 = DAC, \quad B_1 = D(BC + A\dot{C}), \tag{2.3}$$

is called a *reduction* of the regular pair (A, B) (in $\mathcal{J}$). Clearly, such a reduction is not unique, since it depends upon the choice of C and D. However, any two reductions of (A, B) turn out to be equivalent.

More specifically, between admissible pairs (A, B) and $(\bar{A}, \bar{B})$, define a relation $(A, B) \sim (\bar{A}, \bar{B})$ in $\mathcal{J}$ by the conditions $\bar{A} = MAN$ and $\bar{B} = M(BN + A\dot{N})$ for some $M, N \in C^\omega(\mathcal{J}; GL(\mathbb{R}^n))$. This turns out to be an equivalence relation on the set of all admissible pairs (on $\mathcal{J}$). Furthermore, if $(A, B) \sim (\bar{A}, \bar{B})$ then, on $\mathcal{J}$, we have

(i) rank $A(t) =$ rank $\bar{A}(t)$,

(ii) (A, B) is regular if and only if $(\bar{A}, \bar{B})$ is regular, and

(iii) If (A, B) and $(\bar{A}, \bar{B})$ are regular, any reduction (A_1, B_1) of (A, B) is equivalent to any reduction $(\bar{A}_1, \bar{B}_1)$ of $(\bar{A}, \bar{B})$.

This permits the definition of our reduction procedure which, up to equivalence, is independent of any particular choice that must be made at each step. Suppose that the admissible pair (A, B) is regular. Then, any two reductions (A_1, B_1) and $(\bar{A}_1, \bar{B}_1)$ of (A, B) are either both regular or not. If they are, then once again any reductions (A_2, B_2) of and $(\bar{A}_2, \bar{B}_2)$ of (A_1, B_1) and $(\bar{A}_1, \bar{B}_1)$, respectively, will be simultaneously regular, and so on. In line with this, an admissible pair is called *completely regular* (in $\mathcal{J}$) if this reduction procedure can be continued indefinitely. In particular, the complete regularity of a pair (A, B) in $\mathcal{J}$ implies its regularity in $\mathcal{J}$.

For a completely regular pair (A, B) consider a sequence (A_j, B_j), $j = 0, 1, \ldots$, $A_0 = A$, $B_0 = B$, such that (A_j, B_j) is some reduction of (A_{j-1}, B_{j-1}), $j > 0$. Then, $r_j = \max_{t \in \mathcal{J}} [\text{rank } A_j(t)]$, is independent of the specific choice of (A_j, B_j), $j > 0$ and with $r_{-1} = n$ we have $A_j(t) \in \mathcal{L}(\mathbb{R}^{r_{j-1}})$, for $t \in \mathcal{J}$ and $j \geq 0$. Moreover, the $r_j \geq 0$ decrease monotonically with j whence there exists a smallest integer $0 \leq \nu \leq n$ such that $r_\nu = r_{\nu-1}$ and $A_\nu(t) \in GL(\mathbb{R}^{r_{\nu-1}})$, $\forall t \in \mathcal{J} \setminus \mathcal{S}_\nu$ where $\mathcal{S}_\nu \subset \mathcal{J}$ consists only of isolated points. This integer ν is the *index* of the pair (A, B).

THEOREM 2. *Let (1.1) be a DAE with an admissible pair (A, B) of coefficient functions and any $b \in C^k(\mathcal{J}; \mathbb{R}^n)$, $1 \leq k \leq \infty$ or $k = \omega$. Suppose that (A_1, B_1) is any reduction of (A, B)*

defined, say, by the projection Q and the mappings C, D satisfying (2.1), (2.2). Then there exists $u_0 \in C^k(\mathcal{J};\mathbb{R}^n)$ such that $B(t)u_0(t) - b(t) \in$ ext rge $A(t)$, $\forall t \in \mathcal{J}$, and indeed we may use $u_0 = B^T(AA^T + BB^T)^{-1}b$. With this, a differentiable mapping $x : \mathcal{J} \to \mathbb{R}^n$ solves (1.1) if and only if $x = Cx_1 + u_0$ where $x_1 : \mathcal{J} \to \mathbb{R}^r$ is a differentiable solution of the reduced DAE

$$A_1(t)\dot{x}_1 + B_1(t)x = b_1(t), \quad t \in \mathcal{J}, \quad b_1 = D(b - Bu_0 - A\dot{u}_0). \tag{2.4}$$

By recursive application of this result it follows that when (A, B) is completely reducible with index ν, then after ν steps we arrive at a linear ODE

$$A_\nu(t)\dot{x}_\nu + B_\nu(t)x_\nu = b_\nu(t), \tag{2.5}$$

where $A_\nu(t) \in GL(\mathbb{R}^{r_{\nu-1}})$ except perhaps at isolated points. Thus (2.4) is equivalent with an explicit ODE with singularities. A simple example for this is the system

$$\begin{pmatrix} 2t & 2 & 0 \\ 0 & 0 & 2 \\ -2t^3 & -2t^2 & -2t \end{pmatrix} \dot{x} + x = 0$$

given in [1] which has index 2. Here the final reduced ODE is $-2t(1 + 3t^2 + t^4)\dot{x}_2 + (-3 - 8t^2 + 3t^6)\, x_2 = 0$, and hence, is singular at $t = 0$. In [1] it was noted that the DAE has index 2 for $t > 0$ and the singularity at $t = 0$ was interpreted as index 3 at that point.

Of course, the case of (1.1) is of special importance when

$$A_\nu(t) \in GL(\mathbb{R}^{r_{\nu-1}}), \qquad \forall t \in \mathcal{J}. \tag{2.6}$$

The condition (2.6) is independent of the reduction, and if it holds then (2.5) is an explicit, linear ODE without singularities for which the standard existence and uniqueness results are applicable. Thus if the initial condition $x(t_0) = x_0$, $t_0 \in \mathcal{J}$ satisfies a certain consistency condition (see [4]) the resulting initial value problem for (1.1) has a globally defined unique solution in $\mathcal{J}$.

3. FINITE DIFFERENCE APPROXIMATIONS

The global reduction of the DAE (1.1) allows for the construction of convergent finite difference approximations of (1.1). These approximations are applied to the reduced DAE at stage $j = \nu - 1$ of the process which after one further step becomes the ODE (2.5) for which we assume now specifically that (2.6) holds. In other words, the finite difference scheme is applied to a (reduced) DAE of index one. Thus, for ease of notation, we assume here simply that the original DAE (1.1) has index one. As before (A_1, B_1) denotes a reduction of the admissible pair (A, B) defined by the mappings Q, C, D. We also note that the index one assumption is then equivalent, for instance, with the invertibility of $A(t) + Q(t)B(t)$, $\forall t \in \mathcal{J}$, (see [4]).

For a given, sufficiently small step $h > 0$ suppose that $t_i = t_0 + ih \in \mathcal{J}$ for $i = 0, 1, \ldots, m$ and consider first the explicit Euler approximation

$$A(t_i)\frac{1}{h}(x_{i+1} - x_i) + B(t_i)x_i = b(t_i). \tag{3.1}$$

Then the following solvability result holds:

THEOREM 3. *Any solution*

$$x_0, x_1, \ldots, x_m \in \mathbb{R}^n \tag{3.2}$$

of (3.1) satisfies for $i = 0, 1, \ldots, m-1$ the equations

$$Q(t_i)B(t_i)x_i = Q(t_i)b(t_i), \tag{3.3}$$

$$[A(t_i) + Q(t_{i+1})B(t_{i+1})]x_{i+1} = [A(t_i) - hB(t_i)]x_i + hb(t_i) + Q(t_{i+1})b(t_{i+1}). \tag{3.4}$$

Conversely, for sufficiently small h and any given $x_0 \in \mathbb{R}^n$ satisfying (3.3) (at t_0), the solution (3.2) of (3.4) is unique and is also a solution of (3.1).

PROOF. For any solution (3.2) of (3.1) it follows, after multiplication by $Q(t_i)$, that (3.3) holds for $0 \le i \le m-1$. Hence, by adding, for $i = 0, \dots, m-1$, the $(i+1)^{\text{st}}$ equation (3.3) to (3.1) we obtain (3.4). The smoothness of A, B, and Q ensures that, for sufficiently small h, the matrix in the square bracket on the left of (3.4) is nonsingular for $i = 0, \dots, m-1$ since, as noted above, $A(t_i)+Q(t_i)B(t_i)$ is invertible by the index-one assumption. Hence for given x_0 the solution (3.2) of (3.4) is uniquely determined. If (3.3) holds for x_0 (at t_0) then, by induction on i, it follows that this solution (3.2) of (3.4) satisfies (3.3) for $i = 0, \dots, m-1$. In fact, if (3.3) is valid for some i, $0 \le i < m$, then we obtain after multiplication of (3.4) by $Q(t_i)$ that $Q(t_i)Q(t_{i+1})B(t_{i+1})x_{i+1} = Q(t_i)Q(t_{i+1})b(t_{i+1})$. Since $Q(t_i)$ is injective on rge $Q(t_{i+1}) = \text{rge } A(t_{i+1})^{\perp}$ (for $t_{i+1} - t_i$ small enough) this implies that (3.3) holds for $i+1$ in place of i. Now by adding the $i+1^{\text{st}}$ equation (3.3) to (3.4) we see that the solution of (3.4) also solves (3.1). ∎

For any solution of (3.4) it follows from (3.3) that $x_i - u_0(t_i) \in \ker Q(t_i)B(t_i)$ for all i whence

$$x_i - u_0(t_i) = C(t_i)z_i, \qquad \forall i, \tag{3.5}$$

for some sequence $z_i \in \mathbb{R}^r$, $i = 0, \dots, m+1$. Then, using that $B(t_{i+1})u_0(t_{i+1}) - b(t_{i+1})$ belongs to $[\text{rge } A(t_{i+1})]^{\perp}$, we obtain from (3.4), after a simple calculation, that

$$\begin{aligned} A(t_i)C(t_{i+1})\frac{1}{h}(z_{i+1} - z_i) \; + \; & \left[B(t_i)C(t_i) + A(t_i)\frac{1}{h}\left(C(t_{i+1}) - C(t_i)\right)\right] z_i \\ & = c_i \; \equiv \; b(t_i) - A(t_i)\frac{1}{h}\left(u_0(t_{i+1}) - u_0(t_i)\right) - B(t_i)u_0(t_i). \end{aligned} \tag{3.6}$$

After multiplying this by $D(t_i)$ we see that (3.6) is a finite difference approximation of the initial value problem

$$\hat{A}_1(t)\dot{z} + \hat{B}_1(t)z = c(t), \qquad z(t_0) = z_0, \tag{3.7}$$

where z_0 is characterized by the condition $C(t_0)z_0 = x_0 - u_0(t_0)$, and

$$\hat{A}_1(t) = D(t)A(t)C(t+h), \tag{3.8a}$$

$$\hat{B}_1(t) = D(t)\left[B(t)C(t) + A(t)\frac{1}{h}\left(C(t+h) - C(t)\right)\right], \tag{3.8b}$$

$$\bar{b}(t) = b(t) - A(t)\frac{1}{h}\left(u_0(t+h) - u_0(t)\right) - B(t)u_0(t). \tag{3.8c}$$

Since (2.6) was assumed to hold for the ODE (2.5) obtained at the last reduction step, (2.5) is equivalent with the linear, explicit ODE

$$\dot{z} = K(t)z + k(t), \qquad K(t) \equiv \hat{A}_1(t)^{-1}\hat{B}_1(t), \qquad k(t) \equiv \hat{A}_1(t)^{-1}\bar{b}(t). \tag{3.9}$$

Hence, by a standard estimate for Euler's method (see e.g. [7, Theorem 7.5]) we obtain

$$\|z(t_i) - z_i\| \le \kappa_1|h|. \tag{3.10}$$

On the other hand, a comparison of (3.7) and the reduced equation (1.3) shows that for $h \to 0$ the difference of the coefficient functions (3.8) and the corresponding coefficient functions of (2.5) is of order $|h|$ uniformly on $[t_0, t_m]$. Thus a standard application of Gronwall's inequality provides that

$$\max_{t_0 \le t \le t_m} \|x_1(t) - z(t)\| \le \kappa_2|h|,$$

whence altogether we obtain from (1.4) and (3.5) that

$$\|x_i - x(t_i)\| = \|C(t_i)(x_1(t_i) - z_i)\| \le \bar{\kappa}|h|, \qquad \forall i, \tag{3.11}$$

with $\bar{\kappa} = \max_{t_0 \le t \le t_m} ||C(t)||(\kappa_1 + \kappa_2)$. In other words, when x_0 satisfies (3.3) at t_0, then the unique solution of the difference equation (3.4) provides an approximation of the solution of (1.1) with global error $\mathcal{O}(h)$.

The result is easily extended to variable steps in t by using a continuous grid function. It is also conceptually straightforward to derive higher order discretizations or implicit schemes. We illustrate this briefly for the implicit Euler scheme

$$A(t_{i+1})\frac{1}{h}(x_{i+1} - x_i) + B(t_{i+1})x_{i+1} = b(t_{i+1}), \tag{3.12}$$

Now any solution (3.2) satisfies for $i = 0, 1, \ldots, m-1$ the equations

$$Q(t_{i+1})B(t_{i+1})x_{i+1} = Q(t_{i+1})b(t_{i+1}), \tag{3.13}$$

$$[A(t_{i+1}) + Q(t_{i+1})B(t_{i+1}) + hB(t_{i+1})]\, x_{i+1} = A(t_{i+1})x_i + hb(t_{i+1}) + Q(t_{i+1})b(t_{i+1}), \tag{3.14}$$

For small h the matrix on the left of (3.14) is nonsingular and hence, for given x_0 the solution (3.2) of (3.14) is unique. This solution satisfies (3.13) for $i = 0, \ldots, m-1$ as is readily seen by multiplying (3.14) with $Q(t_{i+1})$ and dividing by $h+1$. Thus subtraction of (3.13) from (3.14) shows that the solution of (3.14) also solves (3.12). Moreover, the unique solution of (3.14) represents again an approximation of the solution of (1.1) with global error $\mathcal{O}(h)$. The proof is entirely analogous to that given for (3.4) and will not be repeated here.

If, for $h > 0$ the matrix $A(t_{i+1}) + hB(t_{i+1})$ in (3.12) is nonsingular, then the solution of (3.12) can be computed directly. This observation is well-known to apply also to higher order BDF formulas and is the basis of the widely used DAE solver DASSL (see [3]). However, the nonsingularity assumption requires the matrix pencil $A(t), B(t)$ to be regular for all $t \in \mathcal{J}$ which is not necessarily true for all index one problems. Moreover, since, in any case, the matrix becomes singular for $h = 0$, increasing difficulties are expected for decreasing h. This problem is not shared by either (3.4) or (3.14).

Difference schemes as (3.4) and (3.14) open up surprisingly effective numerical methods for the solution of the systems (1.1). In particular, (3.4) can be used as the base method in an explicit extrapolation integrator. An implementation of the resulting algorithm for general index-one problems (1.1) has been called LTVXE. It has the general form of the extrapolation code EULEX (see [8]) and, as EULEX, it is based on the order and step control mechanism of [9].

Of course, when (1.1) has index exceeding one, the application of this integrator pre-supposes the availability of the DAE arising at stage $\nu - 1$ of the reduction process. Under the assumption that subroutines for all needed derivatives of the coefficients A, B, and b are given, a computational implementation of the reduction process is feasible. This will be discussed elsewhere.

Here we show only one simple example of the method when applied to the index one problem

$$\begin{pmatrix} 1 & t \\ 0 & 0 \end{pmatrix} \dot{x} + \begin{pmatrix} 0 & 0 \\ 1 & t \end{pmatrix} x = \begin{pmatrix} t^2 \\ e^t \end{pmatrix}, \quad x(0) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{3.15}$$

This is an example, called not regular in [3], for which the matrix pencil is singular for all $t \in \mathbb{R}$. Accordingly, as noted above, DASSL cannot be expected to solve (3.15), and, in fact, DASSL failed consistently for any tolerance with either one of the error messages "*the iteration matrix is singular*" or "*the corrector failed to converge repeatedly or with abs(h) = hmin.*"

Our reduction shows readily that (3.15) indeed has has index one and that the exact solution is $x(t) = ((1-t)\, e^t + t^3, e^t - t^2)^\top$. LTVXE performs satisfactorily for all tolerances. For example, a run with LTVXE from $t = 0$ to $t = 8.0$ with tolerance 10^{-8} used 801 steps and produced at $t = 8.0$ the approximate solution $x = (-20354.512,\ 2916.9580)^\top$ which has a relative error of 9.531×10^{-6} under the maximum norm.

REFERENCES

1. S.L. Campbell, Index two linear time-varying singular systems of differential equations, *SIAM J. Sci. Sta. Comput.* **4**, 237–243 (1983).
2. S.L. Campbell, The numerical solution of higher index linear time varying singular systems of differential equations, *SIAM J. Sci. Sta. Comput.* **6**, 334–348 (1985).
3. K.E. Brennan, S.L. Campbell and L. R. Petzold, *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, North-Holland, New York, (1989).
4. P.J. Rabier and W. C. Rheinboldt, Classical and generalized solutions of time-dependent linear differential algebraic equations, Tech. Rept. TR-ICMA-1xx, Inst. for Comp. Math. and Appl., Univ. of Pittsburgh, (October 1993); *Linear Algebra and Applications* (to appear).
5. T. Kato, On the adiabatic theorem of quantum mechanics, *J. Phys. Soc. Japan* **5**, 435–439 (1950).
6. T. Kato, *A Short Introduction to Perturbation Theory for Linear Operators*, Springer-Verlag, New York, NY, (1982).
7. E. Mairer, S.P. Norsett and G. Wanner, *Solving Ordinary Differential Equations I: Nonstiff Problems*, Springer-Verlag, New York, NY, (1987).
8. P. Deuflhard, U. Nowak and U. Poehle, *Solution of Systems of Initial Value Problems by Explicit Euler Discretization with h-Extrapolation*, ELIB-Library, Sub-Library CODELIB, K. Zuse Zenrum f. Informationtechnik, Berlin, Germany, (1988).
9. P. Deuflhard, Order and stepsize control in extrapolation methods, *Num. Math.* **41**, 399–422 (1983).
10. U.M. Ascher, On symmetric schemes and differential algebraic equations, *SIAM J. Sci. Sta. Comput.* **10**, 937–949 (1989).
11. K.D. Clark and L.R. Petzold, Numerical solution of boundary value problems in differential algebraic systems, Technical Report UCRL-98449, Lawrence Livermore National Lab., Numerical Mathematics Group, (April 1988).
12. E. Griepentrog and R. Maerz, *Differential-Algebraic Equations and their Numerical Treatment*, Teubner Texte zur Mathematik, Band 88, Teubner Verlag, Leipzig, Germany, (1986).