SINGULAR IMPLICIT ORDINARY DIFFERENTIAL EQUATIONS AND CONSTRAINTS*

C. W. Gear
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801-2987

L. R. Petzold
Applied Mathematics Division
Sandia National Laboratories
Livermore, CA 94550

Abstract

Implicit ordinary differential equations, also known as differential/algebraic equations, arise naturally in a number of formulations. These are frequently singular and can cause considerable numerical difficulty. We do not know how to handle the general problem, but a common formulation leads to equations that can be integrated in some cases.

INTRODUCTION

Implicit ordinary differential equations arise naturally in many problem formulations in electrical and mechanical systems. For example, a transient finite element analysis leads to the equation

$$My'' + Cy' + Ky = g(t) \tag{1}$$

while a branch analysis of an electrical network gives rise to

$$Ay' + b(y,t) = 0 \tag{2}$$

In this paper we will be particularly interested in the description of a mechanical system subject to constraints using the Lagrangian equations of motion: If the kinetic energy is $T(y, y')$, the external forces are $f(y, y', t)$ and the constraints are $c(y, t) = 0$ we have

$$\frac{d}{dt}\left(\frac{\partial T}{\partial y'}\right)^* - \left(\frac{\partial T}{\partial y}\right)^* + f + \left(\frac{\partial c}{\partial y}\right)^* \lambda = 0 \tag{3}$$

where * is the transpose operator and λ is a Lagrange multiplier. By the usual substitution of $y' = z$, equation (1) or (3) can be reduced to first order for analysis and then all equations can be written in the form

$$F(y', y, t) = 0 \tag{4}$$

for an appropriate dimension of y, although in this paper we will restrict ourselves to the form (2) which is sufficiently general to handle almost all cases that arise in practical problems. (If F in equation (4) is nonlinear in y' and we approach a

*Supported by the U.S. Department of Energy, Grant DOEAC0276ER02383 and by the U.S. DOE Office of Basic Energy Sciences.

point with multiple roots, we may have turning points or singularities which must be resolved by additional conditions, as in the equation $y'^2 + y^2 - 1 = 0$ at $y = 1$ which admits the two solutions $y \equiv 1$ and $y = \cos(t)$.)

If the matrix A in equation (2) is nonsingular, we have a true differential equation which can be rewritten in the explicit form $y' = -A^{-1}b(y, t)$ for analytical (but not necessarily numerical) purposes. However, we are interested in the case that A is singular, as happens in equation (3), for example, because λ' does not occur. Typically $T(y, y')$ is quadratic positive definite in y', so that $\partial T/\partial y'$ is linear in y'. In this case we can replace y' by z in equation (3) and solve for z' to get

$$\begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} y' \\ z' \\ \lambda' \end{bmatrix} + \begin{bmatrix} -z \\ q(y,z,t)+(\frac{\partial c}{\partial y})^*\lambda \\ c(y,t) \end{bmatrix} = 0 \tag{5}$$

A particular example of this is furnished by a simple pendulum. Suppose it swings in a plane with rectangular coordinates (x, y) and velocities (u, v). The kinetic energy is $(u^2 + v^2)/2$ for a unit mass, while the external force is $[0, -1]^T$ for a unit gravitational force in the negative y direction. If the mass is unit distance from the pivot the constraint is $x^2 + y^2 - 1 = 0$ so equation (5) takes the form

$$\begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & 0 \end{bmatrix} \begin{bmatrix} x' \\ y' \\ u' \\ v' \\ \lambda' \end{bmatrix} + \begin{bmatrix} -u \\ -v \\ 2\lambda x \\ 2\lambda y-1 \\ x^2+y^2-1 \end{bmatrix} = 0 \tag{6}$$

In fact, -2λ is the tension in the pendulum rod.

It was known more than a decade ago that some implicit equations, including singular cases, could be solved by a variety of methods including the Backward Differentiation Formula (BDF) methods (see Gear [1]), but more recently it has been found that other problems in this class cause great numerical difficulty. In this paper we will summarize some recent results on this general class of problems and present some new results on those problems that arise from a Lagrange formulation with constraints.

The current state of the art is that we don't know how to solve numerically the most general problems and can show that there are difficulties with known ODE methods. The factor which determines the level of difficulty of the problem is its <u>index</u>. This will be defined below. If the index exceeds 2, almost all problems are difficult and we will show that the index of equation (5) is at least 3. We will explore the pendulum problem in some detail because, although it has index 3, there is something about its structure (which is not yet understood) that causes constant stepsize methods to work for it although they do not work for general index 3 problems.

The type of difficulty that can occur can be seen in a very simple example. Consider the system

$$
\begin{aligned}
x - g(t) &= 0 \\
x' - y &= 0 \\
y' - z &= 0
\end{aligned}
\tag{7}
$$

Although of the form of equation (2), it is not a differential equation. It has the unique solution $(x, y, z) = (g, g', g'')$, independent of initial conditions. If we solve it with the backward Euler method (the BDF of order 1), we get

$$x_n = g_n \tag{8.1}$$

$$y_n = (x_n - x_{n-1})/h_n \tag{8.2}$$

$$z_n = (y_n - y_{n-1})/h_n \tag{8.3}$$

For $n > 3$ the initial values y_0 and z_0 needed to compute y_1, z_1, and z_2 no longer play a role and y_n is an $O(h)$ accurate approximation to $y(t_n) = g_n'$ provided g'' exists. However, unless $h_n = h_{n-1}$, z_n is not an approximation to z'': it is given by

$$z_n = z''(t_n)(1 - \frac{h_n - h_{n-1}}{2h_n})\ (1 + O(h_n + h_{n-1}))$$

This is an example of the general results to be summarized below: non-constant stepsize methods don't work for most problems, while constant stepsize methods work for a limited class of problems including constant coefficient linear problems.

CONSTANT COEFFICIENT LINEAR PROBLEMS

The constant coefficient linear problem

$$Ay' + By + g(t) = 0 \tag{9}$$

can be completely understood by transformation of the matrix pencil (A, B) to Kronecker canonical form (KCF). This is done in equation (9) by substituting $z = Qy$ and premultiplying by P to get

$$(PAQ)z' + (PBQ)z + Pg(t) = 0 \tag{10}$$

We are interested only in the case in which the system (9) is <u>solvable</u>, that is, it has a solution for all sufficient differentiable $g(t)$ and any two solutions with identical initial conditions are the same. It is shown in Sincovec et al. [5] that the necessary and sufficient condition for this is that the pencil (A, B) is nonsingular (that is, that $A + \lambda B$ is nonsingular for almost all λ). In this case, the KCF of the pencil is

$$PAQ = \begin{bmatrix} I_1 & 0 \\ 0 & E \end{bmatrix} \tag{11.1}$$

$$PBQ = \begin{bmatrix} C & 0 \\ 0 & I_2 \end{bmatrix} \tag{11.2}$$

where E is a block diagonal matrix (of the same dimension as I_2) whose blocks each have the form

$$E_i = \begin{bmatrix} 0 & & & \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix}$$

If z is partitioned into $[z^{*}, z_2^{*}]^{*}$, we see that equation (10) takes the form

$$z_1' + Cz_1 + q_1(t) = 0 \tag{12.1}$$

$$Ez_2' + z_2 + q_2(t) = 0 \tag{12.2}$$

and equation (12.2) decomposes into disjoint sets of equations of the form of equation (8). Equation (8) would arise from a 3 by 3 block E_i. The maximum size of the E_i matrices is called the <u>index</u> of the system. The following results are known for the constant coefficient linear problems of index m, and can be found in [5], [2], and [3].

<u>Theorem 2.1</u>. The constant stepsize, k-step BDF method is $O(h^k)$ accurate globally after a maximum of $(m-1)k+1$ steps if $k < 6$. (The results in the initial steps are influenced by initial errors. These errors can be amplified by as much as $O(h^{1-m})$ before they are annihilated.

<u>Theorem 2.2</u>. If the k-step constant stepsize BDF method with $k < 6$ is applied to (9) and the ratio of adjacent stepsizes is bounded, the global error is $O(h_{max}^{q})$ where $q = \min(k, k-m+2)$.

<u>Theorem 2.3</u>. The variable stepsize BDF method does not converge for $m > 3$ if the ratio of adjacent stepsizes is unbounded.

These results indicate that the constant stepsize method is the only one likely to be of use in general linear, constant coefficient problems. Unfortunately, even these methods break down for variable coefficient problems, as we will see in the next section.

<u>TIME-DEPENDENT AND NONLINEAR PROBLEMS</u>

When we study equation (2), it is natural to consider a linearization and associate B with $\partial b/\partial y$. Then we can define the <u>local index</u> of (2) to be the index of $(A, \partial b/\partial y)$. This may vary with y and t. (If it does, we may have singularities, turning points or other difficult phenomena.) In general it doesn't because the index is usually determined by the structure of the problem rather than current variable values.

Index 1 problems present no difficulty--it is proved in [3] that variable stepsize BDF methods work for such problems. The difficulty arises when the index exceeds 1. For this case it is known that even the constant stepsize BDF methods can fail. This can be seen by examining the error equation. We will look at it for the backward Euler method applied to

$$Ay' + B(t)y = g(t) \tag{13}$$

Let $P(t)$, $Q(t)$ transform this to a KCF and let $P_n = P(t_n)$, etc. We get

$$A(y_n - y_{n-1}) + hB_n y_n + hg_n = 0 \tag{14}$$

Substitute the true solution into the lefthand side of equation (14) to get

$$A(y(t_n) - y(t_{n-1})) + hB_n y(t_n) + hg_n = - A \frac{h^2}{2} y_n'' \tag{15}$$

where y_n'' has its components evaluated somewhere in the n-th step. Let $e_n = y_n - y(t_n)$. Subtracting (15) from (14) we get

$$e_n = [A + hB_n]^{-1} A(e_{n-1} + \frac{h^2}{2} y_n'') \tag{16}$$

In other words,

$$e_n = \frac{h^2}{2} \sum_{j=1}^{n} S_j^n y_n'' + S_1^n e_0 \tag{17}$$

where

$$S_j^n = S_n S_{n-1} \cdots S_j$$

and

$$S_j = [A + hB_j]^{-1} A$$

Note that

$$\begin{aligned} S_j &= Q_j Q_j^{-1} S_j Q_j Q_j^{-1} \\ &= Q_j \{ [P_j A Q_j + hP_j B_j Q_j]^{-1} P_j A Q_j \} Q_j^{-1} \\ &= Q_j K_j Q_j{-1} \end{aligned}$$

where

$$K_j = \begin{bmatrix} [I_1 + hC_j]^{-1} & 0 \\ 0 & [E + hI_2]^{-1} E \end{bmatrix}$$

If the method is to converge, we must be able to bound the matrices

$$S_j^n = Q_n K_n Q_n^{-1} Q_{n-1} K_{n-1} Q_{n-1}^{-1} \cdots Q_j K_j Q_j^{-1} \tag{18}$$

Note that if Q_j is independent of j this is trivial because $Q_n^{-1} S_j^n Q_j = K_n \cdots K_j$ is block diagonal so we only have to look at the blocks corresponding to each of the E_i. Each of these blocks takes the form

$$\{[E_i + hI]^{-1} E_i\}^{n-j+1} \tag{19}$$

$$= \begin{bmatrix} 0 & & & \\ h^{-1} & & & \\ -h^{-2} & & & \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ & -h^{-2} & h^{-1} & 0 \end{bmatrix}^{n-j+1}$$

which is zero for $n-j+1 > m-1$ because the displayed matrix has nilpotency m. The problem arises in time-dependent problems from the $Q_{j+1} Q_j^{-1}$ terms in (18) which can be written as

$$Q_{j+1} Q_j^{-1} = I + hQ_j' Q_j^{-1} + O(h^2) \tag{20}$$

The $O(h)$ term in this is multiplied by an h^{1-m} in equation (19) to cause problems when $m > 1$. (This problem is not an artifice of the analysis: it can be observed in examples given in [3].)

LAGRANGE EQUATIONS WITH CONSTRAINTS

The matrix $B = \partial b/\partial y$ for the Lagrange equations given by (5) is

$$B = \begin{bmatrix} 0 & -I & 0 \\ B_{21} & B_{22} & (\frac{\partial c}{\partial y})^* \\ \frac{\partial c}{\partial y} & 0 & 0 \end{bmatrix}$$

An algorithm is given in [3], section 4, for determining the local index, and its application shows that the index of this system is at least 3, and is exactly 3 if and only if

$$\frac{\partial c}{\partial y} (\frac{\partial c}{\partial y})^*$$

is nonsingular. If the constraints c are independent (so that $\partial c/\partial y$ has full rank) this matrix is positive definite, so in most cases the index is exactly 3. If the constraints are not independent, the Lagrange multipliers λ are nonunique so the problem is not solvable in our definition, although the variables of interest, y and z, can be determined. See Lötstedt [4].

It would appear that the Lagrange equations with constraints cannot be solved directly because of their high index, but in at least the pendulum problem (6), the constant stepsize method works as if the problem had constant coefficients. A possible reason for this is the structure of the matrix $Q_{j+1}Q_j^{-1}$ in (18). If this were zero in appropriate blocks, the properties of constant coefficient problems could be obtained. However, that does not appear to be the case for this problem. Matrices P and Q which take this to canonical form are (when $x^2 + y^2 = 1$)

$$P = \begin{bmatrix} 0 & 0 & y & -x & 0 \\ y & -x & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 \\ x & y & 0 & 0 & 0 \\ 0 & 0 & -x & -y & \lambda \end{bmatrix}$$

$$Q = \begin{bmatrix} 0 & y & x & 0 & 0 \\ 0 & -x & y & 0 & 0 \\ y & 0 & 0 & -x & 0 \\ -x & 0 & 0 & -y & 0 \\ 0 & 0 & 0 & 0 & -1/2 \end{bmatrix}$$

$Q_{j+1}Q_j^{-1}$ has the form

$$\begin{bmatrix} \alpha & 0 & 0 & -h\beta & 0 \\ 0 & \alpha & h\beta & 0 & 0 \\ 0 & -h\beta & \alpha & 0 & 0 \\ h\beta & 0 & 0 & \alpha & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

where $\alpha = 1 + O(h^2)$ and $\beta = O(1)$. This does not have zeros in positions which will annihilate the $O(h^{-2})$ term in K_1. The reason why the constant stepsize method works for the pendulum is not yet understood, so the question of how broad a class of Lagrange problems can be handled by constant stepsize methods has yet to be answered.

CONCLUSION

Some types of singular implicit problems present great difficulty, although even some high index problems can be solved for reasons not yet understood. As is pointed out in [3], extrapolation can be used with low order constant stepsize BDF methods for constant coefficient linear problems, which is the reason for exploring their application to nonlinear problems. The same is apparently true for the pendulum problem, that is, preliminary numerical tests are positive. The major question to be answered is how big is the class of problems given by Lagrange equations with constraints that can be handled in the same way.

REFERENCES

[1] Gear, C. W., The simultaneous numerical solution of differential-algebraic equations, IEEE Trans. Circuit Theory TC-18, (1), 1971, 89-95.

[2] Gear, C. W. and L. R. Petzold, Differential/algebraic systems and matrix pencils, Proc. Conference on Matrix Pencils, Pitea, Sweden, March 1982· also, Dept. Rpt. UIUCDCS-82-1086, Dept. Computer Sci., Univ. Illinois, Urbana, 1982.

[3] Gear, C. W. and L. R. Petzold, ODE methods for the solution of differential/algebraic systems, Dept. Rpt. UIUCDCS-82-1103, Dept. Computer Sci., Univ. Illinois, Urbana, 1982.

[4] Lötstedt, Per, A numerical method for the solution of mechanical systems with unilateral constraints, Report TRITA-NA-7920, Royal Inst. Technology, Stockholm, 1979.

[5] Sincovec, R. F., Dembart, B., Epton, M. A., Erisman, A. M., Manke, J. W. and E. L. Yip, Solvability of large-scale descriptor systems, Final Report DOE Contract ET-78-C-01-2876, Boeing Computer Services Co., Seattle, WA.