Numerical Methods for Boundary Value Problems in Differential-Algebraic Equations

Uri M. Ascher
Linda R. Petzold

This paper was prepared for submittal to AIChE National Meeting
Chicago, Illinois
November 11-16, 1990

September, 1990

This is a preprint of a paper intended for publication in a journal or proceedings. Since changes may be made before publication, this preprint is made available with the understanding that it will not be cited or reproduced without the permission of the author.
DISCLAIMER

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.
Numerical Methods for Boundary Value Problems in Differential-Algebraic Equations

Uri M. Ascher
Department of Computer Science
University of British Columbia
Vancouver, British Columbia
Canada V6T 1W5

Linda R. Petzold
Computing & Mathematics Research Division
Lawrence Livermore National Laboratory, L-316
Livermore, California 94550

September 24, 1990

Abstract

Differential-algebraic equation (DAE) boundary value problems arise in a variety of applications, including optimal control and parameter estimation for constrained systems. In this paper we survey these applications and explore some of the difficulties associated with solving the resulting DAE systems.

*The work of this author was partially supported under NSERC Canada Grant OGP0004306

†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.
For finite difference methods, the need to maintain stability in the differential part of the system often necessitates the use of methods based on symmetric discretizations. However, these methods can suffer from instability and loss of accuracy when applied to certain DAE systems. We describe a new class of methods, Projected Implicit Runge-Kutta Methods, which overcomes these difficulties. We give convergence and stability results, and present numerical experiments which illustrate the effectiveness of the new methods.

1 Introduction

Differential-algebraic equation (DAE) boundary value problems arise in a variety of applications, including the determination of optimal control profiles for chemical engineering processes, [15,10,17,18], parameter estimation for differential-algebraic systems in chemical [19,20] and mechanical [7] engineering, and trajectory-prescribed path control for projectiles [6]. The solution of these problems can be numerically quite challenging, in part because often the problems are described by higher-index DAEs\(^1\), for which traditional numerical ODE methods such as collocation can exhibit difficulties such as instability, oscillation and loss of accuracy. In this paper we will explore these problems, and describe a new class of numerical methods which have the possibility of resolving the difficulties in the context of collocation methods such as those used in the ODE BVP solver COLSYS [4].

Although there is now a large body of work addressing the solution of initial-value problems in DAEs (see for example the recent monographs [8, 13,12]), until recently relatively little attention has focused on the solution of DAE boundary value problems. Much of the work to this date has dealt with the solution of semi-explicit index-one systems

\[ y' = f(y, z) \tag{1a} \]

\(^1\)The index of a DAE is a measure of the degree of singularity of the DAE. For systems of ODEs subject to constraints \( y' = f(y, z), \ 0 = g(y, z) \), the index is the number of times the constraints must be differentiated with respect to \( t \) (substituting for \( y' \) from the differential equation) before arriving at an ODE for \( y', z' \). Thus, a standard-form ODE has index 0. Higher-index (generally, this means index 2 or greater) problems pose difficulties for numerical methods. For further details, see [8]. The index is closely related to the order of the path constraints.
$0 = g(y, z)$ \hspace{1cm} (1b)

where $\partial g/\partial z$ is nonsingular. Ascher[2] gives a convergence result for collocation schemes applied to (1), where the collocation methods are applied in such a way that the algebraic components of the system are approximated in a piecewise discontinuous space. Bock, Eich and Schlöder [7] 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 to semi-explicit index-one systems. März and Griepentrog [12] consider fully-implicit index-one DAEs, and give some theoretical results for certain index two systems, 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. Lamour [14] develops a shooting method for index-one DAEs. Clark and Petzold [9] extend the theory for shooting and finite-difference methods for linear boundary value problems in ODEs to the DAE case, including higher-index DAEs. The theory shows how stable and accurate methods for DAE initial-value problems can be extended to boundary value problems. However, the question of how to obtain stable and accurate methods for higher-index DAE systems is not considered.

The determination of optimal control profiles often leads to higher-index DAE boundary value problems. Examples within chemical engineering include problems in reactor design, process start-up, batch process operation, etc. The solution of these DAE optimization problems has been considered recently in [15,10], however much work remains to be done on the theoretical justification of these methods for nonlinear systems, the development of stable and highly-accurate difference formulas for higher-index systems, error estimation and stepsize control [20], etc. Following Logsdon and Biegler [15], the problems take the form

$$\min_{u(t), z(t)} \psi(z(b)) + \int_{a}^{b} \Phi(z(t), u(t)) dt$$

subject to

$$z'(t) = F(z(t), u(t))$$
$$g(u(t), z(t)) \leq 0$$
where $\psi(z(b))$ is the component of the objective function due to the final conditions, $\int_a^b \Phi(z(t), u(t)) dt$ is the component of the objective function due to the integral of state and control vectors, $g$ is the inequality design constraint vector, $z(t)$ is the state profile vector, $u(t)$ is the control profile, $g_f$ is the final conditions inequality constraint, $z_0$ is the initial condition for the state vector, and $z(t)^L$, $z(t)^U$ and $u(t)^L$, $u(t)^U$ are the state and control profile bounds, respectively.

The variational conditions for this problem are

$$
\begin{aligned}
\frac{\partial \Phi}{\partial u} + \frac{\partial F}{\partial u} \Lambda + \frac{\partial g}{\partial u} M = 0 \quad (2a) \\
\frac{\partial \Phi}{\partial z} + \frac{\partial F}{\partial z} \Lambda + \frac{\partial g}{\partial z} M + \dot{\Lambda}(t) = 0 \quad (2b) \\
g(u(t), z(t)) \leq 0 \quad (2c) \\
M(t) g(z(t)) = 0, \quad M(t) \geq 0 \quad (2d) \\
\dot{z}(t) = F(z(t), u(t)), \quad z(a) = z_0 \quad (2e) \\
\Lambda(b) = \left[ \frac{\partial \Phi}{\partial z} + \frac{\partial g_f}{\partial z} M_f \right]_{t_f=b} \quad (2f)
\end{aligned}
$$

where $M(t)$ and $\Lambda(t)$ are adjoint functions of the constraint $g(u(t), z(t)) \leq 0$ and the ODE model, respectively. These conditions form a DAE system, which can be higher index when the constraints (2c) are active. After discretization by collocation methods, the nonlinear programming problem can be solved by techniques such as successive quadratic programming [15,10]. Biegler et al. [15,20] use this approach, but they note difficulties in the solution of index-2 and higher problems, including loss of accuracy of the formulas and difficulties with error estimation and stepsize control. Parameter estimation problems in differential-algebraic equations lead to a set of equations similar to (2), and are efficiently solved by a boundary value DAE numerical approach[19,7].
In this paper, we will consider the solution of index-2 Hessenberg DAE systems. These are systems of the form

\begin{align}
    \mathbf{x}' &= g_1(\mathbf{x}, \mathbf{y}, t) \\
    \mathbf{0} &= g_2(\mathbf{x}, t) \\
    \mathbf{0} &= b(\mathbf{x}(0), \mathbf{x}(1))
\end{align}

The system is index-two if \((\partial g_2/\partial x)(\partial g_1/\partial y)\) is nonsingular. The methods we propose are easily extended to accommodate semi-explicit DAE systems of mixed index one and two. These systems arise naturally when a system of ODEs is subjected to constraints. In (3), the variable \(\mathbf{y}\) plays the role of a Lagrange multiplier. Higher-index Hessenberg DAEs [8] can be brought into this form by differentiating the constraints and adding additional Lagrange multipliers to satisfy the new constraints [11]. Although the index-two systems can be transformed into index-one systems through an additional constraint differentiation, numerical methods applied to the resulting systems no longer preserve the original constraints.

We will first develop an analysis of the conditioning of problems (3), which is an important tool for understanding the stability both of different formulations of DAEs which have the same analytical solutions (such as the formulations which involve differentiation of the constraints discussed above), and of numerical methods applied directly to (3). Because a well-conditioned boundary value problem may have both fast increasing and fast decreasing modes, it is important to be able to use symmetric schemes such as Gaussian collocation. However, previously defined symmetric schemes [13,8] have been shown to be unstable [3] for some well-conditioned DAEs of the form (3). In addition, there is a loss of accuracy which is particularly severe for symmetric methods applied to higher-index DAEs [13]. We will describe a simple modification to implicit Runge-Kutta methods which resolves these difficulties. Application of this idea to collocation methods yields a class of methods which are stable and achieve superconvergence order for (3), and are potentially implementable in a code such as COLSYS [4]. We present

\footnote{Strictly speaking, the instability was shown for fully-implicit index-one systems. However, the example in [3] can be trivially modified to create a Hessenberg index-2 system exhibiting the same instability.}
a numerical example showing the effectiveness of these methods, and finally give some conclusions and directions for future work.

This paper gives an overview of problems, recent results and future plans; for a detailed examination of the methods and their analysis, see [1].

2 Problem conditioning

It is well-known (see e.g. [12], [2]) that DAE problems with index exceeding one are in a sense ill-posed. Hence it is important to investigate the conditioning (stability) of such problems carefully. Such a conditioning analysis enables the evaluation of stability of the various possible formulations of the DAE, as well as of the stability of numerical methods for its solution.

Consider the linear index-two Hessenberg boundary value problem

\[ \begin{align*}
    x' &= G_{11}x + G_{12}y + q_1 \\
    0 &= G_{21}x + q_2 \\
    \beta &= B_0x(0) + B_1x(1)
\end{align*} \tag{4a-4c} \]

where \( G_{11}, G_{12} \) and \( G_{21} \) are smooth functions of \( t \), \( 0 \leq t \leq 1 \), \( G_{11}(t) \in \mathcal{R}^{m_x \times m_x}, G_{12}(t) \in \mathcal{R}^{m_x \times m_y}, G_{21}(t) \in \mathcal{R}^{m_y \times m_x}, m_y \leq m_x \), \( G_{21}G_{12} \) is nonsingular for each \( t \) (hence the DAE is index two), and \( B_0, B_1 \in \mathcal{R}^{(m_\mathcal{E} - m_y) \times m_x} \). All matrices involved are assumed to be uniformly bounded in norm by a constant of moderate size. The inhomogeneities are \( q_1(t) \in \mathcal{R}^{m_x}, q_2(t) \in \mathcal{R}^{m_y}, \beta \in \mathcal{R}^{m_\mathcal{E} - m_y} \).

We seek conditions under which this BVP is guaranteed to be well-conditioned (stable) in an appropriate sense. Since \( G_{21}G_{12} \) is nonsingular, \( G_{12} \) has full rank. Hence there exists a smooth, bounded matrix function \( R(t) \in \mathcal{R}^{(m_\mathcal{E} - m_y) \times m_x} \) whose linearly independent rows form a basis for the nullspace of \( G_{12}^T \). Further, \( R(t) \) can be taken to be orthonormal [1]. Thus, for each \( t, 0 \leq t \leq 1 \),

\[ RG_{12} = 0. \tag{5} \]

We assume, more strongly, that there exists a constant \( K_1 \) of moderate size such that

\[ \|(G_{21}G_{12})^{-1}\| \leq K_1. \tag{6} \]
Then \([1]\) there is a constant \(K_2\) of moderate size such that

\[
\| \left( \begin{array}{c} R \\ G_{21} \end{array} \right)^{-1} \| \leq K_2.
\]  

(7)

Multiplying (4a) by \(R\) we have

\[
Rx' = R(G_{11}x + q_1).
\]  

(8)

Let

\[
v = Rx \quad 0 \leq t \leq 1.
\]  

(9)

Then, using (4b), the inverse transformation is given by

\[
x = \left( \begin{array}{c} R \\ G_{21} \end{array} \right)^{-1} \left( \begin{array}{c} v \\ -q_2 \end{array} \right) \equiv S v + \hat{q}
\]  

(10)

where \(S(t) \in \mathcal{R}^{m_x \times (m_x - m_y)}\) satisfies

\[
RS = I, \quad G_{21}S = 0.
\]  

(11)

Differentiating (9) and substituting (8), we obtain the \textit{essential underlying ODE}

\[
v' = (RG_{11} + R')S[v + Rq_1 + (RG_{11} + R')\hat{q}],
\]  

(12)

which is subject to \(m_x - m_y\) boundary conditions, obtained from (4c) using (10):

\[
(B_0S(0))v(0) + (B_1S(1))v(1) = \beta - B_0\hat{q}(0) - B_1\hat{q}(1).
\]  

(13)

Now, if the ordinary BVP (12), (13) is stable, i.e. if its Green's function is bounded by a constant of moderate size, then a similar conclusion holds for the DAE. We obtain the following stability theorem:

\textbf{Theorem 1} \textit{Let the BVP (4) have smooth, bounded coefficients, and assume that (5) holds and that the essential underlying BVP (12)-(13) is stable. Then there is a constant \(K\) of moderate size such that}

\[
\|x\| \leq K(\|q_1\| + \|q_2\| + |\beta|)
\]  

(14a)

\[
\|y\| \leq K(\|q'_2\| + \|q_1\| + \|q_2\| + |\beta|)
\]  

(14b)
Proof:
Our assumptions guarantee the well-conditioning of the transformation (9), (10). Hence, the inhomogeneities appearing in (12), (13) are bounded in terms of the original ones. The stability of the BVP (12), (13) guarantees a similar bound for $\|v\|$. Conclusion (14a) is then obtained using (10).

Now, given $x$ we obtain $y$ through multiplying (4) by $G_{21}$, yielding

$$y = (G_{21}G_{12})^{-1}G_{21}(x' - G_{11}x - q_1).$$  \hspace{1cm} (15)

Differentiating (4b) we substitute $G_{21}x' = -G_{21}'x - q_2$ in (15). The bound (14b) is obtained from this expression using (14a) and (6). \Box

3 Projected IRK methods

Consider the DAE problem (3). Let $b = (b_1, ..., b_k)^T$, $c = (c_1, ..., c_k)^T$, $A = (a_{ij})_{i,j=1}^k$ be the coefficients of a $k$-stage Implicit Runge-Kutta (IRK) scheme (see, e.g., [8]). We assume that $0 \leq c_1 \leq c_2 \leq ... \leq c_k \leq 1$ and that $A$ is nonsingular (which excludes Lobatto schemes but leaves in all other IRK schemes of practical interest). Denote the internal stage order by $k_I$ ($k_I \geq 1$ for consistency) and the nonstiff order at mesh points by $k_d$ ($k_d \leq 2k$). For collocation schemes, in particular, $k_I = k$ and the $c_i$ are distinct.

Given a mesh

$$\pi : 0 = t_0 < t_1 < ... < t_N = 1$$
$$h_n := t_n - t_{n-1}$$
$$h := \max\{h_n, 1 \leq n \leq N\}$$

(16)

a projected IRK method for (3) samples (3c), requires

$$0 = g_2(x_0, 0)$$

and approximates (3a),(3b) on each mesh subinterval $[t_{n-1}, t_n], 1 \leq n \leq N$, by

$$X'_i = g_1(X_i, Y_i, t_i)$$
$$0 = g_2(X_i, t_i), \quad i = 1, 2, ..., k$$

(17a)
(17b)
\[ x_n = x_{n-1} + h_n \sum_{j=1}^{k} b_j x_j' + G_{12}^n \lambda_n \]  
\[ 0 = g_2(x_n, t_n), \]  
(17c)  
(17d)

where \( t_i = t_{n-1} + h_n c_i \), \( X_i = x_{n-1} + h_n \sum_{j=1}^{k} a_{ij} x_j' \) and \( G_{12}^n = \frac{\partial g_1}{\partial y}(x_n, y_n, t_n) \).

(We are using \( i \) as a local index at each step \( n \). Also, \( y_n \) is the value of the polynomial interpolant of \( Y_i, i = 1, \ldots, k \), at \( t_n \).)

Observe that if we drop the requirement (17d) and set \( \lambda_n = 0 \) then an IRK method is obtained as discussed in [8,13]. Thus, if \( \hat{x}_n \) is the result of one IRK step starting from \( x_{n-1} \), then \( x_n \) is given by

\[ x_n = \hat{x}_n + G_{12}^n \lambda_n \]  
(18)

and can be viewed as the projection of \( \hat{x}_n \) onto the algebraic manifold at the next mesh point \( t_n \).

We now give a basic existence, stability and convergence theorem for the linear case.

**Theorem 2** Given a stable, semi-explicit, linear Hessenberg index two system (4) to be solved numerically by the \( k \)-stage projected IRK method, then for \( h \) sufficiently small

1. The local error in \( x \) is \( O(h_{k+1}^{\min(k_d+1,k_1+1)}) \).
2. There exists a unique projected IRK solution.
3. The projected IRK method is stable, with a moderate stability constant, provided that the BVP has a moderate stability constant \( K \).
4. The global error in \( x \) is \( O(h_{k+1}^{\min(k_d,k_1+1)}) \).
5. The errors in the intermediate variables \( X_i' \) and \( X_i \) are \( O(h_{k+1}^{\min(k_d,k_1)}) \) and \( O(h_{k+1}^{\min(k_d,k_1+1)}) \), respectively.

In the practically important case where the unprojected IRK scheme is a collocation scheme, (17) defines a class of projected collocation methods. For these methods, we can give a much sharper order result, namely
**Theorem 3** Under the assumptions of Theorem 2, the projected collocation method satisfies for $0 \leq t \leq 1$

\begin{align}
|x_\pi(t) - x(t)| &= O(h^{\min(t+1,k_d)}) \\
|x'_\pi(t) - x'(t)| &= O(h^k) \\
y_\pi(t) - y(t)| &= O(h^k).
\end{align}

(19a) \hspace{2cm} (19b) \hspace{2cm} (19c)

Let the coefficient functions and the inhomogeneities in (4) be in $C^{k_d+1}[0,1]$. Then the nonstiff superconvergence order holds for the projected collocation method,

$$|x_n - x(t_n)| = O(h^{k_d}) \quad 0 \leq n \leq N.$$  

(20)

The proofs of these theorems can be found in [1]. The basic approach in the proofs is to show that adequate approximations for the essential underlying ODE (12) are (implicitly) obtained by the projected methods.

Finally, the results from Theorems 1-3 can be combined using standard arguments to yield a convergence theorem for projected collocation methods applied to nonlinear problems.

**Theorem 4** Let $x(t)$, $y(t)$ be an isolated solution of the DAE problem (4) and assume that $g_1$ and $g_2$ have continuous second partial derivatives and that the smoothness assumptions of Theorem 3 hold for the linearized problem in the neighborhood of $x(t)$, $y(t)$. Then there are positive constants $\rho$ and $h_0$ such that for all meshes with $h \leq h_0$

1. There is a unique solution $x_\pi(t), y_\pi(t)$ to the projected collocation equations (17) in a tube $S_\rho(x,y)$ of radius $\rho$ around $x(t)$, $y(t)$.

2. This solution can be obtained by Newton’s method, which converges quadratically provided that the initial guess for $x_\pi(t), y_\pi(t)$ is sufficiently close to $x(t)$, $y(t)$.

3. The error estimates (19)-(20) hold.

4 **Numerical Experiment**

To illustrate how well the projected implicit Runge-Kutta methods work, as compared with their non-projected counterparts, we solved the following
linear problem

\[
x' = \begin{pmatrix} \lambda - \frac{1}{t-2} & 0 \\ \frac{1}{t-2} & -1 \end{pmatrix} x + \begin{pmatrix} (2-t) \lambda \\ \lambda - 1 \end{pmatrix} y + \begin{pmatrix} 3 - t \\ 2 - t \end{pmatrix} e^t \\
0 = (t + 2 \ t^2 - 4) x - (t^2 + t - 2) e^t, \quad \lambda > 0
\]

with initial value \( x_1(0) = 1 \). This problem has the true solution

\[
x = (e^t \ e^t), \quad y = \frac{-e^t}{2 - t}
\]

In Table 1, we present the results of solving this problem, with \( \lambda = 50 \), with the projected and unprojected forms of the 3-stage Gaussian collocation method, with various uniform meshes. The error shown is the error in \( x_1 \) and \( x_2 \). Behavior of the methods for other positive values of \( \lambda \) and for other Gaussian collocation methods was similar.

<table>
<thead>
<tr>
<th>Method</th>
<th>Mesh size</th>
<th>Error(_1)</th>
<th>Error(_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Projected</td>
<td>10</td>
<td>.26e-3</td>
<td>.18e-3</td>
</tr>
<tr>
<td>Projected</td>
<td>20</td>
<td>.71e-7</td>
<td>.59e-7</td>
</tr>
<tr>
<td>Projected</td>
<td>40</td>
<td>.74e-9</td>
<td>.45e-9</td>
</tr>
<tr>
<td>Projected</td>
<td>80</td>
<td>.10e-9</td>
<td>.59e-10</td>
</tr>
<tr>
<td>Unprojected</td>
<td>10</td>
<td>.19e+9</td>
<td>.18e+9</td>
</tr>
<tr>
<td>Unprojected</td>
<td>20</td>
<td>.61e+10</td>
<td>.59e+10</td>
</tr>
<tr>
<td>Unprojected</td>
<td>40</td>
<td>.18e+8</td>
<td>.18e+9</td>
</tr>
<tr>
<td>Unprojected</td>
<td>80</td>
<td>.79e+6</td>
<td>.78e+6</td>
</tr>
</tbody>
</table>

Table 1: Errors for projected vs. unprojected Gaussian collocation

The results clearly show that the projected methods overcome the instability problem and achieve a high rate of convergence.

## 5 Conclusion

We have introduced a new class of numerical methods, *Projected Implicit Runge-Kutta Methods*, for the solution of index-two Hessenberg differential-algebraic systems. The new methods appear to be particularly promising
for boundary value problems, and overcome many of the difficulties associated with previously defined methods for this class of problems. We have developed some important tools for stability analysis and introduced the essential underlying ODE, which enable the understanding of numerical stability behavior for linear systems and numerical methods applied to various formulations of the DAE. Future work is planned to include a nonlinear stability analysis, unified numerical methods for index 0 – 2, and methods for inequality constraints and singular segments. A robust general-purpose code is planned, based on collocation methods. It is expected that the new methods and software will ultimately lead to the solution of a wide variety of applications from control and parameter estimation.

References


