UCRL-JC--104926

DE91 000660

Projected Implicit Runge-Kutta Methods for Differential-Algebraic Boundary Value Problems

Uri 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

August 17, 1990

Abstract

Differential-algebraic boundary value problems arise in the modelling of singular optimal control problems and in parameter estimation for singular systems. A new class of numerical methods for these problems is introduced, and shown to overcome difficulties with previously defined numerical methods.

*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.

1 Introduction

In this paper we describe a new class of numerical methods, *Projected Implicit Runge-Kutta methods* (PIRK), for the solution of index-two Hessenberg systems of initial and boundary value differential-algebraic equations (DAEs)

$$\begin{aligned} \mathbf{x}' &= \mathbf{g}_1(\mathbf{x},\mathbf{y},t) && \text{(1a)} \\ \mathbf{0} &= \mathbf{g}_2(\mathbf{x},t) && \text{(1b)} \\ \mathbf{0} &= \mathbf{b}(\mathbf{x}(0),\mathbf{x}(1)) && \text{(1c)} \end{aligned}$$

The system is index-two if $(\partial \mathbf{g}_2/\partial x)(\partial \mathbf{g}_1/\partial y)$ is nonsingular. These types of systems arise for example in the modelling of singular optimal control problems[5,11], where $\mathbf{y}$ is the control variable in (1), and in parameter estimation for differential-algebraic equations such as multibody systems[6][1]. The new methods appear to be particularly promising for the solution of boundary value problems of the form (1), where the need to maintain stability in the differential part of the system often necessitates the use of methods based on symmetric discretizations. Previously defined numerical methods based on symmetric discretizations have been shown to have severe limitations, including instability, oscillation and loss of accuracy, when applied to (1)[3,7,10] . The new methods overcome these difficulties. Numerical results have so far been very encouraging. However, much work remains to be done before these methods can be made available in the form of a robust general-purpose code such as those now available for ODE boundary value problems[4]. We provide here an overview of our recent results and future plans; for a detailed examination of the methods and analysis, see [1].

2 Problem conditioning

It is well-known (see e.g. [9], [2]) that DAE problems with index exceeding one are in a sense ill-posed. Hence it is important to investigate the

[1] Multibody systems are often formulated initially as index-three DAEs. However, they can easily be converted to the index-two form by techniques introduced by Gear[8]. It can be shown that this reduction does not introduce any conditioning difficulties into the system.

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

$$\mathbf{x}' = G_{11}\mathbf{x} + G_{12}\mathbf{y} + \mathbf{q}_1 \tag{2a}$$
$$\mathbf{0} = G_{21}\mathbf{x} + \mathbf{q}_2 \tag{2b}$$
$$\beta = B_0\mathbf{x}(0) + B_1\mathbf{x}(1) \tag{2c}$$

where G_{11}, G_{12} and G_{21} are smooth functions of t, $0 \le t \le 1$, $G_{11}(t) \in \mathcal{R}^{m_x \times m_x}$, $G_{12}(t) \in \mathcal{R}^{m_x \times m_y}$, $G_{21}(t) \in \mathcal{R}^{m_y \times m_x}$, $m_y \le m_x$, $G_{21}G_{12}$ is nonsingular for each t (hence the DAE is index two), and $B_0, B_1 \in \mathcal{R}^{(m_x - m_y) \times m_x}$. All matrices involved are assumed to be uniformly bounded in norm by a constant of moderate size. The inhomogeneities are $\mathbf{q}_1(t) \in \mathcal{R}^{m_x}, \mathbf{q}_2(t) \in \mathcal{R}^{m_y}, \beta \in \mathcal{R}^{m_x - m_y}$.

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 $R(t) \in \mathcal{R}^{(m_x - m_y) \times m_x}$ 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, $0 \le t \le 1$,

$$RG_{12} = 0. \tag{3}$$

We assume, more strongly, that there exists a constant $\hat{K}$ of moderate size for orthonormal $R(t)$ satisfying (3) such that [1]

$$\left\| \begin{pmatrix} R \\ G_{21} \end{pmatrix}^{-1} \right\| \le \hat{K}. \tag{4}$$

Multiplying (2a) by R we have

$$R\mathbf{x}' = R(G_{11}\mathbf{x} + \mathbf{q}_1). \tag{5}$$

Let

$$\mathbf{v} = R\mathbf{x} \qquad 0 \le t \le 1. \tag{6}$$

Then, using (2b), the inverse transformation is given by

$$\mathbf{x} = \begin{pmatrix} R \\ G_{21} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{v} \\ -\mathbf{q}_2 \end{pmatrix} \equiv S\mathbf{v} + \hat{\mathbf{q}} \tag{7}$$

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

$$RS = I, \qquad G_{21}S = 0. \tag{8}$$

Differentiating (6) and substituting (5), we obtain the *underlying ODE*

$$\mathbf{v}' = [(RG_{11} + R')S]\mathbf{v} + [R\mathbf{q}_1 + (RG_{11} + R')\hat{\mathbf{q}}], \tag{9}$$

which is subject to $m_x - m_y$ boundary conditions, obtained from (2c) using (7):

$$(B_0S(0))\mathbf{v}(0) + (B_1S(1))\mathbf{v}(1) = \beta - B_0\hat{\mathbf{q}}(0) - B_1\hat{\mathbf{q}}(1). \tag{10}$$

Now, if the ordinary BVP (9), (10) 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:

Theorem 1 *Let the BVP (2) have smooth, bounded coefficients, and assume that (4) holds and that the underlying BVP (9)-(10) is stable. Then there is a constant K of moderate size such that*

$$\begin{aligned} \|\mathbf{x}\| &\leq K(\|\mathbf{q}_1\| + \|\mathbf{q}_2\| + |\beta|) && \text{(11a)} \\ \|\mathbf{y}\| &\leq K(\|\mathbf{q}_1'\| + \|\mathbf{q}_2'\| + \|\mathbf{q}_1\| + \|\mathbf{q}_2\| + |\beta|) && \text{(11b)} \end{aligned}$$

Proof:

Our assumptions guarantee the well-conditioning of the transformation (6), (7). Hence, the inhomogeneities appearing in (9), (10) are bounded in terms of the original ones. The stability of the BVP (9), (10) guarantees a similar bound for $\|\mathbf{v}\|$. Conclusion (11a) is then obtained using (7).

Now, given $\mathbf{x}$ we obtain $\mathbf{y}$ through multiplying (2) by G_{21}, yielding

$$\mathbf{y} = (G_{21}G_{12})^{-1}G_{21}(\mathbf{x}' - G_{11}\mathbf{x} - \mathbf{q}_1). \tag{12}$$

The bound (11b) is obtained from this expression using (11a) and (4). □

3 Projected IRK methods

Consider the DAE problem (1). Let $b = (b_1, ..., b_k)^T$, $c = (c_1, ..., c_k)^T$, $\mathcal{A} = (a_{ij})_{i,j=1}^k$ be the coefficients of a k-stage Implicit Runge-Kutta (IRK) scheme (see, e.g., [7]). We assume that $0 \le c_1 \le c_2 \le ... \le c_k \le 1$ and that $\mathcal{A}$ is nonsingular (which excludes Lobatto schemes but leaves in all other IRK schemes of practical interest). Denote the internal stage order by k_I ($k_I \ge 1$ for consistency) and the nonstiff order at mesh points by k_d ($k_d \le 2k$). For collocation schemes, in particular, $k_I = k$ and the c_i are distinct.

Given a mesh

$$\begin{aligned} &\pi : 0 = t_0 < t_1 < ... < t_N = 1 \\ &h_n := t_n - t_{n-1} \\ &h := \max\{h_n, 1 \le n \le N\} \end{aligned} \tag{13}$$

a projected IRK method for (1) samples (1c), requires

$$\mathbf{0} = \mathbf{g}_2(\mathbf{x}_0, 0)$$

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

$$\mathbf{X}_i' = \mathbf{g}_1(\mathbf{X}_i, \mathbf{Y}_i, t_i) \tag{14a}$$
$$\mathbf{0} = \mathbf{g}_2(\mathbf{X}_i, t_i), \qquad i = 1, 2, \ldots, k \tag{14b}$$

$$\mathbf{x}_n = \mathbf{x}_{n-1} + h_n \sum_{j=1}^{k} b_j \mathbf{X}_j' + G_{12}^n \lambda_n \tag{14c}$$
$$\mathbf{0} = \mathbf{g}_2(\mathbf{x}_n, t_n), \tag{14d}$$

where $t_i = t_{n-1} + h_n c_i$, $\mathbf{X}_i = \mathbf{x}_{n-1} + h_n \sum_{j=1}^k a_{ij} \mathbf{X}_j'$ and $G_{12}^n = \frac{\partial \mathbf{g}_1}{\partial \mathbf{y}}(\mathbf{x}_n, \mathbf{y}_n, t_n)$.

Observe that if we drop the requirement (14d) and set $\lambda_n = \mathbf{0}$ then an IRK method is obtained as discussed in [7,10]. Thus, if $\hat{\mathbf{x}}_n$ is the result of one IRK step starting from $\mathbf{x}_{n-1}$, then $\mathbf{x}_n$ is given by

$$\mathbf{x}_n = \hat{\mathbf{x}}_n + G_{12}^n \lambda_n \tag{15}$$

and can be viewed as the projection of $\hat{\mathbf{x}}_n$ 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 (2) to be solved numerically by the k-stage projected IRK method, then for h sufficiently small*

1. *The local error in* $\mathbf{x}$ *is* $O(h_n^{\min(k_d+1,k_I+2)})$.
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* $\mathbf{x}$ *is* $O(h^{\min(k_d,k_I+1)})$.
5. *The errors in the intermediate variables* $\mathbf{X}_i'$ *and* $\mathbf{X}_i$ *are* $O(h^{\min(k_d,k_I)})$ *and* $O(h^{\min(k_d,k_I+1)})$, *respectively.*

In the practically important case where the unprojected IRK scheme is a collocation scheme, (14) 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* $0 \leq t \leq 1$

$$\begin{aligned} |\mathbf{x}_\pi(t) - \mathbf{x}(t)| &= O(h^{\min(k+1,k_d)}) && (16a)\\ |\mathbf{x}_\pi'(t) - \mathbf{x}'(t)| &= O(h^k) && (16b)\\ |\mathbf{y}_\pi(t) - \mathbf{y}(t)| &= O(h^k). && (16c)\end{aligned}$$

Let the coefficient functions and the inhomogeneities in (2) be in $C^{k_d+1}[0,1]$. *Then the nonstiff superconvergence order holds for the projected collocation method,*

$$|\mathbf{x}_n - \mathbf{x}(t_n)| = O(h^{k_d}) \qquad 0 \leq n \leq N. \tag{17}$$

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* $\mathbf{x}(t)$, $\mathbf{y}(t)$ *be an isolated solution of the DAE problem (2) and assume that* $\mathbf{g}_1$ *and* $\mathbf{g}_2$ *have continuous second partial derivatives and that the smoothness assumptions of Theorem 3 hold for the linearized problem in the neighborhood of* $\mathbf{x}(t)$, $\mathbf{y}(t)$. *Then there are positive constants* ρ *and* h_0 *such that for all meshes with* $h \le h_0$

1. *There is a unique solution* $\mathbf{x}_\pi(t), \mathbf{y}_\pi(t)$ *to the projected collocation equations (14) in a tube* $S_\rho(\mathbf{x}, \mathbf{y})$ *of radius* ρ *around* $\mathbf{x}(t)$, $\mathbf{y}(t)$.
2. *This solution can be obtained by Newton's method, which converges quadratically provided that the initial guess for* $\mathbf{x}_\pi(t)$, $\mathbf{y}_\pi(t)$ *is sufficiently close to* $\mathbf{x}(t)$, $\mathbf{y}(t)$.
3. *The error estimates (16)-(17) 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

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

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

$$x = \begin{pmatrix} e^t & e^t \end{pmatrix}, \quad y = \frac{-e^t}{2-t}$$

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 λ and for other Gaussian collocation methods was similar.

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

Method	*Mesh size*	$Error_1$	$Error_2$
Projected	10	.26e-3	.18e-3
Projected	20	.71e-7	.59e-7
Projected	40	.74e-9	.45e-9
Projected	80	.10e-9	.59e-10
Unprojected	10	.19e+9	.18e+9
Unprojected	20	.61e+10	.59e+10
Unprojected	40	.18ee+8	.18e+8
Unprojected	80	.79e+6	.78e+6

Table 1: Errors for projected vs. unprojected Gaussian collocation

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 underlying ODE, which enable the understanding of numerical stability behavior for linear systems. 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

[1] U. Ascher and L. Petzold, *Projected Implicit Runge-Kutta Methods for Differential-Algebraic Equations*, Lawrence Livermore National Laboratory UCRL-JC-104037, May 1990 (submitted to SIAM J. Numer. Anal.)

[2] U. Ascher, *On numerical differential algebraic problems with application to semiconductor device simulation*, SIAM J. Numer. Anal. 26 (1989),

517-538.

[3] U. Ascher, *On symmetric schemes and differential-algebraic equations*, SIAM J. Scient. Stat. Comput. 10 (1989), 937-949.

[4] U. Ascher, R. Mattheij and R. Russell, *Numerical Solution of Boundary Value Problems for Ordinary Differential Equation*, Prentice-Hall, NJ 1988.

[5] J. Betts, T. Bauer, W. Huffman, and K. Zondervan, *Solving the optimal control problem using a nonlinear programming technique part 1: general formulation*, AIAA-84-2037, Proc. AIAA/AAS Astrodynamics Conference, 1984.

[6] H. G. Bock, E. Eich and J. P. Schlöder, *Numerical solution of constrained least squares boundary value problems in differential-algebraic equations*, Universität Heidelberg, 1987.

[7] K. Brenan, S. Campbell and L. Petzold, *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, North-Holland, 1989.

[8] C.W. Gear, *Differential-algebraic equation index transformations*, SIAM J. Scient. Stat. Comput. 9 (1988), 39-47.

[9] E. Griepentrog and R. März, *Differential-Algebraic Equations and their Numerical Treatment*, Teubner-Texte Math. 88, Teubner, Leipzig, GDR 1986.

[10] E. Hairer, C. Lubich and M. Roache, *The Numerical Solution of Differential-Algebraic Systems by Runge-Kutta Methods*, Springer Lecture Notes in Mathematics No. 1409, 1989.

[11] H. Maurer, *Numerical solution of singular control problems using multiple shooting techniques*, J. of Optimization Theory and Applications 18 (1976), 235-257.

END

DATE FILMED

11 / 08 / 90