NUMERICAL ANALYSIS AND MATHEMATICAL MODELLING
BANACH CENTER PUBLICATIONS, VOLUME 24
PWN-POLISH SCIENTIFIC PUBLISHERS
WARSAW 1990

SHOOTING METHODS FOR TRANSFERABLE DAE'S – SOLUTION OF SHOOTING EQUATION

R. LAMOUR

Section of Mathematics, Humboldt University of Berlin,
Berlin, German Democratic Republic

Consider the differential-algebraic equation (DAE)

(1) $$f\big(x'(t), x(t), t\big) = 0, \qquad t\in[a, b],$$

with boundary condition

(2) $$g\big(x(a), x(b)\big) = 0,$$

where f and g fulfil the following assumptions (A):

- f: $B_f \subset \mathbf{R}^n\times\mathbf{R}^n\times[a, b]\to\mathbf{R}^n$, f'_y and f'_x exist and are smooth;
- $\ker\big(f'_y(y, x, t)\big) = N(t)\forall(y, x, t)\in B_f$, $\operatorname{rank}\big(f'_y(y, x, t)\big) = r$, $\dim N(t) = n-r$;
- $Q(t)$ denotes a projection onto $N(t)$, Q is smooth and $P(t):= I-Q(t)$;
- the matrix $G(y, x, t):= f'_y(y, x, t)+f'_x(y, x, t)Q(t)$ is regular $\forall(y, x, t)\in B_f$, i.e. f is transferable;
- g: $B_g\subset\mathbf{R}^n\times\mathbf{R}^n\to M\subset\mathbf{R}^n$, g'_{x_a} and g'_{x_b} exist and $\operatorname{im}(g'_{x_a}, g'_{x_b}) = M$;
- $M\oplus N(a) = \mathbf{R}^n$.

Under these assumptions we know that if x^* is an isolated solution of (1), (2), then (1), (2) is a well-posed equation (see [1]). Let x^* be an isolated solution of (1), (2) and

$$A(t):= f'_y\big(x^{*\prime}(t), x^*(t), t\big), \qquad B(t):= f'_x\big(x^{*\prime}(t), x^*(t), t\big).$$

Consider the matrix-DAE

(3) $$A(t)X'(t, s)+B(t)X(t, s) = 0$$

with the initial condition

(4) $$P(s)\big(X(s, s)-I\big) = 0;$$

then $X(t, s)$ is called the *fundamental matrix* of (1) and

$$X(t, s) := P_s(t)\, Y(t, s)\, P(s),$$

$$Y'(t, s) = \big(P'(t) P_s(t) - P(t) G^{-1}(t) B(t)\big) Y(t, s),$$

$$Y(s, s) = I,$$

P_s denotes the canonical projector

$$P_s := I - QG^{-1} B$$

(see [4]).

In addition to (1), (2) we consider the initial value problem

$$f\big(x'(t), x(t), t\big) = 0, \tag{5}$$

$$P(s)\big(x(s) - z\big) = 0, \tag{6}$$

and we denote the solution of (5), (6) by $x(t; s, z)$. The existence of the solution of (5), (6) is saved by our assumptions in a neighbourhood of x^* (see [1]). The determination of consistent initial values $\big(y(s), x(s)\big)$ is possible by solving the nonlinear system

$$\big(x(s) - z\big) + Q(s)\, y(s) = 0, \quad f\big(y(s), x(s), s\big) = 0. \tag{7}$$

This shows

LEMMA 1. *A, B and D are $(n \times n)$-matrices. Q is a projection and projects onto* $\ker(A)$. *The matrix $G := A + BQ$ is regular and with $P := I - Q$, if $D = I$ or $D = P$, then $\left(\begin{smallmatrix} Q & D \\ A & B \end{smallmatrix}\right)$ is regular and*

$$\begin{pmatrix} Q & D \\ A & B \end{pmatrix} = \begin{pmatrix} 0 & I \\ -I & B \end{pmatrix} \begin{pmatrix} 0 & G \\ I & 0 \end{pmatrix} \begin{pmatrix} 0 & I \\ I & I \end{pmatrix} \begin{pmatrix} P & -Q \\ -Q & P \end{pmatrix} \begin{pmatrix} -I & -D \\ 0 & I \end{pmatrix},$$

$$\begin{pmatrix} Q & D \\ A & B \end{pmatrix}^{-1} = \begin{pmatrix} I - D - (P - DQ) G^{-1} B & (P - DQ) G^{-1} \\ I - QG^{-1} B & QG^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} -I & -D \\ 0 & I \end{pmatrix} \begin{pmatrix} P & -Q \\ -Q & P \end{pmatrix} \begin{pmatrix} -I & I \\ I & 0 \end{pmatrix} \begin{pmatrix} 0 & I \\ G^{-1} & 0 \end{pmatrix} \begin{pmatrix} B & -I \\ I & 0 \end{pmatrix}.$$

Proof. Consider the homogeneous system

(8a)
(8b)
$$\begin{pmatrix} Q & D \\ A & B \end{pmatrix} \begin{pmatrix} y \\ x \end{pmatrix} = 0.$$

(8a) $\to$ $Q(y + Dx) + PDx = 0 \;\to\; Px = 0, \quad Q(y + Dx) = 0;$
(8b) $\to$ $Ay + Bx = (A + BQ)(Py + Qy) + BPx \doteq 0.$

With $Px = 0$ and the nonsingularity of G follows

$$Py+Qy = 0 \quad \to \quad Py = 0, \qquad Qx = 0 \quad \to \quad x = 0 \quad \to \quad Qy = 0 \quad \to \quad y = 0.$$

The product structure and the inverse of the matrix is shown by matrix multiplication. ∎

Remark. The attribute of D is fulfilled for example if

a) $D := P$,

b) $D := I$.

1. Shooting method

The general idea of parallel shooting is well known. We subdivide the interval $[a, b]$ in m subintervals

$$a = t_0 < t_1 < t_2 < \ldots < t_m = b$$

and we try to calculate the values (initial values of integration)

$$z_i^* := x^*(t_i), \qquad i = 0, \ldots, m-1.$$

We get the following system of equations

$$\begin{aligned} &P(t_i)\big(x(t_i;\, t_{i-1}, z_{i-1}) - z_i\big) = 0, \qquad i = 1, \ldots, m-1, \\ &g\big(z_0, x(t_m;\, t_{m-1}, z_{m-1})\big) = 0. \end{aligned} \tag{9}$$

(9) forms a singular system for determination of $z_0, \ldots, z_{m-1}$.

The Jacobian of (9) has a cyclic-$(n \times n)$-block structure like in the case of ODE.

(10)

$$J_1 = \begin{bmatrix} P(t_1)X(t_1, t_0) & -P(t_1) & & & \\ & P(t_2)X(t_2, t_1) & -P(t_2) & & \\ & & & P(t_{m-1})X(t_{m-1}, t_{m-2}) & -P(t_{m-1}) \\ g'_{x_a} & & & & g'_{x_b}X(t_m, t_{m-1}) \end{bmatrix}.$$

Now it is possible to enlarge the system (9) so that the system becomes nonsingular

$$\begin{aligned} P(t_i)\big(x(t_i;\, t_{i-1}, z_{i-1}) - z_i\big) + Q(t_i)\, y_i &= 0, \\ f(y_i, x_i, t_i) &= 0, \end{aligned} \qquad i = 1, \ldots, m-1,$$

$$\begin{aligned} g\big(z, x(t_m;\, t_{m-1}, z_{m-1})\big) + Q(t_0)\, y_0 &= 0, \\ f(y_0, z_0, t_0) &= 0. \end{aligned} \tag{11}$$

THEOREM 1. *For transferable equations* (1) (11) *forms a regular system for determining* (y_i, z_i), $i = 0, \ldots, m-1$.

Proof. The Jacobian J_2 of (11) has a cyclic $(2n \times 2n)$-block structure.

(12)

$$J_2 = \begin{bmatrix} 0 & P(t_1)X(t, t_0)\,Q(t_1) & -P(t_1) & & & & \\ 0 & 0 & f'_{y_1} & f'_{z_1} & & & \\ & & & 0 & P(t_{m-1})X(t_{m-1}, t_{m-2})\,Q(t_{m-1}) & -P(t_{m-1}) & \\ & & & 0 & 0 & f'_{y_{m-1}} & f'_{z_{m-1}} \\ Q(t_0) & g'_{x_a} & & & & 0 & g'_{x_b}X(t_m, t_{m-1}) \\ f'_{y_0} & f'_{z_0} & & & & 0 & 0 \end{bmatrix}.$$

For every block $\begin{pmatrix} Q & D \\ f'_y & f'_z \end{pmatrix}$ the assumptions of Lemma 1 are fulfilled with $D = -P$. This means that J_2 is a nonsingular matrix. ∎

Lemma 1 gives the possibility to reduce (11) without loss of regularity.

$$(13)\quad \begin{aligned} \big(x(t_i;\, t_{i-1}, z_{i-1}) - z_i\big) + Q(t_i)\, y_i &= 0, \quad i = 1, \ldots, m-1, \\ f(y_i, x_i, t_i) &= 0, \\ g\big(z_0, x(t_m;\, t_{m-1}, z_{m-1})\big) + Q(t_0)\, y_0 &= 0, \\ f(y_0, z_0, t_0) &= 0. \end{aligned}$$

The Jacobian J_3 of (13) has the structure

(14)

$$J_3 = \begin{bmatrix} 0 & X(t_1, t_0)\,Q(t_1) & -I & & & & \\ 0 & 0 & f'_{y_1} & f'_{z_1} & & & \\ & & & 0 & X(t_{m-1}, t_{m-2})\,Q(t_{m-1}) & -I & \\ & & & 0 & 0 & f'_{y_{m-1}} & f'_{z_{m-1}} \\ Q(t_0) & g'_{x_a} & & & & 0 & g'_{x_b}X(t_m, t_{m-1}) \\ f'_{y_0} & f'_{z_0} & & & & 0 & 0 \end{bmatrix}$$

is nonsingular because of Theorem 1 and Lemma 1 with $D = -I$.

With respect to (7) the idea of the systems (11) and (13) means to save that the determination of consistent intial values in every subinterval $[t_{i-1}, t_i]$ works with equations with nonsingular Jacobian. Will it be sufficient to add such equations only in one shooting point?

$$(15)\quad \begin{aligned} x(t_i;\, t_{i-1}, z_{i-1}) - z_i &= 0, \quad i = 1, \ldots, m-1,\ i \neq i_0, \\ x(t_{i_0};\, t_{i_0-1}, z_{i_0-1}) - z_{i_0} + Q(t_{i_0})\, y_{i_0} &= 0, \\ f(y_{i_0}, x_{i_0}, t_{i_0}) &= 0, \quad i_0 \neq 0, \\ g\big(z_0, x(t_m;\, t_{m-1}, z_{m-1})\big) &= 0, \\ g\big(z_0, x(t_m;\, t_{m-1}, z_{m-1})\big) + Q(t_0)\, y_0 &= 0, \\ f(y_0, z_0, t_0) &= 0, \quad i_0 = 0. \end{aligned}$$

The Jacobian of (15) has the structure

(16) $J_4 =$

$$\begin{bmatrix} X(t_1, t_0) & -I & & & & & & & \\ & X(t_2, t_0) & -I & & & & & & \\ & & X(t_{i_0}, t_{i_0-1}) & Q(t_i) & I & & & & \\ & & & f'_{y_{i_0}} & f'_{z_{i_0}} & & & & \\ & & & & X(t_{i_0+1}, t_{i_0}) & -I & & & \\ & & & & & & X(t_{m-1}, t_{m-2}) & -I & \\ g'_{x_a} & & & & & & & g'_{x_b} X(t_m, t_{m-1}) \end{bmatrix}$$

THEOREM 2. *The assumptions* (A) *are fulfilled. Then* (15) *represents a system with a nonsingular Jacobian.*

Proof. Consider the equation

(17) $$J_4 u = 0$$

with

$$u = (z_0, z_1, \ldots, z_{i_0-1}, y_{i_0}, z_{i_0}, \ldots, z_{m-1})^T.$$

From (17) we have

(a) $$X(t_i, t_{i-1}) z_{i-1} = z_i, \quad i = 0, \ldots, i_0 - 1,$$

(b) $$Q(t_{i_0}) y_{i_0} - z_{i_0} = -X(t_{i_0}, t_{i_0-1}) z_{i_0-1},$$
$$f'_{y_{i_0}} y_{i_0} + f'_{z_{i_0}} z_{i_0} = 0,$$

(c) $$X(t_i, t_{i-1}) z_{i-1} = z_i, \quad i = i_0, \ldots, m-1,$$

(d) $$g'_{x_a} z_0 + g'_{x_b} X(t_m, t_{m-1}) z_{m-1} = 0.$$

(a), (c) and the qualities of X give

(a′) $$z_{i_0-1} = X(t_{i_0-1}, t_0) z_0,$$

(c′) $$z_{m-1} = X(t_{m-1}, t_{i_0}) z_{i_0}.$$

(a′), (b′) and the representation of the inverse by Lemma 1 with $D = -I$ gives

(b′) $$\begin{bmatrix} y_{i_0} \\ z_{i_0} \end{bmatrix} = \begin{bmatrix} G_{i_0}^{-1} f'_{z_{i_0}} & G_{i_0}^{-1} \\ -P_s(t_{i_0}) & Q(t_{i_0}) G_{i_0}^{-1} \end{bmatrix} \begin{bmatrix} -X(t_{i_0}, t_0) z_0 \\ 0 \end{bmatrix}.$$

(b′), (c′) and (d) give

(d′) $$(g'_{x_a} + g'_{x_b} X(t_m, t_0)) z_0 = 0.$$

For $i_0 = 0$ we get the equations

$$(\mathrm{d}_0) \qquad \begin{aligned} Q(t_0)\, y_0 + (g'_{x_a} + g'_{x_b} X(t_m, t_0)) z_0 &= 0, \\ f'_{y_0}\, y_0 + f'_{z_0} z_0 &= 0. \end{aligned}$$

From (d') ($i_0 \neq 0$) we are able only to conclude that $z_0 \in N(a)$. From (d_0) ($i_0 = 0$) follows because of the qualities of g, the transferability of f, and Lemma 1 that $(y_0, z_0) \equiv 0$. ∎

Now we once more consider system (9) with the Jacobian J_1 (10) and point out that we are only interested in the projections $P(t_i) z_i$ of the initial values z_i ($i = 0, \ldots, m-1$). This allows us to consider equation (9) in the sense that we don't compute the values z_i but only the projections $P(t_i) z_i$.

LEMMA 2. *$\mathfrak{A}$, $\mathfrak{B}$ are $k \times k$ matrices and the matrix pencil ($\mathfrak{A}$, $\mathfrak{B}$) is regular with index 1. Let $\mathfrak{Q}$ be a projector onto* $\ker(\mathfrak{A})$ *and $\mathfrak{P} := I - \mathfrak{Q}$, then exists a nonsingular matrix $\mathfrak{G}$ such that $\mathfrak{A} = \mathfrak{G}\mathfrak{B}$ and $\mathfrak{A}^- := \mathfrak{P}\mathfrak{G}^{-1}$ is a generalised inverse of $\mathfrak{A}$.*

Proof. If we select $\mathfrak{G} := \mathfrak{A} + \mathfrak{B}\mathfrak{Q}$ we know from the assumptions that $\mathfrak{G}$ is nonsingular and $\mathfrak{A} = \mathfrak{G}\mathfrak{P}$.

If $\mathfrak{A}^-$ is a generalised inverse we have to show

(a) $\mathfrak{A}^- \mathfrak{A}\mathfrak{A}^- = \mathfrak{A}^-$,

(b) $\mathfrak{A}\mathfrak{A}^- \mathfrak{A} = \mathfrak{A}$,

(c) $\mathfrak{A}^- \mathfrak{A}$ represents a projector.

With $\mathfrak{A}\mathfrak{P} = \mathfrak{A}$ and $\mathfrak{G}^{-1}\mathfrak{A} = \mathfrak{P}$ we get

(a) $\mathfrak{A}^- \mathfrak{A}\mathfrak{A}^- = \mathfrak{P}\mathfrak{G}^{-1}\mathfrak{A}\mathfrak{P}\mathfrak{G}^{-1} = \mathfrak{P}\mathfrak{G}^{-1} = \mathfrak{A}^-$,

(b) $\mathfrak{A}\mathfrak{A}^- \mathfrak{A} = \mathfrak{A}\mathfrak{P}\mathfrak{G}^{-1}\mathfrak{A} = \mathfrak{A}$,

(c) $\mathfrak{A}^- \mathfrak{A} = \mathfrak{P}\mathfrak{G}^{-1}\mathfrak{A} = \mathfrak{P}$. ∎

LEMMA 3. *Consider a linear system of equations*

$$\mathfrak{A}x = b, \tag{18}$$

$b \in \mathrm{im}(\mathfrak{A})$ and the assumptions of Lemma 2 are fulfilled. Then

$$\mathfrak{P}x = \mathfrak{G}^{-1} b.$$

Proof. A solution x of (18) exists because $b \in \mathrm{im}(\mathfrak{A})$. We use the representation of $\mathfrak{A} = \mathfrak{G}\mathfrak{P}$ and multiply (18) with $\mathfrak{G}^{-1}$. ∎

We use Lemma 2 and 3 with

$$\mathfrak{P} := \mathrm{diag}(P(t_0), \ldots, P(t_{m-1})),$$

$$\mathfrak{A} := J_1,$$

$$\mathfrak{B} = \begin{pmatrix} 0 & -I & & & \\ & 0 & -I & & \\ & & & \ddots & \\ & & & 0 & -I \\ -I & & & & 0 \end{pmatrix},$$

$$\mathfrak{Q} = I - \mathfrak{P}.$$

This leads to the following nonsingular "Jacobian" for determining the shooting values by (9) by Newton's method

$$J_5 = \mathfrak{A} + \mathfrak{B}\mathfrak{Q}$$

$$= \begin{bmatrix} P(t_1)X(t_1, t_0) & -I & & & \\ & P(t_2)X(t_2, t_1) & -I & & \\ & & P(t_{m-1})X(t_{m-1}, t_{m-2}) & -I \\ g'_{x_a} - Q(t_0) & & & g'_{x_b}X(t_m, t_{m-1}) \end{bmatrix}.$$

2. Illustrative example

Consider the DAE

$$\begin{bmatrix} 1 & t \\ 1 & t \end{bmatrix}\begin{bmatrix} x'_1 \\ x'_2 \end{bmatrix} - \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} (t+1)^2 \\ (t+1)^2 - 1 \end{bmatrix}, \quad t \in [1, 2],$$

with the boundary condition

$$x_2(b) = 10.$$

For the integration we use a modified ADAM's method (see [3]) and for solving the nonlinear algebraic equations the procedure NLZYK (see [2]).

We test the methods (11), (13) and (9) with "Jacobian" J_5.

For three shooting points ($m = 3$) and an accuracy of integrating eps $= 1D-4$ and the initial values

$$x(1.0) = (6.000000000, 7.500000000)^T,$$
$$x(1.\bar{3}) = (8.1666666667, 9.6666666667)^T,$$
$$x(1.\bar{6}) = (10.666666667, 12.166666667)^T$$

we get the following results

	(11)	(13)	(9) with J_5	exact solution
$x(1.0)$	3.9946905	3.9946943	3.9947936	4.0
	4.9946905	4.9946943	4.9947936	5.0
$x(1.\bar{3})$	5.4401434	5.4401437	5.4401536	$5.\bar{4}$
	6.4401434	6.4401437	6.4406127	$6.\bar{4}$
$x(1.\bar{6})$	7.1086354	7.1086410	7.1089919	$7.\bar{1}$
	8.1086354	8.1086410	8.1091530	$8.\bar{1}$
$x(2.0)$	9.0000209	9.0000272	9.0003480	9.0
	10.000021	10.000027	10.000348	10.0

	f-calls	g-calls	estimated condition of J_i	length of working area	number of Newton iterations	defect of nl-equation
(11)	7315	7	13.12	463	5	4.298–05
(13)	7289	7	14.83	459	5	4.241–05
(9)$+J_s$	7260	7	11.47	246	5	4.383–04

This reveals that the quality of the results is the same, but with method $(9)+J_s$ we need about half of the storage of the other algorithms. The Jacobian J_5 itself only has a quarter of the elements of the matrices J_2 or J_3.

We need the same storage as in the ODE case.

References

[1] E. Griepentrog, R. März, *Differential-Algebraic Equations and Their Numerical Treatment*, Teubner Texte Band 88, Leipzig 1986.

[2] Hanke, R. Lamour, Winkler, *The program system "RWA" for the solution of two-point boundary-value problems*, Humboldt University of Berlin, Section of Mathematics, Seminarbericht Nr. 67, Berlin 1985.

[3] R. Lamour, *Shooting methods for transferable DAE's – Integration methods*, in preparation.

[4] Lentini, R. März, *The condition of boundary value problems in transferable differential-algebraic equations*, Humboldt University of Berlin, Section of Mathematics, Preprint 136, Berlin 1987.

Presented to the Semester
Numerical Analysis and Mathematical Modelling
February 25 — May 29, 1987