# A Shooting Method for Fully Implicit Index–2 Differential–Algebraic Equations

René Lamour

**Abstract**

A shooting method for two–point–boundary value problems for fully implicit index–1 and –2 differential–algebraic equations is presented. A combination of the shooting equations with a method of the calculation of consistent initial values leads to a system of nonlinear algebraic equations with nonsingular Jacobian. Examples are given.
AMS(MOS) Classification: 65L10
Keywords: DAE, BVP, Index 2, consistent initial value

## 1 Introduction

In this paper we consider the fully implicit index–2 system

$$f(x'(t), x(t), t) = 0, \qquad t \in [a, b], \tag{1.1}$$

with the boundary condition

$$g(x(a), x(b)) = 0. \tag{1.2}$$

Such problems arise as models of electrical networks, chemical reactions or index–reduced systems of mechanical motions. The possibility of the direct solution of given index–2 problems is very useful because

- the index reduction changes the stability behaviour of the DAE
- it is easier to reduce an index–3 system by only one step than by two steps.

The realization of the shooting method is strongly connected with an integration method that integrates index–2 problems well. This is given if

$$\ker(f'_y) = N = \text{const},$$

e.g. for the BDF–method. The presented shooting method links a procedure for the calculation of consistent initial values (this procedure alone is very useful) with the shooting equations, and the Jacobian of the whole method becomes nonsingular.
In Chapter 2 we introduce some projectors which are useful for the description of index–2 DAE's. We define a Green function for the explicit representation of of the solution of a linear index–2 DAE (Chapter 3). The numerical solution of (1.1),(1.2) by a shooting method is presented in Chapter 4 and some remarks to the numerical realization you can find in Chapter 5. Numerical examples complete the paper (Chapter 6).

## 2 Index determination and projectors

We investigate the nonlinear DAE

$$f(x'(t), x(t), t) = 0 \tag{2.1}$$

as an IVP or BVP. For the numerical approximation of (2.1) it is necessary to know which index the DAE has.
Let $x_\star$ be a solution of the considered problem (2.1) and

$$A(t) := f'_y(x'_\star(t), x_\star(t), t) \qquad B(t) := f'_x(x'_\star(t), x_\star(t), t). \tag{2.2}$$

We define the chain of matrix functions [Mä87]

$$\begin{array}{rcl} A_0 & := & A, \qquad\qquad B_0 := B - AP' \\ A_{i+1} & := & A_i + B_i Q_i, \\ B_{i+1} & := & (B_i - A_{i+1}(P_0 P_1 \cdots P_{i+1})' P_0 \cdots P_{i-1}) P_i. \end{array} \tag{2.3}$$

$Q_i$ is defined to be a projector onto $N_i := \ker(A_i(t)), P_i := I - Q_i$ for $i \geq 0$ and $P_0 =: P$. Then the following definition is given.

Definition. *[Mä91] The ordered pair $\{A, B\}$ of continuous matrix functions is said to be* index–$\mu$–tractable *if all matrices $A_j(t), j = 0, \ldots, \mu - 1$, within the chain (2.3) are singular with smooth nullspaces, and $A_\mu(t)$ remains nonsingular.*
*The nonlinear DAE (2.1) is said to be* index–1–tractable locally around $x_\star$ *if the pair of the linearization (2.2) is so, too.*
*The nonlinear DAE (2.1) is said to be* index–$\mu$–tractable locally around $x_\star$ *for $\mu > 1$ if the pair of the linearization (2.2) is so in a neighbourhood of the solution.*

We are interested in the index–2–tractable case under the assumption that

$$\ker(f'_y) = N = const, \text{ i.e. } P' = 0.$$

The following situation is given

$$A_1 = A_0 + B_0 Q\,,\ B_1 = (B_0 - A_1(PP_1)')P, \tag{2.4}$$

$$A_2 = A_1 + B_1 Q_1 = (A + BQ + BPQ_1)(I - P_1(PP_1)'PQ_1). \tag{2.5}$$

**Lemma 2.1** *Denote by $\mathcal{Q}$ an arbitrary projector, and $\mathcal{P} := I - \mathcal{Q}$. Then the matrix*

$$\mathcal{M} := I - \mathcal{Q}Z\mathcal{P}$$

*is nonsingular and its inverse is given by*

$$\mathcal{M}^{-1} = I + \mathcal{Q}Z\mathcal{P}.$$

Proof. We consider the equation

$$\mathcal{M}z = 0.$$

It follows

$$z = \mathcal{Q}Z\mathcal{P}z \Rightarrow \mathcal{P}z = 0,$$

and we have $z = 0$. $\square$

Using Lemma 2.1 it is clear that

$$A_2 \text{ is nonsingular} \Leftrightarrow \tilde{A}_2 := A + BQ + BPQ_1 \tag{2.6}$$

is nonsingular. However for arbitrary projectors $Q$ and $Q_1$ we can test the nonsingularity of $\tilde{A}_2$ pointwise without any knowledge of the derivative of a projector as in $A_2$.

**Lemma 2.2** *The following relations are valid*

$$\begin{array}{lrcl}
 & \textit{if } P' & = & 0: \\
a.) & A_2^{-1}A & = & P_1 P \\
 & \textit{if } P' & = & 0 \textit{ and } Q_1 Q = 0: \\
b.) & A_2^{-1}A & = & P - Q_1 \\
c.) & A_2^{-1}BQ & = & Q \\
d.) & A_2^{-1}BPQ_1 & = & Q_1 - P_1(PQ_1)'PQ_1
\end{array} \tag{2.7}$$

PROOF. For $P' = 0$, $(PQ_1)' = -(PP_1)'$ is fulfilled. We obtain the relations a.) and b.) multiplying (2.5) from the right by $P_1 P$, and c.) by $Q$. For d.) we multiply (2.5) by $A_2^{-1}$ and use the relation b.) and c.). □

Sometimes it is very useful to choose a special structure of the projectors. We focus our interest on the so–called canonical projector $Q_1$ with

$$Q_1 = Q_1 A_2^{-1} BP \tag{2.8}$$

(cf. [Mä91b]). This projector fulfils the condition $Q_1 Q = 0$ and it is really calculable, because

$$Q_1 = Q_1 A_2^{-1} BP = Q_1 \tilde{A}_2^{-1} BP. \tag{2.9}$$

For our further investigations we assume that

$$P' = 0 \tag{2.10}$$

and

$$Q_1 Q = 0, \tag{2.11}$$

where $Q_1$ represents the canonical projector given in (2.9).

# 3 Representation of the solution for a linear index–2 BVP

In this chapter we present a solution of the linear system

$$\begin{aligned}
A(t)x'(t) + B(t)x(t) &= q(t) && (3.1)\\
D_a x(a) + D_b x(b) &= \gamma. && (3.2)
\end{aligned}$$

First we consider the IVP (3.1) with the initial condition

$$P(s)P_1(s)(x(s)-\alpha)=0. \tag{3.3}$$

For the projector $P_1 := I - Q_1$ we prefer now the canonical projector $Q_1$ given in (2.9). (3.1),(3.3) is uniquely solvable for all $q$ with $\{q \in C : Q_1 A_2^{-1} q \in C^1\}$ (cf. [Mä89]).
For better understanding the index–2 case we split the solution $x$ into three parts:

$$u := PP_1x,\ v := PQ_1x \ \text{ and } w := Qx.$$

(With $Q_1Q = 0$ also $PP_1$ and $PQ_1$ are projectors.)
Multiplying (3.1) by $PP_1A_2^{-1}, QP_1A_2^{-1}$ and $PQ_1A_2^{-1}$ we obtain

$$\begin{array}{rcl} u' - (PP_1)'u + PP_1A_2^{-1}Bu & = & PP_1A_2^{-1}q \\ -QQ_1v' - QQ_1(PQ_1)'u + QP_1A_2^{-1}Bu + w & = & QP_1A_2^{-1}q \\ v & = & PQ_1A_2^{-1}q. \end{array} \tag{3.4}$$

The *Fundamental Matrix* of a DAE is given by the solution of the homogeneous IVP

$$AX' + BX = 0 \tag{3.5}$$

$$P(s)P_1(s)(X(s,s)-I) = 0. \tag{3.6}$$

Using (3.4) we have $V := PQ_1X \equiv 0$, $U := PP_1X = PP_1Y$, where $Y$ solves

$$Y' = ((PP_1)' - PP_1A_2^{-1}B)Y, \qquad Y(s,s) = I,$$

and $W := QX = QQ_1(PQ_1)'U - QP_1A_2^{-1}BU$. With (3.6) and $QW = W$, $PP_1U = U$ we create

$$X(t,s) = M(t)Y(t,s)P(s)P_1(s) \tag{3.7}$$

with $M(t) := (I + Q(t)[QQ_1(PQ_1)'(t) - P_1(t)A_2^{-1}B(t)]P(t)P_1(t))$. Using Lemma 2.1 it is easy to verify that

$$PP_1M = PP_1M^{-1} = PP_1.$$

Recall that

$$P(t)P_1(t)Y(t,s)P(s)P_1(s) = Y(t,s)P(s)P_1(s) \tag{3.8}$$

(cf. [Mä89]).

**Remark 1** *Using the special projector $Q = QP_1A_2^{-1}B$ the splitted system (3.4) and also the matrix $M$ of the fundamental matrix $X$ look a little bit easier. However, this projector is very difficult to calculate, because of the derivatives in $A_2$.*

Now we are looking for a representation of the solution of the IVP (3.1),(3.3). Using the fundamental matrix $Y(t,s)$ we have for the component u

$$u(t) = Y(t,s)(PP_1\alpha + \int_s^t Y(s,\tau)PP_1A_2^{-1}q\,d\tau). \tag{3.9}$$

With (3.8) we transform (3.9) into

$$u(t) = Y(t,s)P(s)P_1(s)(\alpha + \int_s^t M(s)Y(s,\tau)PP_1(\tau)h(\tau)d\tau)$$

with $h(t) := PP_1A_2^{-1}q$. For the other components we have

$$\begin{aligned} v &= PQ_1A_2^{-1}q \text{ and} \\ w &= QP_1A_2^{-1}q + QQ_1(PQ_1)'u - QP_1A_2^{-1}Bu + QQ_1v', \end{aligned}$$

and therefore

$$\begin{aligned} x &= u + v + w \\ &= Mu + \bar{q}(t) \end{aligned}$$

with $\bar{q}(t) := (PQ_1 + QP_1)A_2^{-1}q + QQ_1(PQ_1A_2^{-1}q)'$.

$$x(t) = X(t,s)(\alpha + \int_s^t X(s,\tau)h(\tau)d\tau) + \bar{q}(t) \tag{3.10}$$

represents the solution of the IVP (3.1),(3.3). Now we consider the solution $x(t)$ in $t = a$ and $t = b$.

$$\begin{aligned} x(a) &= X(a,a)\alpha + \bar{q}(a) \text{ and} \\ x(b) &= X(b,a)(\alpha + \textstyle\int_a^b X(a,\tau)h(\tau)d\tau) + \bar{q}(b) \end{aligned}$$

with unknown $\alpha$. The boundary condition (3.2) requires

$$\begin{aligned} S\alpha &:= (D_aX(a,a) + D_bX(b,a))\alpha \\ &= \bar{\gamma} - D_bX(b,a)\textstyle\int_a^b X(a,\tau)h(\tau)d\tau =: \tilde{\gamma} \end{aligned} \tag{3.11}$$

with $\bar{\gamma} := \gamma - (D_a\bar{q}(a) + D_b\bar{q}(b))$.

**Theorem 3.1** *[Mä91] Let (3.1), (3.2) be a tractable index–2 equation and the projectors fulfil (2.10) and (2.11). Then, for arbitrary right–hand sides $q$ with $q \in C[a,b]$, $PQ_1A_2^{-1}q \in C^1[a,b]$ and $\gamma \in im(D_a, D_b)$ (3.1), (3.2) have a unique solution iff*

$$ker(S) = im(I - PP_1) \tag{3.12}$$

$$im(S) = im(D_a, D_b). \tag{3.13}$$

Proof. The unique solution of (3.1), (3.2) is related to the solution of the IVP (3.1), (3.3). Then it becomes clear that only $PP_1\alpha$ influences the solution. Hence, we require for $\alpha$

$$\alpha = PP_1\alpha. \tag{3.14}$$

We are looking for solutions of (3.11) in the set $\mathcal{P} := \{z | z \in \text{im}(PP_1)\}$. The right–hand side of (3.11) fulfils $\tilde{\gamma} \in \text{im}(D_a, D_b)$. This means that (3.13) is a necessary condition for the solvability of (3.11). The structure of $X(t,s)$ provides

$$S = SP(a)P_1(a) \text{ or } \ker(S) \supset \text{im}(I - P(a)P_1(a)). \tag{3.15}$$

$\rightarrow$ Let $\alpha \in \mathcal{P}$ be a solution of (3.11), then also $\alpha + \beta \in \mathcal{P}$ solves (3.11) with $\beta \in \ker(S)$. The uniqueness requires that $\mathcal{P} \cap \ker(S) = \{0\} \Rightarrow \ker(S) \subset \text{im}(I - PP_1)$. With (3.15) formulae (3.12) follows.

$\leftarrow$ Let (3.12) be valid, and $\alpha_1$ and $\alpha_2 \in \mathcal{P}$ denote two solutions of (3.11). Then $\alpha_1 - \alpha_2 \in \ker(S)$, but $\ker(S) \cap \mathcal{P} = \{0\}$ and $\alpha_1 = \alpha_2$. $\square$

$S^-$ denotes the generalized reflexive inverse of $S$ with

$$S^- S S^- = S^-, \qquad S S^- S = S$$

and

$$S^- S = P(a) P_1(a). \tag{3.16}$$

This representation of $S^-$ is possible if (3.12) is valid. We multiply (3.11) by $S^-$:

$$P(a) P_1(a) \alpha = S^- \tilde{\gamma}.$$

With (3.10) we have

$$\begin{aligned} x(t) &= X(t,a) S^- \bar{\gamma} \\ &\quad + \textstyle\int_a^t X(t,\tau) h(\tau) d\tau - \int_a^b X(t,a) S^- D_b X(b,\tau) h(\tau) d\tau + \bar{q}(t). \end{aligned}$$

Using (3.16)

$$S^- D_b X(b,a) = P P_1 - S^- D_a X(a,a) \tag{3.17}$$

is valid and, therefore,

$$X(t,\tau) - X(t,a)(P P_1 - S^- D_a X(a,a)) X(a,\tau) = X(t,a) S^- D_a X(a,\tau).$$

Now we introduce the Green's function

$$G(t,s) := \begin{cases} +X(t,a) S^- D_a X(a,s) & s > t \\ -X(t,a) S^- D_b X(b,s) & s < t \end{cases} \tag{3.18}$$

and the following Theorem holds.

**Theorem 3.2** *Let Theorem 3.1 be valid and $S^-$ denotes a reflexive inverse of $S$ with $S^- S = P P_1(a)$. The solution of the BVP (3.1),(3.2) has the representation*

$$x(t) = X(t,a) S^- \bar{\gamma} + \int_a^b G(t,\tau) h(\tau) d\tau + \bar{q}(t).$$

# 4 Numerical solution by shooting method

The solution of BVP's by shooting methods requires that we are able to integrate the considered equation. For DAE's this means that we have to make available consistent initial values. We find different ideas for the calculation of consistent initial values.

**Numerical differentiation** is used in the code DASSL (see also [LPG91]),

**Formel manipulation** is proposed by [Han90],

**Special structure** of the DAE is used by [AP91].

We use a general approach taking advantage of the given subspaces by using special projectors. A further disadvantage of shooting methods for DAE's is the singularity of the Jacobian. This problem we overcome as in the index 1 case (cf. [Lam91]) by the combination of the shooting equation with the equation for the calculation of consistent initial values.

## 4.1 Consistent initial values

We consider the nonlinear DAE

$$f(x'(t), x(t), t) = 0. \tag{4.1}$$

For a better understanding of the index–2 case let us consider the transferable or index–1 case. The assumption $\ker(f'_y) = N(t)$ allows us to transform (4.1) into

$$f((Px)'(t) - P'x(t), x(t), t) = 0 \tag{4.2}$$

(see [GM87]).
**Index 1:** We are looking for consistent initial values for the IVP (4.2) with

$$P(s)(x(s) - \alpha) = 0. \tag{4.3}$$

We split $x = Px + Qx =: u + v$ and denote $y := (Px)' - P'x$ (recall that $Py = y$). Let us define $\eta := y + v$, then (4.2) considered in $t = s$ is written as follows

$$f(P\eta, u + Q\eta, s) = 0 \tag{4.4}$$

with known $u = P\alpha$ and searched $\eta$. The Jacobian of (4.4) is given by

$$f'_y P + f'_x Q \quad (= A_1),$$

which is nonsingular for the index–1 case.

**Index 2:** Now we transfer this technique to the index–2 case. Here we assume the stronger condition that

$$\ker(f'_y) = N \equiv \text{const.} \text{ (i.e. } P = \text{const}, \ P' = 0) \tag{4.5}$$

This assumption does not restrict the class of numerically solvable problems. We know (compare with the example in [GP84]) that only assumption (4.5) saves numerical success for index–2 problems. With (4.5), (4.2) has the structure

$$f((Px)', x, t) = 0. \tag{4.6}$$

We represent (4.6) in a more detailed way

$$\begin{aligned} f((PP_1x)' + (PQ_1x)', x, t) &= 0 \text{ or} \\ f((PP_1x)' - (PP_1)'PP_1x + (PQ_1x)' - (PQ_1)'PP_1x, x, t) &= 0 \end{aligned} \tag{4.7}$$

with the initial condition

$$(PP_1)(s)(x(s) - \alpha) = 0. \tag{4.8}$$

We split x into the components

$$x = PP_1x + PQ_1x + Qx =: u + v + w$$

and we define

$$\begin{aligned} y &:= (PP_1x)' - (PP_1)'PP_1x, \text{ and} \\ \eta &:= y + v + w. \end{aligned} \tag{4.9}$$

Here $PP_1y = y$ is valid. Now (4.7) reads

$$f(PP_1\eta + v' - (PQ_1)'u,\ u + (PQ_1 + Q)\eta,\ s) = 0. \tag{4.10}$$

The trouble in formula (4.10) is caused by the unknown term $v'$. Therefore, we consider (4.10) in a neighbouring point $s + h$ and replace $v'$ by the finite difference

$$v' \sim \frac{v^h - v}{h}.$$

We use the symbol $(\cdot)^h = (\cdot)(s + h)$.

$$\begin{aligned} \bar f \ &:=\ f(PP_1\eta + \frac{(PQ_1\eta)^h - PQ_1\eta}{h} - \\ &\quad (PQ_1)'u,\ u + (PQ_1 + Q)\eta,\ s) = 0 \end{aligned} \tag{4.11}$$

$$\begin{aligned} \bar f^h \ &:=\ f((PP_1\eta)^h + \frac{(PQ_1\eta)^h - PQ_1\eta}{h} - \\ &\quad \{(PQ_1)'u\}^h,\ u^h + \{(PQ_1 + Q)\eta\}^h,\ s + h) = 0 \end{aligned} \tag{4.12}$$

with

$$u^h = u + hu' = u + h(PP_1\eta + (PP_1)'u).$$

**Theorem 4.1** *Let the projector $Q_1$ depend on $u = PP_1x$ and $t$ only, $P' = 0$ and $Q_1Q = 0$. Then the system (4.11),(4.12) has a nonsingular Jacobian in the point $(y_\star + v'_\star - (PQ_1)'u_\star, x_\star, s)$ if $h$ is sufficiently small, where $(.)_\star$ is the part of $x_\star$.*

PROOF. The Jacobian is given by

$$J = \begin{pmatrix} \dfrac{\partial \bar f}{\partial \eta} & \dfrac{\partial \bar f}{\partial \eta^h} \\[2ex] \dfrac{\partial \bar f^h}{\partial \eta} & \dfrac{\partial \bar f^h}{\partial \eta^h} \end{pmatrix}.$$

with

$$\begin{aligned} \frac{\partial \bar f}{\partial \eta} &= A(PP_1 - \tfrac{1}{h}PQ_1) + B(PQ_1 + Q) \\ \frac{\partial \bar f}{\partial \eta^h} &= \tfrac{1}{h}A\{PQ_1\}^h \\ \frac{\partial \bar f^h}{\partial \eta} &= -\tfrac{1}{h}A^hPQ_1 - h\{A(PQ_1)' - B\}^hPP_1 \\ \frac{\partial \bar f^h}{\partial \eta^h} &= \{A(PP_1 + \tfrac{1}{h}PQ_1) + B(PQ_1 + Q)\}^h \end{aligned} \tag{4.13}$$

with $A := f'_y$ and $B := f'_x$. Using the identities given in Lemma 2.2 and

$$\begin{aligned} (P - Q_1)PP_1 &= PP_1 \\ (P - Q_1)PQ_1 &= -QQ_1 \\ P_1P &= (P_1 - Q) \end{aligned} \tag{4.14}$$

for $P' = 0$, we have

$$J = \begin{pmatrix} A_2 & 0 \\ 0 & A_2^h \end{pmatrix} \mu \tag{4.15}$$

with

$$\mu = \begin{pmatrix} I + P_1((1+\frac{1}{h})QQ_1 - (PQ_1)')PQ_1 & -\frac{1}{h}(P-Q_1)\{PQ_1\}^h \\ \frac{1}{h}\{P-Q_1\}^h(PQ_1) - h\{A_2^{-1}(A(PQ_1)' - B)\}^h PP_1 & I + \{P_1((1+\frac{1}{h})QQ_1 - (PQ_1)')PQ_1\}^h \end{pmatrix}.$$

To show the nonsingularity of $\mu$, first we consider the equation

$$\bar{\mu} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = 0 \tag{4.16}$$

with

$$\bar{\mu} = \begin{pmatrix} I + P_1((1+\frac{1}{h})QQ_1 - (PQ_1)')PQ_1 & -\frac{1}{h}(P-Q_1)\{PQ_1\}^h \\ \frac{1}{h}\{P-Q_1\}^h(PQ_1) & I + \{P_1((1+\frac{1}{h})QQ_1 - (PQ_1)')PQ_1\}^h \end{pmatrix}.$$

The result of the first equation of (4.16) after multiplying by $Q_1$ is

$$Q_1\xi_1 = 0 \tag{4.17}$$

and the second equation multiplied by $Q_1^h$ yields

$$Q_1^h\xi_2 = 0. \tag{4.18}$$

Using (4.17) and (4.18) in (4.16) gives

$$\xi_1 = \xi_2 = 0.$$

Now we consider the matrix

$$\tilde{\mu} := \bar{\mu} - \begin{pmatrix} 0 & 0 \\ \bar{h}\{A_2^{-1}(A(PQ_1)' - B)\}^h PP_1 & 0 \end{pmatrix}.$$

$\tilde{\mu}$ depends continuously on $\bar{h}$ and $\tilde{\mu} = \mu$ for $\bar{h} = h$, i.e. if $h$ is sufficiently small, then $\mu$ is nonsingular. $\square$

**Remark 2** *We consider (4.11), (4.12) for a linear DAE*

$$A(t)x'(t) + B(t)x(t) = q(t)$$

*and obtain*

$$\begin{aligned} A(s)(PP_1\eta + \frac{\{PQ_1\eta\}^h - PQ_1\eta}{h} &- (PQ_1)'u) \\ +B(s)(u + (PQ_1 + Q)\eta) &= q(s) \end{aligned} \tag{4.19}$$

$$\begin{aligned} A(s+h)\{PP_1\eta\}^h + \frac{\{PQ_1\eta\}^h - PQ_1\eta}{h} &- \{(PQ_1)'u\}^h) \\ +B(s+h)(u^h + \{(PQ_1 + Q)\eta\}^h) &= q(s+h). \end{aligned} \tag{4.20}$$

*We multiply (4.19) by $A_2^{-1}(s)$ and (4.20) by $A_2^{-1}(s+h)$ and obtain*

$$\begin{aligned}
(P-Q_1)(PP_1\eta + \frac{(PQ_1\eta)^h - PQ_1\eta}{h} &- (PQ_1)'u) \\
+A_2^{-1}(s)B(s)(u+(PQ_1+Q)\eta) &= A_2^{-1}(s)q(s) \qquad (4.21)\\
\{(P-Q_1)(PP_1\eta)\}^h + \frac{(PQ_1\eta)^h - PQ_1\eta}{h} &- \{(PQ_1)'u\}^h) \\
+A_2^{-1}(s+h)B(s+h)\{u+(PQ_1+Q)\eta\}^h) &= \{A_2^{-1}q\}^h. \qquad (4.22)
\end{aligned}$$

*The multiplication of (4.21) and (4.22) by $PQ_1$ and $\{PQ_1\}^h$, respectively, yields*

$$\begin{aligned}
PQ_1\eta &= PQ_1A_2^{-1}q(s)\\
\{PQ_1\eta\}^h &= \{PQ_1A_2^{-1}\}^h q(s+h)
\end{aligned}$$

*or*

$$\lim_{h\to 0}\frac{\{PQ_1\eta\}^h - PQ_1\eta}{h} = v' = (PQ_1A_2^{-1}q)' \qquad (4.23)$$

*Using (4.23) we consider (4.19) for $h\to 0$ and realize exactly the linear DAE because*

$$PP_1\eta + (PQ_1\eta)' - (PQ_1)'u = (Px)'$$

*and*

$$u + (PQ_1+Q)\eta = x.$$

The accuracy of the numerical solution depends essentially on the condition of the matrix $\mu$. We investigate the condition of the matrix $\bar\mu$. Using Lemma 2.1 and $PQ_1(P-Q_1)=0$, the inverse of $\bar\mu$ is given by

$$\bar\mu^{-1} = \begin{pmatrix} I - P_1((1+\frac{1}{h})QQ_1 - (PQ_1)')PQ_1 & -\frac{1}{h}(P-Q_1)\{PQ_1\}^h \\ \frac{1}{h}\{P-Q_1\}^h(PQ_1) & I - \{P_1((1+\frac{1}{h})QQ_1 - (PQ_1)')PQ_1\}^h \end{pmatrix}. \qquad (4.24)$$

We introduce the constants

$$K_1 := \|QQ_1\|_{C[a,b]} \;\; K_2 := \|P_1(PQ_1)'PQ_1\|_{C[a,b]}.$$

Using the Taylor expansion $(PQ_1)^h = PQ_1 + h(PQ_1)' + O(h^2)$ we obtain

$$\begin{aligned}
\|\bar\mu\| &\le (1+K_1+K_2) + \frac{2}{h}K_1 + O(h)\\
\|\bar\mu^{-1}\| &\le (1+K_1+K_2) + \frac{2}{h}K_1 + O(h).
\end{aligned}$$

This proves the
**Corollary** *The condition of $\bar\mu$ is bounded by*

$$cond(\bar\mu) \le ((1+K_1+K_2) + \frac{2}{h}K_1 + O(h))^2.$$

**Remark 3** *The essential part of the estimation shows that $cond(\bar\mu) \sim O(h^{-2})$ in the worst case. This is not surprising because of the numerical differentiation.*

## 4.2 The shooting method

We consider now a boundary value problem

$$
\begin{aligned}
f(x'(t), x(t), t) &= 0, \quad t \in [a, b], && (4.25)\\
g(x(a), x(b)) &= 0. && (4.26)
\end{aligned}
$$

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

$$a = t_0 < t_1 < \cdots < t_m = b$$

and we look for the initial values $z_i := x(t_i), i = 0, \ldots, m-1$.
$x(t; s, z)$ denotes the solution of the IVP (4.25) with

$$PP_1(s)(x(s) - z) = 0.$$

The $z_i$ have to fulfil the boundary condition

$$g(z_0, x(t_m; t_{m-1}, z_{m-1})) = 0 \tag{4.27}$$

and the matching condition

$$(PP_1)_i(z_i - x(t_i; t_{i-1}, z_{i-1})) = 0. \tag{4.28}$$

(The symbol $(\cdot)_i$ reads like $(\cdot)(t_i)$).
The disadvantage of the system (4.27),(4.28) is the singularity of the Jacobian (as in the index–1 case). However we use the same idea that solves this problem in the index 1 case (cf. [Lam91]), too. We combine the shooting equation with the equations (4.11),(4.12) for the determination of the initial values. For this aim we split the variable $z_i$ into the parts

$$z_i = (PP_1)_i z_i + (PQ_1)_i z_i + Q_i z_i =: u_i + v_i + w_i$$

and with $\eta_i := y_i + v_i + w_i$ (cf.(4.9)) we have $z_i = u_i + (PQ_1 + Q)_i \eta_i$. The shooting equations are given by

$$
\begin{aligned}
g(u_0 + (PQ_1 + Q)_0 \eta_0,\ x(t_m; t_{m-1}, u_{m-1})) &= 0 && (4.29)\\
u_i - (PP_1)_i x(t_i; t_{i-1}, u_{i-1}) &= 0 \qquad {\scriptstyle i=1,\ldots,m-1} && (4.30)
\end{aligned}
$$

and the equations for the determination of the initial values in $t_i$ read

$$
\begin{aligned}
\bar{f}_i \ &:= \ f((PP_1)_i \eta_i + \frac{\{PQ_1\eta\}_i^h - (PQ_1)_i \eta_i}{h} - (PQ_1)_i' u_i,\\
&\qquad u_i + (PQ_1 + Q)_i \eta_i,\ t_i) = 0 \qquad\qquad (4.31)\\
\bar{f}_i^h \ &:= \ f(\{PP_1\eta\}_i^h + \frac{\{PQ_1\eta\}_i^h - (PQ_1)_i \eta_i}{h} - \{(PQ_1)_i' u_i\}^h,\\
&\qquad u_i^h + \{(PQ_1 + Q)_i \eta_i\}^h,\ t_i + h) = 0, \qquad i = 0, \ldots, m-1.
\end{aligned}
$$

For the variable $u$ we have to ensure that

$$PP_1 u = u.$$

This is valid for $t_i, i = 1, \ldots, m-1$ by using (4.30). For $u_0$ we extend (4.29) to

$$g(u_0 + (PQ_1 + Q)_0 \eta_0,\ x(t_m; t_{m-1}, u_{m-1})) + K^{-1}(I - PP_1)_0 u_0 = 0, \tag{4.32}$$

where $K$ is a nonsingular matrix with

$$\mathrm{im}(g'_{x_a}, g'_{x_b}) \bigoplus \mathrm{im}(K^{-1}(I - PP_1)_0) = R^n.$$

For the calculation of the unknowns $u_0, \cdots, u_{m-1}$ we have to solve the equations (4.30) and (4.32). But in (4.32) also $\eta_0$ is engaged. This means that we extend our system to the equations for the determination of the initial values (4.31) in the point $t_0$.

**Theorem 4.2** *Let the assumptions of Theorem 3.1 and 4.1 be fulfilled and $Q_1 = Q_1(t)$ only, in this case system (4.32),(4.30) and (4.31) (for $i = 0$) has a nonsingular Jacobian.*

Proof. We order the variables in the following way:

$$\xi = (u_0, \ldots, u_{m-1}, \eta_0, \eta_0^h),$$

then the Jacobian $J$ is given by

$$J = \left( \begin{array}{cccc|cc} G_{a,u} & & & G_{b,u} & G_{a,v} & 0 \\ M_0 & I & & & & \\ & \ddots & \ddots & & & \\ & & M_{m-2} & I & & \\ \hline F_{u_0} & & & & & J_0 \end{array} \right), \tag{4.33}$$

where we have used the following abbreviations

$$\begin{aligned} G_{a,u} &:= g'_{x_a} + K^{-1}(I - PP_1)_0, \qquad G_{b,u} := g'_{x_b} X(t_m, t_{m-1}) \\ G_{a,v} &:= g'_{x_a}(PQ_1 + Q)_0 \\ M_i &:= -(PP_1)_{i+1} X(t_{i+1}, t_i) \\ F_{u_0} &:= \begin{pmatrix} -A_0(PQ_1)'_0 + (BPP_1)_0 \\ \{-A_0(PQ_1)'_0 + (BPP_1)_0\}^h (I + O(h)) \end{pmatrix} \\ J_0 & \quad \text{denotes the matrix (4.13) in } t_0. \end{aligned}$$

We investigate the equation

$$\bar{J}\xi = 0. \tag{4.34}$$

$\bar{J}$ denotes the matrix with the structure of $J$, but the matrices $J_0$ are replaced by $\bar{J}_0 := \begin{pmatrix} A_2 & 0 \\ 0 & A_2^h \end{pmatrix} \bar{\mu}$ (cf. (4.15)). The second to m–th equation of (4.34) is given by

$$u_{i+1} = (PP_1)_{i+1} X(t_{i+1}, t_i) u_i \qquad i = 0, \ldots, m-2. \tag{4.35}$$

This leads to

$$u_{m-1} = (PP_1)_{m-1} X(t_{m-1}, t_0) u_0.$$

The latter 2n–dimensional equations are given by

$$\bar{J}_0 \begin{pmatrix} \eta_0 \\ \eta_0^h \end{pmatrix} = \begin{pmatrix} -A_0(PQ_1)'_0 + (BPP_1)_0 \\ \{-A_0(PQ_1)'_0 + (BPP_1)_0\}^h (I + O(h)) \end{pmatrix} u_0. \tag{4.36}$$

We multiply the first equation of (4.36) by $A_2^{-1}$ and the second one by $(A_2^h)^{-1}$. Using (4.24) and (2.8) we have

$$\eta_0 = ((P-Q_1)(PQ_1)' - A_2^{-1}BPP_1)u_0 \tag{4.37}$$

$$\eta_0^h = \{(P-Q_1)(PQ_1)' - A_2^{-1}BPP_1\}^h(I+O(h))u_0. \tag{4.38}$$

Setting (4.37) in the first equation of (4.34) yields

$$\begin{aligned}
& (g'_{x_a} + K^{-1}(I-PP_1)_0)u_0 + g'_{x_b}X(t_m,t_{m-1})u_{m-1} + g'_{x_a}(PQ_1+Q)_0\eta_0 \\
= \; & (g'_{x_a} + K^{-1}(I-PP_1)_0 + g'_{x_b}X(t_m,t_0) + g'_{x_a}(-Q[Q_1(PQ_1)' - A_2^{-1}B]PP_1)u_0 \\
= \; & (g'_{x_a}(I - Q[Q_1(PQ_1)' - A_2^{-1}B]PP_1) + g'_{x_b}X(t_m,t_0) + K^{-1}(I-PP_1)_0))u_0 \\
= \; & (g'_{x_a}X(t_0,t_0) + g'_{x_b}X(t_m,t_0) + K^{-1}(I-PP_1)_0)u_0.
\end{aligned}$$

Using $u_0 = PP_1u_0$ we obtain

$$(g'_{x_a}X(t_0,t_0) + g'_{x_b}X(t_m,t_0))u_0 = Su_0 = 0,$$

and with (3.12) it follows that $u_0 \in \ker(S)$ or $u_0 = (I-PP_1)z$ and this gives $u_0 = 0$. With (4.35) we have $u_i = 0$ and, using (4.36), $\eta_0$ and $\eta_0^h = 0$. The matrix $J$ is a regular perturbation of $\bar{J}$, i.e. for $h$ sufficiently small also $J$ is nonsingular. □

**Remark 4** *To determine the unknowns*

$$u_0, \ldots, u_{m-1}, \eta_0, \eta_0^h \text{ and } \eta_i, \eta_i^h, {\scriptstyle i=1,\ldots,m-1}$$

*we have to solve the systems* $\{$*(4.32), (4.30),(4.31) with* $i=0\}$ *and (4.31) for* $i = 1,\cdots,m-1$. *All these systems have a nonsingular Jacobian. Consequently, for the solution of the BVP (1.1), (1.2) the represented shooting method is realizable with a common Newton–like method (taking into consideration the structure of the Jacobian, of course).*

However a part of the unknowns of the nonlinear algebraic system are partially projected vectors ($u = PP_1u$). The question is:
Does the Newton method save this condition ?
Or, in other words :
Is the correction $\Delta u$ also a $PP_1$ projection?
To answer this question let us consider the nonlinear system

$$S_g := g(u_0 + (PQ_1+Q)_0\eta_0,\ x(t_m;t_{m-1},u_{m-1})) + K^{-1}(I-PP_1)_0u_0 \tag{4.39}$$

$$S_{i_u} := u_i - (PP_1)_i x(t_i;t_{i-1},u_{i-1}) \qquad {\scriptstyle i=1,\ldots,m-1} \tag{4.40}$$

$$\bar{S}_{0_\eta} := \begin{cases} f((PP_1)_0\eta_0 + & \frac{\{PQ_1\eta\}_0^h - (PQ_1)_0\eta_0}{h} \\ & -(PQ_1)'_0u_0,\ u_0 + (PQ_1+Q)_0\eta_0,\ t_0) \\ f(\{PP_1\eta\}_0^h + & \frac{\{PQ_1\eta\}_0^h - (PQ_1)_0\eta_0}{h} \\ & -\{(PQ_1)'_0u_0\}^h,\ u_0^h + \{(PQ_1+Q)_0\eta_0\}^h,\ t_0+h) \end{cases} \tag{4.41}$$

The Newton–correction $\Delta\xi := (\Delta u_0, ..., \Delta u_{m-1}, \Delta\eta_0, \Delta\eta_0^h) =: (\Delta u, \Delta\bar{\eta}_0^h)$ is given as the solution of

$$J\Delta\xi = -S(\xi). \tag{4.42}$$

Because of the structure of (4.42) (see Remark 4) we have to solve the linear systems

$$\left(\begin{array}{ccccc:cc} G_{a,u} & & & & G_{b,u} & G_{a,v} & 0 \\ M_0 & I & & & & & \\ & \ddots & \ddots & & & & \\ & & M_{m-2} & I & & & \\ \hdashline F_{u_0} & & & & & & J_0 \end{array}\right) \begin{pmatrix} \Delta u \\ \Delta \bar{\eta}_0 \end{pmatrix} = - \begin{pmatrix} S_g \\ S_{1_u} \\ \vdots \\ S_{(m-1)_u} \\ S_{0_\eta} \end{pmatrix} \tag{4.43}$$

We solve (4.43) by elimination of $\Delta\bar{\eta}_0$ in the last equation of (4.43), and using this result in the first equation of (4.43), i.e.

$$\Delta\bar{\eta}_0 = -J_0^{-1}(S_{0_\eta} + F_{u_0}\Delta u_0). \tag{4.44}$$

The first equation of (4.43) is given by

$$G_{a,u}\Delta u_0 + G_{b,u}\Delta u_{m-1} + G_{a,v}\Delta\eta_0 = -S_g \tag{4.45}$$

and it is clear that only $\Delta\eta_0$ influences (4.45). Using (4.37) and the abbreviation

$$\begin{pmatrix} \sigma_{1,0_\eta} \\ \sigma_{2,0_\eta} \end{pmatrix} := J_0^{-1} S_{0_\eta}, \tag{4.46}$$

we have

$$\Delta\eta_0 = -\sigma_{1,0_\eta} - A_2^{-1}F_{1,u_0}\Delta u_0$$

and (4.45) is now represented as

$$(G_{a,u} - G_{a,v}A_2^{-1}F_{1,u_0})\Delta u_0 + G_{b,u}\Delta u_{m-1} = -S_g + G_{a,v}\sigma_{1,0_\eta}. \tag{4.47}$$

The system (4.47) and the matching conditions

$$M_i\Delta u_i + \Delta u_{i+1} = -S_{(i+1)_u}, \qquad i = 0, ..., m-2, \tag{4.48}$$

of (4.43) form a linear system with *block cyclic structure*, as we know it from the shooting methods for ODE's. It is easy to verify that

$$G_{a,u} - G_{a,v}A_2^{-1}F_{1,u_0} = g'_{x_a}X(t_0, t_0) + K^{-1}(I - PP_1)_0. \tag{4.49}$$

From (4.48) and the conditions of the fundamental matrix we have

$$\Delta u_{m-1} = X(t_{m-2}, t_0)\Delta u_0 + (PP_1)_{m-2}z, \tag{4.50}$$

where $z$ represents an expression in $S_{i_u}$ and $M_i$. Using (4.49) and (4.50) we derive from (4.47)

$$\begin{aligned} ((g'_{x_a}X(t_0, t_0) + g'_{x_b}X(t_m, t_0)) + {} & K^{-1}(I - PP_1)_0)\Delta u_0 = \\ & -S_g + G_{a,v}\sigma_{1,0_\eta} - G_{b,u}(PP_1)_{m-2}z, \end{aligned} \tag{4.51}$$

or, shorter, with (3.11)

$$(S + K^{-1}(I - PP_1)_0)\Delta u_0 = d.$$

with

$$d = -S_g + G_{a,v}\sigma_{1,0_\eta} - G_{b,u}(PP_1)_{m-2}z, \tag{4.52}$$

and the solution is given by

$$\Delta u_0 = (S^- + (I - PP_1)_0 K)d \tag{4.53}$$

(see [Lam91], Lemma 2.2), where $S^-$ represents a reflexive inverse of $S$ with

$$S^- S = (PP_1)_0 \text{ and } SS^- = K^{-1}(PP_1)_0 K.$$

For (4.53) we have

$$\begin{aligned} \Delta u_0 &= (S^- SS^- + (I - PP_1)_0 K)d \\ &= ((PP_1)_0 S^- + (I - PP_1)_0 K)d. \end{aligned}$$

With (4.52) and the validity of Theorem 3.1 $d \in \mathrm{im}(S)$, i.e. $d = S\beta$. We have

$$\begin{aligned} (I - PP_1)_0 K d &= K(I - SS^-)d \\ &= K(I - SS^-)S\beta = 0. \end{aligned}$$

With $\Delta u_0$ also the other corrections $\Delta u_i, i = 1, ..., m-1$ are $(PP_1)_i$–projections.
**Corollary** *The Newton method for the solution of the nonsingular system (4.39)–(4.41) does not change the subspace condition of* $u \quad (= PP_1 u)$.

# 5 Numerical realization

The application of a shooting method means that at least the integration of the given problem over an (sub)interval is possible. In the case of DAE's this implies that consistent initial values in the shooting points are available. The rough algorithm for the solution of a TPBVP for DAE's is given as follows:

1. Subdivision of the interval [a,b]
   $a = t_0 < t_1 \cdots < t_{m-1} < t_m = b$;
2. initial guess of the initial values $z_i, i = 1, \cdots, m-1$, at the shooting points;
3. calculation of consistent initial values in the shooting points;
4. solution of the shooting equation;
5. **if** accuracy is high enough, **then**
   print some nice solution pictures
   **else**
   goto 3
   **endif**

To realize this algorithm we have to solve some standard problems of numerical mathematics, but with special structure. The main part is concentrated in step 4 – the solution of the shooting equation. For the nonlinear system we use the very flexible solver NLSOLV developed by the author, which solves systems with arbitrary structure if solvers for the linear

systems with this structure are available.
For the integration a BDF–code (IVPDAE) is used.
For the calculation of the nonlinear algebraic function and its Jacobian we have to calculate in the shooting points the projectors $Q$ and $Q_1$ with (2.8) and we have to determine $K$ in (4.32).

### 5.1 The calculation of the projectors $Q$ and the canonical projector $Q_1$

For a given matrix $A(:= f'_y)$ we have to calculate a projector $Q$ with

$$\mathrm{im}(Q) = \ker(A).$$

Let

$$A = URP_c^T$$

with $P_c$ – a permutation matrix of columns, U – an orthogonal matrix

$$r = \begin{pmatrix} R_1 & R_2 \\ 0 & 0 \end{pmatrix} \}r$$

$R_1$ – nonsingular and $r = \mathrm{rank}(A)$, then

$$Q := P_c \begin{pmatrix} 0 & -R_1^{-1}R_2 \\ 0 & I_{n-r} \end{pmatrix} P_c^T.$$

The UR–decomposition is performed by the Householder method with column pivoting. With the aid of this projector $Q$ we calculate

$$\tilde{A}_1 := A + BQ$$

(see (2.4) for $P' = 0$) and, using the same method, we calculate a projector $\tilde{Q}_1$ with

$$\mathrm{im}(\tilde{Q}_1) = \ker(A_1).$$

With $\tilde{A}_2$ from (2.6) we calculate the canonical projector as

$$Q_1 = \tilde{Q}_1\tilde{A}_2^{-1}BP$$

(see [Mä91]).

### 5.2 The calculation of $K$

In contrast to the ODE's, n additional conditions are not possible in DAE's. Only dim(im($S$)) (see Th.3.1) conditions are allowed and, moreover, we know that

$$\dim(\mathrm{im}(S)) = \dim(\mathrm{im}(PP)_0) =: r.$$

We organize our method in such a way that the first $r$ components of $g$ contain the $r$ conditions, i.e. we have to look for a matrix $K$ so that

$$K^{-1}(I - PP_1)_0$$

projects onto the last $r+1, ..., n$–th components only. We know that there exists a nonsingular matrix $T$ with

$$(I - PP_1)_0 = T \begin{pmatrix} 0 & \\ & I_{n-r} \end{pmatrix} T^{-1} \tag{5.1}$$

If we choose $K = T$, our problem is solved. Determination of $T^{-1}$:
With (5.1) we have

$$\begin{pmatrix} 0 & \\ & I_{n-r} \end{pmatrix} T^{-1} = T^{-1}(I - PP_1)_0 \tag{5.2}$$

and let $T^{-1} =: \begin{pmatrix} t_1 \\ \vdots \\ t_n \end{pmatrix}$ and $t_i^T \in R^n$. With (5.2)

$$\begin{pmatrix} 0 \\ \vdots \\ 0 \\ t_{r+1} \\ \vdots \\ t_n \end{pmatrix} = \begin{pmatrix} t_1 \\ \vdots \\ t_n \end{pmatrix} (I - PP_1)_0$$

is valid, but for us only $t_i,\ i = r+1, ..., n$ is of interest, and for these vectors

$$t_j = t_j(I - PP_1)_0$$

is true if $t_j = j$–th row of $(I - PP_1)_0$. With this selection of $K$ we have finally

$$K^{-1}(I - PP_1)_0 = \begin{pmatrix} 0 & \\ & I_{n-r} \end{pmatrix} (I - PP_1)_0.$$

# 6 Examples

Let us start with the classical pendulum, but in a slightly modified version. We use the representation

$$\begin{aligned} x_1' &= x_3 \\ x_2' &= x_4 \\ x_3' &= -x_1 x_5 \\ x_4' &= -x_2 x_5 + g \\ 0 &= x_1^2 + x_2^2 - 1 \end{aligned} \tag{6.1}$$

and consider the solution with the boundary condition

$$\begin{aligned} x_4(0) &= 0 \\ x_1(0.55) &= 0. \end{aligned} \tag{6.2}$$

That means, we ask for the inital point from which the pendulum needs the time 0.55 to reach the lowest point. (6.1) represents an index–3 DAE so that we have to reduce the index (e.g. by differentiation). The index–2 version is given by

$$\begin{array}{rcl} x_1' &=& x_3 \\ x_2' &=& x_4 \\ x_3' &=& -x_1x_5 \\ x_4' &=& -x_2x_5+g \\ 0 &=& x_1x_3+x_2x_4 \end{array} \tag{6.3}$$

and the index–1 version by

$$\begin{array}{rcl} x_1' &=& x_3 \\ x_2' &=& x_4 \\ x_3' &=& -x_1x_5 \\ x_4' &=& -x_2x_5+g \\ 0 &=& x_3^2+x_4^2+gx_2-(x_1^2+x_2^2)x_5. \end{array} \tag{6.4}$$

What about the boundary conditions ? In the index–$\mu$ case we have to offer $\dim(P\cdots P_{\mu-1})$ boundary conditions. For the index–2 pendulum the following matrices and projectors are given:

$$A=P=\begin{pmatrix} 1&0&0&0&0\\ 0&1&0&0&0\\ 0&0&1&0&0\\ 0&0&0&1&0\\ 0&0&0&0&0 \end{pmatrix}, \qquad B=\begin{pmatrix} 0&0&-1&0&0\\ 0&0&0&-1&0\\ x_5&0&0&0&x_1\\ 0&x_5&0&0&x_2\\ x_3&x_4&x_1&x_2&0 \end{pmatrix},$$

$Q=I-P$,

$$A_1=A+BQ=\begin{pmatrix} 1&0&0&0&0\\ 0&1&0&0&0\\ 0&0&1&0&x_1\\ 0&0&0&1&x_2\\ 0&0&0&0&0 \end{pmatrix}, \qquad Q_1=\begin{pmatrix} 0&0&0&0&0\\ 0&0&0&0&0\\ 0&0&0&0&-x_1\\ 0&0&0&0&-x_2\\ 0&0&0&0&1 \end{pmatrix},$$

$$A_2=A_1+BPQ_1=\begin{pmatrix} 1&0&0&0&x_1\\ 0&1&0&0&x_1\\ 0&0&1&0&x_1\\ 0&0&0&1&x_2\\ 0&0&0&0&-(x_1^2+x_2^2) \end{pmatrix}.$$

The matrix $A_2$ is nonsingular, so that this example is index–2 tractable.

$$Q_{1,s}=Q_1A_2^{-1}BP=\frac{1}{x_1^2+x_2^2}\begin{pmatrix} 0&0&0&0&0\\ 0&0&0&0&0\\ x_1x_3&x_1x_4&x_1^2&x_1x_2&0\\ x_2x_3&x_2x_4&x_2x_1&x_2^2&0\\ -x_3&-x_4&-x_1&-x_2&1 \end{pmatrix},$$

$P_{1,s} = I - Q_{1,s}$,

$$PP_{1,s} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ -\frac{x_1x_3}{x_1^2+x_2^2} & -\frac{x_1x_4}{x_1^2+x_2^2} & \frac{x_2^2}{x_1^2+x_2^2} & -\frac{x_1x_2}{x_1^2+x_2^2} & \\ -\frac{x_2x_3}{x_1^2+x_2^2} & -\frac{x_2x_4}{x_1^2+x_2^2} & -\frac{x_2x_1}{x_1^2+x_2^2} & \frac{x_1^2}{x_1^2+x_2^2} & \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

It is easy to see that $\dim(\mathrm{im}(PP_{1,s})) = 3$. To guarantee the equivalence of (6.1) and (6.3) we extend the boundary condition (6.2) to the condition

$$0 = x_1(0)^2 + x_2(0)^2 - 1.$$

In the index–1 case, $\dim(\mathrm{im}(P)) = 4$ and we use the additional BC (because of the second differentiation of the algebraic equation)

$$0 = x_1(0)x_3(0) + x_2(0)x_4(0).$$

We use $x(0) = (1.0, 0.3, 0.0, 0.0, 1.0)$ as initial values for the single shooting. The theoretical value for $x_2(0.55)$ is equal to 1. The computational results are:

| Index | Accuracy of integration | h of finite difference | $x_1(0)$ | $x_2(0)$ | $x_2(0.55)$ |
|---|---|---|---|---|---|
| 1 | $10^{-10}$ | | 0.948702560 | 0.316169975 | 0.999999998 |
| 1 | $10^{-6}$ | | 0.948707... | 0.31615... | 0.999997... |
| 2 | $10^{-6}$ | $10^{-3}$ | 0.9488... | 0.3158... | 0.99967... |
| 2 | $10^{-6}$ | $10^{-5}$ | 0.9488... | 0.3158... | 0.99966... |

The second example is published by Ascher and Spiteri in [AS93]. Consider

$$\begin{aligned} x_1' &= x_2 + x_1 y \\ x_2' &= -\omega^2 x_1 + x_2 y \\ 0 &= (\tfrac{\pi}{3})^2 x_1^2 + x_2^2 - 1 \end{aligned} \tag{6.5}$$

with the boundary condition

$$x_1(0) = 0.$$

This is an index–2 DAE and the exact solution for $\omega = \frac{\pi}{3}$ is

$$x_1 = \omega^{-1}\sin\omega t,\ x_2 = \cos\omega t,\ y = 0.$$

Let us look for $\omega$ fitting an observed function $r(t)$, where the observations are made on $x_1(t) + x_2(t)$. The necessary conditions for minimizing

$$\frac{1}{2}\int_0^2 (x_1 + x_2 - r)^2 dt \tag{6.6}$$

yield

$$\begin{aligned}
x_1' &= x_2 + x_1 y \\
x_2' &= \omega^2 x_1 + x_2 y \\
\omega' &= 0 \\
\lambda_1' &= y\lambda_1 + \omega^2\lambda_2 - 2(\frac{\pi}{3})^2 x_1\mu - (x_1 + x_2 - r) \\
\lambda_2' &= \lambda_1 - y\lambda_2 - 2x_2\mu - (x_1 + x_2 - r) \\
\nu' &= 2\omega x_1\lambda_2 \\
0 &= (\frac{\pi}{3})^2 x_1^2 + x_2^2 - 1 \\
0 &= x_1\lambda_1 + x_2\lambda_2
\end{aligned}$$

with boundary conditions

$$x_1(0) = 0,\ \nu(0) = 0,\ \nu(2) = 0,$$

$$x_2(2)\lambda_1(2) - (\frac{\pi}{3})^2 x_1(2)\lambda_2(2) = 0.$$

The matrix of the matrix chain $A_2$ is given by

$$A_2 = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & x_2-x_1(1+y) & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & x_1\omega^2-x_2(1+y) & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & x_2+\omega^2*\lambda_2-\lambda_1(1-y)+x_1(1+2(\frac{\pi}{3})^2\mu) & (2(\frac{\pi}{3})^2x_1)+2x_2\omega^2-(2(\frac{\pi}{3})^2x_1y) \\
0 & 0 & 0 & 0 & 1 & 0 & x_1-\lambda_1+\lambda_2(1-y)+x_2(1+2\mu) & (-2(\frac{\pi}{3})^2x_1)+2x_2(1-y) \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 4x_1x_2\omega \\
0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -2
\end{pmatrix}, \quad (6.7)$$

and it is very easy to see that $A_2$ is nonsingular. Projectors in the solution are given by

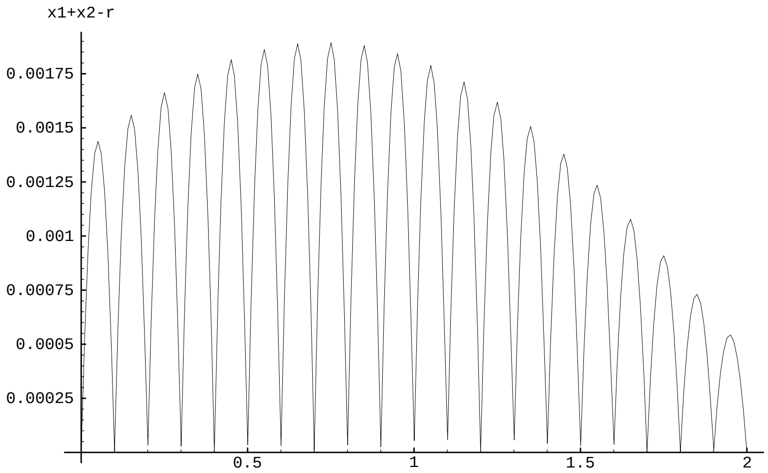
$$PP_1 = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}, PQ_1 = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}.$$

This shows in an impressive that this projector technique splits the different components of the solution of a DAE very effectively.
Different functions $r$ are used. At first, $r$ is chosen as the exact solution of (6.5) and then a

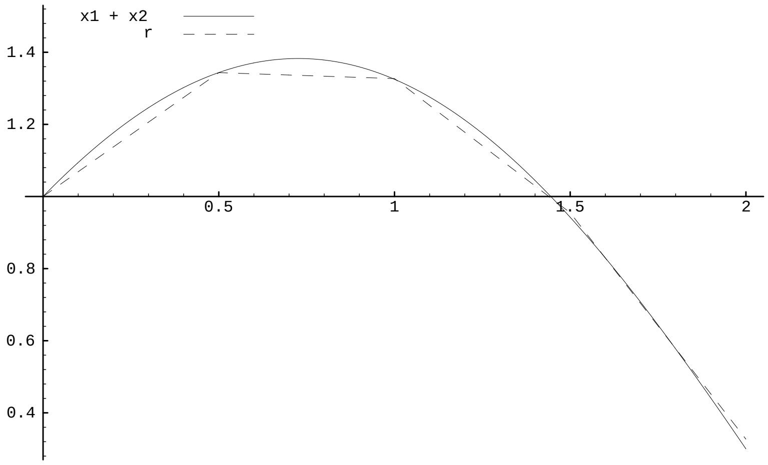
linear equidistant interpolation of this function with a different number of intervals (20, 16, 8, 4) is performed.

| $r(t)$ | | $\omega$ |
|---|---|---|
| $(\frac{3}{\pi})^2 \sin \frac{\pi}{3}t + \cos \frac{\pi}{3}t$ | | $\underline{1.047197}37$ |
| intervals : | 20 | 1.04751092 |
| | 16 | 1.04770058 |
| | 8 | 1.04922216 |
| | 4 | 1.05548500 |

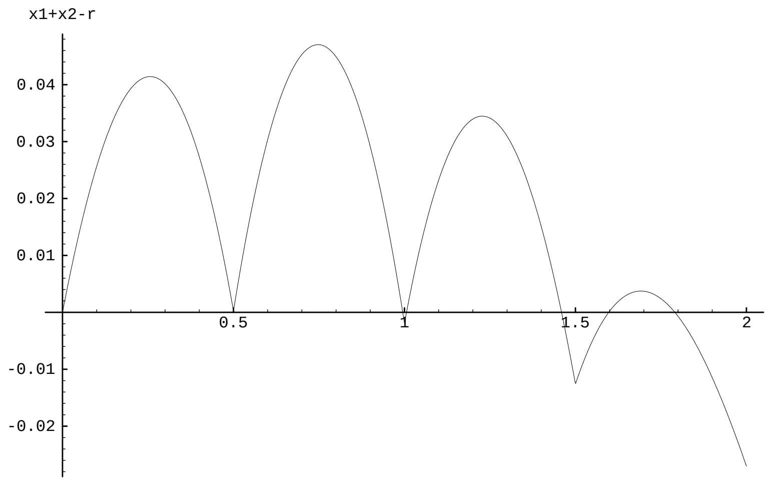
The result for $\omega$ lies in the accuracy of integration and small changes of the function $r$ lead to a moderate changing of $\omega$, too.
For 20 intervals the difference between the given function $r$ and $x_1 + x_2$ is given by



For 4 intervals, $r$ and $x_1 + x_2$ are given by




and their difference by



## References


[AP91] Uri Ascher and Linda Petzold. Stability of computational methods for constrained dynamics systems. TR 91–3, Dept. Computer Science, Univ. of B.C., 1991 *SIAM J. Sci. Stat. Comput.*, 1993. to appear.

[GP84] C.W. Gear and Linda R. Petzold. Ode methods for the solution of differential/algebraic systems. *SIAM J. Numer. Anal.*, 21:716–728, 1984.

[Han90] B. Hansen. Linear time–varying differential–algebraic equations being tractable with the index k. Preprint No. 246, Humboldt-Univ., Sekt. Math., Berlin, 1990.

[Lam91] René Lamour. A well–posed shooting method for transferable dae's. *Numerische Mathematik*, 59, 1991.

[Mä89] R. März. Index-2 differential-algebraic equations. *Results in Math.*, 15:149–171, 1989.

[Mä91b] Roswitha März. On quasilinear index 2 differential–algebraic equations, In E. Griepentrog,, M. Hanke and R. März, editors BerlinSeminar on Differential–Algebraic Equations, Humboldt–Universität zu Berlin, 1992, p 39–60.

[Mä87] Roswitha März. A matrix chain for analysing differential–algebraic equations. Preprint No. 162, der Sektion Mathematik der Humboldt–Universität zu Berlin, 1987.

[Mä91] Roswitha März. Numerical methods for differential–algebraic equations, Acta Numerica (1992),pp 141–198.

[GM87] Eberhard Griepentrog and Roswitha März. Differential–Algebraic Equations and Their Numerical Treatment, Leipzig 1886

q

[LPG91] B. Leimkuhler, L.R. Petzold and C.W. Gear. Approximation methods for the consistent initialization of Differential–Algebraic Equations. SIAM J. Numer. Anal., Vol. 28, No. 1, pp 205–226(1991)

[AS93] Uri M. Ascher and Raymond J. Spiteri Collocation Software for Boundary Value Differential Value Differential–Algebraic Equations Working paper of University of British Columbia 1993