Zentrum für Technomathematik
Fachbereich 3 – Mathematik und Informatik

Collocation methods for solving linear differential-algebraic boundary value problems

Ronald Stöver

Report 99–08

Berichte aus der Technomathematik

Report 99–08 September 1999

Collocation methods for solving linear differential-algebraic boundary value problems

Ronald Stöver

Universität Bremen, Fachbereich 3 - Mathematik und Informatik, Postfach 330 440, D-28334 Bremen, Germany, e-mail: stoever@math.uni-bremen.de

Summary We consider boundary value problems for linear differential-algebraic equations with variable coefficients without any restriction for the index. A well known regularization procedure yields an equivalent index one problem with d differential and $a = n - d$ algebraic equations. Collocation methods based on the regularized BVP approximate the solution x by a continuous piecewise polynomial of degree k and deliver, in particular, consistent approximations at mesh points by using the Radau schemes. Under weak assumptions the collocation problems are uniquely and stably solvable and, if the unique solution x is sufficiently smooth, convergence of order $\min\{k+1, 2k-1\}$ and superconvergence at mesh points of order $2k-1$ is shown. Finally, some numerical experiments illustrating these results are presented.

Key words Differential-algebraic equations – boundary value problems – collocation methods – Radau schemes

Mathematics Subject Classification (1991): 65L10

1 Introduction

In this paper we discuss the numerical solution of linear differential-algebraic boundary value problems (BVPs) with variable coefficients

$$E(t)\dot{x}(t) = A(t)x(t) + f(t) \quad \text{for all } t \in \mathbb{I} \tag{1.1}$$

$$Cx(\underline{t}) + Dx(\overline{t}) = r\,, \tag{1.2}$$

by collocation methods. Here is $\mathbb{I} = [\underline{t}, \overline{t}] \subset \mathbb{R}$ a closed interval, the functions $E, A : \mathbb{I} \longrightarrow \mathbb{R}^{n\times n}$ and $f : \mathbb{I} \longrightarrow \mathbb{R}^n$ are matrix valued respectively vector valued and $C, D \in \mathbb{R}^{d\times n}$, $r \in \mathbb{R}^d$ for some $d \leq n$. The dimension d of the boundary condition is discussed below. A solution $x : \mathbb{I} \longrightarrow \mathbb{R}^n$ is supposed to be continuously differentiable on the whole interval.

There exist several collocation approaches to such problems, see [5], [1] and [3]/[4], [6] and [7] or [9]. While these authors restrict the (differentiation) index of the differential-algebraic equation (DAE) to one or consider only semi explicit problems of index at most two we deal with DAEs of arbitrary index by using the regularization procedure developed in [11].

Up to now only Gauß schemes or Lobatto schemes have been considered in collocation methods for differential-algebraic BVPs causing difficulties of instability, oscillations and loss of order in superconvergence [1] and singular systems of algebraic equations, respectively. To overcome these difficulties several approaches are made, e.g. [5],[6],[7],[9]. Here we use Radau schemes such that consistency of the approximations is fulfilled automatically and none of the above difficulties occur.

The paper is organized as follows. In §2, we give some basic results concerning the regularization, a canonical form for the regularized DAE and the existence and uniqueness of solutions for the BVPs. In §3, we discuss the choice of collocation knots, show the unique solvability of the collocation problems and prove convergence of order $\min\{k+1, 2k-1\}$ and superconvergence of order $2k-1$. Some numerical examples illustrating these results are presented in §4. Finally, we give some conclusions in §5.

2 Basic results

We consider a DAE of index $\nu \geq 1$ and suppose the data to be ν times continuously differentiable, i.e. $E, A \in C^\nu(\mathbb{I}, \mathbb{R}^{n\times n})$, $f \in C^\nu(\mathbb{I}, \mathbb{R}^n)$. Under this assumption there exists [11] a smooth transformation of (1.1) to a DAE of the form

$$\hat{E}(t)\dot{x}(t) = \hat{A}(t)x(t) + \hat{f}(t) \tag{2.1}$$

with

$$\hat{E} = \begin{bmatrix} \hat{E}_1 \\ 0 \end{bmatrix}, \hat{A} = \begin{bmatrix} \hat{A}_1 \\ \hat{A}_2 \end{bmatrix}, \hat{f} = \begin{bmatrix} \hat{f}_1 \\ \hat{f}_2 \end{bmatrix},$$

which has the following properties:

a) The solution sets of (1.1) and (2.1) are the same.
b) The DAE (2.1) is of index one.
c) A value $x_0 \in \mathbb{R}^n$ is consistent if and only if $\hat{A}_2(t_0)x_0 + \hat{f}_2(t_0) = 0$.
d) $\hat{E}_1(t) \in \mathbb{R}^{d\times n}$ has full row rank d for all $t \in \mathbb{I}$.
e) For smooth data of the original DAE we get a smooth regularized equation: $E, A, f \in C^k$ for some $k \geq \nu \Rightarrow \hat{E}, \hat{A}, \hat{f} \in C^{k-(\nu-1)}$
f) We can compute $\hat{E}(t_i), \hat{A}(t_i), \hat{f}(t_i)$ for discrete points $t_i \in \mathbb{I}$ in an efficient and numerically stable way by use of singular value decompositions [13].

By part c) and d) we see that the regularized DAE (2.1) is split into d differential equations

$$\hat{E}_1(t)\dot{x}(t) = \hat{A}_1(t)x(t) + \hat{f}_1(t)$$

and $a = n - d$ algebraic equations

$$0 = \hat{A}_2(t)x(t) + \hat{f}_2(t)\,.$$

Thus it is appropriate to pose d boundary conditions to the DAE.

The main tool in the proofs of §3 is the transformation of (2.1) to a canonical form. We use the equivalence transformation [10]

$$\Big(E_1, A_1\Big) \sim \Big(E_2, A_2\Big) : \Longleftrightarrow \ \exists P \in C(\mathbb{I}, \mathbb{R}^{n\times n}), Q \in C^1(\mathbb{I}, \mathbb{R}^{n\times n})$$
$$\text{pointwise regular :}$$
$$\Big(E_2, A_2\Big) = \Big(PE_1Q, PA_1Q - PE_1\dot{Q}\Big)\,.$$

Proposition 2.1 ([16]) *For $E, A \in C^\nu(\mathbb{I}, \mathbb{R}^{n\times n})$ there exist pointwise regular $P \in C(\mathbb{I}, \mathbb{R}^{n\times n})$, $Q \in C^1(\mathbb{I}, \mathbb{R}^{n\times n})$ such that*

$$P\hat{E}Q = \begin{bmatrix} I_d & 0 \\ 0 & 0 \end{bmatrix}\ ,\ P\hat{A}Q - P\hat{E}\dot{Q} = \begin{bmatrix} 0 & 0 \\ 0 & I_a \end{bmatrix}\ .$$

If $\hat{E}, \hat{A} \in C^k(\mathbb{I}, \mathbb{R}^{n\times n})$ for $k \geq 1$ then this canonical form can be achieved with $P \in C^{k-1}(\mathbb{I}, \mathbb{R}^{n\times n})$, $Q \in C^k(\mathbb{I}, \mathbb{R}^{n\times n})$. For P we have the special structure

$$P = \begin{bmatrix} P_{11} & P_{12} \\ 0 & P_{22} \end{bmatrix} \text{ with } P_{11}(t) \in \mathbb{R}^{d\times d}, P_{12}(t) \in \mathbb{R}^{d\times a}, P_{22}(t) \in \mathbb{R}^{a\times a}.$$

Using the transformation to canonical form we get the following proposition about existence and uniqueness of the solution for the BVP when we transform the boundary condition analogously, i.e.

$$\begin{bmatrix} C_{11} & C_{12} \end{bmatrix} := CQ(\underline{t})\ ,\ \begin{bmatrix} D_{11} & D_{12} \end{bmatrix} := DQ(\overline{t})\ . \tag{2.2}$$

It follows directly by considering the BVP in canonical form.

Proposition 2.2 *A boundary value problem (2.1),(1.2) is uniquely solvable if and only if* $C_{11} + D_{11} \in \mathbb{R}^{d\times d}$ *is regular.*

Of course this condition is only of theoretical character, since P, Q cannot be computed numerically. A more practical condition [8] is the regularity of the shooting matrix $S := CX(\underline{t}, \underline{t}) + DX(\overline{t}, \underline{t}) \in \mathbb{R}^{d\times d}$, where $X(t, \underline{t})$ is a fundamental matrix [8],[16] to the DAE (1.1).

Throughout the paper we use

$$\|y\| := \max_{1\le i\le n} |y_i|\,,\ \|Y\| := \max_{1\le i\le m} \sum_{j=1}^{n} |y_{ij}|$$

as norm for vectors y and matrices Y, respectively.

3 Collocation methods

We want to determine a piecewise polynomial as a numerical approximation to the BVP solution. For that we choose a mesh

$$\pi \ :\ \underline{t} = t_0 < t_1 < \ldots < t_N = \overline{t} \tag{3.1}$$

with mesh widths $h_i := t_{i+1} - t_i \quad (i = 0, \ldots, N-1),\ h := \max h_i$ and use k collocation knots $0 \le \rho_1 < \ldots < \rho_k \le 1$ to subdivide each of the intervals $[t_i, t_{i+1}]$ by collocation points

$$t_{ij} = t_i + h_i\rho_j \quad (j = 1, \ldots, k\,,\ i = 0, \ldots, N-1)\,. \tag{3.2}$$

Then we compute a piecewise polynomial x_π of degree k, i.e. $x_{\pi,i} := x_{\pi|[t_i,t_{i+1}]}$ is a polynom of degree k, such that the following conditions are fulfilled:

a) $E(t_{ij})\dot{x}_{\pi,i}(t_{ij}) = A(t_{ij})x_{\pi,i}(t_{ij}) + f(t_{ij})$ for all i, j,
   i.e. the DAE is satisfied in all collocation points,
b) $x_{\pi,i-1}(t_i) = x_{\pi,i}(t_i)$ for $i = 1, \ldots, N-1$,
   i.e. the polynomial pieces are continuously matched,
c) $0 = \hat{A}_2(t_i)x_{\pi,i}(t_i) + \hat{f}_2(t_i)$ for $i = 0, \ldots, N-1$ and
   $0 = \hat{A}_2(t_N)x_{\pi,N-1}(t_N) + \hat{f}_2(t_N)$,
   i.e. the approximations are consistent in all mesh points t_i,
d) $Cx_{\pi,0}(t_0) + Dx_{\pi,N-1}(t_N) = r$,
   i.e. the boundary condition is satisfied.

Since we need the data E, A, f only at the discrete points t_{ij}, we can work with (2.1) instead of (1.1) by performing the regularization at these points. So without loss of generality we replace a) by

$$\hat{E}(t_{ij})\dot{x}_{\pi,i}(t_{ij}) = \hat{A}(t_{ij})x_{\pi,i}(t_{ij}) + \hat{f}(t_{ij}) \text{ for all } i, j\,.$$

At first sight this is a problem with $N(k+1)n$ unknowns, depending on N polynomial pieces with each based on $k+1$ parameters of dimension n, but

$$\underbrace{Nkn}_{\text{collocation}} + \underbrace{(N-1)n}_{\text{continuity}} + \underbrace{(N+1)a}_{\text{consistency}} + \underbrace{d}_{\text{BC}} = N(k+1)n + Na$$

conditions. When using Gauß knots ρ_j a modification of the collocation method is necessary, e.g. projected collocation (see [5] and [3], [4]) or a perturbation of x_π [6].

The idea is to use schemes such that some conditions are implied by others. By fixing $\rho_k = 1$ (or $\rho_1 = 0$ equivalently) we obtain $t_{ik} = t_i + h_i = t_{i+1}$, thus consistency in t_{i+1} follows already from the collocation condition for t_{ik} and we may skip the conditions c) for $i = 1, \ldots, N$. This yields a problem with $N(k+1)n$ unknowns and $N(k+1)n$ conditions.

The choice $\rho_1 = 0$ and $\rho_k = 1$, e.g. in the Lobatto schemes, means $t_{ik} = t_{i+1} = t_{i+1,0}$, so the collocation conditions in t_{ik} and $t_{i+1,0}$ together with the continuity condition in t_{i+1} cause a redundancy. To overcome these problems [7],[9] as an additional condition the differentiable solution part is supposed to be continuously differentiable instead of just being continuous. But this approach needs the distinction of differential and algebraic parts in the solution and is applicable only to a restricted class of BVP.

So in the following we consider schemes with

$$0 < \rho_1 < \ldots < \rho_k = 1\,, \tag{3.3}$$

e.g. the Radau schemes.

3.1 Solvability of the collocation problems

For the polynomial piece $x_{\pi,i}$ we use a representation as a Lagrange interpolation polynomial according to the points $(t_i, x_i), (t_{i1}, x_{i1}), \ldots, (t_{ik}, x_{ik})$, i.e.

$$x_{\pi,i}(t) = \sum_{l=0}^{k} x_{il} L_l \left(\frac{t - t_i}{h_i} \right) \tag{3.4}$$

with

$$L_l(\tau) := \prod_{j=0, j\neq l}^{k} \frac{\tau - \rho_j}{\rho_l - \rho_j}\,,\ l = 0, \ldots, k\,.$$

Here we use the notation $t_{i0} := t_i$, $\rho_0 := 0$, $x_{i0} := x_i$ and for the collocation conditions we denote $v_{jl} := L_l'(\rho_j) = L_l'\left(\frac{t_{ij}-t_i}{h_i}\right)$ for $j = 1,\ldots,k$ and $l = 0,\ldots,k$.

It is easy to prove [16] that $V := (v_{jl})_{j,l=1,\ldots,k} \in \mathbb{R}^{k\times k}$ is regular and we set $(w_{jl})_{j,l=1,\ldots,k} := V^{-1}$. Finally we introduce $x_N := x_{N-1,k}$.

Summarizing the discussion and using this notation the collocation method reduces to the solution of the system of linear equations (with $j = 1,\ldots,k$ and $i = 0,\ldots,N-1$)

$$\hat{E}(t_{ij})\left(\frac{1}{h_i}\sum_{l=0}^{k} v_{jl}x_{il}\right) - \hat{A}(t_{ij})x_{ij} = \hat{f}(t_{ij}) \tag{3.5}$$

$$x_{ik} = x_{i+1} \tag{3.6}$$

$$-\hat{A}_2(t_0)x_0 = \hat{f}_2(t_0) \tag{3.7}$$

$$Cx_0 + Dx_N = r \tag{3.8}$$

In order to examine this system according to existence and uniqueness of solutions we first consider only the k collocation conditions for each subinterval $[t_i, t_{i+1}]$. Treating only $x_{i1},\ldots,x_{ik}$ as unknowns at this point, (3.5) for $j = 1,\ldots,k$ results in the local system

$$B_i \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ik} \end{bmatrix} = a_i x_i + b_i \tag{3.9}$$

with

$$B_i := \left[\begin{array}{c|c|c|c} \frac{v_{11}}{h_i}\hat{E}_1 - \hat{A}_{11} & \frac{v_{12}}{h_i}\hat{E}_1 & \cdots & \frac{v_{1k}}{h_i}\hat{E}_1 \\ -\hat{A}_{21} & 0 & & 0 \\ \hline \frac{v_{21}}{h_i}\hat{E}_2 & & & \vdots \\ 0 & & & \vdots \\ \hline \vdots & & & \frac{v_{k-1,k}}{h_i}\hat{E}_{k-1} \\ \vdots & & & 0 \\ \hline \frac{v_{k1}}{h_i}\hat{E}_k & \cdots & \frac{v_{k,k-1}}{h_i}\hat{E}_k & \frac{v_{kk}}{h_i}\hat{E}_k - \hat{A}_{1k} \\ 0 & & 0 & -\hat{A}_{2k} \end{array}\right] \in \mathbb{R}^{kn\times kn},$$

$$a_i := \left[\begin{array}{c} -\frac{v_{10}}{h_i}\hat{E}_1 \\ 0 \\ \hline \vdots \\ \hline -\frac{v_{k0}}{h_i}\hat{E}_k \\ 0 \end{array}\right] \in \mathbb{R}^{kn\times n}, \quad b_i := \left[\begin{array}{c} \hat{f}_{11} \\ \hat{f}_{21} \\ \hline \vdots \\ \hline \hat{f}_{1k} \\ \hat{f}_{2k} \end{array}\right] \in \mathbb{R}^{kn\times 1},$$

where $\hat{E}_j := \hat{E}_1(t_{ij})$, $\hat{A}_{1j} := \hat{A}_1(t_{ij})$, $\hat{A}_{2j} := \hat{A}_2(t_{ij})$, $\hat{f}_{1j} := \hat{f}_1(t_{ij})$ and $\hat{f}_{2j} := \hat{f}_2(t_{ij})$.

To prove the regularity of B_i we use the transformation to canonical form introduced in §2 for multiplications of B_i from the left and from the right, respectively, with

$$T_P := \begin{bmatrix} P_1 & & \\ & \ddots & \\ & & P_k \end{bmatrix}, \; T_Q := \begin{bmatrix} Q_1 & & \\ & \ddots & \\ & & Q_k \end{bmatrix} \in \mathbb{R}^{kn \times kn}$$

and $P_j := P(t_{ij})$, $Q_j := Q(t_{ij})$. We also need permutations of the rows and columns done by multiplication with

$$U_k := \begin{bmatrix} I_d & 0 & & & 0 & 0 & & \\ 0 & 0 & & & I_a & 0 & & \\ & I_d & & & & 0 & & \\ & 0 & & & & I_a & & \\ & & \ddots & & & & \ddots & \\ & & & I_d & & & & 0 \\ & & & 0 & & & & I_a \end{bmatrix} \in \mathbb{R}^{kn \times kn} . \tag{3.10}$$

For the following analysis we omit the index i.

Lemma 3.1 *If a transformation to canonical form with $Q \in C^2$ is possible we have the representation*

$$B = T_P^{-1} \, U_k \begin{bmatrix} hV^{-1} \otimes I & 0 \\ 0 & -I \end{bmatrix}^{-1} \Big(I + \Delta \Big) \, U_k^* \, T_Q^{-1}$$

where

$$\Delta := \begin{bmatrix} \Delta^1 & \Delta^2 \\ 0 & 0 \end{bmatrix}, \; \Delta^s := \Big(h \sum_{l=1}^{k} w_{jl} G_{lm}^s \Big)_{j,m=1,\ldots,k} \quad (s = 1, 2)$$

and (for $m = 0, \ldots, k$, $l, j = 1, \ldots, k$)

$$\begin{bmatrix} G_{lm}^1 & G_{lm}^2 \end{bmatrix} := \begin{cases} (\rho_m - \rho_l) v_{lm} (P_{11} \hat{E}_1 \dot{Q})(t_{il}) + \mathcal{O}(h) & l \neq m \\ -(P_{11} \hat{E}_1 \dot{Q})(t_{il}) & l = m \end{cases}$$

For sufficiently small h, the matrix B is regular with

$$B^{-1} = T_Q \, U_k \Big(I - \Delta + \mathcal{O}(h^2) \Big) \begin{bmatrix} hV^{-1} \otimes I & 0 \\ 0 & -I \end{bmatrix} U_k^* \, T_P \, .$$

Proof : With $Q \in C^2$ we can expand
$Q(t_{il}) - Q(t_{ij}) = h(\rho_l - \rho_j)\dot{Q}(t_{ij}) + \mathcal{O}(h^2)$ and this leads to

$$\begin{aligned}
&P_j \begin{bmatrix} \frac{v_{jl}}{h}\hat{E}_j \\ 0 \end{bmatrix} Q_l \\
&= \frac{v_{jl}}{h}\left(P_j \begin{bmatrix} \hat{E}_j \\ 0 \end{bmatrix} Q_j + \begin{bmatrix} P_{11}(t_{ij})\hat{E}_1(t_{ij})\Big(Q(t_{il}) - Q(t_{ij})\Big) \\ 0 \end{bmatrix} \right) \\
&= \begin{bmatrix} \frac{v_{jl}}{h} I & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} G^1_{jl} & G^2_{jl} \\ 0 & 0 \end{bmatrix} \quad \text{for } j \neq l.
\end{aligned}$$

Analogously for $j = l$:

$$\begin{aligned}
&P_j \begin{bmatrix} \frac{v_{jj}}{h}\hat{E}_j - \hat{A}_{1j} \\ -\hat{A}_{2j} \end{bmatrix} Q_j = \frac{v_{jj}}{h} P_j \begin{bmatrix} \hat{E}_j \\ 0 \end{bmatrix} Q_j - P_j \begin{bmatrix} \hat{A}_{1j} \\ \hat{A}_{2j} \end{bmatrix} Q_j \\
&= \begin{bmatrix} \frac{v_{jj}}{h} I & 0 \\ 0 & 0 \end{bmatrix} - \left(\begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} + \begin{bmatrix} (P_{11}\hat{E}_1\dot{Q})(t_{ij}) \\ 0 \end{bmatrix} \right) \\
&= \begin{bmatrix} \frac{v_{jj}}{h} I & 0 \\ 0 & -I \end{bmatrix} + \begin{bmatrix} G^1_{jj} & G^2_{jj} \\ 0 & 0 \end{bmatrix}.
\end{aligned}$$

So by multiplication of B with T_P from the left and T_Q from the right and reordering of the rows and columns using U_k we get

$$U_k^* \, T_P \, B \, T_Q \, U_k =$$

$$\underbrace{\begin{bmatrix} \frac{v_{11}}{h} I & \cdots & \frac{v_{1k}}{h} I & & & \\ \vdots & & \vdots & & & \\ \frac{v_{k1}}{h} I & \cdots & \frac{v_{kk}}{h} I & & & \\ & & & -I & & \\ & & & & \ddots & \\ & & & & & -I \end{bmatrix}}_{= \begin{bmatrix} \frac{1}{h} V \otimes I & 0 \\ 0 & -I \end{bmatrix}} + \underbrace{\begin{bmatrix} G^1_{11} & \cdots & G^1_{1k} & G^2_{11} & \cdots & G^2_{1k} \\ \vdots & & \vdots & \vdots & & \vdots \\ G^1_{k1} & \cdots & G^1_{kk} & G^2_{k1} & \cdots & G^2_{kk} \\ & & & & & \\ & & & & & \\ & & & & & \end{bmatrix}}_{=: \begin{bmatrix} G^1 & G^2 \\ 0 & 0 \end{bmatrix}}.$$

Because of the regularity of V and $(w_{jl})_{j,l} = V^{-1}$ we have

$$\begin{bmatrix} hV^{-1} \otimes I & 0 \\ 0 & -I \end{bmatrix} U_k^* \, T_P \, B \, T_Q \, U_k = I + \Delta$$

and

$$B = T_P^{-1} U_k \begin{bmatrix} hV^{-1} \otimes I & 0 \\ 0 & -I \end{bmatrix}^{-1} \Big(I + \Delta \Big) U_k^* T_Q^{-1},$$

respectively, with the representations (for $s = 1, 2$)

$$\Delta^s = \Big(hV^{-1} \otimes I\Big) G^s = h \begin{bmatrix} w_{11}I \cdots w_{1k}I \\ \vdots \quad\quad \vdots \\ w_{k1}I \cdots w_{kk}I \end{bmatrix} \begin{bmatrix} G^s_{11} \cdots G^s_{1k} \\ \vdots \quad\quad \vdots \\ G^s_{k1} \cdots G^s_{kk} \end{bmatrix}$$
$$= \left(h \sum_{l=1}^{k} w_{jl} G^s_{lm} \right)_{j,m=1,\ldots,k} .$$

Since G_{lm} is bounded for $h \to 0$ and k is fixed, we have

$$\|\Delta\| \le h \max_j \sum_{m=1}^{k} \left\| \sum_{l=1}^{k} w_{jl} G^1_{lm} \right\| + \left\| \sum_{l=1}^{k} w_{jl} G^2_{lm} \right\| = \mathcal{O}(h) .$$

Thus $I + \Delta$ is regular for sufficiently small h and has the inverse

$$\Big(I + \Delta\Big)^{-1} = I - \Delta + \mathcal{O}(h^2) .$$

From this follows

$$\left(\begin{bmatrix} hV^{-1} \otimes I & 0 \\ 0 & -I \end{bmatrix} U_k^* T_P B T_Q U_k \right)^{-1} = \Big(I + \Delta\Big)^{-1} = I - \Delta + \mathcal{O}(h^2)$$

$$\Rightarrow B^{-1} = T_Q \, U_k \Big(I - \Delta + \mathcal{O}(h^2) \Big) \begin{bmatrix} hV^{-1} \otimes I & 0 \\ 0 & -I \end{bmatrix} U_k^* \, T_P$$

and the proof is completed. □

For the continuity conditions $x_{ik} = x_{\pi,i}(t_{i+1}) \stackrel{!}{=} x_{\pi,i+1}(t_{i+1}) = x_{i+1}$ we will use an expression for x_{ik} in dependence of x_i which results from the solution of the local systems (3.9):

$$x_{ik} = \underbrace{\big[\, 0 \cdots 0 \, I \,\big] \, B_i^{-1} a_i}_{=:W_i} \cdot x_i + \underbrace{\big[\, 0 \cdots 0 \, I \,\big] \, B_i^{-1} b_i}_{=:g_i} . \qquad (3.11)$$

The next lemma gives representations for W_i and g_i.

Lemma 3.2 *If a transformation to canonical form with $Q \in C^2$ is possible then the following representations (for $0 \le i \le N-1$) hold:*

$$W_i = Q(t_{i+1}) \begin{bmatrix} I - F_{i1} & -F_{i2} \\ 0 & 0 \end{bmatrix} Q(t_i)^{-1} \quad \text{with } F_{i1} = \mathcal{O}(h_i^2), F_{i2} = \mathcal{O}(h_i),$$

$$g_i = Q(t_{i+1}) \begin{bmatrix} c_i \\ -(P_{22}\hat{f}_2)(t_{i+1}) \end{bmatrix} \quad \text{with } c_i = \mathcal{O}(h_i) .$$

Proof : Using the representation for B_i^{-1} given in Lemma 3.1 we compute $W_i Q(t_i) = [\,0 \ldots 0\, I\,] B_i^{-1} a_i Q(t_i)$.
With $Q_0 := Q(t_i)$ and $Q_0 - Q_j = -\rho_j h_i \dot{Q}(t_{ij}) + \mathcal{O}(h_i^2)$ we have

$$P_j \begin{bmatrix} -\frac{v_{j0}}{h_i}\hat{E}_j \\ 0 \end{bmatrix} Q_0 = -\frac{v_{j0}}{h_i}\left(P_j \begin{bmatrix} \hat{E}_j \\ 0 \end{bmatrix} Q_j + P_j \begin{bmatrix} \hat{E}_j \\ 0 \end{bmatrix} (Q_0 - Q_j) \right)$$

$$\Rightarrow \quad U_k^*\, T_P\, a_i\, Q_0 = U_k^* \begin{bmatrix} P_1 & & \\ & \ddots & \\ & & P_k \end{bmatrix} \begin{bmatrix} -\frac{v_{10}}{h_i}\hat{E}_1 \\ 0 \\ \vdots \\ -\frac{v_{k0}}{h_i}\hat{E}_k \\ 0 \end{bmatrix} Q_0$$

$$= -\frac{1}{h_i} \begin{bmatrix} v_0 \otimes I & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} G_0^1 & G_0^2 \\ 0 & 0 \end{bmatrix}$$

with $v_0 := (v_{j0})_{j=1,\ldots,k}$ and $G_0^s := (G_{j0}^s)_{j=1,\ldots,k}$, $s = 1, 2$. We easily derive that $v_0 = -V \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}^*$ by considering

$$\sum_{l=0}^{k} v_{jl} = \sum_{l=0}^{k} L_l'(\rho_j) = 0$$

for all j, and this leads to

$$\begin{bmatrix} h_i V^{-1} \otimes I & 0 \\ 0 & -I \end{bmatrix} U_k^*\, T_P a_i Q_0 = \underbrace{\begin{bmatrix} I & 0 \\ \vdots & \vdots \\ I & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}}_{=:\tilde{I}} - \underbrace{\begin{bmatrix} \Delta_{10}^1 & \Delta_{10}^2 \\ \vdots & \vdots \\ \Delta_{k0}^1 & \Delta_{k0}^2 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}}_{=:\tilde{\Delta}_i},$$

with $\Delta_{j0}^s = \mathcal{O}(h_i)$ defined as in Lemma 3.1. The next step yields

$$\left(I - \Delta_i + \mathcal{O}(h_i^2)\right)\left(\tilde{I} - \tilde{\Delta}_i\right) = \tilde{I} - \tilde{\Delta}_i - \Delta_i \tilde{I} + \mathcal{O}(h_i^2)$$

$$= \begin{bmatrix} I & 0 \\ \vdots & \vdots \\ I & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} \Delta_{10}^1 & \Delta_{10}^2 \\ \vdots & \vdots \\ \Delta_{k0}^1 & \Delta_{k0}^2 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} \sum \Delta_{1m}^1 & 0 \\ \vdots & \vdots \\ \sum \Delta_{km}^1 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \mathcal{O}(h_i^2) & \mathcal{O}(h_i^2) \\ \vdots & \vdots \\ \mathcal{O}(h_i^2) & \mathcal{O}(h_i^2) \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} * & * \\ \vdots & \vdots \\ * & * \\ I-F_{i1} & -F_{i2} \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} =: \Theta_i \,,$$

with $F_{i1} := \Delta^1_{k0} + \sum_{m=1}^{k} \Delta^1_{km} + \mathcal{O}(h_i^2)$, $F_{i2} := \Delta^2_{k0} + \mathcal{O}(h_i^2)$. Altogether this yields

$$W_i Q(t_i) = [\,0 \ldots 0\, I\,] B_i^{-1} a_i Q_0 = [\,0 \ldots 0\, I\,] \begin{bmatrix} Q_1 & & \\ & \ddots & \\ & & Q_k \end{bmatrix} U_k \Theta_i$$

and hence

$$W_i = Q(t_{i+1}) \begin{bmatrix} I-F_{i1} & -F_{i2} \\ 0 & 0 \end{bmatrix} Q(t_i)^{-1} \,.$$

In order to show that $F_{i2} = \mathcal{O}(h_i^2)$ we use interpolation according to $\rho_0, \ldots, \rho_k$ of the polynomials $p(t) = t$, $q(t) = 1$ and get

$$\sum_{m=0}^{k} L'_m(\rho_l)\rho_m = 1\,, \quad \sum_{m=0}^{k} L'_m(\rho_l) = 0\,.$$

By inserting the definitions of Lemma 3.1 we see

$$\Delta^1_{k0} + \sum_{m=1}^{k} \Delta^1_{km} = h_i \sum_{l=1}^{k} w_{kl} G^1_{l0} + \sum_{m=1}^{k} h_i \sum_{l=1}^{k} w_{kl} G^1_{lm}$$

$$= h_i \sum_{l=1}^{k} w_{kl} (P_{11}\hat{E}_1)(t_{il}) \left(\sum_{m\neq l}^{k} \Big(v_{lm}(\rho_m - \rho_l)\dot{Q}(t_{il}) + \mathcal{O}(h_i) \Big) - \dot{Q}_1(t_{il}) \right)$$

$$= h_i \sum_{l=1}^{k} w_{kl} (P_{11}\hat{E}_1)(t_{il}) \left(\underbrace{\left[\sum_{m=0}^{k} L'_m(\rho_l)(\rho_m - \rho_l) - 1 \right]}_{=0} \dot{Q}_1(t_{il}) + \mathcal{O}(h_i) \right)$$

$$\Rightarrow \quad F_{i1} = \Delta^1_{k0} + \sum_{m=1}^{k} \Delta^1_{km} + \mathcal{O}(h_i^2) = \mathcal{O}(h_i^2)\,.$$

Looking at the definition of Δ_{k0}^2 it is obvious that $F_{i2} = \mathcal{O}(h_i)$.

The representation $g_i = Q(t_{i+1}) \begin{bmatrix} c_i \\ -(P_{22}\hat{f}_2)(t_{i+1}) \end{bmatrix}$ with $c_i = \mathcal{O}(h_i)$ can be derived analogously by inserting the representation for B_i^{-1} given in Lemma 3.1 into $g_i = \begin{bmatrix} 0 \cdots 0 \, I \end{bmatrix} B_i^{-1} b_i$. $\square$

After the solution of the local systems we have the continuity conditions $x_{i+1} = W_i x_i + g_i$, $i = 0, \ldots, N-1$, instead of (3.6) which together with the the consistency condition (3.7) and the boundary condition (3.8) form the global system

$$K_h \begin{bmatrix} x_0 \\ \vdots \\ x_N \end{bmatrix} = g_h \, , \tag{3.12}$$

with $K_h \in \mathbb{R}^{(N+1)n \times (N+1)n}$ and $g_h \in \mathbb{R}^{(N+1)n}$ defined as

$$K_h := \begin{bmatrix} C & & & & D \\ -\hat{A}_2(t_0) & & & & 0 \\ W_0 & -I & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & W_{N-1} & -I \end{bmatrix} , g_h := \begin{bmatrix} r \\ \hat{f}_2(t_0) \\ -g_0 \\ \vdots \\ \vdots \\ -g_{N-1} \end{bmatrix} .$$

We must prove the regularity of K_h and the boundedness of $K_h^{-1} g_h$. For this the following notation is useful: Let $U_N \in \mathbb{R}^{(N+1)n \times (N+1)n}$ be a permutation matrix as U_k in (3.10), and let

$$T_l := \begin{bmatrix} I & 0 & & & \\ 0 & P_{22}(t_0) & & & \\ & & Q(t_1)^{-1} & & \\ & & & \ddots & \\ & & & & Q(t_N)^{-1} \end{bmatrix} , T_r := \begin{bmatrix} Q(t_0) & & & \\ & \ddots & & \\ & & \ddots & \\ & & & Q(t_N) \end{bmatrix}$$

be defined by use of P, Q, which transform the differential-algebraic equation to canonical form. They will be considered for multiplication of K_h from the left and from the right, respectively. Set

$$M_h := \begin{bmatrix} C_{11} & & & D_{11} \\ I & -I & & \\ & \ddots & \ddots & \\ & & I & -I \end{bmatrix} , N_h := \begin{bmatrix} C_{12} & & & D_{12} \\ -F_{02} & 0 & & \\ & \ddots & \ddots & \\ & & -F_{N-1,2} & 0 \end{bmatrix} ,$$

$$D_h := \begin{bmatrix} 0 & & & \\ -F_{01} & 0 & & \\ & \ddots & \ddots & \\ & & -F_{N-1,1} & 0 \end{bmatrix},$$

with $C_{11}, C_{12}, D_{11}, D_{12}$ given in (2.2) and F_{i1}, F_{i2} given in Lemma 3.2, and set $A_h := \begin{bmatrix} M_h & N_h \\ 0 & -I \end{bmatrix}$, $\Delta_h := \begin{bmatrix} D_h & 0 \\ 0 & 0 \end{bmatrix}$.

Lemma 3.3 *We have the representation*

$$K_h = T_l^{-1}\, U_N\, (A_h + \Delta_h)\, U_N^*\, T_r^{-1}\,.$$

For a uniquely solvable BVP (2.1),(1.2) and a smooth transformation function $Q \in C^2$, the matrix K_h is regular for sufficiently small h with

$$K_h^{-1} = T_r\, U_N \Big(I - A_h^{-1}\Delta_h + \mathcal{O}(h^2) \Big) A_h^{-1}\, U_N^*\, T_l\,.$$

Furthermore, $K_h^{-1} g_h$ is bounded by a constant which depends on the data $\hat{E}, \hat{A}, \hat{f}, C, D, r$ and the transformation functions P, Q, but not on h.

Proof : By multiplication with T_l from the left and T_r from the right we get blockwise

$$\begin{bmatrix} I & 0 \\ 0 & P_{22}(t_0) \end{bmatrix} \begin{bmatrix} C \\ -\hat{A}_2(t_0) \end{bmatrix} Q(t_0) = \begin{bmatrix} CQ(t_0) \\ -(P_{22}\hat{A}_2 Q)(t_0) \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} \\ 0 & -I \end{bmatrix}$$

and (according to Lemma 3.2)

$$\begin{aligned} & Q(t_{i+1})^{-1} W_i Q(t_i) \\ &= Q(t_{i+1})^{-1} \left(Q(t_{i+1}) \begin{bmatrix} I - F_{i1} & -F_{i2} \\ 0 & 0 \end{bmatrix} Q(t_i)^{-1} \right) Q(t_i) \\ &= \begin{bmatrix} I - F_{i1} & -F_{i2} \\ 0 & 0 \end{bmatrix}. \end{aligned}$$

Reordering of the rows and columns yields

$$\begin{aligned} & U_N^*\, T_l\, K_h\, T_r\, U_N \\ &= U_N^* \left[\begin{array}{cc|cc|cc|cc} C_{11} & C_{12} & & & & & D_{11} & D_{12} \\ 0 & -I & & & & & 0 & 0 \\ \hline I - F_{01} & -F_{02} & -I & 0 & & & & \\ 0 & 0 & 0 & -I & & & & \\ \hline & & \ddots & & \ddots & & & \\ & & & & I - F_{N-1,1} & -F_{N-1,2} & -I & 0 \\ & & & & 0 & 0 & 0 & -I \end{array}\right] U_N \\ &= A_h + \Delta_h\,. \end{aligned}$$

By Proposition 2.2 the matrix $S := C_{11} + D_{11}$ is regular and so is M_h with inverse

$$M_h^{-1} = \begin{bmatrix} S^{-1} & & \\ & \ddots & \\ & & S^{-1} \end{bmatrix} \begin{bmatrix} I & D_{11} & \cdots & D_{11} \\ \vdots & -C_{11} & \ddots & \vdots \\ \vdots & \vdots & \ddots & D_{11} \\ I & -C_{11} & \cdots & -C_{11} \end{bmatrix} .$$

From this it follows that

$$\begin{aligned} \|M_h^{-1} D_h\| &\le \|S^{-1}\| \max\{\|C_{11}\|, \|D_{11}\|\} \cdot \sum_{i=0}^{N-1} \underbrace{\|F_{i1}\|}_{=\mathcal{O}(h_i^2)} \\ &\le \|S^{-1}\| \max\{\|C_{11}\|, \|D_{11}\|\} \cdot const \cdot h \underbrace{\sum_{i=0}^{N-1} h_i}_{=t_N - t_0} \\ &= \mathcal{O}(h) \quad \text{for } h \to 0 . \end{aligned}$$

By this we see the regularity of A_h and

$$\|A_h^{-1} \Delta_h\| = \left\| \begin{bmatrix} M_h^{-1} & M_h^{-1} N_h \\ 0 & -I \end{bmatrix} \begin{bmatrix} D_h & 0 \\ 0 & 0 \end{bmatrix} \right\| = \|M_h^{-1} D_h\| = \mathcal{O}(h) .$$

Thus $A_h + \Delta_h = A_h \Big(I + A_h^{-1} \Delta_h \Big)$ is regular for sufficiently small h with

$$\Big(A_h + \Delta_h \Big)^{-1} = \Big(I + A_h^{-1} \Delta_h \Big)^{-1} A_h^{-1} = \Big(I - A_h^{-1} \Delta_h + \mathcal{O}(h^2) \Big) A_h^{-1} .$$

Inserting this into $K_h = T_l^{-1} U_N \, (A_h + \Delta_h) \, U_N^* \, T_r^{-1}$ proves the representation of K_h^{-1}.

Using $g_i = Q(t_{i+1}) \begin{bmatrix} c_i \\ -(P_{22} \hat{f}_2)(t_{i+1}) \end{bmatrix}$ (see Lemma 3.2) we get

$$A_h^{-1} \cdot U_N^* \, T_l \, g_h = \begin{bmatrix} M_h^{-1} d \\ e \end{bmatrix}$$

with $e := \big[-(P_{22} \hat{f}_2)(t_0)^* \cdots -(P_{22} \hat{f}_2)(t_N)^* \big]^*$, $d := \big[d_0^* \cdots d_N^* \big]^*$ and

$$d_i := \begin{cases} r + C_{12}(P_{22} \hat{f}_2)(t_0) + D_{12}(P_{22} \hat{f}_2)(t_N) & \text{for } i = 0 \\ -c_{i-1} - F_{i-1,2}(P_{22} \hat{f}_2)(t_{i-1}) & \text{for } i = 1, \ldots, N \end{cases}$$

Since $c_{i-1} = \mathcal{O}(h_{i-1})$, $F_{i-1,2} = \mathcal{O}(h_{i-1})$ we also have $d_i = \mathcal{O}(h_{i-1})$ for $i = 1, \ldots, N$, which gives the boundedness of $M_h^{-1} d$, i.e.

$$\begin{aligned} \|M_h^{-1} d\| &\leq \|S^{-1}\| \Big(\|d_0\| + \max\{\|C_{11}\|, \|D_{11}\|\} \cdot \sum_{i=1}^{N} \|d_i\| \Big) \\ &\leq \|S^{-1}\| \Big(\|d_0\| + \max\{\|C_{11}\|, \|D_{11}\|\} \cdot const \cdot (t_N - t_0) \Big) \, . \end{aligned}$$

Of course e is bounded by $\sup_{t \in \mathbb{I}} \|(P_{22} \hat{f}_2)(t)\|$ and thus

$$\begin{aligned} \|K_h^{-1} g_h\| &= \Big\| T_r \; U_N \Big(I - A_h^{-1} \Delta_h + \mathcal{O}(h^2) \Big) A_h^{-1} \; U_N^* \; T_l \; g_h \Big\| \\ &\leq \Big\| T_r \; U_N \Big(I - A_h^{-1} \Delta_h + \mathcal{O}(h^2) \Big) \Big\| \cdot \max \Big\{ \|M_h^{-1} d\|, \|e\| \Big\} \end{aligned}$$

is bounded independent of h. $\square$

The following theorem about existence and uniqueness of solutions of collocation problems (3.5)-(3.8) can now be proved by combining Lemma 3.1 (solvability of the local systems) and Lemma 3.3 (solvability of the global system). If the data is smooth, i.e. $\hat{E}, \hat{A} \in C^2$, the existence of a transformation to canonical form with $Q \in C^2$ is guaranteed by Proposition 2.1.

Theorem 3.1 *Consider a uniquely solvable BVP (2.1),(1.2) with smooth data $\hat{E}, \hat{A} \in C^2$, $\hat{f} \in C$. For $N \in \mathbb{N}$ and $k \geq 1$ define a mesh π as in (3.1) and collocation points t_{ij} (for $j = 1, \ldots, k$, $i = 0, \ldots, N-1$) as in (3.2) according to knots ρ_j as in (3.3).*

Then for sufficiently small mesh widths $h_0, \ldots, h_{N-1}$, there exists one and only one continuous piecewise polynomial x_π of degree k that satisfies (2.1) at all collocation points t_{ij}, fulfills the boundary condition (1.2) and is consistent at all mesh points t_i.

The collocation methods are stable, i.e. the approximations x_i, x_{ij} are bounded independent of the mesh π, since the x_i are bounded (see Lemma 3.3) and this leads to the boundedness of the x_{ij}, according to a representation

$$x_{ij} = \left(Q(t_{ij}) \begin{bmatrix} I - F_{i1} & -F_{i2} \\ 0 & 0 \end{bmatrix} Q(t_i)^{-1} \right) x_i + Q(t_{ij}) \begin{bmatrix} c_{ij} \\ -(P_{22} \hat{f}_2)(t_{ij}) \end{bmatrix} ,$$

which is similar to $x_{ik} = W_i x_i + g_i$.

3.2 Convergence results

The next aim is to examine convergence of the collocation methods. The expected results (convergence of order k, superconvergence of order $2k-1$) will be proved.

Theorem 3.2 *Let x be the unique solution of a BVP (2.1),(1.2) and let x_π be the unique solution of the corresponding collocation problem with parameter k and sufficiently small mesh widths h_i.*

If x is smooth, i.e. $x \in C^{k+1}(\mathbb{I}, \mathbb{R}^n)$, then

$$\|x - x_\pi\|_\infty = \sup_{t \in \mathbb{I}} \|x(t) - x_\pi(t)\| = \mathcal{O}(h^k) \, .$$

Proof : Interpolation of x analogous to (3.4) means that

$$x(t) = \sum_{l=0}^{k} x(t_{il}) L_l \left(\frac{t - t_i}{h_i} \right) + \underbrace{\frac{x^{(k+1)}(\theta_i(t))}{(k+1)!} \prod_{j=0}^{k} (t - t_{ij})}_{=:\psi_i(t)}$$

for some $\theta_i(t) \in [t_i, t_{i+1}]$. Inserting this representation into the DAE at the collocation points t_{ij} delivers the local system

$$B_i \begin{bmatrix} x(t_{i1}) \\ \vdots \\ x(t_{ik}) \end{bmatrix} = a_i x(t_i) + b_i - \begin{bmatrix} \tau_{i1} \\ \vdots \\ \tau_{ik} \end{bmatrix} ,$$

with B_i, a_i, b_i defined in (3.9) and $\tau_{ij} := \hat{E}(t_{ij}) \dot{\psi}_i(t_{ij})$. Since the collocation problem is uniquely solvable, i.e. B_i is regular, this can be solved and leads to (with W_i, g_i defined in (3.11))

$$x(t_{ik}) = W_i x(t_i) + g_i - \tau_i \, .$$

For the error $\tau_i := \begin{bmatrix} 0 \cdots 0 \, I \end{bmatrix} B_i^{-1} \Big(\tau_{ij} \Big)_{j=1,\ldots,k}$ a representation

$$\tau_i = Q(t_{i+1}) \begin{bmatrix} \varphi_i \\ 0 \end{bmatrix} , \quad \varphi_i = \mathcal{O}(h_i^{k+1})$$

can be derived analogously to that of g_i (see Lemma 3.2) since $\psi_i(t) = \mathcal{O}(h_i^{k+1})$, $\dot{\psi}_i(t) = \mathcal{O}(h_i^k)$. The continuity, boundary and consistency conditions for x lead to the global system (comparable to (3.12))

$$K_h \begin{bmatrix} x(t_0) \\ \vdots \\ x(t_N) \end{bmatrix} = g_h + \tau_h \, , \quad \tau_h := \begin{bmatrix} 0 \\ \tau_0 \\ \vdots \\ \tau_{N-1} \end{bmatrix} .$$

According to the unique solvability of the collocation problem, K_h is regular and so the difference of the global systems for x and x_π, respectively, gives

$$K_h \begin{bmatrix} x(t_0) - x_0 \\ \vdots \\ x(t_N) - x_N \end{bmatrix} = \tau_h \,. \tag{3.13}$$

Due to the higher order of τ_h we have $K_h^{-1}\tau_h = \mathcal{O}(h^k)$ (this can be proved like the boundedness of $K_h^{-1}g_h$ in Lemma 3.3), i.e.

$$\max_i \|x(t_i) - x_i\| = \mathcal{O}(h^k) \,.$$

Looking at the difference in the local systems we obtain

$$\begin{bmatrix} x(t_{i1}) - x_{i1} \\ \vdots \\ x(t_{ik}) - x_{ik} \end{bmatrix} = B_i^{-1} a_i \underbrace{\Big(x(t_i) - x_i\Big)}_{=\mathcal{O}(h^k)} - B_i^{-1} \underbrace{\begin{bmatrix} \tau_{i1} \\ \vdots \\ \tau_{ik} \end{bmatrix}}_{=\mathcal{O}(h_i^k)} \tag{3.14}$$

and hence $\max_j \|x(t_{ij}) - x_{ij}\| = \mathcal{O}(h^k)$.
Considering the difference of the interpolation representations we get

$$x(t) - x_\pi(t) = \sum_{l=0}^{k} \Big(x(t_{il}) - x_{il}\Big) L_l \left(\frac{t - t_i}{h_i}\right) + \psi_i(t)$$

and thus

$$\begin{aligned} \|x(t) - x_\pi(t)\| &\le const \cdot \max_j \|x(t_{ij}) - x_{ij}\| + \|\psi_i(t)\| \\ &= \mathcal{O}(h^k) \,. \quad \square \end{aligned}$$

So far we have only made the assumptions $0 < \rho_1 < \ldots < \rho_k = 1$. Choosing the Radau scheme a higher convergence order in the mesh points t_i, i.e. superconvergence, can be achieved.

Theorem 3.3 *Let x be the unique solution of a BVP (2.1),(1.2) and let x_π be the unique solution of the corresponding collocation problem with parameter k and sufficiently small mesh widths h_i. Use the Radau scheme with $\rho_k = 1$ to construct the collocation points t_{ij}.*

If the data is smooth, i.e. $\hat{E}, \hat{A} \in C^{2k}$, $\hat{f} \in C^{2k-1}$, then

$$\max_{0 \le i \le N} \|x(t_i) - x_i\| = \mathcal{O}(h^{2k-1}) \,.$$

Proof : Since $\hat{E}, \hat{A}$ are smooth, a transformation to canonical form with $P \in C^{2k-1}, Q \in C^{2k}$ exists, see Proposition 2.1. Recalling the consistency of $x_i = x_\pi(t_i)$, we can consider the solution v of the initial value problem $\hat{E}\dot{y} = \hat{A}y + \hat{f}$, $y(t_i) = x_i$ which has, due to the corresponding initial value problem in canonical form, a representation

$$(Q^{-1}v)(t) = \begin{bmatrix} [\, I \; 0 \,] \left(Q(t_i)^{-1} x_i + \int_{t_i}^t (P\hat{f})(s) ds \right) \\ -(P_{22}\hat{f}_2)(t) \end{bmatrix}, t \geq t_i \,.$$

Analogously we have for the solution x_π of the initial value problem $\hat{E}\dot{y} = \hat{A}y + \hat{E}\dot{x}_\pi - \hat{A}x_\pi$, $y(t_i) = x_i$:

$$(Q^{-1}x_\pi)(t) = \begin{bmatrix} [\, I \; 0 \,] \left(Q(t_i)^{-1} x_i + \int_{t_i}^t (P(\hat{E}\dot{x}_\pi - \hat{A}x_\pi))(s) ds \right) \\ (P_{22}\hat{A}_2 x_\pi)(t) \end{bmatrix}$$

for $t_i \leq t \leq t_{i+1}$. Since x_π is consistent in mesh points, the difference of these representations at $t = t_{i+1}$ gives

$$v(t_{i+1}) - x_{i+1} = Q(t_{i+1}) \begin{bmatrix} \int_{t_i}^{t_{i+1}} \phi(s) ds \\ 0 \end{bmatrix},$$

with a smooth function $\phi \in C^{2k-1}([t_i, t_{i+1}], \mathbb{R}^d)$ (note the smoothness of $\hat{E}, \hat{A}, \hat{f}, P$ and x_π) of the form

$$\phi(s) := [\, I \; 0 \,] P(s) \left(\hat{f}(s) - \hat{E}(s)\dot{x}_\pi(s) + \hat{A}(s)x_\pi(s) \right).$$

Because x_π satisfies the DAE at the collocation points t_{ij}, these k points are zeros of ϕ and from this the existence of a smooth function $w \in C^{k-1}$ follows [16], such that $\phi(s) = w(s) \prod_{j=1}^k (s - t_{ij})$ for $s \in [t_i, t_{i+1}]$. A Taylor expansion of w yields a polynom ψ of degree at most $k-2$ with

$$\phi(s) = \psi(s) \prod_{j=1}^k (s - t_{ij}) + \mathcal{O}(h_i^{2k-1}), s \in [t_i, t_{i+1}] \,.$$

From the orthogonality property of the Radau scheme we get

$$\begin{aligned} \int_{t_i}^{t_{i+1}} \phi(s) ds &= \int_{t_i}^{t_{i+1}} \left(\psi(s) \prod_{j=1}^k (s - t_{ij}) + \mathcal{O}(h_i^{2k-1}) \right) ds \\ &= h_i^{k+1} \underbrace{\int_0^1 \psi(t_i + h_i\tau) \prod_{j=1}^k (\tau - \rho_j) d\tau}_{=0} + \mathcal{O}(h_i^{2k}) = \mathcal{O}(h_i^{2k}) \end{aligned}$$

and this means that the local discretization error

$$\phi_i := v(t_{i+1}) - x_{i+1} = Q(t_{i+1}) \begin{bmatrix} \int_{t_i}^{t_{i+1}} \phi(s) ds \\ 0 \end{bmatrix}$$

is of order $\mathcal{O}(h_i^{2k})$.

Considering a fundamental solution matrix $W(\cdot, t_i)$, i.e. a solution of

$$\hat{E}\dot{W} = \hat{A}W \,,\; W(t_i, t_i) = Q(t_i) \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} Q(t_i)^{-1} \,,$$

we see that $x(t) - v(t) = W(t, t_i)(x(t_i) - v(t_i))$ for all $t \geq t_i$, and especially for t_{i+1} we have

$$W(t_{i+1}, t_i)\Big(x(t_i) - x_i\Big) = x(t_{i+1}) - v(t_{i+1}) = x(t_{i+1}) - x_{i+1} - \phi_i \,.$$

This holds for $i = 0, \ldots, N-1$ and builds together with the boundary condition and the consistency condition in t_0 the system

$$\begin{bmatrix} C & & & D \\ -\hat{A}_2(t_0) & & & 0 \\ W(t_1, t_0) & -I & & \\ & \ddots & \ddots & \\ & & W(t_N, t_{N-1}) & -I \end{bmatrix} \begin{bmatrix} x(t_0) - x_0 \\ \vdots \\ x(t_N) - x_N \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -\phi_0 \\ \vdots \\ -\phi_{N-1} \end{bmatrix}$$

comparable to (3.13). From this we derive (as in Lemma 3.3)

$$\max_i \|x(t_i) - x_i\| = \mathcal{O}(h^{2k-1}) \,,$$

since the inhomogeneity is of order $\mathcal{O}(h^{2k})$. □

Inserting this result into (3.14) the following improvement of the convergence order can be shown [16].

Corollary 3.1 *Under the assumptions of Theorem 3.2 it follows that*

$$\|x(t_{ij}) - x_{ij}\| = \mathcal{O}(h_i^{k+1}) + \mathcal{O}(h^{2k-1})$$

for $1 \leq j \leq k, 0 \leq i \leq N-1$ *and*

$$\sup_{t \in \mathbb{I}} \|x(t) - x_\pi(t)\| = \mathcal{O}\left(h^{\min\{k+1, 2k-1\}}\right) \,.$$

4 Numerical examples

In order to illustrate the practicability and effectiveness of the described collocation methods they are applied to three challenging examples. A MATLAB code has been developed including a simple strategy for the generation and refinement of the meshes π. The package DGELDA [13] is used for the regularization. As input the data $\underline{t}, \overline{t}$, C, D, r are needed as well as FORTRAN subroutines for evaluation of E, A, f and its derivatives up to degree $\nu-1$ at discrete points. The parameter $1 \leq k \leq 5$ and a tolerance for the mesh selection must be chosen.

Example 4.1 In [14] a double operational amplifier is modeled as a DAE $E\dot{x}(t) = Ax(t) + f(t)$ with matrices $E, A \in \mathbb{R}^{6\times 6}$ and f depending on an input signal U_{in}. This is an index one problem with $d = 2$, $a = 4$.

Taking a periodic input we can look for a periodic solution x. A necessary and sufficient condition for this is that

$$\hat{E}_1 x(0) - \hat{E}_1 x(T) = 0\,,$$

where T is the periodicity of the input.

We choose parameters $C_1 = C_2 = 10^{-6}$, $R_1 = R_2 = 10^3$, $\alpha = 10^4$ and an input signal $U_{in}(t) = \cos(200\pi t)$ (i.e. $T = 0.01$). The collocation method with $k = 4$ and a tolerance 10^{-9} for the mesh selection computes a periodic solution with an error

$$\max_i \|U_{out}(t_i) - x_{4i}\| \approx 0.24 \cdot 10^{-6}$$

(comparing only the output signals of the analytical solution and the numerical approximation at mesh points).

Example 4.2 Ascher and Spiteri considered in [5] the semi explicit index two problem

$$\begin{bmatrix} 1&0&0\\ 0&1&0\\ 0&0&0 \end{bmatrix} \dot{x}(t) = \begin{bmatrix} \kappa - \frac{1}{2-t} & 0 & (2-t)\kappa \\ \frac{\kappa-1}{2-t} & -1 & \kappa - 1 - \frac{\kappa p(t)}{2+t} \\ t+2-p(t) & t^2-4 & 0 \end{bmatrix} x(t) + \begin{bmatrix} \frac{3-t}{2-t}\mathrm{e}^t \\ \left(2 + \frac{(\kappa+2)p(t)+p'(t)}{t^2-4} - \frac{2tp(t)}{(t^2-4)^2}\right)\mathrm{e}^t \\ -(t^2+t-2)\mathrm{e}^t \end{bmatrix}$$

on $\mathbb{I} = [0, 1]$ with initial condition $x_1(0) = 1$. It consists of $d = 1$ differential and $a = 2$ algebraic equations, so no more boundary conditions are necessary. The solution of this BVP is

$$x(t) = \left[\mathrm{e}^t \left(1 + \frac{p(t)}{t^2 - 4} \right) \mathrm{e}^t - \frac{\mathrm{e}^t}{2 - t} \right]^* .$$

By taking parameter $\kappa = 20$ and the parameter function

$$p(t) = -\left(1 + \mathrm{erf}\Big(\frac{t - 1/3}{\sqrt{2\varepsilon}} \Big) \right) , \varepsilon = 10^{-5}$$

(as done in [5]) a layer region around $t = \frac{1}{3}$ occurs in p and in the second solution component and we can ask whether the collocation method can rebuild this layer region. Figure 4.1 shows the approxi-

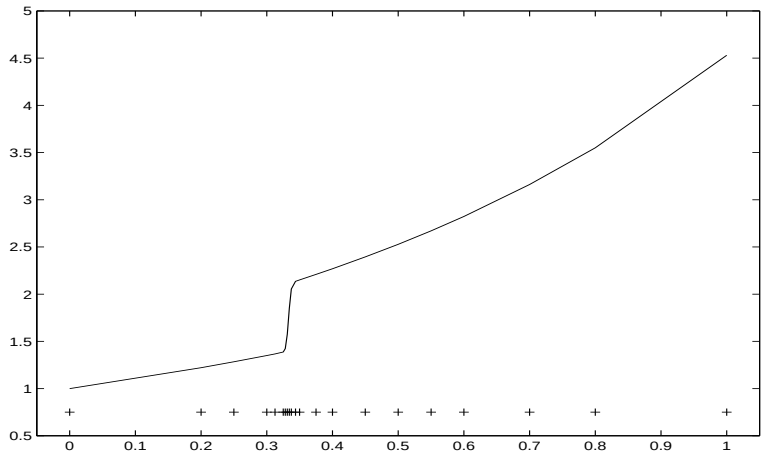

Fig. 4.1. Approximation for x_2 and corresponding mesh

mation for x_2 computed by the collocation method with $k = 4$ and the corresponding mesh which consists of $N = 28$ points and has been built in five refinement steps. The error is

$$\max_i \|x(t_i) - x_i\| \approx 0.3 \cdot 10^{-3} .$$

Example 4.3 A planar truck model is given in [15]. Here we consider its linearization

$$E\dot{x}(t) = Ax(t) + f(t)$$

with constant coefficient matrices $E, A \in \mathbb{R}^{23 \times 23}$ and an inhomogeneity f which depends, among others, on a function u modeling the road

profile. This equation is of index three with $d = 20$, $a = 3$ and badly scaled.

We choose a wavy road profile $u(t) := 0.05 \sin(20\pi t)$ (meaning waves of height 5 cm and length 3 m if the truck is driving at a speed of $30\frac{m}{sec}$) and look for the motion of the driver's seat that is caused by this, i.e. we solve the DAE together with a periodic boundary condition

$$\hat{E}_1 x(0) - \hat{E}_1 x(T) = 0\,, T = 0.1\,.$$

Figure 4.2 shows the solution of the collocation method with $k = 4$ in comparison to the road profile.

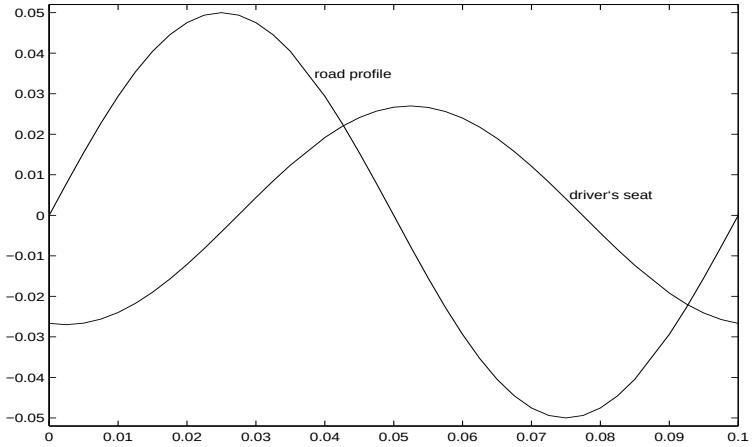

Fig. 4.2.

5 Conclusions

In this paper we have shown that numerical approximations to the solution of linear differential-algebraic BVPs of arbitrary index can be computed in an efficient and stable way via collocation methods applied to the equivalent, regularized BVPs. Since we use Radau schemes, consistency is implied in the collocation conditions, thus we do not need special adaptations, as projected collocation [3],[5] or a perturbation of the approximating polynomial [6], of the collocation methods that work for BVPs for ordinary differential equations. A drawback of these methods may be the unsymmetry caused by the Radau schemes. Similar but symmetric collocation methods are currently under investigation as well as methods for nonlinear BVPs.

Acknowledgements I thank Peter Kunkel for many fruitful discussions. This paper is dedicated to Jule Protzmann.

References

1. Ascher, U. (1989): On numerical differential-algebraic problems with application to semiconductor device simulation. SIAM J. Numer. Anal. 26, 517-538
2. U. Ascher, U., Mattheij, R., Russell, R. (1995): Numerical solution of boundary value problems for ordinary differential equations. SIAM, Philadelphia, second edition
3. Ascher, U., Petzold, L.R. (1991): Projected implicit Runge-Kutta methods for differential-algebraic equations. SIAM J. Numer. Anal. 28, 1097-1120
4. Ascher, U., Petzold, L.R. (1992): Projected collocation for higher-order higher-index differential-algebraic equations. J. Comp. Appl. Math. 43, 243-259
5. Ascher, U., Spiteri, R. (1994): Collocation software for boundary value differential-algebraic equations. SIAM J. Sci. Comput. 15, 938-952
6. Bai, Y. (1991): A perturbed collocation method for boundary value problems in differential-algebraic equations. Appl. Math. Comput. 45, 269-291
7. Bai, Y. (1992): A modified Lobatto collocation for linear boundary value problems of differential-algebraic equations. Computing 49, 139-150
8. Clark, K., Petzold, L.R. (1989): Numerical solution of boundary value problems in differential-algebraic systems. SIAM J. Sci. Stat. Comput. 10, 915-936
9. Degenhardt, A. (1992): Collocation for transferable differential-algebraic equations. Semin.ber., Humboldt-Univ. Berlin, Fachbereich Math 92-1, 83-104
10. Kunkel, P., Mehrmann, V. (1994): Canonical Forms for linear differential-algebraic equations with variable coefficients. J. Comput. Appl. Math. 56, 225-251
11. Kunkel, P., Mehrmann, V. (1996): Local and global invariants of linear differential-algebraic equations and their relation. Electron. Trans. Numer. Anal. 4, 138-157
12. Kunkel, P., Mehrmann, V. (1996): A new class of discretization methods for the solution of linear differential-algebraic equations. SIAM J. Numer. Anal. 33, 1941-1961
13. Kunkel, P., Mehrmann, V., Rath, W., Weickert, J. (1997): A new software package for the solution of linear differential-algebraic equations. SIAM J. Sci. Comput. 18, 115-138
14. Kampowsky, W. Rentrop, P., Schmidt, W. (1992): Classification and numerical simulation of electric circuits. Surv. Math. Ind. 2, 23-65
15. Simeon, B., Grupp, F., Führer, C., Rentrop, P. (1994): A nonlinear truck model and its treatment as a multibody system. J. Comp. Appl. Math. 50,523-532
16. Stöver, R. (1999): Numerische Lösung von linearen differential-algebraischen Randwertproblemen. Doctoral thesis, Universität Bremen

Berichte aus der Technomathematik ISSN 1435-7968

http://www.math.uni-bremen.de/zetem/berichte.html

— Vertrieb durch den Autor —

Reports **Stand: 29. September 1999**

98–01. Peter Benner, Heike Faßbender:
An Implicitly Restarted Symplectic Lanczos Method for the Symplectic Eigenvalue Problem, Juli 1998.

98–02. Heike Faßbender:
Sliding Window Schemes for Discrete Least-Squares Approximation by Trigonometric Polynomials, Juli 1998.

98–03. Peter Benner, Maribel Castillo, Enrique S. Quintana-Ortí:
Parallel Partial Stabilizing Algorithms for Large Linear Control Systems, Juli 1998.

98–04. Peter Benner:
Computational Methods for Linear–Quadratic Optimization, August 1998.

98–05. Peter Benner, Ralph Byers, Enrique S. Quintana-Ortí, Gregorio Quintana-Ortí:
Solving Algebraic Riccati Equations on Parallel Computers Using Newton's Method with Exact Line Search, August 1998.

98–06. Lars Grüne, Fabian Wirth:
On the rate of convergence of infinite horizon discounted optimal value functions, November 1998.

98–07. Peter Benner, Volker Mehrmann, Hongguo Xu:
A Note on the Numerical Solution of Complex Hamiltonian and Skew-Hamiltonian Eigenvalue Problems, November 1998.

98–08. Eberhard Bänsch, Burkhard Höhn:
Numerical simulation of a silicon floating zone with a free capillary surface, Dezember 1998.

99–01. Heike Faßbender:
The Parameterized SR Algorithm for Symplectic (Butterfly) Matrices, Februar 1999.

99–02. Heike Faßbender:
Error Analysis of the symplectic Lanczos Method for the symplectic Eigenvalue Problem, März 1999.

99–03. Eberhard Bänsch, Alfred Schmidt:
Simulation of dendritic crystal growth with thermal convection, März 1999.

99–04. Eberhard Bänsch:
Finite element discretization of the Navier-Stokes equations with a free capillary surface, März 1999.

99–05. Peter Benner:
Mathematik in der Berufspraxis, Juli 1999.

99–06. Andrew D.B. Paice, Fabian R. Wirth:
Robustness of nonlinear systems and their domains of attraction, August 1999.

99–07. Peter Benner, Enrique S. Quintana-Ortí, Gregorio Quintana-Ortí:
Balanced Truncation Model Reduction of Large-Scale Dense Systems on Parallel Computers, September 1999.

99–08. Ronald Stöver:
Collocation methods for solving linear differential-algebraic boundary value problems, September 1999.