Numer. Math. (2002) 91: 475–501
Digital Object Identifier (DOI) 10.1007/s002110100315

Numerische
Mathematik

Symmetric collocation methods for linear differential-algebraic boundary value problems

Peter Kunkel[1]**, Ronald Stöver**[2]

[1] Fachbereich Mathematik, Carl von Ossietzky Universität, Postfach 2503, 26111 Oldenburg, Germany; e-mail: kunkel@math.uni-oldenburg.de
[2] Universität Bremen, Fachbereich 3 - Mathematik und Informatik, Postfach 330 440, 28334 Bremen, Germany; e-mail: stoever@math.uni-bremen.de

Received September 22, 2000 / Revised version received February 7, 2001 /
Published online August 17, 2001 –

Summary. We present symmetric collocation methods for linear differential-algebraic boundary value problems without restrictions on the index or the structure of the differential-algebraic equation. In particular, we do not require a separation into differential and algebraic solution components. Instead, we use the splitting into differential and algebraic equations (which arises naturally by index reduction techniques) and apply Gauß-type (for the differential part) and Lobatto-type (for the algebraic part) collocation schemes to obtain a symmetric method which guarantees consistent approximations at the mesh points. Under standard assumptions, we show solvability and stability of the discrete problem and determine its order of convergence. Moreover, we show superconvergence when using the combination of Gauß and Lobatto schemes and discuss the application of interpolation to reduce the number of function evaluations. Finally, we present some numerical comparisons to show the reliability and efficiency of the new methods.

Mathematics Subject Classification (1991): 65L10

1 Introduction

In this paper, we consider symmetric collocation methods for the solution of linear differential-algebraic boundary value problems (BVPs) with variable

Correspondence to: P. Kunkel

coefficients

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

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

where $\mathbb{I} = [\underline{t}, \overline{t}] \subset \mathbb{R}$ is a closed interval, $E, A \in C^\nu(\mathbb{I}, \mathbb{R}^{n\times n})$, $f \in C^\nu(\mathbb{I}, \mathbb{R}^n)$, $C, D \in \mathbb{R}^{d\times n}$, $r \in \mathbb{R}^d$, $d \leq n$ is the number of inherent differential equations and $\nu \geq 1$ is the well-defined differentiation index (see, e. g., [6]) of the DAE (1.1). A solution x is required to be in $C^1(\mathbb{I}, \mathbb{R}^n)$.

Under these assumptions, the index reduction techniques of [11,13] can be applied to obtain an equivalent DAE of index one. Note that these techniques can be performed numerically at any desired point $t \in \mathbb{I}$. Thus, for the construction and analysis of numerical methods we are allowed to assume that (1.1) already has differentiation index one. Moreover, the reduced systems obtained in this way have the special structure that the differential and the algebraic equations are separated. The methods we present in this paper exploit this special structure. As consequence, their application to higher index problems turns out to be more efficient than that of other collocation methods, although these can be applied to the reduced problem, too (see the discussion in [16] and the numerical comparisons below).

The main problem when using standard symmetric collocation schemes for the discretisation of (1.1), (1.2) is that in general the number of parameters and the number of conditions is unbalanced. For example, one gets an over-determined discrete problem when using Gauß collocation and requiring all approximations at mesh points to be consistent (cp. [4]). On the other side, one gets an under-determined discrete problem when using Lobatto collocation (cp. [5]). The reason for this can be seen in the choice of the discrete solution space. In a correct formulation of (1.1) in terms of a Banach space operator (see, e. g., [8,12]), the differential and algebraic solution components have different smoothness requirements for continuous inhomogeneities. But this is not reflected in the discrete solution space when we look for piecewise polynomial solutions of a certain degree for all components. Thus, in most approaches the DAE (1.1) is required to have separated differential and algebraic components of the unknown function x (e. g., (1.1) is required to be semi-explicit, cp. [2,3]), or that it can easily be transformed into such a form (e. g., by requiring that $\operatorname{kernel} E(t)$ does not depend on t, cp. [5,7]). But this means a significant restriction of the class of treatable problems. One possibility to overcome this restriction is the use of Radau-type collocation (cp. [15,16]). The drawback there is that these schemes are not symmetric thus showing undesirable effects in certain (symmetric) applications.

The approach we will discuss in this paper is based on the observation that a correct Banach space formulation can also be given when we require all solution components to have the same smoothness while the components of the inhomogeneity belonging to the differential and algebraic parts of (1.1) have different smoothness requirements. Since standard index reduction techniques (see, e. g., [11]) yield a reduced system where we can distinguish between these parts, we do not need to restrict the class of treatable problems. In particular, the methods we introduce here combine a Gauß-type scheme with k knots for the differential part with a Lobatto-type scheme with $k+1$ knots for the algebraic part.

The paper is organised as follows. In Sect. 2 we state some basic properties of DAEs that are obtained by index reduction techniques. In Sect. 3 we discuss solvability and convergence properties for the combination of Gauß-type and Lobatto-type schemes including superconvergence for the combination of Gauß and Lobatto schemes. To improve the efficiency of the presented methods we include interpolation techniques in Sect. 4. Finally we present some numerical comparisons in Sect. 5 and give some conclusions in Sect. 6.

2 Basic results

Given a BVP of the form (1.1), (1.2), application of the index reduction techniques of [11,13] yields a DAE

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

with

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

and block-sizes d and $a = n - d$. This equation has index one and is equivalent to (1.1) in the sense that the solution sets are identical. Moreover, the special structure of the reduced DAE allows to distinguish between d differential equations

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

and a algebraic equations

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

For the development of the symmetric collocation methods, we assume without loss of generality that the DAE is in reduced form (2.1). The hats are omitted for simplicity of notation.

The main tool in the proofs of Sect. 3 is the transformation of (1.1) to a canonical form (see [10]). For more details, see [15].

Proposition 2.1 *For $E, A \in C^k(\mathbb{I}, \mathbb{R}^{n\times n})$ as in (2.1), there exist point-wise nonsingular $P \in C^{k-1}(\mathbb{I}, \mathbb{R}^{n\times n})$, $Q \in C^k(\mathbb{I}, \mathbb{R}^{n\times n})$ such that*

$$PEQ = \begin{bmatrix} I_d & 0 \\ 0 & 0 \end{bmatrix} , \quad PAQ - PE\dot{Q} = \begin{bmatrix} 0 & 0 \\ 0 & I_a \end{bmatrix} . \tag{2.2}$$

In particular, P has the special structure

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

Moreover, there exists $T_2 \in C^k(\mathbb{I}, \mathbb{R}^{n\times d})$ with point-wise full column rank and $A_2T_2 = 0$.
If in addition $f \in C^{k-1}(\mathbb{I}, \mathbb{R}^n)$, then $x \in C^{k-1}(\mathbb{I}, \mathbb{R}^n)$ for every solution x of (1.1).

Applying the transformation of Proposition 2.1 to the boundary condition (1.2) yields matrices

$$\begin{bmatrix} C_{11} & C_{12} \end{bmatrix} := CQ(\underline{t}) , \quad \begin{bmatrix} D_{11} & D_{12} \end{bmatrix} := DQ(\overline{t}) . \tag{2.3}$$

In terms of the transformed problem (2.2) (where differential and algebraic parts are decoupled), we can characterise the well-posedness of the considered problems as follows.

Proposition 2.2 *The boundary value problem (1.1), (1.2) is uniquely solvable if and only if $C_{11} + D_{11} \in \mathbb{R}^{d\times d}$ is nonsingular.*

Throughout the paper we use

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

as norms for vectors $y \in \mathbb{R}^n$ and matrices $Y \in \mathbb{R}^{m\times n}$, respectively.

3 Symmetric collocation methods

The aim of the collocation methods is to construct piecewise polynomials as numerical approximations to the BVP solution. For this we choose meshes

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

with mesh widths $h_i := t_{i+1} - t_i \quad (i = 0, \ldots, N-1)$ and a maximum width $h := \max h_i$. We use two schemes (a Gauß-type one and a Lobatto-type one, respectively, see, e. g., [9, Ch. IV] for details on Gauß and Lobatto schemes)

$$0 < \rho_1 < \cdots < \rho_k < 1 , \quad 0 = \sigma_0 < \cdots < \sigma_k = 1 \tag{3.2}$$

to subdivide the intervals $[t_i, t_{i+1}]$ by collocation points (for $i = 0, \ldots, N-1$)

$$t_{ij} = t_i + h_i \rho_j \quad \text{for } j = 1, \ldots, k, \tag{3.3}$$

$$s_{ij} = t_i + h_i \sigma_j \quad \text{for } j = 0, \ldots, k. \tag{3.4}$$

Then we compute a piecewise polynomial x_π of degree k (i. e., $x_{\pi,i} := x_{\pi|[t_i,t_{i+1}]}$ are polynomials of degree k), which is determined by the following set of conditions:

$$E_1(t_{ij})\dot{x}_{\pi,i}(t_{ij}) = A_1(t_{ij})x_{\pi,i}(t_{ij}) + f_1(t_{ij}) \tag{3.5}$$

$$0 = A_2(s_{ij})x_{\pi,i}(s_{ij}) + f_2(s_{ij}) \tag{3.6}$$

for all i, j, i. e., the differential part of the DAE is satisfied at all collocation points t_{ij} and the algebraic part at all collocation points s_{ij}, respectively,

$$T_2(t_i)^* \Big(x_{\pi,i-1}(t_i) - x_{\pi,i}(t_i) \Big) = 0 \tag{3.7}$$

for $i = 1, \ldots, N-1$, i. e, the differential part of x_π is continuous, and

$$Cx_{\pi,0}(t_0) + Dx_{\pi,N-1}(t_N) = r, \tag{3.8}$$

i. e., the boundary condition is fulfilled.

Altogether (3.5)–(3.8) yield

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

conditions. Since each of the N polynomial pieces is described by $k+1$ parameters of dimension n, we have the same number of unknowns. Note also that the consistency of x_π at all mesh points t_i is already implied by the collocation conditions (3.6), since $s_{00} = t_0$ and $s_{ik} = t_{i+1}$ for $i = 0, \ldots, N-1$.

The following proposition shows that not only the differential part (as required by (3.7)) but the whole piecewise polynomial x_π is continuous, if it satisfies the conditions (3.5)–(3.8).

Proposition 3.1 *Let the collocation conditions*

$$0 = A_2(s_{i-1,k})x_{\pi,i-1}(s_{i-1,k}) + f_2(s_{i-1,k})$$

be fulfilled. Then the following conditions are equivalent (for $i = 1, \ldots, N-1$*):*

i) $T_2(t_i)^* \Big(x_{\pi,i-1}(t_i) - x_{\pi,i}(t_i) \Big) = 0\ ,\ 0 = A_2(s_{i0})x_{\pi,i}(s_{i0}) + f_2(s_{i0})$

ii) $x_{\pi,i-1}(t_i) = x_{\pi,i}(t_i)$.

Proof. The claim follows directly from the observation that by construction

$$\begin{bmatrix} T_2(t_i)^* \\ A_2(t_i) \end{bmatrix}$$

is nonsingular. □

In the following we use conditions ii) instead of i). The "missing" collocation condition $0 = A_2(t_0)x_\pi(t_0) + f_2(t_0)$ is considered together with the boundary condition.

We use Lagrange interpolation polynomials according to the points $(s_{i0}, x_{i0}), \ldots, (s_{ik}, x_{ik})$ to represent the pieces $x_{\pi,i}$, i. e.,

$$x_{\pi,i}(t) = \sum_{l=0}^{k} x_{il} L_l\left(\frac{t-t_i}{h_i}\right), \quad L_l(\tau) := \prod_{\substack{j=0\\ j\neq l}}^{k} \frac{\tau - \sigma_j}{\sigma_l - \sigma_j}\,. \tag{3.9}$$

Defining $v_{jl} := L_l'(\rho_j)$ and $u_{jl} := L_l(\rho_j)$ for $l = 0, \ldots, k$, $j = 1, \ldots, k$, we get

$$\dot{x}_{\pi,i}(t_{ij}) = \tfrac{1}{h_i} \sum_{l=0}^{k} v_{jl} x_{il}\,, \quad x_{\pi,i}(t_{ij}) = \sum_{l=0}^{k} u_{jl} x_{il}\,, \quad x_{\pi,i}(s_{ij}) = x_{ij}\,.$$

If we set (for $j, l = 1, \ldots, k$)

$$w_{jl} := \int_0^{\sigma_j} \tilde{L}_l(\tau)d\tau\,, \quad \tilde{L}_l(\tau) := \prod_{\substack{m=1\\ m\neq l}}^{k} \frac{\tau - \rho_m}{\rho_l - \rho_m} \tag{3.10}$$

then we see that $V := (v_{jl})_{j,l}$ is regular with $V^{-1} = (w_{jl})_{j,l}$. Finally we introduce $x_N := x_{N0} := x_{\pi,N-1}(t_N)$.

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

$$\tfrac{1}{h_i} \sum_{l=0}^{k} v_{jl} E_1(t_{ij}) x_{il} - \sum_{l=0}^{k} u_{jl} A_1(t_{ij}) x_{il} = f_1(t_{ij})\,, \tag{3.11}$$

$$-A_2(s_{ij})x_{ij} = f_2(s_{ij})\,, \tag{3.12}$$

$$x_{ik} - x_{i+1,0} = 0\,, \tag{3.13}$$

$$Cx_{00} + Dx_{N0} = r\,, \tag{3.14}$$

$$-A_2(t_0)x_{00} = f_2(t_0)\,. \tag{3.15}$$

3.1 Solvability of the collocation problems

The examination of system (3.11)–(3.15) according to existence and uniqueness of solutions is divided into two steps: First we look at the local systems (for $i = 0, \dots, N-1$)

$$B_i \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ik} \end{bmatrix} = a_i x_{i0} + b_i \tag{3.16}$$

which consist of the collocation conditions (3.11) and (3.12) for $j = 1, \dots, k$. Their solvability is examined in Lemma 3.1. The solutions lead to relations

$$x_{ik} = \underbrace{\left[0 \cdots 0\, I\right] B_i^{-1} a_i}_{=:W_i} \cdot \underbrace{x_{i0}}_{=:x_i} + \underbrace{\left[0 \cdots 0\, I\right] B_i^{-1} b_i}_{=:g_i}\,, \tag{3.17}$$

which yield continuity conditions

$$x_{i+1} = W_i x_i + g_i \tag{3.18}$$

that are used instead of (3.13). Representations for W_i and g_i are given in Lemma 3.2. In the second step we look at the global system

$$K_h \begin{bmatrix} x_0 \\ \vdots \\ x_N \end{bmatrix} = g_h \tag{3.19}$$

representing the continuity conditions (3.18), the boundary condition (3.14) and the consistency condition (3.15) (see (3.22) for the definition of K_h, g_h). Its solvability is examined in Lemma 3.3.

Setting $E_{1j} := E_1(t_{ij})$, $A_{1j} := A_1(t_{ij})$, $A_{2j} := A_2(s_{ij})$, $f_{1j} := f_1(t_{ij})$ and $f_{2j} := f_2(s_{ij})$ for selected fixed i, the local systems (3.16) are given by

$$B_i := \left[\begin{array}{c|c|c|c} \frac{v_{11}}{h_i} E_{11} - u_{11} A_{11} & \frac{v_{12}}{h_i} E_{11} - u_{12} A_{11} & \cdots & \frac{v_{1k}}{h_i} E_{11} - u_{1k} A_{11} \\ -A_{21} & 0 & & 0 \\ \hline \frac{v_{21}}{h_i} E_{12} - u_{21} A_{12} & & & \vdots \\ 0 & & & \vdots \\ \hline \vdots & & & \vdots \\ \vdots & & & \vdots \\ \hline \frac{v_{k1}}{h_i} E_{1k} - u_{k1} A_{1k} & & \cdots & \frac{v_{kk}}{h_i} E_{1k} - u_{kk} A_{1k} \\ 0 & & & -A_{2k} \end{array}\right]$$

$$\in \mathbb{R}^{kn \times kn}\,,$$

$$a_i := \left[\begin{array}{c} -\frac{v_{10}}{h_i}E_{11} + u_{10}A_{11} \\ 0 \\ \hline \vdots \\ \hline -\frac{v_{k0}}{h_i}E_{1k} + u_{k0}A_{1k} \\ 0 \end{array}\right] \in \mathbb{R}^{kn\times n}\,, \; b_i := \left[\begin{array}{c} f_{11} \\ f_{21} \\ \hline \vdots \\ \hline f_{1k} \\ f_{2k} \end{array}\right] \in \mathbb{R}^{kn}\,.$$

In the following lemma we prove the regularity of B_i for sufficiently small h_i using multiplications from the left and from the right, respectively, with

$$T_P := \operatorname{diag}\left(\begin{bmatrix} P_{11}(t_{ij}) & P_{12}(s_{ij}) \\ 0 & P_{22}(s_{ij}) \end{bmatrix}\right)_{j=1,\dots,k}, \; T_Q := \operatorname{diag}\Big(Q(s_{ij})\Big)_{j=1,\dots,k}$$

where P, Q transform the DAE into canonical form (2.2). We also need reordering of the rows and columns done by multiplication with

$$U_k := \begin{bmatrix} I_d & 0 & & & 0 & 0 & & \\ 0 & 0 & & & I_a & 0 & & \\ & I_d & & & & 0 & & \\ & 0 & & & & I_a & & \\ & & \ddots & & & & \ddots & \\ & & & I_d & & & & 0 \\ & & & 0 & & & & I_a \end{bmatrix} \in \mathbb{R}^{kn\times kn}\,. \tag{3.20}$$

Lemma 3.1 *Let the smoothness assumptions* $A_2, P \in C^1$, $Q \in C^2$ *be fulfilled. Define*

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

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

$$\begin{bmatrix} G_{lm}^1 & G_{lm}^2 \end{bmatrix} := \begin{cases} \Big(v_{ll}(\sigma_l - \rho_l) - 1\Big)(P_{11}E_1\dot{Q})(t_{il}) \\ \qquad -\Big(u_{ll} - 1\Big)(P_{11}A_1Q)(t_{il}) + \mathcal{O}(h_i)\,, & l = m\,, \\ v_{lm}(\sigma_m - \rho_l)(P_{11}E_1\dot{Q})(t_{il}) \\ \qquad -u_{lm}(P_{11}A_1Q)(t_{il}) + \mathcal{O}(h_i)\,, & l \neq m\,. \end{cases}$$

Then the representation

$$B_i = T_P^{-1}\, U_k \begin{bmatrix} h_i V^{-1} \otimes I & 0 \\ 0 & -I \end{bmatrix}^{-1} \Big(I + \varDelta_i\Big)\, U_k^*\, T_Q^{-1}$$

holds, and for sufficiently small h_i *the matrix* B_i *is regular with*

$$B_i^{-1} = T_Q\, U_k\, \Big(I - \varDelta_i + \mathcal{O}(h_i^2)\Big) \begin{bmatrix} h_i V^{-1} \otimes I & 0 \\ 0 & -I \end{bmatrix} U_k^*\, T_P\,.$$

Proof. With $A_2, P \in C^1$, $Q \in C^2$ we can expand

$$Q(s_{im}) = Q(t_{il}) + \mathcal{O}(h_i) = Q(t_{il}) + h_i(\sigma_m - \rho_l)\dot{Q}(t_{il}) + \mathcal{O}(h_i^2)\,,$$

$$(P_{12}A_2Q)(s_{il}) = (P_{12}A_2Q)(t_{il}) + \mathcal{O}(h_i) = (P_{11}E_1\dot{Q} - P_{11}A_1Q)(t_{il}) + \mathcal{O}(h_i).$$

This leads to

$$\begin{aligned}
&\begin{bmatrix} P_{11}(t_{il}) \; P_{12}(s_{il}) \end{bmatrix} \begin{bmatrix} \frac{v_{lm}}{h_i} E_{1l} - u_{lm} A_{1l} \\ 0 \end{bmatrix} Q(s_{im}) \\
&= \tfrac{v_{lm}}{h_i}(P_{11}E_1)(t_{il})Q(s_{im}) - u_{lm}(P_{11}A_1)(t_{il})Q(s_{im}) \\
&= \tfrac{v_{lm}}{h_i}(P_{11}E_1Q)(t_{il}) + v_{lm}(\sigma_m - \rho_l)(P_{11}E_1\dot{Q})(t_{il}) \\
&\qquad\qquad -u_{lm}(P_{11}A_1Q)(t_{il}) + \mathcal{O}(h_i) \\
&= \tfrac{v_{lm}}{h_i}\begin{bmatrix} I \; 0 \end{bmatrix} + \begin{bmatrix} G^1_{lm} \; G^2_{lm} \end{bmatrix} \text{ for } m \neq l\,.
\end{aligned}$$

Analogously, we get

$$\begin{aligned}
&\begin{bmatrix} P_{11}(t_{il}) \; P_{12}(s_{il}) \end{bmatrix} \begin{bmatrix} \frac{v_{ll}}{h_i} E_{1l} - u_{ll} A_{1l} \\ -A_{2l} \end{bmatrix} Q(s_{il}) \\
&= \tfrac{v_{ll}}{h_i}(P_{11}E_1)(t_{il})Q(s_{il}) - u_{ll}(P_{11}A_1)(t_{il})Q(s_{il}) - (P_{12}A_2Q)(s_{il}) \\
&= \tfrac{v_{ll}}{h_i}(P_{11}E_1Q)(t_{il}) + v_{ll}(\sigma_l - \rho_l)(P_{11}E_1\dot{Q})(t_{il}) \\
&\qquad -u_{ll}(P_{11}A_1Q)(t_{il}) - (P_{11}E_1\dot{Q} - P_{11}A_1Q)(t_{il}) + \mathcal{O}(h_i) \\
&= \tfrac{v_{ll}}{h_i}\begin{bmatrix} I \; 0 \end{bmatrix} + \begin{bmatrix} G^1_{ll} \; G^2_{ll} \end{bmatrix} \text{ for } m = l\,.
\end{aligned}$$

By multiplication of B_i with T_P from the left and T_Q from the right and reordering of the rows and columns using U_k we obtain

$$U_k^* \, T_P \, B_i \, T_Q \, U_k = \begin{bmatrix} \frac{1}{h_i} V \otimes I & 0 \\ 0 & -I \end{bmatrix} + \begin{bmatrix} G^1 & G^2 \\ 0 & 0 \end{bmatrix}$$

with $G^s := \left(G^s_{lm}\right)_{l,m}$. Since V is regular with $V^{-1} = (w_{jl})_{j,l}$ we have

$$\text{(3.21)} \qquad \begin{bmatrix} h_i V^{-1} \otimes I & 0 \\ 0 & -I \end{bmatrix} U_k^* \, T_P \, B_i \, T_Q \, U_k = I + \Delta_i$$

with Δ_i as given above. Multiplication with the inverses yields the representation of B_i.

Since (for all l, m and $s = 1, 2$) G^s_{lm} is bounded for $h_i \to 0$, we have $\|\Delta_i\| = \mathcal{O}(h_i)$. Thus $I + \Delta_i$ is regular for sufficiently small h_i and has the inverse $(I + \Delta_i)^{-1} = I - \Delta_i + \mathcal{O}(h_i^2)$. By this and (3.21) we see that B_i is regular for sufficiently small h_i and that B_i^{-1} has the given representation. □

Lemma 3.2 *If a transformation to canonical form with $Q \in C^2$ is possible then the following representations for W_i, g_i defined in (3.17) hold:*

$$W_i = Q(t_{i+1})\begin{bmatrix} I-F_{i1} & -F_{i2} \\ 0 & 0 \end{bmatrix} Q(t_i)^{-1} \text{ with } F_{i1} = \mathcal{O}(h_i^2), F_{i2} = \mathcal{O}(h_i),$$
$$g_i = Q(t_{i+1}) \begin{bmatrix} c_i \\ -(P_{22}f_2)(t_{i+1}) \end{bmatrix} \text{ with } c_i = \mathcal{O}(h_i) \,.$$

Proof. Using the representation of B_i^{-1} given in Lemma 3.1 we compute $W_iQ(t_i) = [\,0 \ldots 0\, I\,]B_i^{-1}a_iQ(t_i)$.
With $Q(t_i) = Q(t_{il}) + \mathcal{O}(h_i) = Q(t_{il}) - \rho_l h_i \dot{Q}(t_{il}) + \mathcal{O}(h_i^2)$ we have

$$\begin{aligned} &\left[\, P_{11}(t_{il}) \; P_{12}(s_{il}) \,\right] \begin{bmatrix} -\frac{v_{l0}}{h_i} E_{1l} + u_{l0}A_{1l} \\ 0 \end{bmatrix} Q(t_i) \\ =\;& -\tfrac{v_{l0}}{h_i}(P_{11}E_1Q)(t_{il}) + v_{l0}\rho_l(P_{11}E_1\dot{Q})(t_{il}) + u_{l0}(P_{11}A_1Q)(t_{il}) + \mathcal{O}(h_i) \\ =\;& -\tfrac{v_{l0}}{h_i}\left[\, I \; 0 \,\right] - \left[\, G_{l0}^1 \; G_{l0}^2 \,\right], \end{aligned}$$

hence

$$U_k^* \, T_P \, a_i \, Q(t_i) = -\tfrac{1}{h_i} \begin{bmatrix} v_0 \otimes I & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} G_0^1 & G_0^2 \\ 0 & 0 \end{bmatrix}$$

with $v_0 := (v_{l0})_{l=1,\ldots,k}$ and $G_0^s := (G_{l0}^s)_{l=1,\ldots,k}$ for $s = 1, 2$.
By considering $v_0 = -V \left[\, 1 \; \cdots \; 1 \,\right]^*$, $\Delta_{j0}^s := h_i \sum_{l=1}^k w_{jl}G_{l0}^s = \mathcal{O}(h_i)$ as in Lemma 3.1 and defining

$$\tilde{I} := \begin{bmatrix} (I)_{j=1,\ldots,k} & 0 \\ 0 & 0 \end{bmatrix}, \quad \tilde{\Delta}_i := \begin{bmatrix} (\Delta_{j0}^1)_{j=1,\ldots,k} & (\Delta_{j0}^2)_{j=1,\ldots,k} \\ 0 & 0 \end{bmatrix},$$

this leads to

$$\begin{bmatrix} h_iV^{-1} \otimes I & 0 \\ 0 & -I \end{bmatrix} U_k^* \, T_P a_i Q(t_i) = \tilde{I} - \tilde{\Delta}_i \,.$$

Applying the next factor of the representation of B_i^{-1}, we get

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

$$= \begin{bmatrix} I & 0 \\ \vdots & \vdots \\ I & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} \Delta_{10}^1 & \Delta_{10}^2 \\ \vdots & \vdots \\ \Delta_{k0}^1 & \Delta_{k0}^2 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} \sum \Delta_{1m}^1 & 0 \\ \vdots & \vdots \\ \sum \Delta_{km}^1 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \mathcal{O}(h_i^2) & \mathcal{O}(h_i^2) \\ \vdots & \vdots \\ \mathcal{O}(h_i^2) & \mathcal{O}(h_i^2) \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} * & * \\ \vdots & \vdots \\ * & * \\ I-F_{i1} & -F_{i2} \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}$$

with $F_{i1} := \sum_{m=0}^{k} \Delta^1_{km} + \mathcal{O}(h_i^2)$, $F_{i2} := \Delta^2_{k0} + \mathcal{O}(h_i^2)$.
Altogether this yields

$$W_i Q(t_i) = [\,0 \ldots 0\, I\,] B_i^{-1} a_i Q(t_i) = [\,0 \ldots 0\, I\,] \mathrm{diag}\Big(Q(s_{ij})\Big) U_k \Theta_i$$

and hence (since $s_{ik} = t_{i+1}$)

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

In order to show that $F_{i1} = \mathcal{O}(h_i^2)$, we use interpolation of the polynomials $p(t) = 1$, $q(t) = t$ at the points $\sigma_0, \ldots, \sigma_k$ to obtain

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

By inserting the definitions of Lemma 3.1 we see

$$\sum_{m=0}^{k} \Delta^1_{km} = \sum_{m=0}^{k} h_i \sum_{l=1}^{k} w_{kl} G^1_{lm}$$

$$= h_i \sum_{l=1}^{k} w_{kl} \Bigg(\sum_{\substack{m=0\\ m\neq l}}^{k} \Big[v_{lm}(\sigma_m - \rho_l)(P_{11}E_1\dot{Q}_1)(t_{il}) - u_{lm}(P_{11}A_1Q_1)(t_{il}) \Big]$$

$$+ \Big(v_{ll}(\sigma_l - \rho_l) - 1 \Big)(P_{11}E_1\dot{Q}_1)(t_{il}) - (u_{ll} - 1)(P_{11}A_1Q_1)(t_{il}) + \mathcal{O}(h_i) \Bigg)$$

$$= h_i \sum_{l=1}^{k} w_{kl} \Bigg(\underbrace{\Bigg[\sum_{m=0}^{k} L'_m(\rho_l)(\sigma_m - \rho_l) - 1 \Bigg]}_{=0} (P_{11}E_1\dot{Q}_1)(t_{il})$$

$$- \underbrace{\Bigg[\sum_{m=0}^{k} L_m(\rho_l) - 1 \Bigg]}_{=0} (P_{11}A_1Q_1)(t_{il}) + \mathcal{O}(h_i) \Bigg) = \mathcal{O}(h_i^2)$$

and therefore

$$F_{i1} = \sum_{m=0}^{k} \Delta^1_{km} + \mathcal{O}(h_i^2) = \mathcal{O}(h_i^2)\ .$$

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

$$g_i = Q(t_{i+1}) \begin{bmatrix} c_i \\ -(P_{22}f_2)(t_{i+1}) \end{bmatrix} \text{ with } c_i = \mathcal{O}(h_i)$$

can be derived analogously by inserting the representation for B_i^{-1} given in Lemma 3.1 into $g_i = \begin{bmatrix} 0 \cdots 0 \, I \end{bmatrix} B_i^{-1} b_i$. $\square$

The global system (3.19) is given by $K_h \in \mathbb{R}^{(N+1)n \times (N+1)n}$ and $g_h \in \mathbb{R}^{(N+1)n}$, where

$$
(3.22) \quad K_h := \begin{bmatrix} C & & & & D \\ -A_2(t_0) & & & & 0 \\ W_0 & -I & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & W_{N-1} & -I \end{bmatrix}, \quad g_h := \begin{bmatrix} r \\ f_2(t_0) \\ -g_0 \\ \vdots \\ \vdots \\ -g_{N-1} \end{bmatrix}.
$$

To prove the regularity of K_h and the boundedness of $K_h^{-1} g_h$, we multiply from the left and from the right, respectively, with

$$
T_l := \operatorname{diag}\left(\begin{bmatrix} I & 0 \\ 0 & P_{22}(t_0) \end{bmatrix}, Q(t_1)^{-1}, \ldots, Q(t_N)^{-1} \right), T_r := \operatorname{diag}\Big(Q(t_i)\Big),
$$

where P, Q transform (E, A) to canonical form (2.2). We also use $U_N \in \mathbb{R}^{(N+1)n \times (N+1)n}$, which is defined analogously to U_k in (3.20), to reorder rows and columns. Finally, we set

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

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

with $C_{11}, C_{12}, D_{11}, D_{12}$ given in (2.3) and F_{i1}, F_{i2} given in Lemma 3.2, and $A_h := \begin{bmatrix} M_h & N_h \\ 0 & -I \end{bmatrix}$, $\Delta_h := \begin{bmatrix} D_h & 0 \\ 0 & 0 \end{bmatrix}$.

Lemma 3.3 *The matrix K_h of the global system (3.19) given in (3.22) has the representation*

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

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 E, A, f, C, D, r and the transformation functions P, Q, but not on the maximum mesh width h.

Proof. By multiplication with T_l from the left and T_r from the right we get block-wise

$$\begin{bmatrix} I & 0 \\ 0 & P_{22}(t_0) \end{bmatrix} \begin{bmatrix} C \\ -A_2(t_0) \end{bmatrix} Q(t_0) = \begin{bmatrix} C_{11} & C_{12} \\ 0 & -I \end{bmatrix},$$
$$\begin{bmatrix} I & 0 \\ 0 & P_{22}(t_0) \end{bmatrix} \begin{bmatrix} D \\ 0 \end{bmatrix} Q(t_N) = \begin{bmatrix} D_{11} & D_{12} \\ 0 & 0 \end{bmatrix},$$
$$Q(t_{i+1})^{-1} W_i Q(t_i) = \begin{bmatrix} I - F_{i1} & -F_{i2} \\ 0 & 0 \end{bmatrix},$$

if we use the representation of W_i given in Lemma 3.2. Reordering of the rows and columns yields

$$U_N^* \, T_l \, K_h \, T_r \, U_N = A_h + \Delta_h \, ,$$

and by multiplying with the inverses we get the representation of K_h.

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

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

Using Lemma 3.2, it follows that

$$\|M_h^{-1} D_h\| \le \|S^{-1}\| \max\{\|C_{11}\|, \|D_{11}\|\} \cdot \sum_{i=0}^{N-1} \underbrace{\|F_{i1}\|}_{=\mathcal{O}(h_i^2)} = \mathcal{O}(h).$$

Since M_h is regular, the same holds for A_h. We obtain $\|A_h^{-1} \Delta_h\| = \|M_h^{-1} D_h\| = \mathcal{O}(h)$, thus $A_h + \Delta_h$ is regular for sufficiently small h and

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

This proves the regularity of K_h and the representation of K_h^{-1}.

Using

$$g_i = Q(t_{i+1}) \begin{bmatrix} c_i \\ -(P_{22} f_2)(t_{i+1}) \end{bmatrix}, \quad c_i = \mathcal{O}(h_i), \quad F_{i2} = \mathcal{O}(h_i)$$

(see Lemma 3.2) together with the representations of K_h^{-1} and M_h^{-1}, the boundedness of $K_h^{-1}g_h$ independent of h follows along the lines of the proof of Lemma 3.3 in [16]. □

The existence and uniqueness of solutions of collocation problems (3.5)–(3.8) is equivalent to the unique solvability of the local systems (3.16) and the global system (3.19). Thus existence and uniqueness follows by combining Lemma 3.1 (concerning the local systems) and Lemma 3.3 (concerning the global system). For smooth data, i. e., $E, A \in C^2$, the existence of a transformation to canonical form with $P \in C^1, Q \in C^2$ is guaranteed by Proposition 2.1.

Theorem 3.1 *Consider a uniquely solvable BVP (2.1),(1.2) with smooth data $E, A \in C^2$, $f \in C$. For $N \in \mathbb{N}$ and $k \geq 1$ define a mesh π as in (3.1) and for $i = 0, \ldots, N-1$ collocation points $t_{ij}, j = 1, \ldots, k$ as in (3.3) and $s_{ij}, j = 0, \ldots, k$ as in (3.4), respectively, according to knots ρ_j, σ_j as in (3.2).*

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 the collocation conditions (3.5),(3.6), fulfills the boundary condition (3.8) and is consistent at all mesh points t_i.

A collocation method is said to be stable, if the approximations x_i, x_{ij} remain bounded (independent of π) for decreasing mesh widths h_i (see, e. g., [1]). In this sense, the symmetric collocation methods (3.11)–(3.15) are stable, since the x_i are bounded (see Lemma 3.3) and the x_{ij} satisfy the relation

$$x_{ij} = \left(Q(s_{ij}) \begin{bmatrix} I - F_{ij1} & -F_{ij2} \\ 0 & 0 \end{bmatrix} Q(t_i)^{-1} \right) x_i + Q(s_{ij}) \begin{bmatrix} c_{ij} \\ -(P_{22}f_2)(s_{ij}) \end{bmatrix},$$

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

3.2 Convergence results

In this section we examine the collocation methods concerning convergence. Assuming a smooth solution of the BVP, we prove convergence of order k and for special schemes order $k+1$ together with superconvergence of order $2k$ at mesh points.

Theorem 3.2 *Consider a uniquely solvable BVP (2.1),(1.2) with a smooth solution $x \in C^{k+1}(\mathbb{I}, \mathbb{R}^n)$. Let π be a mesh as in (3.1) with sufficiently small mesh widths h_i and use schemes ρ_j, σ_j as in (3.2). Let x_π be the* *unique solution of the corresponding symmetric collocation method. Then we have*

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

Proof. Interpolation of x analogous to (3.9) yields

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

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

$$B_i \begin{bmatrix} x(s_{i1}) \\ \vdots \\ x(s_{ik}) \end{bmatrix} = a_i x(t_i) + b_i - \begin{bmatrix} \tau_{i1} \\ \vdots \\ \tau_{ik} \end{bmatrix}, \quad \tau_{ij} := \begin{bmatrix} (E_1\dot\psi_i - A_1\psi_i)(t_{ij}) \\ 0 \end{bmatrix}$$

with B_i, a_i, b_i defined in (3.16). Obviously we have $\psi_i(t_{ij}) = \mathcal{O}(h_i^{k+1})$ and $\dot\psi_i(t_{ij}) = \mathcal{O}(h_i^k)$, thus $\tau_{ij} = \mathcal{O}(h_i^k)$.
Since the collocation problem is uniquely solvable for sufficiently small h_i (i. e., B_i is regular), we can solve for $x(s_{ik}) = x(t_{i+1})$. We get that (with W_i, g_i defined in (3.17))

$$x(t_{i+1}) = W_i x(t_i) + g_i - \tau_i\,.$$

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 given in Lemma 3.2. The continuity, boundary and consistency conditions for x lead to the global system (comparable to (3.19))

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

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

$$K_h \begin{bmatrix} x(t_0) - x_0 \\ \vdots \\ x(t_N) - x_N \end{bmatrix} = \tau_h \,. \tag{3.23}$$

Due to $\tau_h = \mathcal{O}(h^{k+1})$ 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, using order $k+1$ instead of $g_i = \mathcal{O}(h_i)$), i. e.,

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

Looking at the difference in the local systems we obtain

$$\begin{bmatrix} x(s_{i1}) - x_{i1} \\ \vdots \\ x(s_{ik}) - x_{ik} \end{bmatrix} = B_i^{-1} a_i \underbrace{\Big(x(t_i) - x_i\Big)}_{=\mathcal{O}(h^k)} - B_i^{-1} \underbrace{\begin{bmatrix} \tau_{i1} \\ \vdots \\ \tau_{ik} \end{bmatrix}}_{=\mathcal{O}(h_i^k)} \tag{3.24}$$

and hence $\max_j \|x(s_{ij}) - x_{ij}\| = \mathcal{O}(h^k)$.
From this the convergence order k for any $t \in \mathbb{I}$ can be derived easily by looking at the differences of the interpolation representations for x and x_π, respectively. □

For special choices of the schemes in (3.2), this result can be improved to a higher convergence order at mesh points t_i, so-called superconvergence.

Theorem 3.3 *Consider a BVP (2.1),(1.2) with unique solution x. Let π be a mesh as in (3.1). Use Gauß knots $0 < \rho_0 < \ldots < \rho_k < 1$ and Lobatto knots $0 = \sigma_0 < \ldots < \sigma_k = 1$ to construct the collocation points t_{ij}, s_{ij}. Suppose furthermore that the mesh widths h_i are sufficiently small, such that the corresponding symmetric collocation method has a unique solution x_π.*

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

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

Proof. By Proposition 2.1, there exist $P \in C^{2k}, Q \in C^{2k+1}$ transforming the DAE to canonical form (2.2). Since x_i is consistent, the initial value problem $E\dot{y} = Ay + f$, $y(t_i) = x_i$ is uniquely solvable and the solution v has a representation (using the transformation (2.2) to canonical form)

$$(Q^{-1}v)(t) = \begin{bmatrix} [\, I \;\, 0 \,] \Big(Q(t_i)^{-1}x_i + \int_{t_i}^{t} (Pf)(s)ds\Big) \\ -(P_{22}f_2)(t) \end{bmatrix}, \quad t \ge t_i\,.$$

The approximation x_π is the solution of the initial value problem $E\dot{y} = Ay + (E\dot{x}_\pi - Ax_\pi)$, $y(t_i) = x_i$, and has the form

$$(Q^{-1}x_\pi)(t) = \begin{bmatrix} [\, I \; 0 \,] \left(Q(t_i)^{-1}x_i + \int_{t_i}^{\mathrm{t}} (P(E\dot{x}_\pi - Ax_\pi))(s)ds \right) \\ (P_{22}A_2x_\pi)(t) \end{bmatrix}$$

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

$$v(t_{i+1}) - x_{i+1} = Q(t_{i+1}) \begin{bmatrix} \int_{t_i}^{t_{i+1}} \phi_d(s)ds + \int_{t_i}^{t_{i+1}} \phi_a(s)ds \\ 0 \end{bmatrix}, \tag{3.25}$$

with functions

$$\phi_d := P_{11}(f_1 - E_1\dot{x}_\pi + A_1x_\pi)\,, \qquad \phi_a := P_{12}(f_2 + A_2x_\pi)\,.$$

Due to the smoothness of the data, we have $\phi_d, \phi_a \in C^{2k}$. Since x_π satisfies the collocation conditions, the collocation points $t_{i1}, \ldots, t_{ik}$ are zeros of ϕ_d and $s_{i0}, \ldots, s_{ik}$ are zeros of ϕ_a, respectively. From this follows (see, e. g., [15]) the existence of smooth functions $w_d \in C^k, w_a \in C^{k-1}$ with

$$\phi_d(s) = w_d(s)\prod_{j=1}^{k}(s - t_{ij})\,, \qquad \phi_a(s) = w_a(s)\prod_{j=0}^{k}(s - s_{ij})\,.$$

Taylor expansion yields $w_d = \psi_d + \mathcal{O}(h_i^k)$, $w_a = \psi_a + \mathcal{O}(h_i^{k-1})$ with polynomials ψ_d of degree $\le k-1$ and ψ_a of degree $\le k-2$, respectively. By inserting this into (3.25) and using the orthogonality properties of the Gauß and Lobatto schemes (see, e. g., [9, Ch. IV]), we obtain

$$\begin{aligned} \int_{t_i}^{t_{i+1}} \phi_d(s)ds &= \int_{t_i}^{t_{i+1}} \left[\psi_d(s)\prod_{j=1}^{k}(s - t_{ij}) + \mathcal{O}(h_i^{2k})\right]ds \\ &= h_i^{k+1} \underbrace{\int_0^1 \psi_d(t_i + h_i\tau)\prod_{j=1}^{k}(\tau - \rho_j)d\tau}_{=0} + \mathcal{O}(h_i^{2k+1})\,, \\ \int_{t_i}^{t_{i+1}} \phi_a(s)ds &= \int_{t_i}^{t_{i+1}} \left[\psi_a(s)\prod_{j=0}^{k}(s - s_{ij}) + \mathcal{O}(h_i^{2k})\right]ds \\ &= h_i^{k+2} \underbrace{\int_0^1 \psi_a(t_i + h_i\tau)\prod_{j=0}^{k}(\tau - \sigma_j)d\tau}_{=0} + \mathcal{O}(h_i^{2k+1})\,. \end{aligned}$$

Altogether we have

$$\phi_i := v(t_{i+1}) - x_{i+1} = Q(t_{i+1}) \begin{bmatrix} \int_{t_i}^{t_{i+1}} \phi_d(s)ds + \int_{t_i}^{t_{i+1}} \phi_a(s)ds \\ 0 \end{bmatrix}$$
$$= \mathcal{O}(h_i^{2k+1}) .$$

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

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

we see that $x(t) - v(t) = W(t, t_i)(x(t_i) - v(t_i))$ for all $t \geq t_i$. Setting $t = t_{i+1}$, we particularly get

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

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

$$\begin{bmatrix} C & & & D \\ -A_2(t_0) & & & 0 \\ W(t_1, t_0) & -I & & \\ & \ddots & \ddots & \\ & & W(t_N, t_{N-1}) & -I \end{bmatrix} \begin{bmatrix} x(t_0) - x_0 \\ \vdots \\ x(t_N) - x_N \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -\phi_0 \\ \vdots \\ -\phi_{N-1} \end{bmatrix}$$

comparable to (3.23). From this we derive (as in Lemma 3.3)

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

since now the inhomogeneity is of order $\mathcal{O}(h^{2k+1})$. □

To show a higher convergence order for a special choice of the schemes, we need a simple lemma.

Lemma 3.4 *For Gauß knots* $0 < \rho_1 < \ldots < \rho_k < 1$ *and Lobatto knots* $0 = \sigma_0 < \ldots < \sigma_k = 1$ *we have*

$$\int_0^{\sigma_j} \prod_{l=1}^{k} (\tau - \rho_l) d\tau = 0 , \quad j = 0, \ldots, k .$$

Proof. The Gauß and Lobatto knots are defined via the zeros of the Legendre polynomials and their derivatives, respectively. The claim follows directly from the Legendre differential equation, see, e. g., [9, Ch. IV]. □

Corollary 3.1 *Under the assumptions of Theorem 3.3 it follows that*

$$\max_j \|x(s_{ij}) - x_{ij}\| = \mathcal{O}(h_i^{k+2}) + \mathcal{O}(h^{2k}) \quad \textit{for } k \geq 2$$

and

$$\|x - x_\pi\|_\infty = \mathcal{O}(h^{k+1})\,.$$

Proof. Looking at (3.24) in the proof of Theorem 3.2 and using $x(t_i) - x_i = \mathcal{O}(h^{2k})$ due to Theorem 3.3, it is obvious that we must show $B_i^{-1}(\tau_{ij})_j = \mathcal{O}(h_i^{k+2})$ to prove the first assertion. For this we exploit the special choice of the knots.

The transformation to canonical form yields

$$\begin{aligned} P_{11}(E_1\dot{\psi}_i - A_1\psi_i) &= (P_{11}E_1Q)\tfrac{\mathrm{d}}{\mathrm{dt}}(Q^{-1}\psi_i) - (P_{11}A_1Q - P_{11}E_1\dot{Q})(Q^{-1}\psi_i) \\ &= [\,I\;0\,]\tfrac{\mathrm{d}}{\mathrm{dt}}(Q^{-1}\psi_i) + (P_{12}A_2Q)(Q^{-1}\psi_i) \\ &= \dot{\varphi} + \mathcal{O}(h_i^{k+1})\,, \end{aligned}$$

when defining $\varphi := [\,I\;0\,](Q^{-1}\psi_i)$. For smooth data $E, A \in C^{2k+1}$, $f \in C^{2k}$ we get a smooth solution $x \in C^{2k}$ (see Proposition 2.1), thus the interpolation error ψ_i is smooth. Since $Q \in C^{2k+1}$ by Proposition 2.1, it follows that $\varphi \in C^{2k}$, in particular $\varphi \in C^{k+2}$ for $k \geq 2$. By interpolation of $\dot{\varphi}$ at the points t_{il} and by a Taylor expansion of the interpolation error we obtain

$$\begin{aligned} \sum_{l=1}^{k} \tilde{L}_l\Big(\frac{t - t_i}{h_i}\Big)\dot{\varphi}(t_{il}) &= \dot{\varphi}(t) - \frac{\dot{\varphi}^{(k)}(\theta(t))}{k!}\prod_{l=1}^{k}(t - t_{il}) \\ &= \dot{\varphi}(t) - c\prod_{l=1}^{k}(t - t_{il}) + \mathcal{O}(h_i^{k+1}) \end{aligned}$$

with the constant $c := \frac{1}{k!}\varphi^{(k+1)}(t_i)$ and Lagrange polynomials $\tilde{L}_l$ as in (3.10). Inserting the definition of w_{jl} given in (3.10) leads to

$$\begin{aligned} \sum_{l=1}^{k} w_{jl}\dot{\varphi}(t_{il}) &= \int_0^{\sigma_j}\sum_{l=1}^{k}\tilde{L}_l(\tau)\dot{\varphi}(t_{il})d\tau = \tfrac{1}{h_i}\int_{t_i}^{s_{ij}}\sum_{l=1}^{k}\tilde{L}_l\Big(\frac{t - t_i}{h_i}\Big)\dot{\varphi}(t_{il})dt \\ &= \tfrac{1}{h_i}\int_{t_i}^{s_{ij}}\dot{\varphi}(t)dt - \tfrac{c}{h_i}\int_{t_i}^{s_{ij}}\prod_{l=1}^{k}(t - t_{il})dt + \mathcal{O}(h_i^{k+1}) \\ &= \frac{\varphi(s_{ij}) - \varphi(t_i)}{h_i} - ch_i^k\int_0^{\sigma_j}\prod_{l=1}^{k}(\tau - \rho_l)d\tau + \mathcal{O}(h_i^{k+1}) \\ &= \mathcal{O}(h_i^{k+1})\,, \end{aligned}$$

since $s_{ij}, t_i = s_{i0}$ are zeros of ψ_i and thus of φ, and the second term is zero by Lemma 3.4, respectively. Altogether we have (recalling $V^{-1} = (w_{jl})_{j,l}$)

$$U_k^* T_P\Big(\tau_{ij}\Big)_j = \begin{bmatrix} \Big([P_{11}(E_1\dot{\psi}_i - A_1\psi_i)](t_{ij})\Big)_j \\ 0 \end{bmatrix} = \begin{bmatrix} \Big(\dot{\varphi}(t_{ij}) + \mathcal{O}(h_i^{k+1})\Big)_j \\ 0 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} h_i V^{-1} \otimes I & 0 \\ 0 & -I \end{bmatrix} U_k^* T_P\Big(\tau_{ij}\Big)_j = h_i \begin{bmatrix} \Big(\sum_{l=1}^k w_{jl}\dot{\varphi}(t_{il})\Big)_j \\ 0 \end{bmatrix} + \mathcal{O}(h_i^{k+2}) = \mathcal{O}(h_i^{k+2})$$

$$\Rightarrow B_i^{-1}\Big(\tau_{ij}\Big)_j = T_Q U_k\Big(I - \Delta_i + \mathcal{O}(h_i^2)\Big) \begin{bmatrix} h_i V^{-1} \otimes I & 0 \\ 0 & -I \end{bmatrix} U_k^* T_P\Big(\tau_{ij}\Big)_j = \mathcal{O}(h_i^{k+2})\,.$$

The convergence order $k+1$ for any $t \in \mathbb{I}$ follows now by considering the difference of the interpolation representations for x and x_π (cp. end of proof for Theorem 3.2). □

4 Collocation with interpolation

A drawback of the symmetric methods may be the number of evaluations of the data E, A, f needed to construct the matrices B_i. Since we have two schemes ρ_j, σ_j and two sets of collocation points t_{ij}, s_{ij}, we need $2Nk+1$ evaluations instead of only $Nk+1$ for conventional collocation.

To overcome this drawback, we can, for smooth $E, A, f \in C^{k+1}$, interpolate the data using the collocation points s_{ij}:

$$E_1(t) = \underbrace{\sum_{m=0}^{k} L_m\left(\frac{t-t_i}{h_i}\right) E_1(s_{im})}_{=:p_E(t)} + \underbrace{\frac{E_1^{(k+1)}(\theta(t))}{(k+1)!}\prod_{m=0}^{k}(t-s_{im})}_{=:\psi_E(t)}$$

and $A_1 = p_A + \psi_A$, $f_1 = p_f + \psi_f$ analogously. If we replace $E_1(t_{ij})$, $A_1(t_{ij})$, $f_1(t_{ij})$ by $p_E(t_{ij})$, $p_A(t_{ij})$, $p_f(t_{ij})$ in the collocation condition (3.11), we obtain the following problem (with $i = 0, \dots, N-1$ and $j = 1, \dots, k$), for which data evaluations at the points s_{ij} are sufficient:

$$\sum_{l=0}^{k}\left[\frac{v_{jl}}{h_i}\sum_{m=0}^{k} u_{jm}E_1(s_{im}) - u_{jl}\sum_{m=0}^{k} u_{jm}A_1(s_{im})\right]\tilde{x}_{il}$$

$$= \sum_{m=0}^{k} u_{jm}\ f_1\ (s_{im}) \tag{4.1}$$

$$-A_2(s_{ij})\tilde{x}_{ij} = f_2(s_{ij}) \tag{4.2}$$

$$\tilde{x}_{ik} - \tilde{x}_{i+1,0} = 0 \tag{4.3}$$

$$C\tilde{x}_{00} + D\tilde{x}_{N0} = r \tag{4.4}$$

$$-A_2(t_0)\tilde{x}_{00} = f_2(t_0) \tag{4.5}$$

For this problem we prove results analogous to Theorem 3.1 (unique solvability), Theorem 3.2 (convergence order k) and Theorem 3.3 (superconvergence of order $2k$).

Theorem 4.1 *Consider a uniquely solvable BVP (2.1),(1.2) with solution x and smooth data $E, A \in C^{k+2}$, $f \in C^{k+1}$, $k \geq 1$. For $N \in \mathbb{N}$ define a mesh π as in (3.1) and collocation points s_{ij} for $i = 0, \dots, N-1$, $j = 0, \dots, k$ as in (3.4) according to knots σ_j as in (3.2). Use knots ρ_j as in (3.3) to compute $v_{jm} = L'_m(\rho_j)$ and $u_{jm} = L_m(\rho_j)$ (see (3.9) for definition of L_m).*

i) *For sufficiently small mesh widths $h_0, \dots, h_{N-1}$, there exists one and only one continuous piecewise polynomial $\tilde{x}_\pi$ of degree k that satisfies the interpolated collocation conditions (4.1), the collocation conditions (4.2), fulfills the boundary condition (4.4) and is consistent at all mesh points t_i.*

ii) *If the mesh widths are sufficiently small, the symmetric collocation method using interpolation is of convergence order k, i. e.,*

$$\|x - \tilde{x}_\pi\| = \mathcal{O}(h^k)\ .$$

iii) *If we use Lobatto knots $0 = \sigma_0 < \ldots < \sigma_k = 1$ and Gauß knots $0 < \rho_1 < \ldots < \rho_k < 1$ and if the data fulfills the smoothness conditions $E, A \in C^{2k+1}$, $f \in C^{2k}$, then the symmetric collocation method using interpolation is superconvergent of order $2k$, i. e.,*

$$\max_{0 \leq i \leq N} \|x(t_i) - \tilde{x}_i\| = \mathcal{O}(h^{2k})$$

for sufficiently small h.

Proof. As in Sect. 3, we start by considering local systems

$$\tilde{B}_i \begin{bmatrix} \tilde{x}_{i1} \\ \vdots \\ \tilde{x}_{ik} \end{bmatrix} = \tilde{a}_i \tilde{x}_{i0} + \tilde{b}_i$$

built of the collocation conditions (4.1),(4.2) (for $j = 1, \dots, k$). Due to the interpolation errors $\psi_{E,A,f}(t_{ij}) = \mathcal{O}(h_i^{k+1})$ we have

$$\tilde{B}_i = B_i + \mathcal{O}(h_i^k)\,, \quad \tilde{a}_i = a_i + \mathcal{O}(h_i^k)\,, \quad \tilde{b}_i = b_i + \mathcal{O}(h_i^{k+1})$$

with B_i, a_i, b_i of the local system (3.16). Applying Lemma 3.1, we see that $\tilde{B}_i$ is regular for sufficiently small h_i and $\tilde{B}_i^{-1} = B_i^{-1} + \mathcal{O}(h_i^{k+2})$ since $B_i^{-1} = \mathcal{O}(h_i)$. This yields continuity conditions

$$\tilde{x}_{i+1,0} = \tilde{x}_{ik} = \tilde{W}_i \tilde{x}_{i0} + \tilde{g}_i$$

with $\tilde{W}_i = W_i + \mathcal{O}(h_i^{k+1})$, $\tilde{g}_i = g_i + \mathcal{O}(h_i^{k+2})$. Thus we get a global system

$$\tilde{K}_h \begin{bmatrix} \tilde{x}_0 \\ \vdots \\ \tilde{x}_N \end{bmatrix} = \tilde{g}_h$$

with $\tilde{K}_h = K_h + \mathcal{O}(h^{k+1})$, $\tilde{g}_h = g_h + \mathcal{O}(h^{k+2})$ and K_h, g_h of the global system (3.19). Here we apply Lemma 3.3 to achieve that $\tilde{K}_h$ is regular for sufficiently small h with

$$\tilde{K}_h^{-1} = \Big(K_h(I + \mathcal{O}(h^k))\Big)^{-1} = (I + \mathcal{O}(h^k))K_h^{-1}\,.$$

From this it follows that

$$\tilde{K}_h^{-1}\tilde{g}_h = (I + \mathcal{O}(h^k))K_h^{-1}(g_h + \mathcal{O}(h^{k+2})) = K_h^{-1}g_h + \mathcal{O}(h^k)$$

is bounded independent of h, because the same holds for $K_h^{-1}g_h$. Since the unique solvability of the collocation problem with interpolation is equivalent to the regularity of $\tilde{B}_i$ ($i = 0, \dots, N-1$) and $\tilde{K}_h$, assertion i) is proved.

Convergence order k can be proved as in Theorem 3.2.

To prove superconvergence we argue analogously to the proof of Theorem 3.3. Here we define three functions

$$\tilde{\phi}_d := P_{11}(p_f - p_E\dot{\tilde{x}}_\pi + p_A\tilde{x}_\pi)\,, \qquad \tilde{\phi}_a := P_{12}(f_2 + A_2\tilde{x}_\pi)\,,$$
$$\tilde{\phi}_\psi := P_{11}(\psi_f - \psi_E\dot{\tilde{x}}_\pi + \psi_A\tilde{x}_\pi)$$

and obtain a local discretisation error

$$v(t_{i+1}) - \tilde{x}_{i+1} = Q(t_{i+1}) \begin{bmatrix} \int_{t_i}^{t_{i+1}} \tilde{\phi}_d(s)ds + \int_{t_i}^{t_{i+1}} \tilde{\phi}_a(s)ds + \int_{t_i}^{t_{i+1}} \tilde{\phi}_\psi(s)ds \\ 0 \end{bmatrix}.$$

Due to the collocation conditions, $\tilde{\phi}_d$ has zeros t_{ij} and $\tilde{\phi}_a$ has zeros s_{ij}, respectively. The s_{ij} are also zeros of $\tilde{\phi}_\psi$, since they are zeros of the interpolation errors. □

5 Numerical examples

To illustrate the practicability and effectiveness of the described symmetric collocation methods we present three representative examples. The results are compared to that of Radau collocation [16] and COLDAE [3].

A MATLAB code for the construction and solution of local systems (3.16) and global systems (3.19) has been developed, including a simple strategy for the generation and refinement of the meshes π. The package DGELDA [14] is used for the regularisation of the data E, A, f at discrete points t_{ij}, s_{ij}, thus FORTRAN subroutines for the evaluation of E, A, f and its derivatives up to order $\nu - 1$ at discrete points are needed. Furthermore, the data $\underline{t}, \bar{t}, C, D, r$ are needed as input, and the parameter $1 \leq k \leq 5$ and a tolerance for the mesh selection must be chosen.

As discussed in Sect. 4, the symmetric methods need $2Nk+1$ evaluations of the data E, A, f instead of $Nk + 1$ for Radau collocation, since two sets t_{ij}, s_{ij} of collocation points are used. Besides this, the computational effort is the same for symmetric and Radau collocation, respectively, because the local and global systems have the same dimensions and structures. If data evaluations are expensive, we can apply collocation with interpolation (i. e., we solve (4.1)–(4.5)). For the following three examples, we report only the results of symmetric collocation without interpolation (i. e., solutions of (3.11)–(3.15)), since we obtained comparably accurate results when we worked with interpolation.

Example 5.1 In order to demonstrate the potential drawbacks of the asymmetric Radau methods, we consider the ordinary boundary value problem ([1], p. 394)

$$\varepsilon u''(t) = -2tu'(t)\,, \quad u(-1) = -1\,,\, u(1) = 1$$

with small parameter $0 < \varepsilon \ll 1$. The solution is $u(t) = \mathrm{erf}(t/\sqrt{\varepsilon})$, where the Gaussian error function is defined by

$$\mathrm{erf}(t) = \frac{2}{\sqrt{\pi}} \int_0^t e^{-s^2}\, ds.$$

With Radau collocation, we can compute approximations only for moderate values of ε, i. e., $\varepsilon \geq 10^{-3}$. For $\varepsilon = 10^{-3}$, Fig. 1 shows the errors $u(t_i) - u_\pi(t_i)$ of Radau and symmetric collocation, respectively, according to $k = 5$ collocation points per subinterval, five subintervals in the initial meshes and a tolerance 10^{-4} for the mesh refinement. While the mesh that is generated by the Radau method is much coarser in the right subinterval $[0, 1]$ than in the left half $[-1, 0]$, the result of the symmetric collocation method is a symmetric mesh and a symmetric approximation.

For $\varepsilon = 10^{-4}, 10^{-5}, 10^{-6}$, the Radau method failed, but we got approximations by use of symmetric collocation or COLDAE.

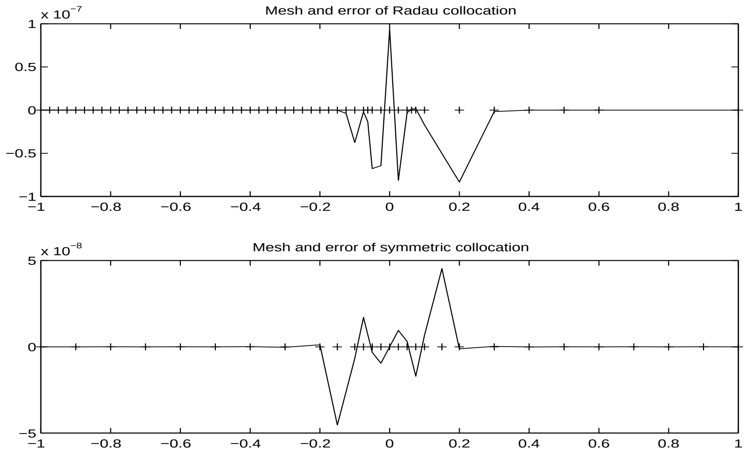

Fig. 1.

Example 5.2 The second example is

$$\begin{bmatrix} 0 & 0 & 0 \\ 1 & -t & 0 \\ -1 & t & 1 \end{bmatrix} \dot{x} = \begin{bmatrix} -1 & t & 0 \\ 0 & 0 & 0 \\ 0 & t^2 & 1 \end{bmatrix} x + \begin{bmatrix} \mathrm{e}^{t/2} \\ 0 \\ 0 \end{bmatrix}, \; t \in [-5, 0]$$

$$\begin{bmatrix} 1 \; 7 \; 0 \end{bmatrix} x(-5) + \begin{bmatrix} 0 \; 4 \; 1 \end{bmatrix} x(0) = 6 \,.$$

This is an index-two problem with $d = 1$ differential and $a = 2$ algebraic equations. The solution is

$$x(t) = \mathrm{e}^{t/2}(1 - \tfrac{t}{2}\,,\, -\tfrac{1}{2}\,,\, t^2 + 4t + 8)^* \,.$$

For $k = 1, ..., 5$ collocation points per subinterval and uniform meshes with appropriate numbers N of subintervals, we computed approximations using symmetric collocation, the Radau method and COLDAE.

Since this index-two problem is not semi-explicit, COLDAE can not be applied directly. The index reduction technique due to [11] is used to obtain an index-one formulation. But this is not semi-explicit either, thus we need to transform it into the semi-explicit index-two problem

$$\dot{x} = y\,, \quad 0 = \hat{E}y - \hat{A}x - \hat{f}\,,$$

which is of doubled dimension. Furthermore, the consistency condition $\hat{A}_2(t_0)x(t_0) + \hat{f}_2(t_0) = 0$ at $t_0 = -5$ must be considered as an additional boundary condition. In other words, this problem can not be attacked by COLDAE without applying the index reduction and even by doing this, more computational work in comparison to Radau or symmetric collocation is needed.

Table 1. Errors according to uniform meshes for Example 4.2

		Symmetric Coll.		Radau Collocation		COLDAE	
k	N	err_i	order	err_i	order	err_i	order
1	50	0.26e-2		0.17		0.28e-2	
	100	0.65e-3	2.0	0.82e-1	1.0	0.71e-3	2.0
	200	0.16e-3	2.0	0.41e-1	1.0	0.18e-3	2.0
2	20	0.16e-4		0.74e-3		0.18e-4	
	40	0.10e-5	4.0	0.90e-4	3.0	0.11e-5	4.0
	80	0.64e-7	4.0	0.11e-4	3.0	0.71e-7	4.0
3	10	0.39e-6		0.13e-4		0.44e-6	
	20	0.61e-8	6.0	0.38e-6	5.1	0.68e-8	6.0
	40	0.95e-10	6.0	0.12e-7	5.0	0.11e-9	6.0
4	6	0.17e-7		0.43e-6		0.19e-7	
	12	0.68e-10	8.0	0.34e-8	7.0	0.77e-10	7.9
	24	0.26e-12	8.0	0.26e-10	7.0	0.29e-12	8.0
5	4	0.13e-8		0.28e-7		0.14e-8	
	8	0.12e-11	10.0	0.50e-10	9.1	0.14e-11	10.0

In Table 1 the errors $err_i(N) := \max_{0\leq i\leq N} \|x(t_i) - x_i\|$ and the corresponding orders $\log\left(\frac{err_i(N/2)}{err_i(N)}\right) / \log(2)$ are given. We clearly see that the theoretical superconvergence results ($2k$ for symmetric collocation and COLDAE, $2k-1$ for the Radau method) can be verified for this example. We also recognize that not only the orders but also the absolute values err_i are approximately the same for symmetric collocation and COLDAE, while the results of the Radau method are less accurate.

Example 5.3 For the third example we transform a DAE given in [3, Example 1] and obtain

$$E(t) = \begin{bmatrix} 1 & -t & 0 \\ t & 1 & -t \\ p(t)-2 & -t(p(t)-2) & 0 \end{bmatrix},$$

$$A(t) = \begin{bmatrix} \kappa - \frac{1}{2-t} & \frac{2}{2-t} - \kappa t & (2-t)\kappa \\ \frac{\kappa-1}{2-t} - t - 1 & -t\frac{\kappa-1}{2-t} - 1 & t + \kappa - \frac{\kappa p(t)}{2+t} \\ \kappa t(t^2-3) - \frac{p(t)-2}{2-t} & 2\frac{p(t)-2}{2-t} - 4\kappa & \kappa\Big(p(t)(2-t) - t^3 + 6t - 4\Big) \end{bmatrix},$$

$$f(t) = \begin{bmatrix} \frac{3-t}{2-t} \\ 2 + \frac{(\kappa+2)p(t)+\dot{p}(t)}{t^2-4} - 2\frac{tp(t)}{(t^2-4)^2} \\ (p(t)-2)\frac{3-t}{2-t} - \kappa(t^2+t-2) \end{bmatrix} \mathrm{e}^t,$$

with $t \in [0,1]$, parameter $\kappa \in \mathbb{R}$ and a smooth function $p \in C^1(\mathbb{I},\mathbb{R})$. The boundary condition is $x_1(0) = 1$.

This problem is of index two and consists of $d = 1$ differential and $a = 2$ algebraic equations. We set $\kappa = 20$ and choose

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

Thus a layer region around $t = \frac{1}{3}$ occurs in p and also in the solution

$$x(t) = \tfrac{\mathrm{e}^t}{t^2+1} \begin{bmatrix} 1 + t - \frac{t^2}{2-t} + \frac{tp(t)}{t^2-4} \\ 1 - t - \frac{t}{2-t} + \frac{p(t)}{t^2-4} \\ -\frac{t^2+1}{2-t} \end{bmatrix}.$$

We examine this problem using $k = 4$ collocation points per subinterval and five subintervals in the initial meshes. The tolerances for mesh refinement are chosen such that comparable numbers N of subintervals in the final meshes occur. In Table 2 we report these numbers N together with the errors $err := \max \|x(t) - x_\pi(t)\|$ measured at 101 equidistant points $t \in \mathbb{I}$.

Table 2. Errors for Example 5.3

Symmetric Collocation			Radau Collocation			COLDAE		
tol	N	err	tol	N	err	tol	N	err
$3 \cdot 10^{-9}$	69	0.83e-8	10^{-7}	68	0.67e-7	10^{-4}	66	0.11e-5
10^{-11}	161	0.21e-9	10^{-9}	144	0.10e-8	10^{-5}	160	0.40e-6

As in Example 5.2, COLDAE cannot be applied to this problem directly. It has to be regularised and the reduced BVP must be transformed into a semi-explicit index-two problem of doubled dimension. Thus the application of COLDAE is more expensive regarding the computational work. Moreover, the results of COLDAE are less accurate when we compare approximations obtained with similar numbers of subintervals.

6 Conclusions

In this paper, we have developed symmetric collocation methods for the solution of linear differential-algebraic boundary value problems as they occur by index reduction. Thus, in combination with index reduction, we can solve BVPs of arbitrary index. The key point was to use a Gauß-type scheme for the differential part and a Lobatto-type scheme with one more knot for the algebraic part. We showed that the results known for differential equations also hold in the case of differential-algebraic equations including superconvergence for the combination of Gauß and Lobatto schemes. In order to reduce the number of function evaluations that are needed when using two different schemes, we introduced interpolation and showed that the convergence properties are not influenced by this modification. Finally, we showed the applicability and accuracy of these methods in comparison to other approaches.

References

1. Ascher, U., Mattheij, R., Russell, R. (1995): Numerical solution of boundary value problems for ordinary differential equations. SIAM, Philadelphia, second edition
2. Ascher, U., Petzold, L.R. (1992): Projected collocation for higher-order higher-index differential-algebraic equations. J. Comp. Appl. Math. **43**, 243–259
3. Ascher, U., Spiteri, R. (1994): Collocation software for boundary value differential-algebraic equations. SIAM J. Sci. Comput. **15**, 938–952
4. Bai, Y. (1991): A perturbed collocation method for boundary value problems in differential-algebraic equations. Appl. Math. Comput. **45**, 269–291
5. Bai, Y. (1992): A modified Lobatto collocation for linear boundary value problems of differential-algebraic equations. Computing **49**, 139–150
6. Brenan, K.E., Campbell, S.L., Petzold, L.R. (1989): Numerical solution of initial-value problems in differential-algebraic equations. North-Holland, New York: Elsevier
7. Degenhardt, A. (1992): Collocation for transferable differential-algebraic equations. Semin.ber., Humboldt-Univ. Berlin, Fachbereich Mathematik 92-1, 83–104
8. Griepentrog, E., März, R. (1986): Differential-algebraic equations and their numerical treatment. Leipzig: Teubner Verlag
9. Hairer, E., Wanner, G. (1991): Solving Ordinary Differential Equations II — Stiff and Differential-Algebraic Problems. Berlin: Springer
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): A new class of discretization methods for the solution of linear differential-algebraic equations. SIAM J. Numer. Anal. **33**, 1941–1961
12. Kunkel, P., Mehrmann, V. (1996): Generalized inverses of differential-algebraic operators. SIAM J. Matrix Anal. Appl. **17**, 426–442
13. Kunkel, P., Mehrmann, V. (1996): Local and global invariants of linear differential-algebraic equations and their relations. Electron. Trans. Numer. Anal. **4**, 138–157
14. 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
15. Stöver, R. (1999): Numerische Lösung von linearen differential-algebraischen Randwertproblemen. Doctoral thesis, Universität Bremen
16. Stöver, R. (2001): Collocation methods for solving linear differential-algebraic boundary value problems. Numer. Math. **88**, 771–795