Least-Squares Collocation for Higher Index Differential-Algebraic Equations

Michael Hanke[a,*], Roswitha März[b], Caren Tischendorf[b], Ewa Weinmüller[c], Stefan Wurm[c]

[a] *KTH Royal Institute of Technology, School of Engineering Sciences, Department of Mathematics, S-100 44 Stockholm, Sweden*
[b] *Humboldt University of Berlin, Institute of Mathematics, D-10099 Berlin, Germany*
[c] *Vienna University of Technology, Institute for Analysis and Scientific Computing, A-1040 Wien, Austria*

Abstract

Differential-algebraic equations with higher index give rise to essentially ill-posed problems. Therefore, their numerical approximation requires special care. In the present paper, we state the notion of ill-posedness for differential-algebraic equations more precisely. Based on this property, we construct a regularization procedure using a least-squares collocation approach by discretizing the pre-image space. Numerical experiments show that the resulting method has excellent convergence properties and is not much more computationally expensive than standard collocation methods used in the numerical solution of ordinary differential equations or index-1 differential-algebraic equations. Convergence is shown for a limited class of higher index differential-algebraic equations.

Keywords: differential-algebraic equation, higher index, essentially ill-posed problem, collocation, boundary value problem

1. Introduction

In the present paper, we consider boundary value problems (BVPs) for linear differential-algebraic equations (DAEs)

$$A(t)(Dx)'(t) + B(t)x(t) = q(t), \quad t \in [a,b], \tag{1}$$

$$G_a x(a) + G_b x(b) = \gamma. \tag{2}$$

Here, $[a,b]$ denotes a finite interval, $q : [a,b] \to \mathbb{R}^m$ is a vector-valued function, $B : [a,b] \to \mathbb{R}^{m\times m}$, $A : [a,b] \to \mathbb{R}^{m\times k}$ are at least continuous but sufficiently smooth matrix-valued functions, and $D : [a,b] \to\in \mathbb{R}^{k\times m}$ is at least continuously differentiable.

*Corresponding author
Email addresses: hanke@nada.kth.se (Michael Hanke), maerz@math.hu-berlin.de (Roswitha März), caren@math.hu-berlin.de (Caren Tischendorf), ewa.weinmueller@tuwien.ac.at (Ewa Weinmüller), stefan.wurm@tuwien.ac.at (Stefan Wurm)

Moreover, $G_a, G_b \in \mathbb{R}^{l\times m}$ and $\gamma \in \mathbb{R}^l$, and l is the dynamical degree of freedom of the DAE. We are interested in solutions $x : [a, b] \to \mathbb{R}^m$ satisfying (1)–(2) in a sense which will be defined later.
Here, for clarity of the presentation, we focus on DAEs featuring variables partitioned into differentiated and algebraic components by assuming a constant matrix function D of the special form,

$$D = [\, I \;\; 0\,], \quad \operatorname{rank} D = k. \tag{3}$$

In particular, this is the case for all semi-explicit DAEs. The first k components of the unknown function x are the *differentiated* components and the subsequent $m - k$ components, the first derivatives of which are not involved, are traditionally called the *algebraic* components. We refer to Subsection 6.1 for more general DAEs.

Collocation methods using piecewise polynomial ansatz functions are well-established and robust numerical methods to approximate BVPs in explicit ordinary differential equations and index-1 DAEs, which are well-posed in their natural Banach spaces, see [1, 15] for the respective comprehensive surveys.

There are different possibilities on how to make the collocation ansatz. Below we describe only one of these possibilities which we address later on and which is, for instance, implemented in COLDAE [2]. Basically, as an ansatz for the differentiated components, we use continuous piecewise polynomial functions of a certain degree and for the algebraic components, generally discontinuous piecewise polynomial functions, whose degree is lower by one.

Let $n \in \mathbb{N}$ and consider the following partition of the interval $[a, b]$:

$$a = t_0 < t_1 < \cdots < t_n = b.$$

For $K \geq 0$, let us denote by $\mathcal{P}_K$ the set of all polynomials of degree less or equal to K.

We fix a certain integer $N \geq 1$ and approximate the differentiated solution components $x_1, \ldots, x_k$ by continuous, piecewise polynomial functions of degree N with possible breakpoints at $t_1, \ldots, t_{n-1}$, while we approximate the algebraic components $x_{k+1}, \ldots, x_m$ by possibly discontinuous piecewise polynomial functions of degree $N-1$ with possible jumps at $t_1, \ldots, t_{n-1}$. Consequently, we search for the numerical approximation x_n in the function set X_n,

$$\begin{aligned} X_n = \{p \in L^2(a,b)^m :\ & p_\kappa \in C[a,b],\ \ p_\kappa|_{[t_{j-1},t_j)} \in \mathcal{P}_N,\ \kappa = 1,\ldots,k,\ \ j = 1,\ldots,n, \\ & p_\kappa|_{[t_{j-1},t_j)} \in \mathcal{P}_{N-1},\ \kappa = k+1,\ldots,m,\ \ j = 1,\ldots,n\}. \end{aligned} \tag{4}$$

By construction, $p \in X_n$ implies $Dp \in C[a, b]^k$. Since X_n has dimension $Nmn + k$, $Nmn + k$ conditions are necessary to uniquely determine $p \in X_n$.
The collocation points t_{ji} are specified by choosing N values

$$0 < \rho_1 < \cdots < \rho_N < 1,$$

and setting $t_{ji} := t_{j-1} + \rho_i h_j$, $j = 1, \ldots, n$, $i = 1, \ldots, N$, where $h_j = t_j - t_{j-1}$. This choice excludes Lobatto and Radau collocation points. Note that in COLDAE, Gaussian points are used [2].
In order to determine the discrete solution $p \in X_n$, the classical collocation method is applied directly to the DAE system,

$$A(t_{ji})(Dp)'(t_{ji}) + B(t_{ji})p(t_{ji}) = q(t_{ji}), \quad i = 1, \ldots, N, j = 1, \ldots, n \tag{5}$$

$$G_a p(a) + G_b p(b) = \gamma. \tag{6}$$

It follows immediately that (5) – (6) consists of $Nmn + l$ conditions to determine x_n. For index-1 DAEs, we have $l = k$ and therefore, the above BVP is well-balanced. In case of higher index DAEs, l is less than k and for a balanced system, extra boundary conditions have to be prescribed.

This kind of collocation directly applied to the given BVP works well for index-1 DAEs and also for a limited class of index-2 DAEs via *projected collocation* [2, 15]. Here, we are interested in a direct treatment of general higher index DAEs, without any preliminary and incorporated index reduction procedures as applied, for instance, in [13, 21].

We now report on experiments with direct collocation for two simple academic examples of index 2 and index 3. Both examples are known to cause serious difficulties in the numerical integration depending on the involved parameters. All computations have been carried out in MATLAB.[1] The ansatz polynomials for the $\kappa-$th component of p on the subinterval $[t_{j-1}, t_j)$ have been represented as

$$p_\kappa|_{[t_{j-1},t_j)}(t) = z_{n,\kappa,0} + h \sum_{s=1}^{N} z_{n,\kappa,s} \bar{\psi}_s \left(\frac{t - t_{j-1}}{h} \right), \ \kappa = 1, \ldots, k,$$

$$p_\kappa|_{[t_{j-1},t_j)}(t) = \sum_{s=1}^{N} z_{n,\kappa,s} \psi_s \left(\frac{t - t_{j-1}}{h} \right), \ \kappa = k+1, \ldots, m,$$

in which

$$\psi_s(\tau) = \prod_{\lambda=1,\ldots,N,\, \lambda \neq s} \frac{\tau - \rho_\lambda}{\rho_s - \rho_\lambda}, \quad \tau \in [0,1], \quad s = 1, \ldots, N,$$

$$\bar{\psi}_s(\tau) = \int_0^\tau \psi_s(\sigma) \mathrm{d}\sigma, \quad \tau \in [0,1], \quad s = 1, \ldots, N.$$

Example 1.1. *Consider the DAE system*

$$\begin{aligned} x_1'(t) + \lambda x_1(t) - x_2(t) - x_3(t) &= q_1(t), \\ x_2'(t) + (\eta t(1 - \eta t) - \eta) x_1(t) + \lambda x_2(t) - \eta t x_3(t) &= q_2(t), \\ (1 - \eta t) x_1(t) + x_2(t) &= q_3(t), \quad t \in [0,1], \end{aligned}$$

[1] MATLAB Release 2013b, The MathWorks, Inc., Natick, Massachusetts, United States.

with right hand side q chosen in such a way that

$$x_1(t) = e^{-t}\sin t,$$
$$x_2(t) = e^{-2t}\sin t,$$
$$x_3(t) = e^{-t}\cos t,$$

is a solution. This DAE arises from problem [11, (5.1)] *for $\beta = 1$. It is a semi-explicit system in Hessenberg form with index 2 and dynamical degree of freedom $l = 1$ for all parameter values. Hence, one boundary condition is appropriate. We choose it to be*

$$x_1(0) = 0.$$

The DAE fits formally into the form (1)–(2) with

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix},\ D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix},\ B = \begin{bmatrix} \lambda & -1 & -1 \\ \eta t(1-\eta t) - \eta & \lambda & -\eta t \\ 1 - \eta t & 1 & 0 \end{bmatrix}.$$

Because of $k = 2$ and $l = 1$, the collocation system becomes underdetermined. In order to obtain a unique solution we add the following consistent initial condition:

$$x_2(0) = 0.$$

□

Example 1.2. *We address the DAE system*

$$\begin{aligned} x_2'(t) + x_1(t) &= q_1(t), \\ t\eta x_2'(t) + x_3'(t) + (\eta+1)x_2(t) &= q_2(t), \\ t\eta x_2(t) + x_3(t) &= q_3(t), \quad t \in [0,1]. \end{aligned}$$

It can be cast into the form (1)–(2) by setting

$$A = \begin{bmatrix} 1 & 0 \\ t\eta & 1 \\ 0 & 0 \end{bmatrix},\ D = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},\ B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1+\eta & 0 \\ 0 & t\eta & 1 \end{bmatrix},$$

where a simple permutation of the variables results in the required form of D, cf. [17, p. 168]. *In* [17] *this problem is used to illustrate the failure of backward differentiation formulas, see also* [14, Section 8.3]. *As in the previous example, the exact solution is chosen to be*

$$x_1(t) = e^{-t}\sin t,$$
$$x_2(t) = e^{-2t}\sin t,$$
$$x_3(t) = e^{-t}\cos t,$$

and this determines q. This DAE has index 3 and the dynamical degree of freedom $l = 0$ for all η. This means that the solution is uniquely defined without any boundary conditions. In order to have a unique solution for the collocation system, the conditions

$$x_2(0) = 0, \quad x_3(0) = 1$$

were posed. □

Table 1: Error of the collocation solution for Example 1.1 with $\eta = -25$, $\lambda = -1$, and $N = 4$. The collocation points ρ_i are chosen to be Gauss-Legendre nodes. Note that x_3 is the algebraic variable

n	$\|x_3 - p_3\|_\infty$
20	4.67e+6
40	8.62e+3
80	5.26e+2
160	5.61e+1
320	6.74e+0
640	8.74e–1

.

Table 2: Error of the collocation solution for Example 1.2 with $\eta = -2$ and $N = 3$. The collocation points are distributed uniformly and x_1 is the algebraic component

n	$\|x_1 - p_1\|_\infty$
20	3.74e+006
40	9.84e+016
80	3.51e+038
160	2.04e+082
320	2.98e+170
640	3.06e+307

.

Typical results of the direct collocation are provided in Tables 1 and 2. Obviously, this classical collocation method is useless in the context of higher index DAEs. In Example 1.1 the error is large given the smooth exact solution, while in Example 1.2 the error grows unboundedly. Figure 1 shows the error of x_3 from Example 1.1; an exponential error growth can be observed.

For all higher index DAEs, the dynamical degree of freedom l is less than k, and in order to ensure a balanced discrete system extra conditions are necessary. This gives rise to the question if it wouldn't be reasonable to augment the system by additional collocation equations. It would be in the spirit of collocation methods if, instead of artificial boundary conditions, additional collocation points were introduced by enlarging the set of ρ_i's. Clearly, this idea results in an overdetermined and not necessarily consistent discrete problem which we propose to solve in a least-squares sense.

Let the nodes $0 < \rho_1 < \cdots < \rho_N < 1$ be given as specified above. We now add *extra collocation points* σ_i, $i = 1, \ldots, N + 1$,

$$\sigma_i = \begin{cases} \rho_1/2, & i = 1, \\ (\rho_{i-1} + \rho_i)/2, & 2 \le i \le N, \\ (\rho_N + 1)/2, & i = N + 1, \end{cases}$$

and compute $p \in X_n$ as least-squares solution of the overdetermined discrete system of

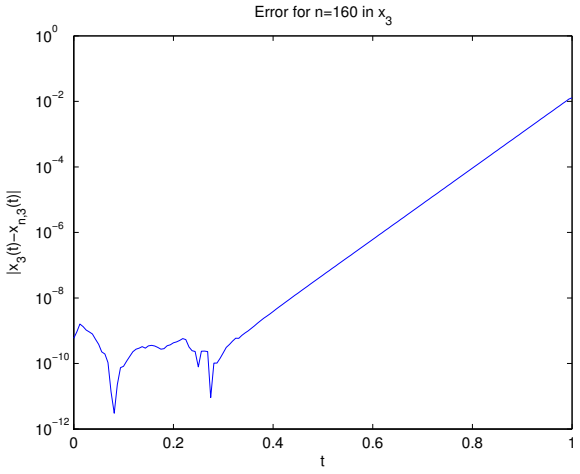

Figure 1: Example 1.1: Error plot of x_3 with $n = 160$

.

Table 3: Collocation results for Example 1.1. The details correspond to those from Table 1

n	Standard: $\|x_3 - p_3\|_\infty$	Least-squares:$\|x_3 - p_3\|_\infty$
20	4.67e+6	4.67e–07
40	8.62e+3	6.91e–08
80	5.26e+2	7.72e–09
160	5.61e+1	9.79e–10
320	6.74e+0	2.47e–10
640	8.74e–1	8.47e–10

$Nnm + l + (N + 1)nm$ equations

$$A(t)(Dp)'(t) + B(t)p(t) = q(t), \quad t \in S_j, \ j = 1, \ldots, n,$$
$$G_a p(a) + G_b p(b) = \gamma,$$

where S_j is the extended set of collocation points,

$$S_j = \{t_{j-1} + \rho_i h_j, \ i = 1, \ldots, N\} \cup \{t_{j-1} + \sigma_i h_j, \ i = 1, \ldots, N + 1\}. \tag{7}$$

Typical results of the overdetermined least-squares collocation are shown in Tables 3 and 4. Now, we observe a convergent behavior of the method. In particular, in both cases the order of convergence is approximately $N - 1$. Figure 2 shows the error of x_3 in Example 1.1.

The numerical experiments presented here, immediately suggest the following questions: What is the reason for the excellent behavior of this least-squares collocation?

Table 4: Collocation results for Example 1.2. The details correspond to those from Table 2

n	Standard: $\|x_1 - p_1\|_\infty$	Least-squares:$\|x_1 - p_1\|_\infty$
20	3.74e+006	3.26e–4
40	9.84e+016	7.52e–5
80	3.51e+038	1.81e–5
160	2.04e+082	4.42e–6
320	2.98e+170	1.11e–6
640	3.06e+307	1.06e–6

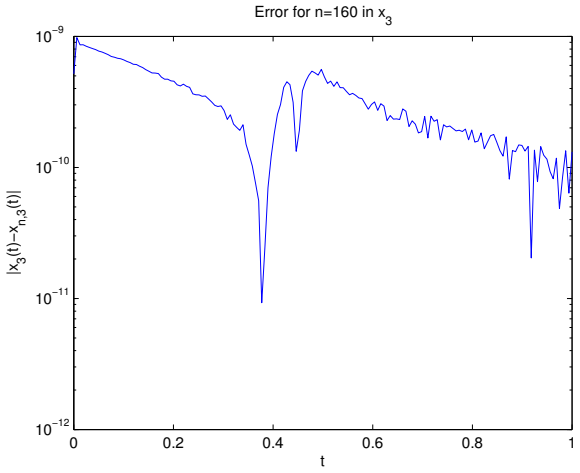

Figure 2: Example 1.1: Error plot of x_3 with $n = 160$ using least-squares variant of the collocation method

Which further methods can become efficient approaches in context of higher index DAEs? Which classes of problems can be treated by them?
At the time being, we are able to provide only very first and incomplete answers to these questions and we hope that the present paper will encourage specialists in the area to deal with this matter.

BVPs for higher index DAEs are known to be *essentially ill-posed* in naturally given topologies ([9, 10, 16], see also [14, 15, 18]). This is the case even when the boundary conditions are accurately stated. Therefore, we try to answer the questions by imbedding the least-squares collocation into the context of regularization methods for ill-posed problems. This seems to be a promising idea, since collocation serves as a well-known regularization approach. Generally, there are two ways to proceed:

- Least-squares collocation by discretization of the image space;
- Least-squares approach for discretized pre-image spaces.

Least-squares collocation for DAEs by discretization of the image space has been investigated already in [9]. Conditions for convergence were derived via the theory of reproducing kernel Hilbert spaces and a practical implementation was presented. However, this method turned out to be rather inefficient. In view of the observations in our experiments, we try to pursue the second approach following ideas from [12]. To our knowledge, the method for higher index DAEs discussed here has not been considered before. In Subsection 6.2, we mention a few different approaches being related to our method in the broadest sense.

The paper is organized as follows. In Section 2, we provide the basic functional analytic background. Essential ingredients for proving convergence of the regularization methods are estimates of certain (in-)stability constants. In Section 3 we succeed to derive such thresholds for DAEs of a special structure but it seems that these results can be improved and generalized.
Section 4 contains convergence estimations for the proposed least-squares collocation method. We report on further numerical experiments in Section 5. Additional references and conclusions are collected in Section 6. For reader's convenience, useful facts on Legendre polynomials are recapitulated in Appendix.

2. Functional Analytic Background

In this section we consider BVPs (1) – (2), but now, instead of assuming the special condition (3) we only require that the DAE has a properly stated leading term, that is,

$$\ker A(t) \oplus \operatorname{im} D(t) = \mathbb{R}^k, \quad t \in [a, b].$$

2.1. Differential-Algebraic Operators

Let $x : [a, b] \to \mathbb{R}^m$ be a vector-valued function such that the product Dx is differentiable in the generalized sense. Then, we define the *differential-algebraic operator* (DA operator) T acting on such functions by

$$(Tx)(t) = A(t)(D(t)x(t))' + B(t)x(t), \quad \text{a.e. } t \in [a, b].$$

For a well-defined operator T, we must specify the image space Y, the pre-image space X, as well as the domain of definition $\operatorname{dom} T$. Since in the regularization context, we are interested in a Hilbert space setting, two natural choices are:

- $X = L^2(a,b)^m,\ Y = L^2(a,b)^m,\ \operatorname{dom} T = \{x \in X : Dx \in H^1(a,b)^k\}$,
- $X = H^1_D(a,b) = \{x \in L^2(a,b)^m : Dx \in H^1(a,b)^k\} = \operatorname{dom} T$, $Y = L^2(a,b)^m$.

In the first case, the linear operator T is densely defined and closed [18] while it is bounded in the second case, which is more convenient in our context. The function space $H^1_D(a,b)$ equipped with its natural inner product is a Hilbert space [18, Lemma 6.9].

Next, we adopt the notion of the tractability index of an DA operator from [18, Section 4.2]. This is reasonable for both settings since this notion is tied to the coefficients A, B, D only.

Definition 2.1. *The DA operator $T : X \to Y$ is said to be fine with tractability index $\mu \in \mathbb{N}$ and characteristic values*

$$r_0 \le \cdots \le r_{\mu-1} < r_\mu = m, \quad l = m - \sum_{i=0}^{\mu-1}(m - r_i), \tag{8}$$

if there is an admissible matrix function sequence with (8) and the coefficients A, B, D are as smooth as required for the existence of completely decoupling projectors.

Let us by Π_{can} denote the canonical projector function of the associated DAE. The projector $\Pi_{can}(t)$ acts in $\mathbb{R}^m$ and its rank is l for all $t \in [a,b]$, where l is the dynamical degree of freedom of the associated DAE.

Note that for constant coefficients A, B, D the operator T is fine, exactly if the matrix pencil $\{AD, B\}$ is regular. Then the tractability index coincides with the Kronecker index and the characteristic values describe the structure of the Weierstraß–Kronecker form. Moreover, Π_{can} represents the spectral projector of the pencil onto the eigenspace corresponding to the finite eigenvalues along the one corresponding to the infinite eigenvalues.

Theorem 2.2. *Let the bounded DA operator $T : H^1_D(a,b) \to L^2(a,b)^m$ be fine with index $\mu \in \mathbb{N}$ and characteristic values (8). Then the following statements hold:*

(1) $\ker T$ *has finite dimension,* $\dim \ker T = l = \operatorname{rank} \Pi_{can}$.

(2) T *is surjective, thus Fredholm, exactly if* $\mu = 1$.

(3) *If* $\mu > 1$, *then* $\operatorname{im} T$ *is a nonclosed, proper subset of* $L^2(a,b)^m$.

(4) *If* $\mu > 1$ *and the coefficients A, B, D are smooth enough, then the inclusion* $C^\infty[a,b]^m \subset \operatorname{im} T$ *holds, so that T is densely solvable.*

Proof. The assertions (1) – (4) can be verified in the same way as the corresponding parts of [18, Theorem 4.2] by replacing the Banach space setting by the Hilbert space setting. Also, keep in mind that we now deal with generalized derivatives and that the respective equations are satisfied a.e. in $[a, b]$. □

Example 2.3 (Continuation of Example 1.1)**.** *The image of the operator T,*

$$(Tx)(t) = \begin{bmatrix} x_1'(t) + \lambda x_1(t) - x_2(t) - x_3(t) \\ x_2'(t) + (\eta t(1-\eta t) - \eta)x_1(t) + \lambda x_2(t) - \eta t x_3(t) \\ (1-\eta t)x_1(t) + x_2(t) \end{bmatrix}, \quad \text{a.e. } t \in (0,1),$$

defined on $H_D^1(0,1) = \{x_1, x_2 \in H^1(0,1),\ x_3 \in L^2(0,1)\}$ becomes

$$\operatorname{im} T = \{q \in L^2(0,1)^3 : q_3 \in H^1(0,1)\}.$$

□

Example 2.4 (Continuation of Example 1.2)**.** *The image of the operator T,*

$$(Tx)(t) = \begin{bmatrix} x_2'(t) + x_1(t) \\ t\eta x_2'(t) + x_3'(t) + (\eta+1)x_2(t) \\ \eta t x_2(t) + x_3(t) \end{bmatrix}, \quad \text{a.e. } t \in (0,1),$$

defined on $H_D^1(0,1) = \{x_2, x_3 \in H^1(0,1),\ x_1 \in L^2(0,1)\}$ becomes

$$\operatorname{im} T = \{q \in L^2(0,1)^3 : q_3 \in H^1(0,1),\ q_2 - q_3' \in H^1(0,1)\}.$$

□

The Moore-Penrose generalized inverse of the operator T is bounded if and only if the index does not exceed 1. Consequently, for higher index problems, an equation of the kind $Tx = q$, with $q \in L^2(a,b)$, becomes essentially ill-posed. Nevertheless, the nullspace of T is finite dimensional and its dimension l depends on the index. Therefore, augmenting $Tx = q$ by suitably chosen boundary conditions makes the problem uniquely solvable. We emphasize, that the unique solvability of the resulting BVP in DAEs does not change its ill-posedness.
Finally, we introduce the operator $\mathcal{T} : H_D^1(a,b) \to L^2(a,b)^m \times \mathbb{R}^l =: \mathcal{Y}$ associated with the BVP (1) – (2) by

$$\mathcal{T}x = \begin{bmatrix} Tx \\ G_a x(a) + G_b x(b) \end{bmatrix}, \quad x \in H_D^1(a,b).$$

To ensure that $G_a x(a) + G_b x(b)$ is well-defined, we restrict the boundary conditions by assuming

$$\ker G_a = \ker D(a), \quad \ker G_b = \ker D(b). \tag{9}$$

Then, for instance, $G_a x(a) = G_a D(a)^+ D(a)x(a)$ is well-defined together with $D(a)x(a)$ by Sobolev's imbedding theorem.

Theorem 2.5. *Let the bounded DA operator $T : H^1_D(a,b) \to L^2(a,b)^m$ be fine with index $\mu \in \mathbb{N}$ and characteristic values (8) and let the boundary conditions be restricted by (9). Then the following statements hold:*

(1) *The BVP $\mathcal{T}x = (q,\gamma)$ is uniquely solvable for each $\gamma \in \mathbb{R}^l$ and each right-hand side $q \in \operatorname{im} T$, if and only if the condition*

$$\ker(G_a X(a,a) + G_b X(b,a)) = \ker \Pi_{can}(a) \tag{10}$$

holds. Here, $X(t,a)$ denotes the maximal fundamental solution matrix of the associated DAE, normalized at point a.

(2) *If (10) is valid, then the equation $\mathcal{T}x = (q,\gamma)$ is well-posed if $\mu = 1$ and otherwise essentially ill-posed.*

(3) *If (10) is valid, then $\mathcal{T}$ is injective.*

(4) *If $\mu = 1$ and (10) is valid, then there exists a constant bound $c_{\mathcal{T}} > 0$ such that*

$$\|\mathcal{T}x\|_Y \geq c_{\mathcal{T}} \|x\|_{H^1_D}, \quad x \in H^1_D(a,b).$$

Proof. The arguments used in [15, Theorem 2.1] can be adopted accordingly to verify (1) – (3). Result (4) follows from the fact that $\mathcal{T}$ is bijective. Then, $\mathcal{T}^{-1}$ is bounded and $c_{\mathcal{T}} = 1/\|\mathcal{T}^{-1}\|$. □

Below, we primarily try to approximate ill-posed BVPs for higher index DAEs using a regularization method. The following lemma concerning transformations and refactorizations of DAEs, might be helpful to trace back certain questions to DAEs in special form.

Lemma 2.6. *Let the bounded DA operator $T : H^1_D(a,b) \to L^2(a,b)^m$ be fine with index $\mu \in \mathbb{N}$ and characteristic values (8) and let the boundary conditions satisfy conditions (9) and (10).*
Let the matrix functions $L, K \in C([a,b], L(\mathbb{R}^m))$ and $H \in C^1([a,b], L(\mathbb{R}^k))$ be pointwise nonsingular. Moreover,

$$\begin{aligned} LAH =: \tilde{A}, \quad H^{-1}DK =: \tilde{D}, \quad LBK - LARH(H^{-1}R)'DK =: \tilde{B}, \\ \tilde{T}\tilde{x} := \tilde{A}\tilde{D}\tilde{x}' + \tilde{B}\tilde{x}, \quad \tilde{x} \in H^1_{\tilde{D}}(a,b), \\ G_a K(a) =: \tilde{G}_a, \quad GbK(b) =: \tilde{G}_b, \\ \tilde{\mathcal{T}}\tilde{x} := (\tilde{T}\tilde{x}, \tilde{G}_a\tilde{x}(a) + \tilde{G}_b\tilde{x}(b)), \quad \tilde{x} \in H^1_{\tilde{D}}(a,b). \end{aligned}$$

Then, the DA operator $\tilde{T}$ inherits from T the index and all characteristic values and the conditions (9) and (10) are valid accordingly for the transformed boundary conditions. In addition, the following statements hold: $H^1_{\tilde{D}}(a,b) = K^{-1}(H^1_D(a,b))$, $\tilde{\mathcal{T}}$ is injective as $\mathcal{T}$ is, and there are positive constants c_l, c_u such that

$$c_l \frac{\|\mathcal{T}x\|^2_Y}{\|x\|^2_{H^1_D}} \leq \frac{\|\tilde{\mathcal{T}}\tilde{x}\|^2_Y}{\|\tilde{x}\|^2_{H^1_{\tilde{D}}}} \leq c_u \frac{\|\mathcal{T}x\|^2_Y}{\|x\|^2_{H^1_D}}, \quad x \in H^1_D(a,b), \quad \tilde{x} = K^{-1}x.$$

Proof. Owing to [14, Section 2.3], transformations and refactorizations do not alter the characteristic values and the index. The maximal fundamental solution matrix normalized at point a and the canonical projector transform by

$$\tilde{X}(t,a) = K(t)^{-1}X(t,a)K(a), \quad \tilde{\Pi}_{can}(t) = K(t)^{-1}\Pi_{can}(t)K(t).$$

Since the original boundary conditions (2) are accurately stated in the sense of condition (10), so are the transformed boundary conditions

$$\tilde{G}_a\tilde{x}(a) + \tilde{G}_b\tilde{x}(b) = \gamma,$$

and hence, the operator $\tilde{\mathcal{T}}$ is also injective. Next, $x \in H^1_D(a,b)$ implies $\tilde{x} = K^{-1}x \in H^1_{\tilde{D}}(a,b)$, and vice versa. Moreover, we have

$$\|\mathcal{T}x\|_Y^2 = \|L^{-1}\underbrace{LTK}_{\tilde{T}}\tilde{x}\|_Y^2 + |\tilde{G}_a\tilde{x}(a) + \tilde{G}_b\tilde{x}(b)|^2 \geq c_1\|\tilde{\mathcal{T}}\tilde{x}\|_Y^2,$$

and

$$\begin{aligned}\|x\|_{H^1_D}^2 &= \|K\tilde{x}\|_{H^1_D}^2 = \|K\tilde{x}\|_{L^2}^2 + \|(DK\tilde{x})'\|_{L^2}^2 = \|K\tilde{x}\|_{L^2}^2 + \|(H\tilde{D}\tilde{x})'\|_{L^2}^2 \\ &= \|K\tilde{x}\|_{L^2}^2 + \|H(\tilde{D}\tilde{x})' + H'\tilde{D}\tilde{x}\|_{L^2}^2 \leq c_2(\|\tilde{x}\|_{L^2}^2 + \|(\tilde{D}\tilde{x})'\|_{L^2}^2) = c_2\|\tilde{x}\|_{H^1_{\tilde{D}}}^2,\end{aligned}$$

with a suitable constant c_2. Thus,

$$\frac{\|\mathcal{T}x\|_Y^2}{\|x\|_{H^1_D}^2} \geq \frac{c_1}{c_2}\frac{\|\tilde{\mathcal{T}}\tilde{x}\|_Y^2}{\|\tilde{x}\|_{H^1_{\tilde{D}}}^2}, \quad x \in H^1_D(a,b),\ \tilde{x} = K^{-1}x.$$

The remaining part of the inequality follows using the opposite transformation and refactorization. □

2.2. Regularization by Discretization in Pre-Image Space

The following presentation follows suggestions made in [12]. Let $X, \mathcal{Y}$ be Hilbert spaces and $\mathcal{T} : X \to \mathcal{Y}$ be a bounded injective linear operator. Certainly, we primarily have in mind the operator representing the BVP for a higher index DAE, see the previous subsection.

Let us denote by $\mathcal{T}^\dagger$ the Moore-Penrose generalized inverse of $\mathcal{T}$. Note that $\mathcal{T}^\dagger$ is continuous only if $\operatorname{im}\mathcal{T}$ is closed in $\mathcal{Y}$, however, this is not the case for higher index DAEs, see Theorem 2.2. If $y \in \operatorname{dom}(\mathcal{T}^\dagger)$, set

$$x^\dagger := \mathcal{T}^\dagger y.$$

If $\mathcal{T}$ is injective and densely solvable, which will be the case in our application below, the operators $\mathcal{T}^\dagger$ and $\mathcal{T}^{-1}$ coincide, and, in particular, $\operatorname{dom}\mathcal{T}^\dagger = \operatorname{dom}\mathcal{T}^{-1} = \operatorname{im}\mathcal{T}$.

Let $X_n \subset X$ be a sequence of finite dimensional subspaces of X such that $X_n \subset X_{n+1}$ and $\bigcup_{n=1}^\infty X_n$ is dense in X. Let $P_n : X \longrightarrow X_n$ denote the orthoprojector. Consequently, for any $x \in X$, we have

$$P_n x \longrightarrow x \text{ for } n \longrightarrow \infty.$$

The approximations x_n of $x^\dagger$ are defined by

$$x_n \in \operatorname{argmin}\{\|\mathcal{T}z - y\|^2 : \ z \in X_n\}. \tag{11}$$

Let us now define the following values:

$$\alpha_n = \|(I - P_n)x^\dagger\|, \quad \beta_n = \|\mathcal{T}(I - P_n)x^\dagger\|, \quad \gamma_n = \inf_{z\in X_n, z\neq 0} \frac{\|\mathcal{T}z\|}{\|z\|} = \frac{1}{\|(\mathcal{T}P_n)^\dagger\|}.$$

Note that $\gamma_n > 0$ if and only if $\mathcal{T}|_{X_n}$ is invertible, but this is the case, since $\mathcal{T}$ is injective. Owing to formula [12, (2.11)], we have

$$\|x_n - x^\dagger\| \le \frac{\beta_n}{\gamma_n} + \alpha_n.$$

As discussed in [12], it may happen that the speed of convergence for β_n is much faster than that for α_n. In the application which we have in mind, there is no reason to expect such a behavior. Therefore, the use of the general estimate

$$\beta_n \le \|\mathcal{T}\|\alpha_n$$

is well justified. The constant α_n measures the approximation quality of $x^\dagger$ by elements of X_n while γ_n measures to which extend the discrete operator $\mathcal{T}|_{X_n}$ becomes unstable. So it is reasonable to assume that

$$\alpha_n = O(n^{-k_1}), \quad \gamma_n \ge c\, n^{-k_2},$$

with a positive constant c and constants $k_{1,2} \ge 0$, such that

$$\|x_n - x^\dagger\| = O(n^{k_2-k_1}).$$

We intend to use this general approach and then to introduce the collocation by replacing the norm in $\mathcal{Y}$ (typically the L^2-norm) by a discrete version. In the context of regularization, this discretization will be interpreted as a perturbation y^{δ_n} of the exact right-hand side y such that

$$\|y - y^{\delta_n}\| \le \delta_n.$$

Again, formula [12, (2.11)] provides an estimate, namely

$$\|x_n^{\delta_n} - x^\dagger\| \le \frac{\delta_n + \beta_n}{\gamma_n} + \alpha_n. \tag{12}$$

In fact, the primary task is to provide such an *instability threshold* $\gamma_n \ge c\, n^{-k_2}$. For index-1 DAEs, the inverse $\mathcal{T}^{-1}$ is bounded, so it simply follows that $\gamma_n \ge c_{\mathcal{T}} > 0$, see Theorem 2.5. For higher index DAEs, $\mathcal{T}^{-1}$ becomes unbounded. We will look into this challenging matter in Section 3.

2.3. Collocation Using Discrete Norms

We turn back to a uniquely solvable BVP, $\mathcal{T}x = y$ with $y = (q, \gamma) \in \operatorname{im}\mathcal{T}$, as described in Theorem 2.5. In the practical implementation of the method (11) the norm in $L^2(a, b)$ must be replaced by a discrete approximation using, e.g., a numerical interpolation formula. To this end, let the equidistant partition

$$a = t_0 < t_1 < \cdots < t_n = b,$$

be given and let the linear space of piecewise polynomial functions X_n be defined as before in (4). Moreover, let the $M = N + \nu$, $\nu \geq 0$ interpolation nodes,

$$0 < \tau_1 < \tau_2 < \cdots < \tau_M < 1,$$

be fixed. Finally, let us denote the resulting sets of collocation points on the subinterval $[t_{j-1}, t_j)$ by

$$S_j = \{t_{j-1} + \tau_i h,\ i = 1, \ldots, M\}, \quad j = 1, \ldots, n. \tag{13}$$

Instead of minimizing $\|\mathcal{T}p - y\|^2$ over $p \in X_n$, we have in mind to seek a least-squares solution of the overdetermined discrete system comprising $Mnm + l$ equations,

$$A(t)(Dp)'(t) + B(t)p(t) = q(t), \quad t \in S_j,\ j = 1, \ldots, n, \tag{14}$$

$$G_a p(a) + G_b p(b) = \gamma. \tag{15}$$

In this context, we assume q to be sufficiently smooth, so that for every $t \in S_j$ the function value $q(t)$ is well-defined and interpolation makes sense.
We denote by q_n the interpolating piecewise polynomial defined by

$$q_n|_{[t_{j-1},t_j)} \in \mathcal{P}^m_{M-1}, \quad q_n(t) = q(t),\ t \in S_j,\ j = 1, \ldots, n. \tag{16}$$

Set $y_n = (q_n, \gamma)$ such that

$$\delta_n = \|y - y_n\|_Y = \|q - q_n\|_{L^2} = O(h^M). \tag{17}$$

Regarding (12), we may turn to the equation $\mathcal{T}x = y_n$ and the problem to be solved reads:

$$x_n^{\delta_n} \in \operatorname{argmin}\{\|Tp - y_n\|^2_{L^2} + |G_a p(a) + G_b p(b) - \gamma|^2 :\ p \in X_n\}. \tag{18}$$

Let the entries of the coefficients A and B be polynomials of at most degree $N_{A,B}$. Then, for each $p \in X_n$, the expression $Tp = A(Dp)' + Bp$ is a piecewise polynomial function and $Tp|_{[t_{j-1},t_j)} \in \mathcal{P}^m_{N+N_{A,B}}$ on each subinterval. Choosing $\nu \geq 1 + N_{A,B}$, that is,

$$M - 1 \geq N + N_{A,B}$$

we ensure $\{Tp - q_n\}|_{[t_{j-1},t_j)} = \{A(Dp)' + Bp - q_n\}|_{[t_{j-1},t_j)} \in \mathcal{P}^m_{M-1}$. Note that we have $N_{A,B} = 1$ in both Examples 1.1 and 1.2. The experiments in Section 1 are realized with $M = 2N + 1$.

Let us now consider an arbitrary piecewise function w with values in $\mathbb{R}^m$ defined on our partition, such that $w|_{[t_{j-1},t_j)} \in \mathcal{P}^m_{M-1}$ on each subinterval. We represent w as

$$w(t) = \sum_{i=1}^{M} w(t_{j-1} + \tau_i h) l_{ji}(t), \quad t \in [t_{j-1}, t_j), \ j = 1, \dots, n,$$

where the corresponding Lagrange basis polynomial is denoted by l_{ji}.
Consequently, we have

$$\begin{aligned}
\|w\|^2_{L^2} &= \int_a^b |w(t)|^2 \mathrm{d}t = \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} |w(t)|^2 \mathrm{d}t \\
&= \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \langle \sum_{i=1}^{M} w(t_{j-1} + \tau_i h) l_{ji}(t), \sum_{i=1}^{M} w(t_{j-1} + \tau_i h) l_{ji}(t) \rangle \mathrm{d}t \\
&= \sum_{j=1}^{n} \sum_{i,\kappa=1}^{M} \int_{t_{j-1}}^{t_j} l_{ji}(t) l_{j\kappa}(t) \mathrm{d}t \ \langle w(t_{j-1} + \tau_i h), w(t_{j-1} + \tau_i h) \rangle.
\end{aligned}$$

The integrals

$$L_{i\,\kappa} = \frac{1}{h} \int_{t_{j-1}}^{t_j} l_{ji}(t) l_{j\kappa}(t) \mathrm{d}t$$

are independent of j and h, and the resulting matrix $L = (L_{i\kappa})$ is symmetric and positive definite. Now, we obtain

$$\begin{aligned}
\|w\|^2_{L^2} &= h \sum_{j=1}^{n} \sum_{i,\kappa=1}^{M} L_{i\kappa} \ \langle w(t_{j-1} + \tau_i h), w(t_{j-1} + \tau_i h) \rangle \\
&= h \sum_{j=1}^{n} \sum_{i,\kappa=1}^{M} L_{i\kappa} \ w(t_{j-1} + \tau_i h)^T w(t_{j-1} + \tau_i h) \\
&= h \sum_{j=1}^{n} W_j^T (L \otimes I_m) W_j = h \, W^T \, \mathrm{diag}(L \otimes I_m, \cdots, L \otimes I_m) \ W =: h \, W^T \mathcal{L} \, W,
\end{aligned}$$

where

$$W_j = \begin{bmatrix} w(t_{j-1} + \tau_1 h) \\ \vdots \\ w(t_{j-1} + \tau_M h) \end{bmatrix} \in \mathbb{R}^{mM}, \quad W = \begin{bmatrix} W_1 \\ \vdots \\ W_n \end{bmatrix} \in \mathbb{R}^{mMn}.$$

Finally, we find constants c_L, $\bar{c}_L$ depending only on L such that

$$c_L h^{\frac{1}{2}} |W|_2 \le \|w\|_{L^2} \le \bar{c}_L h^{\frac{1}{2}} |W|_2. \tag{19}$$

Here, we denote the Euclidean norm of $W \in \mathbb{R}^{mMn}$ by $|W|_2$. From the above, the next statement follows.

Proposition 2.7. *Let $\mathcal{W}_n$ denote the linear set of all piecewise defined function w with values in $\mathbb{R}^m$ such that $w|_{[t_{j-1},t_j)} \in \mathcal{P}^m_{M-1}$ on each subinterval of the above partition and let $|w|_2 := |W|_2$ for $w \in \mathcal{W}_n$.*
Then, the relation $\|w\|^2_{L^2} = h\, W^T \mathcal{L}\, W$ holds for all $w \in \mathcal{W}_n$ and the norms $\|\cdot\|_{L^2}$ and $|\cdot|_2$ are equivalent on $\mathcal{W}_n$.

In a way, this justifies our collocation approach aiming at finding an approximation x^δ_n by providing the least-squares solution of the overdetermined collocation scheme (14)–(15). More precisely: Consider $w = A(Dp)' + Bp - q_n \in \mathcal{W}_n$. Regarding the interpolation condition (16), expression

$$\sum_{j=1}^{n} \sum_{t\in S_j} |A(t)(Dp)'(t) + B(t)p(t) - q(t)|^2 + |G_a p(a) + G_b p(b) - \gamma|^2 \tag{20}$$

coincides with $|w|_2^2 + |G_a p(a) + G_b p(b) - \gamma|^2$. This means, that instead of minimizing

$$\|w\|^2_{L^2} + |G_a p(a) + G_b p(b) - \gamma|^2, \tag{21}$$

cf. (18), we use the equivalent norm $|w|_2$ for $\|w\|_{L_2}$. In this context, $\|w\|_{L_2}$ can be interpreted as a weighted form of $|w|_2$. Experiments using both norms indicate no significant differences, see Section 5.

3. Estimating the Instability Thresholds for Special Cases

Let X_n, related to the partition

$$a = t_0 < t_1 < \cdots < t_n = b,$$

be defined by (4). The instability threshold

$$\gamma_n = \inf_{p\in X_n, p\neq 0} \frac{\|\mathcal{T} p\|_Y}{\|p\|_{H^1_D}}$$

plays the key role in the analysis of method (11) and hence, of the least-squares approach (14)–(15). Note again that an operator $\mathcal{T}$ corresponding to a well-posed BVP, i.e., an index-1 DAE completed by accurately stated boundary condition, has a bounded inverse such that $\gamma_n = (\|\mathcal{T}^{-1}\|)^{-1}$ is a constant. In all higher index cases $\mathcal{T}^{-1}$ is no longer bounded, see Theorem 2.5, and γ_n depends on n. Not surprisingly, the behavior of γ_n turns out to be dominated by the largest inherent Jordan chain part which appoints the index, see Subsection 3.1.

In the next lemma, we provide a tool to possibly trace back the problem to a single subinterval. To this end, we introduce values $\gamma_{n,\,j}$ such that, for $j = 1, \ldots, n$,

$$\|T^{[j]} p^{[j]}\|_{L^{2\,[j]}} \geq \gamma_{n,\,j}\, \|p^{[j]}\|_{H^{1\,[j]}_D}, \quad p_1, \ldots, p_k \in \mathcal{P}_N,\ p_{k+1}, \ldots, p_m \in \mathcal{P}_{N-1},$$

where the superscript $[j]$ indicates restrictions to the subinterval $[t_{j-1}, t_j)$. In general, one has $\gamma_{n,\,j} \geq 0$.

Lemma 3.1. *If the DA operator T is regular with index μ and injective, then*

$$\gamma_n \geq \min\{\gamma_{n,1}, \ldots, \gamma_{n,n}\} > 0.$$

Proof. Since T is regular and injective, the dynamical degree of freedom is $l = 0$ and T coincides with $\mathcal{T}$. Furthermore, as T is regular and injective on each subinterval of $[a, b]$, we have $\gamma_{n,j} > 0$ for all j.

Recall that X_n is defined by (4). Dropping the continuity conditions yields

$$\begin{aligned}\tilde{X}_n = \{p \in & L^2(a,b)^m : \ p_\kappa|_{[t_{j-1},t_j)} \in \mathcal{P}_N, \ \kappa = 1, \ldots, k, \\ & p_\kappa|_{[t_{j-1},t_j)} \in \mathcal{P}_{N-1}, \ \kappa = k+1, \ldots, m, \ j = 1, \ldots, n\} \supset X_n.\end{aligned}$$

Consequently,

$$\begin{aligned}\gamma_n^2 &= \inf_{p \in X_n, p \neq 0} \frac{\|Tp\|_{L^2}^2}{\|p\|_{H_D^1}^2} \geq \inf_{p \in \tilde{X}_n, p \neq 0} \frac{\|Tp\|_{L^2}^2}{\|p\|_{H_D^1}^2} = \inf_{p \in \tilde{X}_n, p \neq 0} \frac{\sum_{j=1}^n \|T^{[j]} p^{[j]}\|_{L^{2\,[j]}}^2}{\sum_{j=1}^n \|p^{[j]}\|_{H_D^{1\,[j]}}^2} \\ &\geq \inf_{p \in \tilde{X}_n, p \neq 0} \frac{\sum_{j=1}^n \gamma_{n,j}^2 \, \|p^{[j]}\|_{H_D^{1\,[j]}}^2}{\sum_{j=1}^n \|p^{[j]}\|_{H_D^{1\,[j]}}^2} \geq \min\{\gamma_{n,1}^2, \ldots, \gamma_{n,n}^2\} > 0.\end{aligned}$$

□

3.1. DAEs in Weierstraß–Kronecker Form

The DAE

$$\begin{bmatrix} I & 0 \\ 0 & J \end{bmatrix} x'(t) + \begin{bmatrix} W & 0 \\ 0 & I \end{bmatrix} x(t) = q(t), \quad t \in [a, b],$$

with an arbitrary matrix W of size $l \times l$ and a nilpotent Jordan block matrix J of size $(m - l) \times (m - l)$ is said to be in *Weierstraß–Kronecker form*. The Kronecker index of this DAE is defined as the maximal size of the Jordan chains in J and coincides here with both the differentiation index and the tractability index. J has the form

$$J = \begin{bmatrix} J_1 & & \\ & \ddots & \\ & & J_s \end{bmatrix}, \quad J_i = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix} \in L(\mathbb{R}^{l_i}), \quad l_1 + \cdots + l_s = m - l.$$

We factorize,

$$J_i = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix} =: A_i D_i,$$

$$\begin{bmatrix} I & 0 \\ 0 & J \end{bmatrix} = \begin{bmatrix} I & & & \\ & A_1 & & \\ & & \ddots & \\ & & & A_s \end{bmatrix} \begin{bmatrix} I & & & \\ & D_1 & & \\ & & \ddots & \\ & & & D_s \end{bmatrix} =: AD,$$

such that A has full column-rank $k = l + \sum_{i=1}^{s}(l_i - 1)$ and D has full row-rank k. This allows us to formulate the DAE with a proper leading term as

$$A(Dx)'(t) + \begin{bmatrix} W & 0 \\ 0 & I \end{bmatrix} x(t) = q(t), \quad t \in [a, b].$$

The coefficient matrix D has rows being unit vectors. A permutation of the variables would meet condition (3), but here, it is more transparent to modify the set X_n accordingly, see (4). According to the special DAE structure, the DA operator T decomposes the system in the following way:

$$\begin{aligned} Tx = (T_{[dyn]}u,\, T_{[1]}v_{[1]}, \cdots, T_{[s]}v_{[s]}), \quad x = (u, v_{[1]}, \cdots, v_{[s]}) \in (H^1)^l \times H^1_{D_1} \times \cdots \times H^1_{D_s}, \\ T_{[dyn]}u = u' + Wu, \\ T_{[i]}v_{[i]} = A_i(D_i v_{[i]})' + v_{[i]}, \quad i = 1, \ldots, s, \end{aligned}$$

and due to the accurately stated boundary conditions, a well-posed ordinary BVP results,

$$u' + Wx = q_{[dyn]}, \quad B_a u(a) + B_b u(b) = \gamma.$$

Therefore, the associated composed operator $\mathcal{T}_{[dyn]}$ has a bounded inverse and we have $\mathcal{T}x = (\mathcal{T}_{[dyn]}u, T_{[1]}v_{[1]}, \cdots, T_{[s]}v_{[s]})$. Thus,

$$\begin{aligned} \|\mathcal{T}x\|_Y^2 = \|\mathcal{T}_{[dyn]}u\|^2_{L^{2,l}\times\mathbb{R}^l} + \sum_{i=1}^{s} \|T_{[i]}v_{[i]}\|^2_{L^{2,l_i}}, \\ \|x\|^2_{H^1_D} = \|u\|^2_{H^{1,l}} + \sum_{i=1}^{s} \|v_{[i]}\|^2_{H^1_{D_i}}. \end{aligned}$$

Define $\gamma_{n,[u]} := (\|\mathcal{T}^{-1}_{[dyn]}\|)^{-1}$ and introduce the thresholds $\gamma_{n,[i]}$ associated with the other parts in such a way that

$$\|T_{[i]}p_{[i]}\|_{L^{2,l_i}} \geq \gamma_{n,[i]} \, \|p_{[i]}\|_{H^1_{D_i}} \quad \text{for all} \quad p_{[i]} \in X_{n,[i]}, \quad i = 1, \ldots, s.$$

In turn, each $T_{[i]}$ is injective on its part, and hence $\gamma_{n,[i]} > 0$. Finally, we obtain

$$\begin{aligned} \gamma_n^2 &= \inf_{p\in X_n, p\neq 0} \frac{\|\mathcal{T}p\|_Y^2}{\|p\|^2_{H^1_D}} = \inf_{p\in X_n, p\neq 0} \frac{\|\mathcal{T}_{[dyn]}p_{[u]}\|^2_{L^{2,l}\times\mathbb{R}^l} + \sum_{i=1}^{s}\|T_{[i]}p_{[i]}\|^2_{L^{2,l_i}}}{\|p_{[u]}\|^2_{H^{1,l}} + \sum_{i=1}^{s}\|p_{[i]}\|^2_{H^1_{D_i}}} \\ &\geq \inf_{p\in X_n, p\neq 0} \frac{\gamma^2_{n,[u]}\|p_{[u]}\|^2_{H^{1,l}} + \sum_{i=1}^{s}\gamma^2_{n,[i]}\|p_{[i]}\|^2_{H^1_{D_i}}}{\|p_{[u]}\|^2_{H^{1,l}} + \sum_{i=1}^{s}\|p_{[i]}\|^2_{H^1_{D_i}}} \geq \gamma^2_{n,[u]} + \sum_{i=1}^{s}\gamma^2_{n,[i]} > 0. \end{aligned}$$

This result is summarized in the following proposition.

Proposition 3.2. *If the DAE has Weierstraß–Kronecker form and the boundary conditions are accurately stated, then the following estimate holds:*

$$\gamma_n \geq (\gamma^2_{n,[u]} + \sum_{i=1}^{s}\gamma^2_{n,[i]})^{\frac{1}{2}} > 0.$$

Note that the values $\gamma_{n,[i]}$ corresponding to Jordan chains of size $l_i = 1$ are independent of n, namely $\gamma_{n,[i]} = 1$, since then $T_{[i]}v_{[i]} = v_{[i]}$. It turns out that only those $\gamma_{n,[i]}$ associated with $l_i > 1$ may depend on n.

Clearly, one expects that $\gamma_{n,[i]}$ behaves the worse the larger the size l_i. In Section 3.3, we show an estimate for the case of a Jordan chain DAE.

3.2. A Class of DAEs with Variable Coefficients

We now turn to more general BVPs of type (1) – (2), with D being in the special form (3). Let the boundary condition be accurately stated so that condition (10) is satisfied. Let the DAE be transformable into Weierstraß–Kronecker form, i.e., there are continuous, pointwise nonsingular matrix functions $L, K \in C([a,b], L(\mathbb{R}^m))$ and a continuously differentiable, pointwise nonsingular refactorization matrix $H \in C^1([a,b], L(\mathbb{R}^k))$ such that (cf. Subsection 3.1)

$$LAH =: \tilde{A} = \begin{bmatrix} I & & & \\ & \tilde{A}_1 & & \\ & & \ddots & \\ & & & \tilde{A}_s \end{bmatrix}, \quad H^{-1}DK =: \tilde{D} = \begin{bmatrix} I & & & \\ & \tilde{D}_1 & & \\ & & \ddots & \\ & & & \tilde{D}_s \end{bmatrix},$$

$$\tilde{A}\tilde{D} = \begin{bmatrix} I & 0 \\ 0 & \tilde{J} \end{bmatrix}, \quad LBK - LARH(H^{-1}R)'DK =: \tilde{B} = \begin{bmatrix} \tilde{W} & 0 \\ 0 & I \end{bmatrix}.$$

The set of ansatz functions X_n defined by (4) for the original DAE transforms into $K^{-1}X_n$. Thereby, the continuous components are appropriately transformed into continuous ones. However, the correct set of ansatz functions $\tilde{X}_n$ for the transformed problem consists of piecewise polynomials of degree up to N for the components indicated by $\tilde{D}$ and up to degree $N-1$ for the remaining components. If K is a constant matrix function, then we obtain $\tilde{X}_n = K^{-1}X_n$ and hence, by Lemma 2.6,

$$\gamma_n = \inf_{p \in X_n, p \neq 0} \frac{\|\mathcal{T}p\|_Y^2}{\|p\|_{H^1_D}^2} \geq \frac{1}{c_u} \inf_{p \in \tilde{X}_n, p \neq 0} \frac{\|\tilde{\mathcal{T}}p\|_Y^2}{\|p\|_{H^1_D}^2} = \frac{1}{c_u}\,\tilde{\gamma}_n.$$

Again, the problem can be traced back to purely Jordan chain DAEs which will be considered next.

3.3. A Pure Jordan Chain DAE

We consider the constant coefficient DAE

$$A(Dx)'(t) + Bx(t) = q(t),$$

where $B = I \in \mathbb{R}^{\mu\times\mu}$, $D = \mathrm{diag}(1,\ldots,1,0) \in \mathbb{R}^{\mu\times\mu}$, and $A = -J \in \mathbb{R}^{\mu\times\mu}$,

$$J = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix}.$$

The DAE is regular with index μ and it has dynamical degree of freedom $l = 0$ hence $T = \mathcal{T}$. In order to slightly simplify the matters, we assume the interval $[a, b]$ to be normalized to $[0, 1]$.

The operator T is defined on $X = H_D^1(0,1) = \{x \in L^2(0,1)^\mu | x_1, \ldots, x_{\mu-1} \in H^1(0,1)\}$ and $\operatorname{im} T \subseteq Y = L^2(0,1)^\mu$. Obviously, T is injective. Moreover, for $\mu > 1$, the problem is ill-posed, cf. Theorem 2.5.

On $[0, 1]$, we introduce an equidistant partition with n subintervals of length $h = n^{-1}$ such that $t_j = jh,\ j = 0, \ldots, n$. The finite dimensional subspace X_n, cf. (4), consists of all piecewise vector polynomials $(p_1, p_2, \ldots, p_\mu)$ such that

$$p_i|_{[t_{j-1},t_j)} \in \mathcal{P}_N, \quad p_i \text{ is continuous}, \quad i = 1, \ldots, \mu - 1, \quad p_\mu|_{[t_{j-1},t_j)} \in \mathcal{P}_{N-1}.$$

The aim of this Section is to derive estimates for the asymptotic behavior of γ_n. Owing to Lemma 3.1, we restrict our attention to a single subinterval. We consider only the first subinterval $[0, h]$ and indicate this by using index h in norms, etc. This subinterval is representative for all subintervals of the grid on $[0, 1]$, since the DAE has constant coefficients.

Our first aim is to obtain a representation of the norms $\|p\|_{H_D^1,\,h}$ and $\|Tp\|_{L^2,\,h}$, in order to derive the estimates for $\gamma_{n,\,h}$. From

$$\begin{aligned}
\|Tp\|_{L^2,\,h}^2 &= \int_0^h (p_1^2(t) + \sum_{i=2}^{\mu} [p_i(t) - p_{i-1}'(t)]^2)\,\mathrm{d}t \\
&= \int_0^h (\sum_{i=1}^{\mu} p_i^2(t) + \sum_{i=2}^{\mu} (p_{i-1}')^2(t) - 2\sum_{i=2}^{\mu} p_i(t)p_{i-1}'(t))\,\mathrm{d}t \\
&= \|p\|_{H_D^1}^2 - 2\sum_{i=2}^{\mu} \int_0^h p_i(t)p_{i-1}'(t)\,\mathrm{d}t
\end{aligned}$$

we conclude

$$\gamma_{n,\,h}^2 = \inf_{\substack{p \in X_n \\ p \neq 0}} \frac{\|\mathcal{T}p\|_{L^2,\,h}^2}{\|p\|_{H_D^1,\,h}^2} = 1 - \sup_{\substack{p \in X_n \\ p \neq 0}} \frac{1}{\|p\|_{H_D^1,\,h}^2} \sum_{i=2}^{\mu} \int_0^h 2p_i(t)p_{i-1}'(t)\,\mathrm{d}t. \tag{22}$$

In the following lemma a relation between the norms of a polynomial and its derivative is given.

Lemma 3.3. *Let $N \geq 1$ and $h > 0$. Then,*

$$z_h := \inf_{\substack{p \in \mathcal{P}_N \\ p' \neq 0}} \frac{\int_0^h p^2(t)\,dt}{\int_0^h (p')^2(t)\,dt} = \frac{1}{\lambda_N} h^2 > 0,$$

where λ_N is the spectral radius of the matrix $\bar{K} = (k_{ij})_{i,j \in \{0,\ldots,N\}}$ given in equation (25).

Remark 3.4. *It can be shown that the following estimates for the eigenvalue λ_N hold:*

$$\lambda_N \le 2(4N^2-1)\begin{cases}\left(1-\cos\frac{1}{N+1}\pi\right)^{-1}, & N \text{ even},\\ \left(1-\cos\frac{1}{N+2}\pi\right)^{-1}, & N \text{ odd}.\end{cases}$$

The upper bound has an asymptotic expansion $\lambda_N \le \frac{16}{\pi^2}N^4 + O(N^2)$. Therefore, the constant $1/\lambda_N$ is of reasonable size.

Proof. To verify Lemma 3.3, we use the Legendre polynomials φ_κ as a basis in the representation of $p \in \mathcal{P}_N$, see Appendix A for the definition and some properties of the Legendre polynomials. For

$$p(t) = \sum_{\kappa=0}^{N} c_\kappa \varphi_\kappa(t),$$

we obtain

$$\int_0^h p^2(t)\,\mathrm{d}t = \int_0^h \left(\sum_{\kappa=0}^{N} c_\kappa\varphi_\kappa(t)\right)^2 \mathrm{d}t = \sum_{\kappa=0}^{N} c_\kappa^2 \|\varphi_\kappa\|^2$$

and

$$\int_0^h (p')^2(t)\,\mathrm{d}t = \int_0^h \left(\sum_{\kappa=0}^{N} c_\kappa\varphi_\kappa'(t)\right)^2 \mathrm{d}t = \sum_{\kappa,\lambda=0}^{N} c_\kappa c_\lambda \int_0^h \varphi_\kappa'\varphi_\lambda'\,\mathrm{d}t.$$

We introduce

$$\tilde{M} = \operatorname{diag}(\|\varphi_0\|^2,\dots,\|\varphi_N\|^2) = h\operatorname{diag}(1,1/3,\dots,1/(2N+1)) =: hM \tag{23}$$

and $\tilde{K} := (\tilde{k}_{\kappa\lambda})_{\kappa,\lambda=0,\dots,N}$ with

$$\tilde{k}_{\kappa\lambda} = \int_0^h \varphi_\kappa'\varphi_\lambda' dt = \begin{cases}\frac{4}{\|\varphi_{\tilde\kappa-1}\|^2} + \frac{4}{\|\varphi_{\tilde\kappa-3}\|^2} + \cdots, & \text{if } \kappa{+}\lambda \text{ is even},\\ 0, & \text{otherwise},\end{cases}$$

where $\tilde\kappa = \min(\kappa,\lambda)$. Then

$$\sup_{\substack{p\in\mathcal{P}_N\\ p'\neq 0}} \frac{\int_0^h (p')^2(t)\,\mathrm{d}t}{\int_0^h p^2(t)\,\mathrm{d}t} = \sup_{c\in\mathbb{R}^{N+1}} \frac{c^\top \tilde{K} c}{c^\top \tilde{M} c},$$

where $\tilde{K}$ has the following structure:

$$\tilde{K} = h^{-1}K, \quad K = \begin{bmatrix} 0 & 0 & 0 & 0 & \cdots & 0\\ 0 & k_1 & 0 & k_1 & \cdots & \\ 0 & 0 & k_2 & 0 & \cdots & \\ 0 & k_1 & 0 & k_3 & \cdots & \\ \vdots & \vdots & \vdots & \vdots & \ddots & \\ 0 & & & & & k_N \end{bmatrix}.$$

Here, k_ι are constants which depend only on κ, λ but not on h. Clearly, K is symmetric and positive semi-definite. Furthermore,

$$\sup_{\substack{p\in\mathcal{P}_N\\ p'\neq 0}} \frac{\int_0^h (p')^2(t)\,\mathrm{d}t}{\int_0^h p^2(t)\,\mathrm{d}t} = \frac{1}{h^2} \sup_{c\in\mathbb{R}^{N+1}} \frac{c^\top K c}{c^\top M c} = \frac{1}{h^2} \sup_{d\in\mathbb{R}^{N+1}} \frac{d^\top M^{-1/2} K M^{-1/2} d}{d^\top d}. \tag{24}$$

The last term in (24) is the Rayleigh quotient corresponding to the symmetric and positive semi-definite matrix

$$\bar{K} = M^{-1/2} K M^{-1/2} \tag{25}$$

such that

$$\inf_{\substack{p\in\mathcal{P}_N\\ p'\neq 0}} \frac{\int_0^h p^2(t)\,\mathrm{d}t}{\int_0^h (p')^2(t)\,\mathrm{d}t} = \frac{1}{\lambda_N} h^2$$

with λ_N being the largest eigenvalue of the matrix $\bar{K}$ in (25). □

Lemma 3.5. *Let $\mu \geq 2$ and let $h > 0$ be sufficiently small. Then,*

$$\sup_{\substack{p\in X_n\\ p\neq 0}} \frac{1}{\|p\|^2_{H^1_D,h}} \sum_{i=2}^{\mu} \int_0^h 2p_i(t)p'_{i-1}(t)\,dt \leq \frac{1}{\sqrt{1+\psi(z_h)}},$$

with

$$z_h = \frac{1}{\lambda_N} h^2, \quad \psi(z_h) = \begin{cases} z_h & \text{if } \mu = 2, \\ z_h^2 - 2z_h^3 & \text{if } \mu = 3, \\ z_h^{\mu-1} - (\mu-2) z_h^{\mu} & \text{if } \mu \geq 4, \end{cases}$$

where λ_N is the largest eigenvalue of the matrix $\bar{K}$ in (25).

Proof. We denote

$$s := \sup_{\substack{p\in X_n\\ p\neq 0}} \frac{1}{\|p\|^2_{H^1_D,h}} \sum_{i=2}^{\mu} \int_0^h 2p_i(t)p'_{i-1}(t)\,\mathrm{d}t.$$

1. For $\mu = 2$, we have

$$s = \sup_{\substack{p\in X_n\\ p\neq 0}} \frac{\int_0^h 2p_2(t)p'_1(t)\,\mathrm{d}t}{\int_0^h \left(p_1^2(t) + p_2^2(t) + (p'_1)^2(t)\right)\mathrm{d}t}.$$

The definition of z_h implies

$$\int_0^h p_1^2(t)\,\mathrm{d}t \geq z_h \int_0^h (p'_1)^2(t)\,\mathrm{d}t$$

and, consequently,

$$
\begin{aligned}
&-\int_0^h 2\sqrt{1+z_h}p_2(t)p_1'(t)\,\mathrm{d}t + \int_0^h (p_1^2(t)+p_2^2(t)+(p_1')^2(t))\,\mathrm{d}t \\
&= \int_0^h p_1^2(t)\,\mathrm{d}t - z_h\int_0^h (p_1')^2(t)\,\mathrm{d}t + \int_0^h \left(p_2(t) - \sqrt{1+z_h}\,p_1'(t)\right)^2 \mathrm{d}t \geq 0.
\end{aligned}
$$

Finally,

$$
s \leq \frac{1}{\sqrt{1+z_h}} = \frac{1}{\sqrt{1+\psi(z_h)}}
$$

and the result follows.

2. For $\mu = 3$, we have

$$
s = \sup_{\substack{p\in X_n \\ p\neq 0}} \frac{\int_0^h 2p_2(t)p_1'(t) + 2p_3(t)p_2'(t)\,\mathrm{d}t}{\int_0^h (p_1^2(t)+p_2^2(t)+p_3^2(t)+(p_1')^2(t)+(p_2')^2(t))\,\mathrm{d}t}.
$$

Again, the definition of z_h implies

$$
\int_0^h p_1^2(t)\,\mathrm{d}t \geq z_h\int_0^h (p_1')^2(t)\,\mathrm{d}t \quad \text{and} \quad \int_0^h p_2^2(t)\,\mathrm{d}t \geq z_h\int_0^h (p_2')^2(t)\,\mathrm{d}t. \tag{26}
$$

It is not difficult to verify that for $\psi(z_h) = z_h^2 - 2z_h^3$ the estimate

$$
1 - \frac{1+\psi(z_h)}{1-\frac{1}{z_h}\psi(z_h)} \geq -z_h \tag{27}
$$

holds, if $z_h > 0$ is sufficiently small. This is the case if we choose a sufficiently small h, cf. Lemma 3.3. Using (26) and (27), we now conclude

$$
\begin{aligned}
&-\int_0^h 2\sqrt{1+\psi(z_h)}p_2(t)p_1'(t)\,\mathrm{d}t \\
&+\int_0^h (p_1^2(t) + (1-\tfrac{1}{z_h}\psi(z_h))p_2^2(t) + (p_1')^2(t))\,\mathrm{d}t \\
&\quad = \int_0^h p_1^2(t)\,\mathrm{d}t + \left(1 - \frac{1+\psi(z_h)}{1-\frac{1}{z_h}\psi(z_h)}\right)\int_0^h (p_1')^2(t)\,\mathrm{d}t \\
&\quad + \int_0^h \Big(\sqrt{1-\tfrac{1}{z_h}\psi(z_h)}p_2(t) - \frac{\sqrt{1+\psi(z_h)}}{\sqrt{1-\frac{1}{z_h}\psi(z_h)}}\,p_1'(t)\Big)^2\,\mathrm{d}t \\
&\quad \geq \int_0^h p_1^2(t)\,\mathrm{d}t - z_h\int_0^h (p_1')^2(t)\,\mathrm{d}t \geq 0
\end{aligned} \tag{28}
$$

and

$$
\begin{aligned}
&- \int_0^h 2\sqrt{1+\psi(z_h)}p_3(t)p_2'(t)\,\mathrm{d}t \\
&+ \int_0^h \left(\tfrac{1}{z_h}\psi(z_h)p_2^2(t) + p_3^2(t) + (p_2')^2(t)\right)\mathrm{d}t \\
&\quad = \tfrac{1}{z_h}\psi(z_h)\int_0^h p_2^2(t)\,\mathrm{d}t - \psi(z_h)\int_0^h (p_2')^2(t)\,\mathrm{d}t \\
&\quad + \int_0^h \left(p_3(t) - \sqrt{1+\psi(z_h)}\,p_2'(t)\right)^2 \mathrm{d}t \\
&\quad \geq \frac{\psi(z_h)}{z_h}\left(\int_0^h p_2^2(t)\,\mathrm{d}t - z_h\int_0^h (p_2')^2(t)\,\mathrm{d}t\right) \geq 0
\end{aligned} \tag{29}
$$

for a sufficiently small $h > 0$. Taking a sum of the inequalities (28) and (29) yields

$$
s \leq \frac{1}{\sqrt{1+\psi(z_h)}}
$$

for a sufficiently small $h > 0$.

3. For $\mu \geq 4$, we first introduce

$$
\nu_i(z) := \begin{cases} 1 & \text{if } i = 1, \\ z^{i-1} - (i-1)z^i & \text{if } 2 \leq i \leq \mu - 1, \\ 0 & \text{if } i = \mu. \end{cases}
$$

If $z_h > 0$ is sufficiently small then, for $\psi(z_h) = z^{\mu-1} - (\mu-2)z^\mu$, we obtain

$$
z_h\nu_{i-1}(z_h) \geq \frac{\psi(z_h) + \nu_i(z_h)}{1 - \nu_i(z_h)} \quad \text{for } i = 2, ..., \mu. \tag{30}
$$

From (30) and

$$
\int_0^h p_i^2(t)\,\mathrm{d}t \geq z_h\int_0^h (p_i')^2(t)\,\mathrm{d}t \quad \text{for } i = 2, ..., \mu,
$$

it follows

$$
\begin{aligned}
&- \int_0^h 2\sqrt{1+\psi(z_h)}p_i(t)p_{i-1}'(t)\,\mathrm{d}t \\
&+ \int_0^h \left(\nu_{i-1}(z_h)p_{i-1}^2(t) + (1-\nu_i(z_h))p_i^2(t) + (p_{i-1}')^2(t)\right)\mathrm{d}t \\
&\quad = \nu_{i-1}(z_h)\int_0^h p_{i-1}^2(t)\,\mathrm{d}t + \left(1 - \frac{1+\psi(z_h)}{1-\nu_i(z_h)}\right)\int_0^h (p_{i-1}')^2(t)\,\mathrm{d}t \\
&\quad + \int_0^h \Big(\sqrt{1-\nu_i(z_h)}p_i(t) - \frac{\sqrt{1+\psi(z_h)}}{\sqrt{1-\nu_i(z_h)}}\,p_{i-1}'(t)\Big)^2\,\mathrm{d}t \\
&\quad \geq \nu_{i-1}(z_h)\int_0^h p_{i-1}^2(t)\,\mathrm{d}t - z_h\nu_{i-1}(z_h)\int_0^h (p_{i-1}')^2(t)\,\mathrm{d}t \geq 0
\end{aligned}
$$

for $i = 2, ..., \mu$. Forming the sum over all inequalities, $i = 2, ..., \mu$, results in

$$
\begin{aligned}
&-\sqrt{1+\psi(z_h)}\sum_{i=2}^{\mu}\int_0^h 2p_i(t)p'_{i-1}(t)\,\mathrm{d}t + \sum_{i=1}^{\mu}\int_0^h p_i^2(t)\,\mathrm{d}t + \sum_{i=2}^{\mu}\int_0^h (p'_{i-1})^2(t)\,\mathrm{d}t \\
&\quad = -\sum_{i=2}^{\mu}\int_0^h 2\sqrt{1+\psi(z_h)}p_i(t)p'_{i-1}(t)\,\mathrm{d}t \\
&\quad + \sum_{i=2}^{\mu}\int_0^h (\nu_{i-1}(z_h)p_{i-1}^2(t) + (1-\nu_i(z_h))p_i^2(t) + (p'_{i-1})^2(t))\,\mathrm{d}t \geq 0
\end{aligned}
$$

and finally

$$
s \leq \frac{1}{\sqrt{1+\psi(z_h)}}
$$

if $h > 0$ is sufficiently small.

□

The following theorem follows immediately by combining (22), Lemma 3.5, and Lemma 3.1.

Theorem 3.6. *Let the DA operator T associated with a pure Jordan chain have index $\mu \geq 2$ and let $h > 0$ be sufficiently small. Then,*

$$
\gamma_n^2 \geq \gamma_{n,h}^2 = \inf_{\substack{p\in X_n \\ p\neq 0}} \frac{\|Tp\|_{L^2,h}^2}{\|p\|_{H_D^1,h}^2} \geq 1 - \frac{1}{\sqrt{1+\psi(z_h)}},
$$

with

$$
z_h = \frac{1}{\lambda_N}h^2 \quad \text{and} \quad \psi(z_h) = \begin{cases} z_h & \text{if } \mu = 2, \\ z_h^2 - 2z_h^3 & \text{if } \mu = 3, \\ z_h^{\mu-1} - (\mu-2)z_h^{\mu} & \text{if } \mu \geq 4, \end{cases}
$$

where λ_N is the largest eigenvalue of the matrix $\bar{K}$ given in (25). In particular, there is a positive constant c such that

$$
\gamma_n \geq ch^{\mu-1}.
$$

We emphasize that the estimates in Theorem 3.6 are not sharp. In the case of linear ansatz functions, the estimates can be further improved to provide a threshold γ_n being independent of the index μ.

Theorem 3.7. *Let the Jordan chain operator have arbitrary index $\mu \geq 2$ and $N = 1$. Then, for sufficiently small $h > 0$, the following estimate holds:*

$$
\gamma_n^2 \geq \gamma_{n,h}^2 = \inf_{\substack{p\in X_n \\ p\neq 0}} \frac{\|Tp\|_{L^2,h}^2}{\|p\|_{H_D^1,h}^2} = 1 - \frac{1}{\sqrt{1+h^2/\lambda_1}} \geq \frac{1}{2\lambda_1}h^2,
$$

where $\lambda_1 = 12$ is the largest eigenvalue of the matrix $\bar{K}$ specified (25) *and $N = 1$.*

Proof. Let

$$s := \sup_{\substack{p \in X_n \\ p \neq 0}} \frac{1}{\|p\|^2_{H^1_D, h}} \sum_{i=2}^{\mu} \int_0^h 2 p_i(t) p'_{i-1}(t) \, \mathrm{d}t.$$

and let us denote by φ_0 and φ_1 the zero and first order Legendre polynomials scaled to $[0, h]$. Then, we have the following representation for $p \in X_n$ and $t \in [0, h]$:

$$\begin{aligned} p_i(t) &= c_{i0}\varphi_0(t) + c_{i1}\varphi_1(t), \quad i = 1, \ldots, \mu - 1, \\ p_\mu(t) &= c_{\mu 0}\varphi_0(t). \end{aligned}$$

Denote $c := (c_{10}, \ldots, c_{\mu 0}, c_{11}, \ldots, c_{\mu-1,1})$. Since $p'_i(t) = c_{i1} \frac{2}{\|\varphi_0\|^2_h} \varphi_0(t)$, $i = 1, \ldots, \mu - 1$, it holds

$$\begin{aligned} S_1(c) &= \sum_{i=2}^{\mu} \int_0^h 2 p_i(t) p'_{i-1}(t) \, \mathrm{d}t = \sum_{i=2}^{\mu} 4 c_{i0} c_{i-1,1}, \\ S_2(c) &= \|p\|^2_{H^1_D, h} = h \sum_{i=1}^{\mu} c^2_{i0} + \left(\frac{h}{3} + \frac{4}{h} \right) \sum_{i=1}^{\mu-1} c^2_{i1}. \end{aligned}$$

In order to determine s, we compute the derivatives of $g(c) = S_1(c)/S_2(c)$ and obtain

$$\begin{aligned} &\frac{\partial S_1(c)}{\partial c_{i0}} = 4 c_{i-1,1}, i = 2, \ldots, \mu, \quad \frac{\partial S_1(c)}{\partial c_{10}} = 0, \\ &\frac{\partial S_1(c)}{\partial c_{i1}} = 4 c_{i+1,0}, i = 1, \ldots, \mu - 1, \\ &\frac{\partial S_2(c)}{\partial c_{i0}} = 2 h c_{i0}, i = 1, \ldots, \mu, \quad \frac{\partial S_2(c)}{\partial c_{i1}} = 2 \left(\frac{h}{3} + \frac{4}{h} \right) c_{i1}, i = 1, \ldots, \mu - 1. \end{aligned}$$

This yields

$$\begin{aligned} \frac{\partial g(c)}{\partial c_{i1}} &= \frac{1}{S_2(c)^2} \left(4 c_{i+1,0} S_2(c) - 2 \left(\frac{h}{3} + \frac{4}{h} \right) c_{i1} S_1(c) \right), i = 1, \ldots, \mu - 1, \\ \frac{\partial g(c)}{\partial c_{i0}} &= \frac{1}{S_2(c)^2} \begin{cases} 4 c_{i-1,1} S_2(c) - 2 h c_{i0} S_1(c), & i = 2, \ldots, \mu, \\ -2 h c_{10} S_1(c), & i = 1. \end{cases} \end{aligned} \tag{31}$$

Setting the expressions in (31) to zero, we arrive at, $i = 2, \ldots, \mu$,

$$2 c_{i,0} S_2(c) = \left(\frac{h}{3} + \frac{4}{h} \right) c_{i-1,1} S_1(c), \quad 2 c_{i-1,1} S_2(c) = h c_{i0} S_1(c).$$

Finally, we have

$$c_{i-1,1} = \frac{1}{2} h \frac{S_1(c)}{S_2(c)} c_{i0} \tag{32}$$

and

$$c_{i,0} S_2(c) = (1 + h^2/12) \frac{S_1^2(c)}{S_2(c)}. \tag{33}$$

If $S_1(c) = 0$, then $S_2(c) = 0$ follows and this contradicts $S_2(c) \neq 0$. Therefore, it holds $S_1(c) \neq 0$ and we obtain $c_{10} = 0$ from (31). There exists an index i_0 with $c_{i_0 0} \neq 0$. Otherwise, (32) would provide $c_{i1} = 0$ for all $i = 1, \ldots, \mu - 1$ and hence, $S_2(c) = S_1(c) = 0$. Using (33) for $i = i_0$, we conclude

$$\frac{S_1^2(c)}{S_2^2(c)} = \frac{1}{4(h^2 + 12)h^2}$$

which results in

$$s = \left(1 + \frac{h^2}{12}\right)^{-1/2}.$$

Finally, we apply the inequality $1 - (1 + w)^{-1/2} \geq \frac{1}{2}w$ which holds for all sufficiently small $w > 0$ and set $w = h^2/\lambda_1$. This completes the proof. □

Remark 3.8. *It should be again emphasized that the estimates in the previous theorem do not depend on the index. In order to further investigate this issue, we carried out numerical computations. By expressing the term*

$$\frac{\|Tp\|^2_{L^2,h}}{\|p\|^2_{H^1_D,h}}$$

via Legendre polynomials, the determination of $\gamma_{n,h}$ can be reduced to a matrix eigenvalue problem. We solved this problem numerically for $N \leq 5$ and $\mu \leq 6$ and observed inequalities of the form

$$\gamma_{n,h} \geq c\, h^{\min(N,\mu-1)},$$

which is consistent with statements formulated in Theorem 3.6 *for $N > \mu$ and in Theorem* 3.7 *for $N = 1$.*

4. Convergence Estimations

Now, the convergence properties of the least-squares method can be derived. Let the function set X_n, related to the uniform partition

$$a = t_0 < t_1, \ldots < t_n = b$$

with stepsize $h = (b - a)/n$ and $N \geq 1$, be given by (4). Following ideas developed in Subsection 2.2, we need the estimates for the approximation errors α_n and β_n. We have

$$\alpha_n = \|x^\dagger - P_n x^\dagger\|_{H^1_D}, \quad P_n x^\dagger = \operatorname{argmin}\{\|x^\dagger - p\|_{H^1_D} : p \in X_n\},$$

and $\beta_n = \|\mathcal{T}(x^\dagger - P_n x^\dagger)\|_{L^2} \leq \|\mathcal{T}\|\alpha_n$. Let the solution $x^\dagger$ be sufficiently smooth so that the interpolation function $p_{int} \in X_n$ for $x^\dagger$ is well-defined by N interpolation nodes on each subinterval of the partition and, additionally, by $Dp_{int}(a) = Dx^\dagger(a)$. Then, standard interpolation results provide the estimate

$$\alpha_n \leq \|p_{int} - x^\dagger\|_{H^1_D} \leq c_\alpha\, h^N.$$

Theorem 4.1. *Let the BVP (1)–(2) with $\mu \geq 1$ satisfy the assumptions of Theorem* 2.5(1)*. Let the unique solution $x^\dagger$ of the BVP be sufficiently smooth.*
Moreover, let a constant c exist so that the instability threshold γ_n satisfies the inequality

$$\gamma_n \geq c\, h^{\min(N,\mu-1)} \tag{34}$$

for all sufficiently small stepsizes $h > 0$. Then the following statements hold:

(1) *The approximate solution*

$$x_n = \operatorname{argmin}\{\|A(Dp)' + Bp - q\|_{L^2}^2 + |G_a p(a) + G_b p(b) - \gamma|^2 : p \in X_n\}$$

satisfies the inequality $\|x_n - x^\dagger\|_{H_D^1} = O(h^{\max(0,N-\mu+1)})$.

(2) *If, additionally, the entries of the coefficients A and B are polynomials at most of degree $N_{A,B}$ and M is chosen in such a way that $M \geq 1 + N + N_{A,B}$, then the least-squares collocation solution x_n^δ of the overdetermined system (14)–(15) satisfies*

$$\|x_n^{\delta_n} - x^\dagger\|_{H_D^1} = O(h^{\max(0,N-\mu+1)}).$$

In particular, in both cases the discrete solutions remain bounded in $H_D^1(a,b)$ and the choice of N such that $\mu - 1 < N$, ensures $x_n \to x^\dagger$ and $x_n^{\delta_n} \to x^\dagger$, in $H_D^1(a,b)$, respectively.

Proof. The results in (1) and (2) follow from the estimate (12) for $\delta_n = 0$ and for $\delta_n = O(h^M)$, respectively. See also (17) and Proposition 2.7. □

We emphasize that assumption (34) is analytically verified for several problem classes in Section 3. Moreover, it is also justified by numerical eigenvalue computations, see Remark 3.8.
Note that Examples 1.1 and 1.2 do not belong to the class of DAEs described in Subsection 3.2 with a constant transformation matrix K. Nevertheless, they also show the orders predicted in Theorem 4.1.

5. Numerical Experiments

First, in Subsection 5.1, we intend to illustrate the results derived in Sections 3.3. To this end, the convergence order for the cases $\mu = 3$ and $\mu = 4$ is numerically evaluated. Then, we reconsider Examples 1.1 and 1.2. All experiments have been carried out in MATLAB.

5.1. Jordan Systems

Here, we consider the DAEs investigated in Section 3.3, namely,

$$A(Dx)'(t) + Bx(t) = q(t), \quad t \in [0,1],$$

with $B = I \in \mathbb{R}^{\mu\times\mu}$, $D = \operatorname{diag}(1,\dots,1,0) \in \mathbb{R}^{\mu\times\mu}$, and $A = -J \in \mathbb{R}^{\mu\times\mu}$, where

$$J = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix},$$

for $\mu = 3$ and $\mu = 4$. The right-hand side q is chosen such that the exact solution of the system is

$$x(t) = \begin{cases} \begin{bmatrix} e^{-t}\sin t \\ e^{-2t}\sin t \\ e^{-t}\cos t \end{bmatrix}, & \mu = 3, \\ \begin{bmatrix} e^{-t}\sin t \\ e^{-2t}\sin t \\ e^{-t}\cos t \\ e^{-2t}\cos t \end{bmatrix}, & \mu = 4. \end{cases}$$

Following variants of the least-squares collocation method have been used:

- The degree N of the ansatz function varied between 1 and 6.
- The collocation points, ρ_i, $i = 1,\dots,N$, were either uniformly distributed on $[0,1]$ or Gaussian points.
- The number ν of additional collocation points per subinterval was either $N+1$ or 1.

  $N+1$: The additional collocation points were the midpoints between two subsequent collocation points ρ_i, including $\rho = 0$ and $\rho = 1$. This corresponds to the choice (7).

  1 : The additional collocation point was located in the center of each subinterval. If this point already belongs the set specified by ρ_i, the midpoint between ρ_i and ρ_{i+1}, was chosen.
- The least-square solution was computed using both criteria, (21) and (20), which is indicated in the tables by L^2 and $\mathbb{R}$, respectively (cf. Proposition 2.7).

To determine the order of convergence, the problems were solved on grids with constant stepsizes, corresponding to $n = 2, 4, \dots, 64$ subintervals. The error was measured in the $H^1_D(0,1)$-norm. Because of the special structure of the matrices, the $L^2(0,1)$-norm of the error behaves similarly. The results are presented in Tables 5 and 6.

We can draw the following conclusions:

- The collocation solutions stay bounded. Even for high index and polynomials of low degrees, we do not observe the divergence behavior which was typical for the standard collocation approach.

Table 5: DAE in Jordan form with index $\mu = 3$: Numerical estimates for the convergence order. The column headers mean: Theory = estimate according to Proposition 4.1, L^2 = least-square collocation according to (18) and $\mathbb{R}$ = least-square collocation according to (20). The N collocation points are either uniform (Nu) or Gaussian (Ng).

		$\nu = N + 1$		$\nu = 1$	
	Theory	L^2	$\mathbb{R}$	L^2	$\mathbb{R}$
1u	0	0.2	0.1	0.2	0.1
1g	0	0.2	0.1	0.2	0.1
2u	0	1.0	1.0	1.0	1.0
2g	0	1.0	1.0	1.0	1.0
3u	1	2.1	2.1	2.1	2.1
3g	1	2.1	2.1	2.0	2.0
4u	2	3.0	3.0	3.2	3.1
4g	2	3.0	3.0	3.1	3.1
5u	3	4.0	4.1	4.2	4.2
5g	3	4.1	4.1	4.2	4.2
6u	4	4.5	4.8	4.8	5.0
6g	4	5.1	5.0	5.1	5.1

- The order of convergence is mostly one order higher than predicted by the theory. We do not have a satisfactory explanation for that since some of our stability estimates are in fact sharp.
- There is no significant difference between Gaussian and uniform collocation points.
- The order of convergence is independent of the number ν of additional collocation points.
- If the $L^2(0, 1)$-norm in the image space is replaced by the Euclidean norm in $\mathbb{R}^{\mu Mn}$ using the norm equivalence (19), the order of convergence is retained. This property is slightly surprising, since the Euclidean norm can be interpreted as only a first order approximation to the norm of $L^2(0, 1)$.
- For $\mu = 4$ and $N = 2$, non-integer orders of convergence arise.
- If the number of subintervals n is further increased, we observe growing errors, especially for polynomials of higher degrees. As expected, the conditioning of the least-squares problem becomes worse if the index and the polynomial degree increase.

5.2. Examples 1.1 and 1.2 Revisited

In this subsection, we collect the results of the experiments specified in Subsection 5.1 and carried out for Examples 1.1 and 1.2. We calculated the numerical orders of

Table 6: DAE in Jordan form with index $\mu = 4$: Numerical estimates for the convergence order. The column headers mean: Theory = estimate according to Proposition 4.1, L^2 = least-square collocation according to (18) and $\mathbb{R}$ = least-square collocation according to (20). The N collocation points are either uniform (Nu) or Gaussian (Ng).

		$\nu = N + 1$		$\nu = 1$	
	Theory	L^2	$\mathbb{R}$	L^2	$\mathbb{R}$
1u	0	0.1	0.0	0.0	0.0
1g	0	0.1	0.0	0.0	0.0
2u	0	0.4	0.4	0.4	0.4
2g	0	0.4	0.4	0.4	0.3
3u	0	1.1	1.1	1.1	1.1
3g	0	1.1	1.1	1.1	1.1
4u	1	2.1	2.1	2.1	2.1
4g	1	2.1	2.1	2.1	2.1
5u	2	2.6	2.7	3.2	3.1
5g	2	2.9	2.9	3.1	3.1
6u	3	3.5	3.7	4.3	3.6
6g	3	4.3	4.3	4.1	4.2

Table 7: Example 1.1: Error of the collocation solution for $\eta = -25$, $\lambda = -1$, and $N = 4$. The collocation points ρ_i are Gauss-Legendre nodes and the error is measured in the $H_D^1(0,1)$-norm.

n	L^2	order	$\mathbb{R}$	order
20	1.09e-007	0.0	1.36e-007	0.0
40	1.03e-008	3.3	1.70e-008	2.9
80	1.08e-009	3.2	1.89e-009	3.1
160	1.37e-010	3.0	1.79e-010	3.4
320	8.64e-011	0.7	2.04e-011	3.1
640	2.11e-010	-1.3	2.17e-011	-0.1

convergence using the norm of $L^2(0,1)$, cf. (21) and $\mathbb{R}^{3Mn}$, cf. (20), in the image space. All results can be found in Tables 7 and 8.
Again, we can see that the numerically estimated order of convergence is higher than expected in view of the theory.

6. Further References and Conclusions

In this section, we indicate how to treat more general classes of DAEs. Moreover, we list references which are seemingly related to our approach and finally, provide the conclusions.

Table 8: Example 1.2: Error of the collocation solution for $\eta = -2$ and $N = 3$. The collocation points ρ_i are uniformly distributed and the error is measured in the $H^1_D(0,1)$-norm.

n	L^2	order	$\mathbb{R}$	order
20	8.43e-005	0.0	8.65e-005	0.0
40	1.94e-005	2.0	1.97e-005	2.1
80	4.65e-006	2.0	4.70e-006	2.0
160	1.14e-006	2.0	1.15e-006	2.0
320	2.83e-007	2.0	2.83e-007	2.0
640	9.55e-008	1.6	7.04e-008	2.0

6.1. More General DAEs

Standard form DAEs,

$$Ex' + Fx = q,$$

can be reformulated into DAEs with properly stated leading terms by appropriate factorizations $E = AD$, cf. [14, Sections 2.7 and 7.2] and [18, Theorem 3.1].
General regular DAEs with properly stated leading term,

$$A(Dx)' + Bx = q,$$

can be rewritten in enlarged form,

$$Au' + Bx = q, \quad u - Dx = 0,$$

such that the latter meets the above request for partitioned variables, see [14, Section 3.11] and [15].

6.2. Further References

Least-squares collocation methods for DAEs have been already developed in different contexts. In [8], a method which corresponds to a discretization of the image space was introduced. The discrete solution was represented using a basis constructed via the reproducing kernels of certain Sobolev spaces, similarly to the approach in [9].
In [4], the pre-image space is discretized in a way similar to the one used in this paper. However, the number M of collocation points per subinterval is chosen to be *lower* than the degree N of the piecewise polynomials. The additional degrees of freedom are fixed by choosing the polynomial of minimal norm among all ansatz function satisfying the collocation conditions. No proofs of convergence are given.
In the continuous Sobolev gradient method discussed in [19], the expression $\|Tx - q\|_{L^2}$ is minimized by solving the Hilbert space valued ODE,

$$\varphi'(\tau) = -T^*T\varphi(\tau) + T^*q, \quad \varphi(0) = x_0,$$

and using finite difference schemes. If $q \in \operatorname{im} T$ then $\varphi(t)$ converges to some x_* with $Tx_* = q$, where x_* is the nearest solution to the initial guess x_0. The convergence of the discretization is not studied.

Adding to a given DAE the differentiated version of the constraint may result in a lower-index overdetermined DAE, e.g. [14, Proposition 10.8]. Mainly in multibody dynamics simulation, it is quite common to consider overdetermined DAEs comprising the originally given DAE and derivatives of the constraints. Then, the resulting overdetermined DAE is integrated by applying a special Gauß–Newton method at each integration step.
A first implementation of this idea is realized in the code ODASSL [6, 7]. We emphasize, that our approach is entirely different. In particular, we do not at all apply derivative array systems. In connection with Taylor series methods, the derivative array system is interpreted as an underdetermined system solved by least-squares techniques, e.g. [3, 5]. Likewise, those ideas are also basically different from ours.

6.3. Conclusions

In the present paper, we develope a new method for the numerical solution of boundary value problems in linear higher-index DAEs. The motivation for this method originates from the fact that higher-index DAEs are problems which are essentially ill-posed problems in natural topologies. We provide the corresponding functional analytic setting.
The basic idea of the proposed numerical method is the approximation of such a problem by a least-squares method where both, the image and the pre-image space are discretized. In the context of DAEs, this idea results in an extremely simple algorithm whose computational complexity is comparable to standard polynomial collocation methods for systems of ordinary differential equations. In particular, neither analytical preprocessing nor special structures of the DAE is necessary. In the numerical experiments, the method behaves in a robust and stable way, showing fast convergence. In our opinion, treating the DAEs as ill-posed problems is a fruitful approach and this idea deserves further research interest.

Here, only partial stability and convergence results are provided. Having a more complete theory available, would enable the construction of efficient codes for higher-index DAE systems.

Appendix A. Basic Facts About Legendre Polynomials

Legendre polynomials are defined by the following recursion formula:

$$\begin{aligned} P_0(x) &= 1, \\ P_1(x) &= x, \\ (n+1)P_{n+1}(x) &= (2n+1)xP_n(x) - nP_{n-1}(x). \end{aligned}$$

They have the following properties for all $n, m \geq 0$:

$$\begin{aligned} \operatorname{lin}\{P_0, \ldots, P_n\} &= \mathcal{P}_n, \\ \int_{-1}^{1} P_n(x) P_m(x) dx &= \frac{2}{2n+1} \delta_{nm}, \\ \frac{d}{dx} P_{n+1}(x) &= (2n+1) P_n(x) + (2(n-2)+1) P_{n-2}(x) + \cdots \\ &= \frac{2P_n(x)}{\|P_n\|^2} + \frac{2P_{n-2}(x)}{\|P_{n-2}\|^2} + \cdots \end{aligned}$$

Here, we denoted by $\|\cdot\|$ the norm in $L^2(-1, 1)$. More details and proofs can be found in, e.g., [22, Chapter IV.1], [23], [20, p. 63].
We are interested in using these polynomials on the interval $[0, h]$. Therefore, let

$$\varphi_n(t) = P_n\Big(\frac{2}{h} t - 1\Big).$$

Then,

$$\begin{aligned} \int_0^h \varphi_n(t) \varphi_m(t) dt &= \frac{h}{2n+1} \delta_{nm}, \\ \frac{d}{dt} \varphi_{n+1}(t) &= 2\left(\frac{\varphi_n(t)}{\|\varphi_n\|_h^2} + \frac{\varphi_{n-2}(t)}{\|\varphi_{n-2}\|_h^2} + \cdots \right) \end{aligned}$$

follows. In the last line, we denoted by $\|\cdot\|_h$ the norm in $L^2(0, h)$.

References

[1] U.M. Ascher, R.M.M. Mattheij, and R.D. Russell. *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*. Prentice Hall, Englewood Cliffs, New Jersey, 1988.

[2] U.M. Ascher and R. Spiteri. Collocation software for boundary value differential-algebraic equations. *SIAM J. Sci. Comput.*, 15:938–952, 1994.

[3] A. Barrlund. Constrained least squares methods for linear timevarying DAE systems. *Numer. Math.*, 60:145–161, 1991.

[4] M. V. Bulatov, N. P. Rakhvalov, and L. S. Solovarova. Numerical solution of differential-algebraic equations using the spline collocation-variation method. *Computational Mathematics and Mathematical Physics*, 53(3):284–295, 2013.

[5] S. L. Campbell. The numerical solution of higher index linear time varying singular systems of differential equations. *SIAM J. Sci. and Stat. Comput.*, 6:334–348, 1985.

[6] E. Eich-Soellner and C. Führer. *Numerical Methods in Multibody Dynamics*. B. G. Teubner Stuttgart, 1998.

[7] C. Führer. *Differential-algebraische Gleichungssysteme in mechanischen Mehrkörpersystemen. Theorie, numerische Ansätze und Anwendungen*. PhD thesis, Technische Universität München, 1988.

[8] V. K. Gorbunov, V. V. Petrischev, and V. Y. Sviridov. Development of normal spline method for linear integro-differential equations. In P. M. A. Sloot, D. Abramson, A. Bogdanov, J. J. Dongarra, A. Zomaya, and Y. Gorbachev, editors, *International Conference Computational Science — ICCS 2003*, volume 2658 of *Lecture Notes in Computer Science*, pages 492–499. Springer Berlin Heidelberg, 2003.

[9] M. Hanke. On a least-squares collocation method for linear differential-algebraic equations. *Numer. Math.*, 54:79–90, 1988.

[10] M. Hanke. Linear differential-algebraic equations in spaces of integrable functions. *J. Differential Equations*, 79:14–30, 1989.

[11] M. Hanke, E. Izquierdo Macana, and R. März. On asymptotics in case of linear index-2 differential-algebraic equations. *SIAM J. Numer. Anal.*, 35:1326–1346, 1998.

[12] B. Kaltenbacher and J. Offtermatt. A convergence analysis of regularization by discretization in preimage space. *Math. Comp.*, 81(280):2049–2069, 2012.

[13] P. Kunkel, V. Mehrmann, and R. Stöver. Symmetric collocation methods for unstructured nonlinear differential-algebraic equations of arbitrary index. *Numer. Math.*, 98:277–304, 2004.

[14] R. Lamour, R. März, and C. Tischendorf. *Differential-Algebraic Equations: A Projector Based Analysis*. Differential-Algebraic Equations Forum. Springer-Verlag Berlin Heidelberg New York Dordrecht London, 2013. Series Editors: A. Ilchman, T. Reis.

[15] R. Lamour, R. März, and E. Weinmüller. *Surveys in Differential-Algebraic Equations III*, chapter Boundary-Value Problems for Differential-Algebraic Equations: A Survey, pages 177–309. Differential-Algebraic Equations Forum. Springer Heidelberg, 2015. ed. by A. Ilchmann and T. Reis.

[16] R. März. On correctness and numerical treatment of boundary value problems in DAEs. *Zhurnal Vychisl. Matem. i Matem. Fiziki*, 26(1):50–64, 1986.

[17] R. März. Numerical methods for differential-algebraic equations. *Acta Numerica*, 1:141–198, 1992.

[18] R. März. *Surveys in Differential-Algebraic Equations II*, chapter Differential-Algebraic Equations from a Functional-Analytic Viewpoint: A Survey, pages 163–285. Differential-Algebraic Equations Forum. Springer Heidelberg, 2015. ed. by A. Ilchmann and T. Reis.

[19] R. Nittka and M. Sauter. Sobolev gradients for differential algebraic equations. *Electronic Journal of Differential Equations*, 2008(42):1–31, 2008.

[20] T.J. Rivlin. *An introduction to the approximation of functions*. Dover publications, 1969.

[21] R. Stöver. *Numerische Lösung von linearen differential-algebraischen Randwertproblemen*. PhD thesis, Universität Bremen, January 1999. Doctoral thesis, Logos Verlag Berlin.

[22] P. K. Suetin. *Classical Orthogonal Polynomials (in Russian)*. Nauka Moskva, 2nd edition, 1979.

[23] P.K. Suetin (originator). Legendre polynomials. In M. Hazewinkel, editor, *Encyclopedia of Mathematics*. Springer, 2011. ISBN 1402006098, `https://www.encyclopediaofmath.org/index.php/Legendre_polynomials`.