MATHEMATICS OF COMPUTATION
Volume 88, Number 318, July 2019, Pages 1647–1683
https://doi.org/10.1090/mcom/3393
Article electronically published on November 5, 2018

LEAST-SQUARES COLLOCATION FOR HIGHER-INDEX LINEAR DIFFERENTIAL-ALGEBRAIC EQUATIONS: ESTIMATING THE INSTABILITY THRESHOLD

MICHAEL HANKE, ROSWITHA MÄRZ, AND CAREN TISCHENDORF

Abstract. Differential-algebraic equations with higher-index give rise to essentially ill-posed problems. The overdetermined least-squares collocation for differential-algebraic equations which has been proposed recently is not much more computationally expensive than standard collocation methods for ordinary differential equations. This approach has displayed impressive convergence properties in numerical experiments, however, theoretically, till now convergence could be established merely for regular linear differential-algebraic equations with constant coefficients. We present now an estimate of the instability threshold which serves as the basic key for proving convergence for general regular linear differential-algebraic equations.

1. Introduction

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

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

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

Here, $x : [a,b] \to \mathbb{R}^m$ denotes the unknown vector-valued function, $[a,b] \subset \mathbb{R}$ is a finite interval, the right-hand side $y : [a,b] \to \mathbb{R}^m$ is a sufficiently smooth 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. We focus on DAEs featuring partitioned variables by assuming a constant matrix function $D \in \mathbb{R}^{k\times m}$ and the leading term of the special form,

$$D = [I\ 0], \quad \operatorname{rank} D = k, \quad \operatorname{rank} A(t) = k, \quad t \in [a,b]. \tag{3}$$

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 are the *nondifferentiated* ones, traditionally called the *algebraic* components. We emphasize that no derivatives of the algebraic components appear in the DAE.

If the matrix $D \in \mathbb{R}^{k\times m}$ is slightly more general in the sense that there is a permutation matrix $M_{perm} \in \mathbb{R}^{m\times m}$ such that $DM_{perm} = [I\ 0]$, our analysis

Received by the editor October 9, 2017, and, in revised form, March 16, 2018, and May 18, 2018.

2010 *Mathematics Subject Classification.* Primary 65L80; Secondary 65L08, 65L10.

Key words and phrases. Differential-algebraic equation, higher-index, essentially ill-posed problem, collocation, boundary value problem, initial value problem.

The work of the third author was supported by the German Research Foundation (DFG CRC/Transregio 154, Subproject C02).

applies analogously, however, the description would become even more involved. More general DAEs showing a time-varying matrix function D can be treated by reformulating them into an appropriate semi-explicit form, for instance,

$$\begin{aligned} A(t)u'(t) + B(t)x(t) &= y(t), \\ u(t) - D(t)x(t) &= 0, \quad t \in [a\ b]; \end{aligned}$$

see [3, Subsection 5.1] for more details.

Moreover, $G_a, G_b \in \mathbb{R}^{l \times m}$ and $r \in \mathbb{R}^l$. Thereby, l is the dynamical degree of freedom of the DAE, that is, the number of free parameters of the general solution of the DAE (e.g., [6, Section 2], [5, Section 2.6]), which can be fixed by initial and boundary conditions. Initial value problems (IVPs) are incorporated by $G_b = 0$. We suppose $0 \leq l \leq k < m$. If $l = 0$, then there are no free parameters and no boundary condition will be given.

As in [3], we put the problem in a Hilbert space setting and consider generalized solutions $x \in H^1_D$,

$$\begin{aligned} H^1_D &:= H^1_D((a,b),\mathbb{R}^m) := \{x \in L^2 : Dx \in H^1\}, \\ L^2 &:= L^2((a,b),\mathbb{R}^m), \\ H^1 &:= H^1((a,b),\mathbb{R}^k), \end{aligned}$$

satisfying the condition (2) as well as the DAE (1) for a.e. $t \in (a,b)$. To ensure that the expression $G_a x(a) + G_b x(b)$ is well-defined for all $x \in H^1_D$, we restrict the boundary conditions by assuming

$$\ker G_a \supseteq \ker D, \quad \ker G_b \supseteq \ker D. \tag{4}$$

Then $G_a x(a) + G_b x(b) = G_a D^+ Dx(a) + G_b D^+ Dx(b)$ is well-defined together with $Dx(a), Dx(b)$. The latter expressions are well-defined since $Dx \in H^1$ and the evaluation of functions from H^1 at a certain point makes sense. Actually, condition (4) implies that the boundary condition (2) applies to the first k components of x only.

The so-called *classical* or *standard* 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 [1,6] for the respective comprehensive surveys. In contrast, first attempts to treat linear higher-index DAEs via a simple[1] overdetermined polynomial collocation system and a least-squares linear solver – referred to as *least-squares collocation* – are reported in [3].

Here, we follow the ansatz from [3]. We approximate the differentiated components by continuous piecewise polynomial functions of a certain degree and the algebraic components by generally discontinuous piecewise polynomial functions, whose degree is lower by one. More precisely, we consider the partition of the interval $[a,b]$,

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

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

[1] We simply use more collocation points than in standard methods, but we do not apply derivative array systems.

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

$$\begin{aligned} X_\pi = \{p \in H_D^1 : p_\kappa|_{[t_{j-1},t_j)} \in \mathcal{P}_N,\ \kappa = 1, \dots, k,\ \ j = 1, \dots, n, \\ p_\kappa|_{[t_{j-1},t_j)} \in \mathcal{P}_{N-1},\ \kappa = k+1, \dots, m,\ \ j = 1, \dots, n\}. \end{aligned} \tag{5}$$

Since X_π has dimension $Nmn + k$, $Nmn + k$ conditions are necessary to uniquely determine $p \in X_\pi$. The standard collocation methods work with N collocation points on each subinterval. In contrast, as first proposed in [3], we specify $M > N$ least-squares collocation points by choosing values

$$0 < \tau_1 < \cdots < \tau_M < 1, \tag{6}$$

and setting

$$S_j := \{t_{j-1} + \tau_i h_j,\ i = 1, \dots, M\}, \quad h_j = t_j - t_{j-1}, \quad j = 1, \dots, n.$$

In order to determine the discrete solution $p \in X_\pi$, we solve the overdetermined system directly using the original BVP,

$$A(t)(Dp)'(t) + B(t)p(t) = y(t), \quad t \in S_j, \quad j = 1, \dots, n, \tag{7}$$

$$G_a p(a) + G_b p(b) = r \tag{8}$$

in the least-squares sense. This means that we seek a minimal element of the sum of squares

(9)

$$\phi_{\pi,M}(p) := \sum_{j=1}^{n} \frac{h_j}{M} \sum_{t \in S_j} |A(t)(Dp)'(t) + B(t)p(t) - y(t)|^2 + |G_a p(a) + G_b p(b) - r|^2,$$

which represents a nonnegative functional defined on X_π and which will be regarded as approximation of the functional

$$\begin{aligned} \phi_\pi(p) :=& \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} |A(t)(Dp)'(t) + B(t)p(t) - y(t)|^2 \mathrm{dt} + |G_a p(a) + G_b p(b) - r|^2 \\ =& \|A(Dp)' + Bp - y\|_{L^2}^2 + |G_a p(a) + G_b p(b) - r|^2,\ p \in X_\pi. \end{aligned} \tag{10}$$

IVPs and BVPs are treated in the same way, more precisely, IVPs are treated as BVPs. In the present context, it is a secondary matter whether we have initial conditions or boundary conditions. The crux of the matter is the treatment of the DAE itself.

Any element $p \in X_\pi$ is given by $mnN + k$ parameters. The choice $M = N$ corresponds to standard polynomial collocation methods yielding $mnN + l$ equations. These methods work well for regular ODEs and index-1 DAEs, in which $l = k = m$ and $l = k < m$, respectively (see [1,6]). Then the system (7),(8) is uniquely solvable. In contrast, since the dynamical degree l of higher-index DAEs is always less than k, the system (7),(8) with $M = N$ becomes necessarily underdetermined, notwithstanding that the given BVP is uniquely solvable. For uniqueness of the discrete solution, one has to add $k - l > 0$ additional equations. Nevertheless, as it is well known, even if we do so in a reasonable manner, the classical collocation method fails to work well; see, e.g., Example 1.1 below.

Since we are interested in higher-index DAEs, we always put $M > N$ and we do not apply the standard collocation, except for the comparison experiment in Example 1.1. As a matter of course, the choice $M > N$ goes along with an overdetermined system (7),(8) comprising more equations than unknowns. Naturally, we treat such a system in a least-squares sense. At this place we emphasize that the resulting numerical procedure, that is, the overdetermined least-squares collocation, is in essence as cheap as the classical polynomial collocation, we merely replace the linear solver by a least-squares solver. This is a quite simple idea, but we are impressed by the numerical results, a few of which are reported in [3]. Possibly, a direct discretization method and corresponding software for higher-index DAEs will be created herefrom in the near future. This would require much less effort and could become much more convenient than the existing methods each of which utilizes derivative array systems. However, so far, we are merely in the early stages. The mathematics of this computational project appears to be quite challenging. The very first ideas reported in [3] apply to constant coefficient DAEs only. In the present paper we offer a basic proof for BVPs involving linear arbitrary-index DAEs with time-varying coefficients.

To demonstrate the great potential of the overdetermined least-squares collocation we next adopt one of the experiments from [3]. The related DAE is known to cause serious difficulties and failures in the numerical integration by Runge-Kutta and BDF methods depending on the movement of characteristic subspaces; see [7, p. 168], also [5, Section 8.3], for details.

Example 1.1. We address the DAE system

$$\begin{aligned} x_2'(t) + x_1(t) &= y_1(t),\\ t\eta x_2'(t) + x_3'(t) + (\eta+1)x_2(t) &= y_2(t),\\ t\eta x_2(t) + x_3(t) &= y_3(t), \quad t \in [0,1]. \end{aligned}$$

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.

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. The most sensible component concerning numerical computations is the algebraic one x_1. Let

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

serve as an exact solution and this determines y. In order to have a unique solution also for the classical collocation system, the conditions

$$p_2(0) = 0, \quad p_3(0) = 1,$$

were posed.

TABLE 1. Collocation results for Example 1.1. The table shows the error $\|x - p\|_{H^1_D}$.

n	Standard	Least-squares
20	2.81e+006	1.47e-4
40	5.59e+016	3.52e-5
80	1.45e+038	8.59e-6
160	6.07e+081	2.12e-6
320	6.29e+169	5.28e-7

TABLE 2. Collocation results for Example 1.1. The table shows the error $\|x_i - p_i\|_{L^\infty}$ for the three solution components

n	Standard			Least-squares		
	$i = 1$	$i = 2$	$i = 3$	$i = 1$	$i = 2$	$i = 3$
20	5.56e+006	3.03e+004	5.99e+004	2.09e-4	1.10e-06	2.18e-06
40	1.55e+017	4.23e+014	8.41e+014	5.03e-5	1.31e-07	2.65e-07
80	5.70e+038	7.76e+035	1.55e+036	1.23e-5	1.60e-08	3.20e-08
160	3.36e+082	2.29e+079	4.57e+079	3.06e-6	1.98e-09	4.00e-09
320	4.93e+170	1.68e+167	3.35e+167	7.68e-7	2.50e-10	5.00e-10

Table 1 displays the errors for $\eta = -2$ and $N = 3$ and equidistant partitions π. The left column displays the results from standard collocation with $M = N$ uniformly distributed collocation points on each subinterval and the right column shows the results from least-squares collocation with $M = 2N + 1$ uniformly distributed least-squares collocation points. The improvement is impressive!

In order to obtain a deeper insight, we present in Table 2 the error of the individual solution components in $L^\infty(0, 1)$.

The computations have been carried out in MATLAB.[2] □

In the present paper, we provide estimates of the so-called *instability threshold*[3] for arbitrary-index linear DAEs with variable coefficients. We consider these estimates in Section 4, most notably Theorem 4.1, as the main result of the present paper. It considerably generalizes the results from [3] merely given for constant-coefficient DAEs. This leads to sufficient convergence conditions for the least-squares method applied to systems (1)–(2). Moreover, a conjecture made in [3] concerning constant coefficient DAEs, is proven.

The paper is organized as follows. In Section 2 we summarize properties of differential-algebraic operators representing (1)–(2) in a natural Hilbert space setting. It turns out that such operators are essentially ill-posed if a higher-index DAE is involved. The associated abstract least-squares method is introduced and discussed in Section 3 leading to corresponding convergence results (Theorem 3.1). The necessary estimates of the instability threshold supposed in Theorem 3.1 are verified in Section 4. The abstract least-squares method as formulated in Hilbert spaces requires the evaluation of certain integrals. We address this question in

[2] MATLAB Release 2016a, The MathWorks, Inc., Natick, Massachusetts, United States.

[3] See Section 3 for the precise definition.

Section 5 and deduce sufficient convergence conditions for the least-squares collocation. Finally, we present some numerical examples in Section 6. Conclusions will be drawn in Section 7.

2. Differential-algebraic operators acting on H_D^1

In this subsection we represent first the DAE (1) and then the BVP (1) – (2) as operator equations

$$Tx = y \quad \text{and} \quad \mathcal{T}x = (y, r). \tag{11}$$

For this aim we define the *differential-algebraic operator* (DA operator) $T : H_D^1 \to L^2$,

$$(Tx)(t) = A(t)(Dx)'(t) + B(t)x(t), \quad \text{a.e. } t \in (a, b), \quad x \in H_D^1.$$

The function space H_D^1 equipped with its natural inner product,

$$(x, \bar{x})_{H_D^1} := (x, \bar{x})_{L^2} + ((Dx)', (D\bar{x})')_{L^2}, \quad x, \bar{x} \in H_D^1,$$

is a Hilbert space [8, Lemma 6.9] and the DA operator T is bounded since $A(\cdot)$ and $B(\cdot)$ are assumed to be at least continuous.

Next, we recall the notion of the tractability index of the DA operator T from [3]; see also [8, Section 4.2]. This notion is tied to the coefficients A, B, D only. In essence, the DA operator is regular with tractability index μ if the DAE represented as operator equation (11) is so.

The tractability index is specified by means of certain sequences of continuous matrix functions $G_{i+1} := G_i + B_i Q_i$, built pointwise on $[a, b]$ using special projector functions Q_i onto $\ker G_i$ and starting from $G_0 := AD, G_1 = G_0 + BQ_0$. Denoting $r_j := \operatorname{rank} G_j$, the construction yields $r_0 \leq r_1 \leq \cdots \leq r_i \leq r_{i+1}$. The matrix function sequence $G_0, \ldots, G_\kappa$ is *admissible* if it is well-defined and the ranks $r_0, \ldots, r_\kappa$ are constant in time [8, Definition 4.1].

Definition 2.1. The DA operator $T : H_D^1 \to L^2$ is said to be *regular with tractability index $\mu \in \mathbb{N}$ and characteristic values*

$$r_0 \leq \cdots \leq r_{\mu-1} < r_\mu = m, \quad l := m - \sum_{i=0}^{\mu-1}(m - r_i), \tag{12}$$

if there is an admissible matrix function sequence $G_0, \ldots, G_\mu$ with (12). If, additionally, the coefficients A, B, D are as smooth as required for the existence of completely decoupling projectors, then the DA operator T is said to be *fine.*

Let Π_{can} denote the canonical projector function of the associated fine DAE; see [5, Definition 2.37]. The projector $\Pi_{\text{can}}(t)$ acts in $\mathbb{R}^m$ and its rank is l given in (12) for all $t \in [a, b]$. The number l actually accounts for 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 ones corresponding to the infinite eigenvalues [5, Theorem 1.33].

We quote [3, Theorem 2.2] concerning the characteristic properties of T.

Theorem 2.2. *Let the bounded DA operator $T : H^1_D \to L^2$ be fine with tractability index $\mu \in \mathbb{N}$ and characteristic values* (12)*. Then the following statements hold:*

(a) $\ker T$ *has finite dimension,* $\dim \ker T = l = \operatorname{rank} \mathit{\Pi}_{\mathrm{can}}$.

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

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

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

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

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

is defined on $H^1_D = \{x \in L^2 : x_2, x_3 \in H^1((0,1),\mathbb{R})\}$, with $m = 3,\ k = 2$. Its image becomes

$$\operatorname{im} T = \{y \in L^2 : y_3 \in H^1((0,1),\mathbb{R}),\ y_2 - y_3' \in H^1((0,1),\mathbb{R})\} \subset L^2.$$

This operator T is injective. The canonical projector function is simply $\mathit{\Pi}_{\mathrm{can}} = 0$, which corresponds to $l = 0$. □

Finally, we introduce the operator $\mathcal{T} : H^1_D \to L^2 \times \mathbb{R}^l =: Z$ associated with the BVP (1) – (2) by

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

The product space $Z = L^2 \times \mathbb{R}^l$ equipped with its natural inner product

$$(z,\bar{z})_Z := (y,\bar{y})_{L^2} + \langle r,\bar{r}\rangle, \quad z = (y,r),\ \bar{z} = (\bar{y},\bar{r}) \in Z,$$

is again a Hilbert space. Here, $\langle\cdot,\cdot\rangle$ denotes the Euclidean scalar product of $\mathbb{R}^l$. We quote [3, Theorem 2.4] to provide properties of the operator $\mathcal{T}$.

Theorem 2.4. *Let the bounded DA operator $T : H^1_D \to L^2$ be fine with index $\mu \in \mathbb{N}$ and characteristic values* (12) *and let the boundary conditions be restricted by* (4)*. Then the following statements hold:*

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

$$\ker(G_a X(a,a) + G_b X(b,a)) = \ker \mathit{\Pi}_{\mathrm{can}}(a) \tag{13}$$

holds. Here, $X(t,a)$ *denotes the maximal fundamental solution matrix of the associated DAE, normalized at point* a.[4]

(b) *If* (13) *is valid, then the equation* $\mathcal{T}x = (y,r)$ *is well-posed if* $\mu = 1$ *and otherwise essentially ill-posed.*[5]

(c) *If* (13) *is valid, then* $\mathcal{T}$ *is injective.*

(d) *If* $\mu = 1$ *and* (13) *is valid, then there exists a constant bound* $c_{\mathcal{T}} > 0$ *such that*

$$\|\mathcal{T}x\|_{L^2\times\mathbb{R}^l} \geq c_{\mathcal{T}} \|x\|_{H^1_D}, \quad x \in H^1_D(a,b).$$

[4] X is the unique solution of the IVP $A(t)(DX)'(t) + B(t)X(t) = 0,\ t \in (a,b),\ X(a) = \mathit{\Pi}_{\mathrm{can}}(a)$.

[5] The operator equation $\mathcal{T}x = (y,r)$ is said to be essentially ill-posed, if the range $\operatorname{im}\mathcal{T} \subset L^2 \times \mathbb{R}^l$ is a nonclosed subset.

3. Basic convergence assertions

Now we turn to convergence properties of the least-squares method applied to the operator equation representing the BVP (1) – (2). Let $\mathcal{T}$ be injective and $(y, r) \in \operatorname{im} \mathcal{T}$ be given, $x_* = \mathcal{T}^{-1}(y, r)$.

Let the function set X_π, related to the partition

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

with maximal stepsize h and minimal stepsize $h_{\min}$ and the degree $N \geq 1$, be given by (5) as before.

Regarding convergence properties for $h \to 0$ we have in mind a sequence of partitions

$$\pi_s : a = t_{0,s} < \cdots < t_{n_s,s} = b,$$

with maximal and minimal stepsizes $h_{(s)}, h_{\min,s}, \;\; n_s \to \infty, \;\; h_{(s)} \to 0$. The degree N is uniform for all corresponding function sets X_{π_s}. For easier reading we drop the extra integer s but we thoroughly assure the reader that the indicated constants do not depend on the partitions and stepsizes in fact.

Following ideas developed in [4] (see also [3, Section 2.2]), the approximate solution

$$p_\pi = \operatorname{argmin}\{\|A(Dp)' + Bp - y\|^2_{L^2} + |G_a p(a) + G_b p(b) - r|^2 : p \in X_\pi\} \tag{14}$$

satisfies the inequality

$$\|p_\pi - x_*\|_{H^1_D} \leq \frac{\beta_\pi}{\gamma_\pi} + \alpha_\pi, \tag{15}$$

with x_* denoting the unique solution of the BVP and

$$\begin{aligned}
\alpha_\pi &:= \|x_* - \mathfrak{P}_\pi x_*\|_{H^1_D}, \quad \mathfrak{P}_\pi x_* = \operatorname{argmin}\{\|x_* - p\|_{H^1_D} : p \in X_\pi\},\\
\beta_\pi &:= \|\mathcal{T}(x_* - \mathfrak{P}_\pi x_*)\|_Z \leq \|\mathcal{T}\|\alpha_\pi,\\
\gamma_\pi &:= \inf_{p \in X_\pi, p \neq 0} \frac{\|\mathcal{T}p\|_Z}{\|p\|_{H^1_D}} = \inf_{p \in X_\pi, p \neq 0} \frac{(\|Tp\|^2_{L^2} + |G_a p(a) + G_b p(b)|^2)^{1/2}}{\|p\|_{H^1_D}}.
\end{aligned}$$

Aiming for convergence properties, one needs upper estimates for the approximation errors α_π and β_π, and a positive estimate from below for the *instability threshold* γ_π.

Let the solution x_* be sufficiently smooth so that the interpolation function $p_{\text{int}} \in X_\pi$ for x_* is well-defined by N interpolation nodes on each subinterval of the partition π and, additionally, by $Dp_{\text{int}}(a) = Dx_*(a)$. Then, standard interpolation results provide the estimates

$$\alpha_\pi \leq \|p_{\text{int}} - x_*\|_{H^1_D} \leq c_\alpha h^N, \quad \beta_\pi \leq c_\beta h^N, \tag{16}$$

where c_α and c_β are constants independent of the special partition π. The most challenging task in this context is providing an appropriate estimate for the threshold γ_π. In [3], for equidistant partitions π with sufficiently small stepsizes, the estimate

$$\gamma_\pi \geq c_\gamma h^{\min(N, \mu - 1)} \tag{17}$$

with a constant $c_\gamma > 0$, is conjectured owing to numerous numerical experiments and a strong proof for the cases $N \geq \mu - 1$ as well as $N = 1$ for constant-coefficient DAEs.

Theorem 4.1 below in Section 4 verifies the inequality

$$\gamma_\pi \geq c_\gamma h^{\mu-1} \tag{18}$$

for general regular linear DAEs, and this can be seen as the main result of the present paper. In addition, the stronger estimate (17) is shown for a special class of DAEs including all regular constant-coefficient DAEs (Theorem 4.7). So far it remains open if the stronger estimate is valid in more general cases.

As a consequence of the estimates (15) and (16) as well as the Theorems 4.1 and 4.7 we obtain the following.

Theorem 3.1. *Let the BVP* (1)–(2)*, with index $\mu \geq 1$, satisfy the assumptions of Theorem* 2.4 *and let* (13) *be true. Denote by x_* the unique solution. Moreover, assume the coefficients A, B of the BVP to be sufficiently smooth.*

Let X_π be given by (5)*. Then the following statements are valid for all partitions π with sufficiently small stepsize h and uniformly bounded ratios $\frac{h}{h_{\min}} \leq \rho$:*

(a) *The least-squares solutions p_π defined by* (14) *satisfy*

$$\|p_\pi - x_*\|_{H^1_D} \leq ch^{N-\mu+1}.$$

Hence, the choice of N such that $N \geq \mu$ ensures convergence in H^1_D, that is, $p_\pi \to x_*$ for $h \to 0$.

(b) *Moreover, if the coefficients A and B are constant, the solutions p_π fulfill*

$$\|p_\pi - x_*\|_{H^1_D} \leq ch^{\max(0,N-\mu+1)}$$

and the discrete solutions remain bounded in H^1_D also if $N < \mu - 1$.

One has $\rho = 1$ when restricting to equidistant partitions. In general, the constant c depends on the bound ρ.

For providing the approximation p_π in practice, one needs to evaluate the integral

$$\|\underbrace{A(Dp)' + Bp - y}_{=:f}\|^2_{L^2} = \|f\|^2_{L^2} = \int_a^b |f(t)|^2 \mathrm{dt}.$$

We choose $M > N \geq 1$ and represent $\int_a^b |f(t)|^2 \mathrm{dt} = F^T \mathcal{L} F + \mathcal{R}$, with a specific[6] symmetric, positive definite matrix $\mathcal{L}$ and

$$F := \begin{bmatrix} F_1 \\ \vdots \\ F_n \end{bmatrix} \in \mathbb{R}^{mMn}, \quad F_j := \Big(\frac{h_j}{M}\Big)^{1/2} \begin{bmatrix} f(t_{j-1} + \tau_1 h_j) \\ \vdots \\ f(t_{j-1} + \tau_M h_j) \end{bmatrix} \in \mathbb{R}^{mM},$$

and further (cf. (10), (9))

$$\begin{aligned} \phi_\pi(p) &= F^T \mathcal{L} F + |G_a p(a) + G_b p(b) - r|^2 + \mathcal{R}, \\ \phi_{\pi,M}(p) &= F^T F + |G_a p(a) + G_b p(b) - r|^2. \end{aligned}$$

We keep in mind that the function f, the vector F, and the remainder term $\mathcal{R}$ depend on p. Later on, neglecting the remainder term $\mathcal{R}$, we minimize the expression

$$F^T \mathcal{L} F + |G_a p(a) + G_b p(b) - r|^2 \tag{19}$$

instead of the functional ϕ_π.

[6] $\mathcal{L}$ is built from basic Lagrangian polynomials; see [3].

Note that the Euclidean norm of $F \in \mathbb{R}^{mMn}$ equals $(F^T F)^{1/2}$. The expression $(F^T \mathcal{L} F)^{1/2}$ indicates a further $\mathbb{R}^{mMn}$-norm which is equivalent to the Euclidean norm. By minimizing expression (19) or $\phi_{\pi,M}$ the matter is traced back to least-squares collocation in two versions which differ only by the norms used for $F \in \mathbb{R}^{mMn}$.

In Section 5 we will provide conditions that allow us to control the remainder term $\mathcal{R}$ such that convergence properties similar to those described in Theorem 3.1 remain true even for the discretized version (19).

4. Estimating the instability threshold

In this section we show the inequality

$$\gamma_\pi \geq c_\gamma h^{\mu-1}$$

to be valid for general regular linear DAEs with sufficiently smooth coefficients. We summarize this main result in more detail in the following theorem. The proof is performed in Subsection 4.3 below. It applies special properties of piecewise polynomials given in Subsection 4.1 and basic facts concerning DAEs, which are collected in Subsection 4.2. With the highly desirable sharper estimation (17) in mind we address the case $1 \leq N < \mu - 1$ in Subsection 4.4. In particular, we verify the stronger inequality (17) for arbitrary regular DAEs with constant coefficients which has been conjectured in [3] and proved for $N = 1$.

Theorem 4.1. *Let the bounded DA operator $T : H_D^1 \to L^2$ be fine with index $\mu \in \mathbb{N}$ and characteristic values* (12) *and let the boundary conditions be restricted by* (4)*. Let the condition* (13) *be valid.*

Let X_π be given by (5) *as before, and $N \geq 1$.*

Then the following statements are valid for all partitions π with sufficiently small maximal stepsizes h and uniformly bounded ratios $\frac{h}{h_{\min}} \leq \rho$:

(a) *If $\mu = 1$, then there is a constant $c_\gamma > 0$ such that $\gamma_\pi \geq c_\gamma$.*
(b) *If $\mu = 2$, then there is a constant $c_\gamma > 0$ such that $\gamma_\pi \geq c_\gamma h_{\min} \geq c_\gamma \frac{1}{\rho} h$.*
(c) *If $\mu \geq 2$ and the coefficients A and B are sufficiently smooth,*[7] *then there is a constant $c_\gamma > 0$ such that $\gamma_\pi \geq c_\gamma h_{\min}^{\mu-1} \geq c_\gamma \frac{1}{\rho^{\mu-1}} h^{\mu-1}$.*

Naturally, we are interested in a constant $c_\gamma > 0$ as large as possible. Details will be disclosed in the proof below; see also Remarks 4.5 and 4.6. At the moment we only point out that c_γ depends on N: the larger N is the smaller c_γ becomes.

Theorem 4.1 provides two different constants for index-2 DAEs, one in item (b), say $c_{\gamma,b}$, given by a very special proof, and another one in item (c), say $c_{\gamma,c}$, arising as a particular case within a much more general context. The ratio $c_{\gamma,b}/c_{\gamma,c}$ is independent of N; cf., Remark 4.6.

4.1. **An auxiliary estimation concerning piecewise polynomials.**

The following lemma is a straightforward consequence of [3, Lemma 3.3].[8]

[7] See Subsection 4.3 below for details.

[8] These inequalities are closely related to the so-called inverse estimates in the finite element context.

Lemma 4.2. *Let the function $q : [a, b] \to \mathbb{R}^m$ be polynomial with degree $\leq K$, $K \geq 0$, in each of its components and on each subinterval of the partition $\pi : a = t_0 < \ldots < t_n = b$. Then the relations*

$$\|q^{(i)}\|^2_{L^2} \leq c^*_i \frac{1}{h^{2i}_{\min}} \|q\|^2_{L^2}, \quad \|q^{(i)}\|^2_{H^1_D} \leq C^*_i \frac{1}{h^{2i}_{\min}} \|q\|^2_{H^1_D}, \quad , i = 1, \cdots, K,$$
$$\|q^{(K+1)}\|^2_{L^2} = 0, \quad \|q^{(K+1)}\|^2_{H^1_D} = 0,$$

are valid with constants

$$c^*_i = 4^i \lambda_K \cdots \lambda_{K-i+1}, \quad C^*_i = 4^i \max\{\lambda_K \cdots \lambda_{K-i+1}, \lambda_{K-1} \cdots \lambda_{K-i}\},$$

where the $\lambda_j > 0$ are certain matrix eigenvalues; see [3, Lemma 3.3].

Proof. For $K = 0$ the statement is trivially satisfied. Set $K \geq 1$. We have $q_i|_{[t_{j-1}, t_j)} \in \mathcal{P}_K$ and therefore

$$\begin{aligned} \|q\|^2_{L^2} &= \int_a^b |q(t)|^2 \mathrm{dt} = \sum_{j=1}^n \sum_{i=1}^m \int_{t_{j-1}}^{t_j} q_i(t)^2 \mathrm{dt} = \sum_{j=1}^n \sum_{i=1}^m \int_0^{h_j} q_i(t_{j-1} + s)^2 \mathrm{ds} \\ &\geq \sum_{j=1}^n \sum_{i=1}^m \frac{h_j^2}{4\lambda_K} \int_0^{h_j} q_i'(t_{j-1} + s)^2 \mathrm{ds} = \sum_{j=1}^n \frac{h_j^2}{4\lambda_K} \int_{t_{j-1}}^{t_j} |q'(t)|^2 \mathrm{dt} \\ &\geq \frac{h^2_{\min}}{4\lambda_K} \sum_{j=1}^n \int_{t_{j-1}}^{t_j} |q'(t)|^2 \mathrm{dt} = \frac{h^2_{\min}}{4\lambda_K} \int_a^b |q'(t)|^2 \mathrm{dt} = \frac{h^2_{\min}}{4\lambda_K} \|q'\|^2_{L^2}. \end{aligned}$$

Then owing to $q_i'|_{[t_{j-1}, t_j)} \in \mathcal{P}_{K-1}$ we obtain $\|q'\|^2_{L^2} \geq \frac{h^2_{\min}}{4\lambda_{K-1}} \|q''\|^2_{L^2}$ and further $\|q\|^2_{L^2} \geq \frac{h^2_{\min}}{4\lambda_K} \frac{h^2_{\min}}{4\lambda_{K-1}} \|q''\|^2_{L^2}$, and so on. □

4.2. **Preliminaries in matters of DAEs.** To verify the statements of Theorem 4.1 we apply results of the projector based DAE analysis. We collect here just the necessary ingredients and refer to [5, 8] for details. Let the DA operator $T : H^1_D \to L^2$ corresponding to the DAE (1) be fine with tractability index $\mu \geq 2$ and the characteristic values (12). Then there are an admissible sequence of matrix-valued functions starting from $G_0 := AD$ and ending up with a nonsingular G_μ, see [5, Definition 2.6], as well as associated projector valued functions

$$P_0 := D^+ D \quad \text{and} \quad P_1, \ldots, P_{\mu-1} \in \mathcal{C}([a, b], \mathbb{R}^{m \times m})$$

which provide a fine decoupling of the DAE. We have then the further projector valued functions

$$\begin{aligned} Q_i = I - P_i, \ i = 0, \ldots, \mu - 1, \\ \mathit{\Pi}_0 := P_0, \ \mathit{\Pi}_i := \mathit{\Pi}_{i-1} P_i \in \mathcal{C}([a, b], \mathbb{R}^{m \times m}), \ i = 1, \ldots, \mu - 1, \\ D\mathit{\Pi}_i D^+ \in \mathcal{C}^1([a, b], \mathbb{R}^{k \times k}), \ i = 1, \ldots, \mu - 1. \end{aligned}$$

By means of the projector functions we decompose the unknown x and decouple the DAE itself into their characteristic parts; see [5, Section 2.4].

The component $u = D\mathit{\Pi}_{\mu-1} x = D\mathit{\Pi}_{\mu-1} D^+ D x$ satisfies the explicit regular ODE residing in $\mathbb{R}^k$,

$$u' - (D\mathit{\Pi}_{\mu-1} D^+)' u + D\mathit{\Pi}_{\mu-1} G_\mu^{-1} B \mathit{\Pi}_{\mu-1} D^+ u = D\mathit{\Pi}_{\mu-1} G_\mu^{-1} y, \tag{20}$$

and the components $v_i = \Pi_{i-1}Q_ix = \Pi_{i-1}Q_iD^+Dx,\ i = 1,\dots,\mu-1$ satisfy the triangular subsystem involving several differentiations,

$$\begin{bmatrix} 0 & \mathcal{N}_{12} & \cdots & \mathcal{N}_{1,\mu-1} \\ & 0 & \ddots & \vdots \\ & & \ddots & \mathcal{N}_{\mu-2,\mu-1} \\ & & & 0 \end{bmatrix} \begin{bmatrix} (Dv_1)' \\ \\ \vdots \\ (Dv_{\mu-1})' \end{bmatrix} + \begin{bmatrix} I & \mathcal{M}_{12} & \cdots & \mathcal{M}_{1,\mu-1} \\ & I & \ddots & \vdots \\ & & \ddots & \mathcal{M}_{\mu-2,\mu-1} \\ & & & I \end{bmatrix} \begin{bmatrix} v_1 \\ \\ \vdots \\ v_{\mu-1} \end{bmatrix} = \begin{bmatrix} \mathcal{L}_1 \\ \\ \vdots \\ \mathcal{L}_{\mu-1} \end{bmatrix} y. \tag{21}$$

The coefficients $\mathcal{N}_{i,j}$, $\mathcal{M}_{i\,j}$, and $\mathcal{L}_i$ are subsequently given. Finally, one has for $v_0 = Q_0x$ the representation

$$v_0 = \mathcal{L}_0y - \mathcal{H}_0D^+u - \sum_{j=1}^{\mu-1}\mathcal{M}_{0\,j}v_j - \sum_{j=1}^{\mu-1}\mathcal{N}_{0\,j}(Dv_j)'. \tag{22}$$

The subspace $\operatorname{im} D\Pi_{\mu-1}$ is an invariant subspace for the ODE (20). The components $v_0, v_1, \dots, v_{\mu-1}$ remain within their subspaces $\operatorname{im} Q_0$, $\operatorname{im}\Pi_{\mu-2}Q_1,\dots$, $\operatorname{im}\Pi_0Q_{\mu-1}$, respectively. The structural decoupling is associated with the decomposition

$$x = D^+u + v_0 + v_1 + \cdots + v_{\mu-1}.$$

All coefficients in (20) – (22) are continuous and explicitly given in terms of an admissible matrix function sequence as

$$\begin{aligned}
\mathcal{N}_{01} &:= -Q_0Q_1D^+, \\
\mathcal{N}_{0j} &:= -Q_0P_1\cdots P_{j-1}Q_jD^+, && j = 2,\dots,\mu-1, \\
\mathcal{N}_{i,i+1} &:= -\Pi_{i-1}Q_iQ_{i+1}D^+, && i = 1,\dots,\mu-2, \\
\mathcal{N}_{ij} &:= -\Pi_{i-1}Q_iP_{i+1}\cdots P_{j-1}Q_jD^+, && j = i+2,\dots,\mu-1,\ i = 1,\dots,\mu-2, \\
\mathcal{M}_{0j} &:= Q_0P_1\cdots P_{\mu-1}\mathcal{M}_jD\Pi_{j-1}Q_j, && j = 1,\dots,\mu-1, \\
\mathcal{M}_{ij} &:= \Pi_{i-1}Q_iP_{i+1}\cdots P_{\mu-1}\mathcal{M}_jD\Pi_{j-1}Q_j, && j = i+1,\dots,\mu-1,\ i = 1,\dots,\mu-2, \\
\mathcal{L}_0 &:= Q_0P_1\cdots P_{\mu-1}G_\mu^{-1}, \\
\mathcal{L}_i &:= \Pi_{i-1}Q_iP_{i+1}\cdots P_{\mu-1}G_\mu^{-1}, && i = 1,\dots,\mu-2, \\
\mathcal{L}_{\mu-1} &:= \Pi_{\mu-2}Q_{\mu-1}G_\mu^{-1}, \\
\mathcal{H}_0 &:= Q_0P_1\cdots P_{\mu-1}\mathcal{K}\Pi_{\mu-1},
\end{aligned}$$

in which

$$\mathcal{K} := (I-\Pi_{\mu-1})G_\mu^{-1}B_{\mu-1}\Pi_{\mu-1} + \sum_{\lambda=1}^{\mu-1}(I-\Pi_{\lambda-1})(P_\lambda - Q_\lambda)(D\Pi_\lambda D^+)'D\Pi_{\mu-1},$$

$$\mathcal{M}_j := \sum_{k=0}^{j-1}(I-\Pi_k)\{P_kD^+(D\Pi_kD^+)' - Q_{k+1}D^+(D\Pi_{k+1}D^+)'\}D\Pi_{j-1}Q_lD^+,$$
$$j = 1,\dots,\mu-1.$$

It should be added at this point, that the coefficients of the ODE (20) are uniquely determined in the scope of the fine decoupling. This justifies our speaking about *the inherent explicit regular ODE* (IERODE) of the given DAE.

We introduce the function space (cf., also [8])

$$\begin{aligned} Y := \big\{y \in L^2 : v_{\mu-1} &:= \mathcal{L}_{\mu-1}y, \quad Dv_{\mu-1} \in H^1, \\ v_{\mu-j} &:= \mathcal{L}_{\mu-j}y - \sum_{i=1}^{j-1} \mathcal{N}_{\mu-j,\mu-j+i}(Dv_{\mu-j+i})' - \sum_{i=1}^{j-1} \mathcal{M}_{\mu-j,\mu-j+i}v_{\mu-j+i}, \\ Dv_{\mu-j} &\in H^1 \quad \text{for} \quad j = 2, \dots, \mu-1\big\} \end{aligned}$$

and its norm

$$\|y\|_Y := \big(\|y\|_{L^2}^2 + \sum_{i=1}^{\mu-1}\|(Dv_i)'\|_{L^2}^2\big)^{1/2}, \quad y \in Y.$$

Both, the space and its norm are very special and strongly depend on the decoupling coefficients which in turn depend on the given data A, D, B. For this reason, those function spaces are termed *factitious* in [8].

Proposition 4.3. *Let the DA operator* $T : H_D^1 \to L^2$ *be fine with characteristic values* (12) *and index* $\mu \geq 2$. *Then the following results:*

(a) $\operatorname{im} T = Y$.
(b) *The space* Y *equipped with the norm* $\|\cdot\|_Y$ *is complete.*
(c) *Let the operator* $\mathcal{T}$ *corresponding to the BVP satisfy the conditions* (4) *and* (13). *Then there is a constant* c_Y *such that the inequality*

$$\|x\|_{H_D^1} \leq c_Y\ (\|y\|_Y^2 + |r|^2)^{1/2} \quad \textit{for} \quad y \in Y, r \in \mathbb{R}^l, x = \mathcal{T}^{-1}(y, r)$$

becomes valid.

Proof. (a) and (b) can be checked by a straightforward use of the above decoupling formulas analogously to the case of the Banach space setting in [8].

(c) The operator T is bounded also with respect to the new image space $(Y, \|\cdot\|_Y)$. Namely, for each $x \in H_D^1$ one has $\|Tx\|_{L^2} \leq c_T\|x\|_{H_D^1}$ and further, owing to the decoupling,

$$\begin{aligned} Dv_i &= D\Pi_{i-1}Q_ix = D\Pi_{i-1}Q_iD^+Dx, \\ (Dv_i)' &= (D\Pi_{i-1}Q_iD^+)'Dx + D\Pi_{i-1}Q_iD^+(Dx)', \quad i = 1, \dots, \mu-1, \end{aligned}$$

which leads to $\|Tx\|_Y \leq c_T^Y\|x\|_{H_D^1}$. Therefore, in the new setting, the operator $\mathcal{T} : H_D^1 \to Y \times \mathbb{R}^l$ is a homeomorphism, and hence, its inverse is bounded. □

Next we focus our interest on elements $Tp = A(Dp)' + Bp$, $p \in X_\pi$. Tp belongs to Y, basically it is continuous on the intervals of the partition π and has possible jumps at the gridpoints. To this end, let $\mathcal{C}_\pi^\kappa$ denote the linear space of functions being bounded and piecewise of class $\mathcal{C}^\kappa$ with jumps and breakpoints only at the gridpoints of π. Denote

$$\begin{aligned} Y_\pi &:= \{y \in L^2 : D\mathcal{L}_{\mu-i}y \in \mathcal{C}_\pi^{\mu-i}([a,b],\mathbb{R}^k), \quad i = 1,\dots,\mu-1\},\\ Y_\pi^0 &:= \{y \in \mathcal{C}_\pi^0([a,b],\mathbb{R}^m) : D\mathcal{L}_{\mu-i}y \in \mathcal{C}_\pi^{\mu-i}([a,b],\mathbb{R}^k), \quad i = 1,\dots,\mu-1\}. \end{aligned}$$

Lemma 4.4. *Let the DA operator T be fine with index $\mu > 1$ and let its coefficients A and B be sufficiently smooth such that*

$$D\mathcal{N}_{\mu-i,\mu-i+j},\ D\mathcal{M}_{\mu-i,\mu-i+j}D^+ \in \mathcal{C}^{\mu-i}([a,b],\mathbb{R}^{k\times k}),\\ j = 1,\dots,i-1,\ i = 2,\dots,\mu-1.$$

(a) *Then the inclusion $Y_\pi \subset Y$ follows and further the inequality*

$$\|y\|_Y^2 \le \|y\|_\pi^2 := \|y\|_{L^2}^2 + \sum_{i=1}^{\mu-1}\sum_{s=0}^{\mu-i} d_{i,s}\ \|(D\mathcal{L}_{\mu-i}y)^{(s)}\|_{L^2}^2, \quad y \in Y_\pi,$$

with constants $d_{l,s}$ basically given by the coefficients A and B. In particular, for $\mu = 2$ it results that $d_{1,0} = 0, d_{1,1} = 1$, i.e.,

$$\|y\|_Y^2 \le \|y\|_\pi^2 := \|y\|_{L^2}^2 + \|(D\mathcal{L}_1 y)'\|_{L^2}^2, \quad y \in Y_\pi.$$

For $\mu \ge 3$, the coefficients $d_{i,s}$ with $s > 0$ are strictly positive. The coefficient $d_{1,\mu-1}$ in front of the highest derivative term reads

$$d_{1,\mu-1} = 2\|D\Pi_0 Q_1 \cdots Q_{\mu-1}D^+\|_\infty^2.$$

(b) *If, additionally,*

$$D\mathcal{L}_{\mu-i}[A\ B] \in \mathcal{C}^{\mu-i}, \quad i = 1,\dots,\mu-1,$$

then the inclusion $T(X_\pi) \subset Y_\pi^0 \subset Y$ is also valid.

Proof. (a) We show that for each arbitrary $y_+ \in Y_\pi$ there exists a $x_+ \in H_D^1$ such that $Tx_+ = y_+$. We first provide a solution $u_+ \in H^1$ of the IVP

$$u' - (D\Pi_{\mu-1}D^+)'u + D\Pi_{\mu-1}G_\mu^{-1}B\Pi_{\mu-1}D^+u = D\Pi_{\mu-1}G_\mu^{-1}y_+,\ u(a) = 0.$$

We put (cf. (21)) $v_{+,\mu-1} = \mathcal{L}_{\mu-1}y_+$ yielding $Dv_{+,\mu-1} = D\mathcal{L}_{\mu-1}y_+ \in \mathcal{C}_\pi^{\mu-1}$ and then consecutively for $j = 2,\dots,\mu-1$,

$$v_{+,\mu-j} = \mathcal{L}_{\mu-j}y_+ - \sum_{i=1}^{j-1}\big[\mathcal{M}_{\mu-j,\mu-j+i}D^+Dv_{+,\mu-j+i} + \mathcal{N}_{\mu-j,\mu-j+i}(Dv_{+,\mu-j+i})'\big]$$

yielding $Dv_{+,\mu-j} \in \mathcal{C}_\pi^{\mu-j}$. Finally we determine $v_{+,0}$ according to (22). The resulting function $x_+ = D^+u_+ + v_{+,0} + v_{+,1} + \cdots + v_{+,\mu-1}$ belongs to H_D^1 and satisfies the DAE a.e. on $[a,b]$. This proves that $y_+ \in Y$, and hence $Y_\pi \subset Y$.

Next we provide the norm-inequality. For $\mu = 2$ the assertion is evident. We turn to the case $\mu \geq 3$.

Let $y \in Y_\pi$ be given. Regarding the definition of the function space Y which is closely related to the decoupled system (21) we state that $(Dv_{\mu-1})^{(i)} = (D\mathcal{L}_{\mu-1}y)^{(i)}$, $i = 1, \dots, \mu-1$, and derive by straightforward technical computations consecutively for $j = 2, \dots, \mu-1$,

$$\begin{aligned}(Dv_{\mu-j})' = (D\mathcal{L}_{\mu-j}y)' - \sum_{i=1}^{j-1}\big[&(D\mathcal{M}_{\mu-j,\mu-j+i}D^+)'Dv_{\mu-j+i}\\ &+ \big(D\mathcal{M}_{\mu-j,\mu-j+i}D^+ + (D\mathcal{N}_{\mu-j,\mu-j+i})'\big)(Dv_{\mu-j+i})'\\ &+ D\mathcal{N}_{\mu-j,\mu-j+i}(Dv_{\mu-j+i})''\big]\\ = (D\mathcal{L}_{\mu-j}y)' - \big[&D\mathcal{N}_{\mu-j,\mu-j+1}(D\mathcal{L}_{\mu-j+1}y)''\\ &+ \cdots + (-1)^{j-1}D\mathcal{N}_{\mu-j,\mu-j+1}\cdots D\mathcal{N}_{\mu-2,\mu-1}(D\mathcal{L}_{\mu-1}y)^{(j)}\big]\\ &+ \sum_{i=1}^{j-1}\sum_{s=0}^{i}\mathcal{E}_{j,i,s}(D\mathcal{L}_{\mu-j+i}y)^{(s)}.\end{aligned}$$

Regarding the definition of the coefficients $\mathcal{N}_{k,k+1}$ and the basic properties of the involved projector functions we obtain

$$\begin{aligned}(Dv_{\mu-j})' = (D\mathcal{L}_{\mu-j}y)' &+ \sum_{i=1}^{j-1} D\Pi_{\mu-j-1}Q_{\mu-j}\cdots Q_{\mu-j+i}D^+(D\mathcal{L}_{\mu-j+i}y)^{(i+1)}\\ &+ \sum_{i=1}^{j-1}\sum_{s=0}^{i}\mathcal{E}_{j,i,s}(D\mathcal{L}_{\mu-j+i}y)^{(s)},\end{aligned}$$

where the matrix functions $\mathcal{E}_{j,i,s}$ are given by derivatives of the coefficients $\mathcal{M}_{k,l}$ and by $\mathcal{N}_{k,l}$, and their derivatives. Each of the involved coefficients is sufficiently smooth, at least continuous, thus uniformly bounded on $[a, b]$. The coefficients $\mathcal{E}_{j,i,s}$ vanish in case of constant A, B.

The highest involved derivative term is $(D\mathcal{L}_{\mu-1}y)^{(\mu-1)}$, and it can be found exclusively in

$$\begin{aligned}(Dv_1)' = (D\mathcal{L}_1y)' &+ \sum_{i=1}^{\mu-2} D\Pi_0Q_1\cdots Q_{i+1}D^+(D\mathcal{L}_{i+1}y)^{(i+1)}\\ &+ \sum_{i=1}^{\mu-2}\sum_{s=0}^{i}\mathcal{E}_{\mu-1,i,s}(D\mathcal{L}_{i+1}y)^{(s)}.\end{aligned}$$

We estimate

$$\begin{aligned}\|(Dv_{\mu-j})'\|_{L^2} \leq& \|(D\mathcal{L}_{\mu-j}y)'\|_{L^2}\\ &+\sum_{i=1}^{j-1}\|D\Pi_{\mu-j-1}Q_{\mu-j}\cdots Q_{\mu-j+i}D^+\|_\infty\|(D\mathcal{L}_{\mu-j+i}y)^{(i+1)}\|_{L^2}\\ &+\sum_{i=1}^{j-1}\sum_{s=0}^{i}\underbrace{\|\mathcal{E}_{j,i,s}\|_\infty}_{=:e_{j,i,s}}\|(D\mathcal{L}_{\mu-j+i}y)^{(s)}\|_{L^2}\end{aligned}$$

and thus

$$\begin{aligned}&\sum_{j=1}^{\mu-1}\|(Dv_{\mu-j})'\|_{L^2}\\ &\quad\leq\sum_{j=1}^{\mu-1}\|(D\mathcal{L}_{\mu-j}y)'\|_{L^2}\\ &\quad+\sum_{j=2}^{\mu-1}\sum_{i=1}^{j-1}\|D\Pi_{\mu-j-1}Q_{\mu-j}\cdots Q_{\mu-j+i}D^+\|_\infty\|(D\mathcal{L}_{\mu-j+i}y)^{(i+1)}\|_{L^2}\\ &\quad+\sum_{j=2}^{\mu-1}\sum_{i=1}^{j-1}\sum_{s=0}^{i}e_{j,i,s}\,\|(D\mathcal{L}_{\mu-j+i}y)^{(s)}\|_{L^2}.\end{aligned}$$

We rearrange the last formula to the form

$$\sum_{j=1}^{\mu-1}\|(Dv_{\mu-j})'\|_{L^2}\leq\sum_{j=1}^{\mu-1}\sum_{s=0}^{\mu-j}\tilde{d}_{j,s}\|(D\mathcal{L}_{\mu-j}y)^{(s)}\|_{L^2},$$

with the coefficients

$$\begin{aligned}\tilde{d}_{1,\mu-1} =& \|D\Pi_0Q_1\cdots Q_{\mu-1}D^+\|_\infty,\\ \tilde{d}_{1,\mu-2} =& \|D\Pi_1Q_2\cdots Q_{\mu-1}D^+\|_\infty+e_{\mu-1,\mu-2,\mu-2},\\ \tilde{d}_{1,\mu-3} =& \|D\Pi_2Q_3\cdots Q_{\mu-1}D^+\|_\infty+e_{\mu-2,\mu-3,\mu-3}+e_{\mu-1,\mu-2,\mu-3},\\ &\cdots\\ \tilde{d}_{1,2} =& \|D\Pi_{\mu-3}Q_{\mu-2}Q_{\mu-1}D^+\|_\infty+\sum_{j=2}^{\mu-1}e_{j,j-1,2},\\ \tilde{d}_{1,1} =& 1+\sum_{j=2}^{\mu-1}e_{j,j-1,1},\quad \tilde{d}_{1,0}=\sum_{j=2}^{\mu-1}e_{j,j-1,0},\\ \tilde{d}_{2,\mu-2} =& \|D\Pi_0Q_1\cdots Q_{\mu-2}D^+\|_\infty,\\ &\cdots\\ \tilde{d}_{2,1} =& 1+\sum_{j=3}^{\mu-1}e_{j,j-2,1},\quad \tilde{d}_{2,0}=\sum_{j=3}^{\mu-1}e_{j,j-2,0},\\ &\cdots\\ \tilde{d}_{\mu-1,1} =& 1,\quad \tilde{d}_{\mu-1,0}=0.\end{aligned}$$

Observe that the coefficients $\tilde{d}_{j,s}$ are strictly positive for each $s > 0$. Finally we derive

$$\begin{aligned}\sum_{j=1}^{\mu-1}\|(Dv_{\mu-j})'\|_{L^2}^2 &\leq \Big(\sum_{j=1}^{\mu-1}\|(Dv_{\mu-j})'\|_{L^2}\Big)^2\\ &\leq \Big(\sum_{j=1}^{\mu-1}\sum_{s=0}^{\mu-j}\tilde{d}_{j,s}\|(D\mathcal{L}_{\mu-j}y)^{(s)}\|_{L^2}\Big)^2\\ &\leq \sum_{j=1}^{\mu-1}\sum_{s=0}^{\mu-j}d_{j,s}\|(D\mathcal{L}_{\mu-j}y)^{(s)}\|_{L^2}^2,\end{aligned}$$

where $d_{1,\mu-1} := 2\tilde{d}_{1,\mu-1}^2$ and $d_{i,s} := 2(S-1)\tilde{d}_{i,s}^2$ if $(i,s) \neq (1,\mu-1)$. Thereby, $S := \sum_{i=1}^{\mu-1}\sum_{s=0}^{\mu-i} 1 = \frac{1}{2}\mu(\mu+1)-1$ denotes the maximal number of summands.

(b) For each arbitrary $p \in X_\pi$ and the corresponding $y = Tp = A(Dp)' + Bp$ it holds that $y \in C_\pi^0$ and $\mathcal{L}_{\mu-j}y = \mathcal{L}_{\mu-j}A(Dp)' + \mathcal{L}_{\mu-j}Bp \in \mathcal{C}_\pi^{\mu-j}$, thus $Tp \in Y_\pi^0$. □

4.3. **Proof of Theorem 4.1.**

Part (a). The first assertion is a consequence of the boundedness of the inverse operator $\mathcal{T}^{-1}$.

Part (b). In the case of $\mu = 2$ one has simply

$$Y = \{y \in L^2 : v_1 = \mathcal{L}_1 y, Dv_1 \in H^1\} = \{y \in L^2 : D\Pi_0 Q_1 G_2^{-1} y \in H^1\}$$

and $\|y\|_Y^2 = \|y\|_{L^2}^2 + \|(D\Pi_0 Q_1 G_2^{-1}y)'\|_{L^2}^2$.

Consider an arbitrary $p \in X_\pi$ and set $q := Tp = A(Dp)' + Bp$, $r := G_a p(a) + G_b p(b)$. Owing to the decoupling we find that

$$\begin{aligned}D\Pi_0 Q_1 G_2^{-1} q &= D\Pi_0 Q_1 p = D\Pi_0 Q_1 D^+ Dp,\\ (D\Pi_0 Q_1 G_2^{-1} q)' &= (D\Pi_0 Q_1 D^+)' Dp + D\Pi_0 Q_1 D^+ (Dp)',\\ \|(D\Pi_0 Q_1 G_2^{-1} q)'\|_{L^2}^2 &\leq 2\|(D\Pi_0 Q_1 D^+)'\|_\infty^2 \, \|Dp\|_{L^2}^2 + 2\|D\Pi_0 Q_1 D^+\|_\infty^2 \, \|(Dp)'\|_{L^2}^2,\end{aligned}$$

and Lemma 4.2 implies

$$\begin{aligned}\|(D\Pi_0 Q_1 G_2^{-1} q)'\|_{L^2}^2 &\leq 2\Big(\|(D\Pi_0 Q_1 D^+)'\|_\infty^2 + \|D\Pi_0 Q_1 D^+\|_\infty^2 \, \frac{c_1^*}{h_{\min}^2}\Big)\|Dp\|_{L^2}^2\\ &\leq \frac{1}{h_{\min}^2}\big(2c_1^*\|D\Pi_0 Q_1 D^+\|_\infty^2 + O(h^2)\big)\|Dp\|_{L^2}^2\\ &\leq \frac{1}{h_{\min}^2} 3c_1^*\|D\Pi_0 Q_1 D^+\|_\infty^2 \|Dp\|_{L^2}^2 \qquad (23)\end{aligned}$$

for sufficiently small $h > 0$ where $c_1^*\|D\Pi_0 Q_1 D^+\|_\infty^2 > 0$. On the other hand we decompose

$$Dp = D\Pi_1 p + D\Pi_0 Q_1 p = D\Pi_1 p + D\Pi_0 Q_1 G_2^{-1} q.$$

Taking into account that the component $D\Pi_1 p$ satisfies the IERODE and the boundary condition we obtain

$$\|Dp\|_{L^2}^2 \leq \; 2K\,(\|q\|_{L^2}^2 + |r|^2) + 2\|D\Pi_0 Q_1 G_2^{-1}\|_\infty^2 \|q\|_{L^2}^2 \leq 2(K+d)(\|q\|_{L^2}^2 + |r|^2)$$

with $d :=\|D\Pi_0 Q_1 G_2^{-1}\|_\infty^2 = \|D\mathcal{L}_1\|_\infty^2 > 0$. It happens that $K = 0$ if the IERODE is absent owing to $\Pi_1 = 0$. K has moderate size if the related BVP is well-conditioned (cf. [6]). Inserting into (23) we arrive at

$$\|(D\Pi_0 Q_1 G_2^{-1} q)'\|_{L^2}^2 \leq \frac{6}{h_{\min}^2}\tilde{g}_1(\|q\|_{L^2}^2 + |r|^2),$$

with $\tilde{g}_1 := c_1^*\|D\Pi_0 Q_1 D^+\|_\infty^2 (K+d) > 0$, and further, for sufficiently fine partitions,

$$\|q\|_Y^2 + |r|^2 \leq \|q\|_{L^2}^2 + |r|^2 + \frac{6\tilde{g}_1}{h_{\min}^2}(\|q\|_{L^2}^2 + |r|^2) \leq \frac{9\tilde{g}_1}{h_{\min}^2}(\|q\|_{L^2}^2 + |r|^2).$$

Finally, regarding this and applying Proposition 4.3(c) we obtain

$$\begin{aligned}\gamma_\pi^2 &= \inf_{p\in X_\pi, p\neq 0} \frac{\|Tp\|_{L^2}^2 + |G_a p(a) + G_b p(b)|^2}{\|p\|_{H_D^1}^2} \\ &= \inf_{p\in X_\pi, p\neq 0} \underbrace{\frac{\|Tp\|_Y^2 + |G_a p(a) + G_b p(b)|^2}{\|p\|_{H_D^1}^2}}_{\geq c_Y^{-2}} \underbrace{\frac{\|Tp\|_{L^2}^2 + |G_a p(a) + G_b p(b)|^2}{\|Tp\|_Y^2 + |G_a p(a) + G_b p(b)|^2}}_{\geq h_{\min}^2/9\tilde{g}_1} \\ &\geq \frac{1}{9c_Y^2\tilde{g}_1} h_{\min}^2 = c_\gamma^2 h_{\min}^2.\end{aligned}$$

Part (c). Let the coefficients A and B be smooth enough to ensure that (cf. (21))

$$D\mathcal{N}_{\mu-i,\mu-i+s},\ D\mathcal{M}_{\mu-i,\mu-i+s}D^+ \in \mathcal{C}^{\mu-i}, \quad i = 2,\dots,\mu-1,\ s = 1,\dots,i-1,$$

$$D\mathcal{L}_{\mu-i},\ D\mathcal{L}_{\mu-i}A,\ D\mathcal{L}_{\mu-i}B \in \mathcal{C}^{\mu-i}, \quad i = 1,\dots,\mu-1.$$

By construction, for $i = 1,\dots,\mu-1$, the matrix function $D\mathcal{L}_{\mu-i}$ has the nullspace

$$\ker D\mathcal{L}_{\mu-i} = \ker \mathcal{L}_{\mu-i} = \ker G_\mu Q_{\mu-i} P_{\mu-i+1}\cdots P_{\mu-1} G_\mu^{-1}$$

of constant dimension $r_{\mu-i}$. By this, the pointwise Moore-Penrose inverse $(D\mathcal{L}_{\mu-i})^+$ is as smooth as $D\mathcal{L}_{\mu-i}$, and so are the orthoprojector function

$$U_{\mu-i} := (D\mathcal{L}_{\mu-i})^+ D\mathcal{L}_{\mu-i}$$

as well as the matrix functions

$$\mathfrak{A}_{\mu-i} := U_{\mu-i}A, \quad \mathfrak{B}_{\mu-i} := U_{\mu-i}B.$$

Given the partition $\pi : a = t_0 < t_1 < \cdots < t_n = b$ with midpoints $t_{j-1/2} := t_{j-1} + h_j/2,\ j = 1,\dots,n$, we introduce the auxiliary functions

$$\begin{aligned} U_{\pi,\,\mu-i}(t) &:= \sum_{s=0}^{\mu-i}\frac{1}{s!}(t-t_{j-1/2})^s U_{\mu-i}^{(s)}(t_{j-1/2}), \\ \mathfrak{A}_{\pi,\,\mu-i}(t) &:= U_{\pi,\,\mu-i}(t)\sum_{\rho=0}^{\mu-i}\frac{1}{\rho!}(t-t_{j-1/2})^\rho\, \mathfrak{A}_{\mu-i}^{(\rho)}(t_{j-1/2}), \\ \mathfrak{B}_{\pi,\,\mu-i}(t) &:= U_{\pi,\,\mu-i}(t)\sum_{\rho=0}^{\mu-i}\frac{1}{\rho!}(t-t_{j-1/2})^\rho\, \mathfrak{B}_{\mu-i}^{(\rho)}(t_{j-1/2}), \quad t \in [t_{j-1}, t_j), \\ & j = 1,\dots,n, \quad i = 1,\dots,\mu-1, \end{aligned}$$

the components of which are piecewise polynomial. By straightforward computations it can be checked that

$$(24)\quad \begin{aligned} &\mathfrak{A}^{(s)}_{\pi,\mu-i}(t_{j-1/2}) = \mathfrak{A}^{(s)}_{\mu-i}(t_{j-1/2}) = (U_{\mu-i}A)^{(s)}(t_{j-1/2}),\\ &\mathfrak{B}^{(s)}_{\pi,\mu-i}(t_{j-1/2}) = \mathfrak{B}^{(s)}_{\mu-i}(t_{j-1/2}) = (U_{\mu-i}B)^{(s)}(t_{j-1/2}),\\ &\qquad\qquad s=0,\ldots,\mu-i,\quad i=1,\ldots,\mu-1,\quad j=1,\ldots,n, \end{aligned}$$

and, furthermore, for $h\to 0$,

$$(25)\quad \begin{aligned} &\frac{1}{h^{\mu-i}}\,\|\mathfrak{A}_{\mu-i}-\mathfrak{A}_{\pi,\,\mu-i}\|_\infty := \frac{1}{h^{\mu-i}}\max_{a\le t\le b}|\mathfrak{A}_{\mu-i}(t)-\mathfrak{A}_{\pi,\,\mu-i}(t)| \to 0,\\ &\frac{1}{h^{\mu-i}}\,\|\mathfrak{B}_{\mu-i}-\mathfrak{B}_{\pi,\,\mu-i}\|_\infty := \frac{1}{h^{\mu-i}}\max_{a\le t\le b}|\mathfrak{B}_{\mu-i}(t)-\mathfrak{B}_{\pi,\,\mu-i}(t)| \to 0,\\ &\qquad\qquad i=1,\ldots,\mu-1. \end{aligned}$$

Next, the projector functions from Subsection 4.2 providing a fine decoupling also provide the decompositions

$$\begin{aligned} I &= P_1\cdots P_{\mu-1} + (I-P_1\cdots P_{\mu-1})\\ &= P_1\cdots P_{\mu-1} + Q_1P_2\cdots P_{\mu-1} + \cdots + Q_{\mu-2}P_{\mu-1} + Q_{\mu-1},\\ I &= G_\mu P_1\cdots P_{\mu-1}G_\mu^{-1} + G_\mu(I-P_1\cdots P_{\mu-1})G_\mu^{-1}\\ &= G_\mu P_1\cdots P_{\mu-1}G_\mu^{-1} + G_\mu Q_1P_2\cdots P_{\mu-1}G_\mu^{-1}\\ &\qquad + \cdots + G_\mu Q_{\mu-2}P_{\mu-1}G_\mu^{-1} + G_\mu Q_{\mu-1}G_\mu^{-1}\\ &= G_\mu P_1\cdots P_{\mu-1}G_\mu^{-1} + B_1\mathcal{L}_1 + \cdots + B_{\mu-2}\mathcal{L}_{\mu-2} + B_{\mu-1}\mathcal{L}_{\mu-1}. \end{aligned}$$

By this we define the additional bounded operator $T_\pi : H^1_D \longrightarrow L^2$,

$$T_\pi x := G_\mu P_1\cdots P_{\mu-1}G_\mu^{-1}Tx + \sum_{i=1}^{\mu-1} B_{\mu-i}\mathcal{L}_{\mu-i}[\mathfrak{A}_{\pi,\mu-i}(Dx)' + \mathfrak{B}_{\pi,\mu-i}x],\quad x\in H^1_D,$$

and investigate the difference

$$\begin{aligned} Tx - T_\pi x &= \sum_{i=1}^{\mu-1} B_{\mu-i}\mathcal{L}_{\mu-i}[(A-\mathfrak{A}_{\pi,\mu-i})(Dx)' + (B-\mathfrak{B}_{\pi,\mu-i})x]\\ &= \sum_{i=1}^{\mu-1} B_{\mu-i}\mathcal{L}_{\mu-i}[(\mathfrak{A}_{\mu-i}-\mathfrak{A}_{\pi,\mu-i})(Dx)' + (\mathfrak{B}_{\mu-i}-\mathfrak{B}_{\pi,\mu-i})x]. \end{aligned}$$

Regarding the relations (25) we know that there is a constant $C_{L^2}>0$ such that

$$(26)\qquad \|Tx - T_\pi x\|_{L^2} \le hC_{L^2}\|x\|_{H^1_D},\quad x\in H^1_D.$$

Next we estimate the difference $Tp - T_\pi p$ for $p\in X_\pi$ in the Y-norm. Since both Tp and $T_\pi p$ belong to the space Y^0_π we can use Lemma 4.4 and obtain

$$\begin{aligned} \|Tp - T_\pi p\|^2_Y &\le \|Tp - T_\pi p\|^2_\pi\\ &= \|Tp - T_\pi p\|^2_{L^2} + \sum_{i=1}^{\mu-1}\sum_{s=0}^{\mu-i} d_{i,s}\|(D\mathcal{L}_{\mu-i}(Tp - T_\pi p))^{(s)}\|^2_{L^2}. \end{aligned}$$

Owing to the properties of the projector functions it holds that $\mathcal{L}_{\mu-i}B_{\mu-i}\mathcal{L}_{\mu-i} = \mathcal{L}_{\mu-i}$ and, for $i \neq j$, $\mathcal{L}_{\mu-j}B_{\mu-i}\mathcal{L}_{\mu-i} = 0$. This implies

$$\begin{aligned} D\mathcal{L}_{\mu-i}(Tp - T_\pi p) &= D\mathcal{L}_{\mu-i}\{(\mathfrak{A}_{\mu-i} - \mathfrak{A}_{\pi,\mu-i})(Dp)' + (\mathfrak{B}_{\mu-i} - \mathfrak{B}_{\pi,\mu-i})p\} \\ &= \underbrace{D\mathcal{L}_{\mu-i}[(\mathfrak{A}_{\mu-i} - \mathfrak{A}_{\pi,\mu-i})\ (\mathfrak{B}_{\mu-i} - \mathfrak{B}_{\pi,\mu-i})]}_{=:W_{\mu-i}} \begin{bmatrix} (Dp)' \\ p \end{bmatrix} =: W_{\mu-i}\tilde{p}. \qquad (27) \end{aligned}$$

The matrix function $W_{\mu-i}$ is again of class $\mathcal{C}^{\mu-i}$, and $\tilde{p}$ is piecewise polynomial. Further, the expressions

$$\frac{1}{h^{\mu-i-s}}\|W_{\mu-i}^{(s)}\|_\infty, \quad s = 0,\dots,\mu-i,\ i = 1,\dots,\mu-1,$$

become arbitrarily small if h tends to zero. Deriving

$$\begin{aligned} (D\mathcal{L}_{\mu-i}(Tp - T_\pi p))^{(s)} &= (W_{\mu-i}\tilde{p})^{(s)} \\ &= W_{\mu-i}^{(s)}\tilde{p} + sW_{\mu-i}^{(s-1)}\tilde{p}^{(1)} + \ldots + sW_{\mu-i}^{(1)}\tilde{p}^{(\mu-i-1)} + W_{\mu-i}\tilde{p}^{(\mu-i)} \end{aligned}$$

and using Lemma 4.2 we estimate

$$\begin{aligned} \|(D\mathcal{L}_{\mu-i}(Tp - T_\pi p))\|_{L^2} &\leq \|W_{\mu-i}\|_\infty\|\tilde{p}\|_{L^2} = \|W_{\mu-i}\|_\infty\|p\|_{H_D^1}, \\ \|(D\mathcal{L}_{\mu-i}(Tp - T_\pi p))'\|_{L^2} &\leq \|W'_{\mu-i}\|_\infty\|\tilde{p}\|_{L^2} + \|W_{\mu-i}\|_\infty\|\tilde{p}'\|_{L^2} \\ &= \|W'_{\mu-i}\|_\infty\|p\|_{H_D^1} + \|W_{\mu-i}\|_\infty\|p'\|_{H_D^1} \\ &\leq \|W'_{\mu-i}\|_\infty\|p\|_{H_D^1} + \|W_{\mu-i}\|_\infty\frac{\sqrt{C_1^*}}{h_{\min}}\|p\|_{H_D^1} \\ &\leq \Big(\|W'_{\mu-i}\|_\infty + \|W_{\mu-i}\|_\infty\frac{\sqrt{C_1^*}\rho}{h}\Big)\|p\|_{H_D^1} \end{aligned}$$

and, analogously, for $s = 2,\dots,\mu-i$,

$$\|(D\mathcal{L}_{\mu-i}(Tp - T_\pi p))^{(s)}\|_{L^2} \leq \Big(\|W_{\mu-i}^{(s)}\|_\infty + \ldots + \|W_{\mu-i}\|_\infty\frac{\sqrt{C_s^*}\rho^s}{h^s}\Big)\|p\|_{H_D^1}.$$

Consequently, it follows that the inequalities

$$\|(D\mathcal{L}_{\mu-i}(Tp - T_\pi p))^{(s)}\|_{L^2} \leq \varepsilon_{i,s}\|p\|_{H_D^1}, \quad p \in X_\pi,$$

and, finally,

$$\|(Tp - T_\pi p)\|_Y \leq \varepsilon_Y\|p\|_{H_D^1}, \quad p \in X_\pi, \qquad (28)$$

are valid with values $\varepsilon_{i,s}$, ε_Y being arbitrarily small if h is sufficiently small.

Owing to Proposition 4.3 it holds that

$$\inf_{p\in X_\pi, p\neq 0} \frac{(\|Tp\|_Y^2 + |G_a p(a) + G_b p(b)|^2)^{1/2}}{\|p\|_{H_D^1}} \geq \frac{1}{c_Y}.$$

On the other hand, regarding (28) we can estimate

$$\|T_\pi p\|_Y = \|Tp - (Tp - T_\pi p)\|_Y \geq \|Tp\|_Y - \|Tp - T_\pi p\|_Y \geq \|Tp\|_Y - \varepsilon_Y\|p\|_{H_D^1},$$

and

$$
\begin{aligned}
(\|T_\pi p\|_Y^2 + |G_a p(a) + G_b p(b)|^2)^{1/2} &\geq \frac{1}{\sqrt{2}}\big(\|T_\pi p\|_Y + |G_a p(a) + G_b p(b)|\big)\\
&\geq \frac{1}{\sqrt{2}}\big(\|Tp\|_Y + |G_a p(a) + G_b p(b)| - \varepsilon_Y \|p\|_{H_D^1}\big)\\
&\geq \frac{1}{\sqrt{2}}\big((\|Tp\|_Y^2 + |G_a p(a) + G_b p(b)|^2)^{1/2} - \varepsilon_Y \|p\|_{H_D^1}\big)\\
&\geq \frac{1}{\sqrt{2}}\Big(\frac{1}{c_Y} - \varepsilon_Y\Big)\|p\|_{H_D^1}.
\end{aligned}
$$

Since ε_Y becomes arbitrarily small for sufficiently fine partitions, it results that

$$
\inf_{p\in X_\pi, p\neq 0} \frac{\|T_\pi p\|_Y^2 + |G_a p(a) + G_b p(b)|^2}{\|p\|_{H_D^1}^2} \geq \frac{1}{8c_Y^2}. \tag{29}
$$

We summarize what we have obtained so far as follows:

$$
\begin{aligned}
\gamma_\pi &= \inf_{p\in X_\pi, p\neq 0} \frac{(\|Tp\|_{L^2}^2 + |G_a p(a) + G_b p(b)|^2)^{1/2}}{\|p\|_{H_D^1}}\\
&= \inf_{p\in X_\pi, p\neq 0} \underbrace{\frac{(\|T_\pi p\|_Y^2 + |G_a p(a) + G_b p(b)|^2)^{1/2}}{\|p\|_{H_D^1}}}_{\geq(\sqrt{8}c_Y)^{-1}}\\
&\qquad\times \underbrace{\left(\frac{\|Tp\|_{L^2}^2 + |G_a p(a) + G_b p(b)|^2}{\|T_\pi p\|_Y^2 + |G_a p(a) + G_b p(b)|^2}\right)^{1/2}}_{=:\mathfrak{E}(p)}.
\end{aligned}
$$

Next we provide an estimate of the expression $\mathfrak{E}(p)$.

For each given nontrivial $p \in X_\pi$, the corresponding $q = T_\pi p$ belongs to Y_π^0, and the inequality (29) implies

$$
\|p\|_{H_D^1}^2 \leq 8c_Y^2(\|q\|_Y^2 + |G_a p(a) + G_b p(b)|^2). \tag{30}
$$

Denote further, for $i = 1, \ldots, \mu - 1$,

$$
q_{\mu-i} = \mathfrak{A}_{\pi,\mu-i}(Dp)' + \mathfrak{B}_{\pi,\mu-i}p,
$$

such that

$$
\mathcal{L}_{\mu-i}q = \mathcal{L}_{\mu-i}q_{\mu-i}, \quad U_{\mu-i}q = U_{\mu-i}q_{\mu-i}.
$$

Owing to Lemma 4.4 we can estimate

$$
\begin{aligned}
\|q\|_Y^2 \leq \|q\|_\pi^2 &= \|q\|_{L^2}^2 + \sum_{i=1}^{\mu-1}\sum_{s=0}^{\mu-i} d_{i,s}\ \|(D\mathcal{L}_{\mu-i}q)^{(s)}\|_{L^2}^2\\
&= \|q\|_{L^2}^2 + \sum_{i=1}^{\mu-1}\sum_{s=0}^{\mu-i} d_{i,s}\ \|(D\mathcal{L}_{\mu-i}q_{\mu-i})^{(s)}\|_{L^2}^2.
\end{aligned}
$$

Deriving

$$
(D\mathcal{L}_{\mu-i}q_{\mu-i})^{(s)} = (D\mathcal{L}_{\mu-i})^{(0)}(q_{\mu-i})^{(s)} + \ldots + (D\mathcal{L}_{\mu-i})^{(s)}(q_{\mu-i})^{(0)}
$$

yields

$$\begin{aligned}\|(D\mathcal{L}_{\mu-i}q_{\mu-i})^{(s)}\|_{L^2} \leq & \|(D\mathcal{L}_{\mu-i})^{(0)}\|_\infty \|(q_{\mu-i})^{(s)}\|_{L^2} \\ & + \ldots + \|(D\mathcal{L}_{\mu-i})^{(s)}\|_\infty \|(q_{\mu-i})^{(0)}\|_{L^2}.\end{aligned}$$

Since each component of $q_{\mu-i}$ is piecewise polynomial, we obtain by Lemma 4.2 the further inequalities

$$\begin{aligned}\|(D\mathcal{L}_{\mu-i}q_{\mu-i})^{(s)}\|_{L^2} &\leq \frac{1}{h^s_{\min}}\left(\sqrt{c^*_s}\|D\mathcal{L}_{\mu-i}\|_\infty + O(h)\right) \|q_{\mu-i}\|_{L^2}, \\ \|(D\mathcal{L}_{\mu-i}q_{\mu-i})^{(s)}\|^2_{L^2} &\leq \frac{1}{h^{2s}_{\min}}\left(c^*_s\|D\mathcal{L}_{\mu-i}\|^2_\infty + O(h)\right) \|q_{\mu-i}\|^2_{L^2},\end{aligned}$$

and

$$\begin{aligned}\|q\|^2_Y &\leq \|q\|^2_{L^2} + \sum_{i=1}^{\mu-1}\sum_{s=0}^{\mu-i} d_{i,s}\frac{1}{h^{2s}_{\min}}\left(c^*_s\|D\mathcal{L}_{\mu-i}\|^2_\infty + O(h)\right) \|q_{\mu-i}\|^2_{L^2} \\ &= \|q\|^2_{L^2} + \sum_{i=1}^{\mu-1}\frac{1}{h^{2\mu-2i}_{\min}}\big(\underbrace{d_{i,\mu-i}c^*_{\mu-i}\|D\mathcal{L}_{\mu-i}\|^2_\infty}_{=:g_{\mu-i}} + O(h)\big) \|q_{\mu-i}\|^2_{L^2}.\end{aligned}$$

So far we have the relation

$$\begin{aligned}&\|q\|^2_Y + |G_a p(a) + G_b p(b)|^2 \\ &\leq \|q\|^2_{L^2} + |G_a p(a) + G_b p(b)|^2 + \sum_{i=1}^{\mu-1}\frac{2}{h^{2\mu-2i}_{\min}} g_{\mu-i}\, \|q_{\mu-i}\|^2_{L^2}, \qquad (31)\end{aligned}$$

with

$$g_{\mu-1} = d_{1,\mu-1}c^*_{\mu-1}\|D\mathcal{L}_{\mu-1}\|^2_\infty > 0; \qquad (32)$$

see Lemma 4.4 for $d_{1,\mu-1}$ and Lemma 4.2 for $c^*_{\mu-1}$.

We have $q = T_\pi p$, $q = Tp - (Tp - T_\pi p)$, and regarding (26), (30),

$$\begin{aligned}\|q\|^2_{L^2} &\leq 2\|Tp\|^2_{L^2} + 2h^2 C^2_{L^2}\|p\|^2_{H^1_D} \\ &\leq 2\|Tp\|^2_{L^2} + 2h^2 C^2_{L^2} 8c^2_Y(\|q\|^2_Y + |G_a p(a) + G_b p(b)|^2). \qquad (33)\end{aligned}$$

Using (27) we find for $i = 1, \ldots, \mu - 1$ that

$$\begin{aligned}U_{\mu-i}q_{\mu-i} = U_{\mu-i}q &= U_{\mu-i}Tp - U_{\mu-i}(Tp - T_\pi p) \\ &= U_{\mu-i}Tp + (D\mathcal{L}_{\mu-i})^+ W_{\mu-i}\tilde{p}, \qquad (34)\end{aligned}$$

in which the expressions $\omega_{\mu-i} := \frac{1}{h^{\mu-i}}\|W_{\mu-i}\|^2_\infty$ become arbitrarily small for sufficiently fine partitions.

Next, if the projector functions $U_{\mu-i}$ are constant ones, we know that

$$q_{\mu-i} = U_{\mu-i}q_{\mu-i} = U_{\mu-i}q = U_{\mu-i}Tp + (D\mathcal{L}_{\mu-i})^+ W_{\mu-i}\tilde{p},$$

thus

$$\begin{aligned}\|q_{\mu-i}\|^2_{L^2} &\leq 2\|Tp\|^2_{L^2} + 2\|(D\mathcal{L}_{\mu-i})^+\|^2_\infty h^{2\mu-2i}\omega^2_{\mu-i}\|p\|^2_{H^1_D} \\ &\leq 2\|Tp\|^2_{L^2} + 2\|(D\mathcal{L}_{\mu-i})^+\|^2_\infty h^{2\mu-2i}\omega^2_{\mu-i} 8c^2_Y(\|q\|^2_Y + |G_a p(a) + G_b p(b)|^2).\end{aligned}$$

Now we obtain from (31) the inequality

$$(1-\Omega)(\|q\|_Y^2 + |G_ap(a)+G_bp(b)|^2) \\ \leq 2\|Tp\|_{L^2}^2 + |G_ap(a)+G_bp(b)|^2 + 2\sum_{i=1}^{\mu-1}\frac{2}{h_{\min}^{2\mu-2i}} g_{\mu-i}\ \|Tp\|_{L^2}^2,$$

with

$$\Omega := 2h^2C_{L^2}^2 8c_Y^2 + 2\sum_{i=1}^{\mu-1}\Big(\frac{h}{h_{\min}}\Big)^{2\mu-2i} g_{\mu-i}\|(D\mathcal{L}_{\mu-i})^+\|_\infty^2 \omega_{\mu-i}^2 8c_Y^2$$

being arbitrarily small together with $\omega_1,\ldots,\omega_{\mu-1}$ for sufficiently fine partitions. Supposing the partitions to be fine enough such that $\Omega \leq 1/2$ we arrive at the inequalities

$$\begin{aligned}
&\|T_\pi p\|_Y^2 + |G_ap(a)+G_bp(b)|^2 \\
&\quad\leq 4(\|Tp\|_{L^2}^2 + |G_ap(a)+G_bp(b)|^2) + 4\sum_{i=1}^{\mu-1}\frac{2}{h_{\min}^{2\mu-2i}} g_{\mu-i}\ \|Tp\|_{L^2}^2 \\
&\quad\leq 4(\|Tp\|_{L^2}^2 + |G_ap(a)+G_bp(b)|^2) + \frac{1}{h_{\min}^{2\mu-2}}(8g_{\mu-1}+O(h^2))\|Tp\|_{L^2}^2 \\
&\quad\leq \frac{1}{h_{\min}^{2\mu-2}}(8g_{\mu-1}+O(h^2))(\|Tp\|_{L^2}^2 + |G_ap(a)+G_bp(b)|^2) \\
&\quad\leq \frac{1}{h_{\min}^{2\mu-2}}(9g_{\mu-1})(\|Tp\|_{L^2}^2 + |G_ap(a)+G_bp(b)|^2),
\end{aligned}$$

and hence,

$$\mathfrak{E}(p) \geq \frac{1}{3\sqrt{g_{\mu-1}}} h_{\min}^{\mu-1}.$$

If the projector functions $U_{\mu-i}$ vary with t, the situation is slightly more involved. Then, regarding the properties (24) we express, for $t\in[t_{j-1},t_j),\ j=1,\ldots,n$,

$$\begin{aligned}
\mathfrak{A}_{\pi,\mu-i}(t) &= \sum_{s=0}^{\mu-i}\frac{1}{s!}(t-t_{j-1/2})^s\ \mathfrak{A}_{\pi,\mu-i}^{(s)}(t_{j-1/2}) + \mathfrak{R}_{\mathfrak{A},\mu-i}(t) \\
&= \sum_{s=0}^{\mu-i}\frac{1}{s!}(t-t_{j-1/2})^s\ (U_{\mu-i}A)^{(s)}(t_{j-1/2}) + \mathfrak{R}_{\mathfrak{A},\mu-i}(t),
\end{aligned}$$

and, analogously,

$$\begin{aligned}
\mathfrak{B}_{\pi,\mu-i}(t) &= \sum_{s=0}^{\mu-i}\frac{1}{s!}(t-t_{j-1/2})^s\ \mathfrak{B}_{\pi,\mu-i}^{(s)}(t_{j-1/2}) + \mathfrak{R}_{\mathfrak{B},\mu-i}(t) \\
&= \sum_{s=0}^{\mu-i}\frac{1}{s!}(t-t_{j-1/2})^s\ (U_{\mu-i}B)^{(s)}(t_{j-1/2}) + \mathfrak{R}_{\mathfrak{B},\mu-i}(t).
\end{aligned}$$

Since the components of $\mathfrak{R}_{\mathfrak{A},\mu-i}$ and $\mathfrak{R}_{\mathfrak{B},\mu-i}$ are piecewise smooth as polynomials it results that

$$\|\mathfrak{R}_{\mathfrak{A},\mu-i}\|_\infty = O(h^{\mu-i+1}), \quad \|\mathfrak{R}_{\mathfrak{B},\mu-i}\|_\infty = O(h^{\mu-i+1}), \quad i=1,\ldots,\mu-1.$$

This leads to the relations

$$\begin{aligned} U_{\pi,\mu-i}\mathfrak{A}_{\pi,\mu-i} &= \mathfrak{A}_{\pi,\mu-i} + U_{\pi,\mu-i}\mathfrak{R}_{\mathfrak{A},\mu-i}, \\ U_{\pi,\mu-i}\mathfrak{B}_{\pi,\mu-i} &= \mathfrak{B}_{\pi,\mu-i} + U_{\pi,\mu-i}\mathfrak{R}_{\mathfrak{B},\mu-i}, \end{aligned}$$

and

$$\begin{aligned} q_{\mu-i} &= \mathfrak{A}_{\pi,\mu-i}(Dp)' + \mathfrak{B}_{\pi,\mu-i}p \\ &= U_{\pi,\mu-i}q_{\mu-i} - U_{\pi,\mu-i}(\mathfrak{R}_{\mathfrak{A},\mu-i}(Dp)' + \mathfrak{R}_{\mathfrak{B},\mu-i}p) \\ &= \underbrace{U_{\mu-i}q_{\mu-i}}_{=U_{\mu-i}q} + (U_{\pi,\mu-i} - U_{\mu-i})q_{\mu-i} - U_{\pi,\mu-i}(\mathfrak{R}_{\mathfrak{A},\mu-i}(Dp)' + \mathfrak{R}_{\mathfrak{B},\mu-i}p). \end{aligned}$$

It follows that

$$(I - (U_{\pi,\mu-i} - U_{\mu-i}))q_{\mu-i} = U_{\mu-i}q - U_{\pi,\mu-i}(\mathfrak{R}_{\mathfrak{A},\mu-i}(Dp)' + \mathfrak{R}_{\mathfrak{B},\mu-i}p)$$

and regarding (34)

$$\begin{aligned} (I-(U_{\pi,\mu-i} - U_{\mu-i}))q_{\mu-i} &= U_{\mu-i}q - U_{\pi,\mu-i}(\mathfrak{R}_{\mathfrak{A},\mu-i}(Dp)' + \mathfrak{R}_{\mathfrak{B},\mu-i}p) \\ &= U_{\mu-i}Tp - (D\mathcal{L}_{\mu-i})^+W_{\mu-i}\tilde{p} - U_{\pi,\mu-i}(\mathfrak{R}_{\mathfrak{A},\mu-i}(Dp)' + \mathfrak{R}_{\mathfrak{B},\mu-i}p). \end{aligned}$$

For sufficiently fine partitions we estimate $\|(I-(U_{\pi,\mu-i}-U_{\mu-i}))^{-1}\|_\infty \leq 2$. This gives

$$\|q_{\mu-i}\|_{L^2} \leq 2\|Tp\|_{L^2} + h^{\mu-i}\beta_{\mu-i}\|p\|_{H^1_D},$$

with $\beta_{\mu-i} := 2\|(D\mathcal{L}_{\mu-i})^+\|_\infty\ \omega_{\mu-i} + 2\max\{\frac{1}{h^{\mu-i}}\|\mathfrak{R}_{\mathfrak{A},\mu-i}\|_\infty, \frac{1}{h^{\mu-i}}\|\mathfrak{R}_{\mathfrak{A},\mu-i}\|_\infty\}$ being arbitrarily small for h tending to zero.

Also regarding (30) we arrive at

$$\|q_{\mu-i}\|^2_{L^2} \leq 4\|Tp\|^2_{L^2} + 2h^{2(\mu-i)}\beta^2_{\mu-i}8c^2_Y(\|q\|^2_Y + |G_ap(a) + G_bp(b)|^2).$$

Inserting this result into (31) and also using (33) we arrive at the inequalities

$$\begin{aligned} (1-\tilde{\Omega})(\ &\|T_\pi p\|^2_Y + |G_ap(a) + G_bp(b)|^2\ ) \\ &\leq 2\|Tp\|^2_{L^2} + |G_ap(a) + G_bp(b)|^2 + \sum_{i=1}^{\mu-1}\frac{8}{h_{\min}^{2\mu-2i}}g_{\mu-i}\ \|Tp\|^2_{L^2} \\ &\leq \frac{1}{h_{\min}^{2\mu-2}}(8g_{\mu-1} + O(h^2))\,(\|Tp\|^2_{L^2} + |G_ap(a) + G_bp(b)|^2\ ) \\ &\leq \frac{1}{h_{\min}^{2\mu-2}}9g_{\mu-1}\,(\|Tp\|^2_{L^2} + |G_ap(a) + G_bp(b)|^2\ ) \end{aligned}$$

with

$$\tilde{\Omega} := \left(2h^2C^2_{L^2} + \sum_{i=1}^{\mu-1}\beta^2_{\mu-i}\left(\frac{h}{h_{\min}}\right)^{2\mu-2i}g_{\mu-i}\right)8c^2_Y.$$

Since $\tilde{\Omega} \leq 1/2$ for sufficiently fine partitions, we finally obtain

$$\|T_\pi p\|^2_Y + |G_ap(a) + G_bp(b)|^2 \leq \frac{2}{h_{\min}^{2\mu-2}}9g_{\mu-1}(\|Tp\|^2_{L^2} + |G_ap(a) + G_bp(b)|^2)$$

yielding

$$\mathfrak{C}(p) \geq \frac{1}{3\sqrt{2g_{\mu-1}}}h_{\min}^{\mu-1}.$$

Summarizing all we know means

$$\gamma_\pi \geq \frac{1}{\sqrt{8}} \frac{1}{c_Y} \frac{1}{3\sqrt{2g_{\mu-1}}} h_{\min}^{\mu-1} = \frac{1}{12 c_Y \sqrt{g_{\mu-1}}} h_{\min}^{\mu-1}.$$

□

Remark 4.5. The proof renders details concerning the constant c_γ. In the index-1 case one has simply $c_\gamma = c_Y^{-1}$ with c_Y from Proposition 4.3(b). In the higher-index case the constant c_γ provided in Theorem 4.1(c) is inversely proportional to the value c_Y from Proposition 4.3(b) and also to $\sqrt{g_{\mu-1}}$, with

$$g_{\mu-1} = d_{1,\mu-1} c_{\mu-1}^{*} \|D\mathcal{L}_{\mu-1}\|_\infty^2 > 0;$$

see Lemma 4.4 for $d_{1,\mu-1}$ and Lemma 4.2 for $c_{\mu-1}^*$. We stress again that $c_{\mu-1}^*$ increases with the polynomial degree N.

Remark 4.6. For index-2 DAEs Theorem 4.1 offers one constant c_γ in item (b) and another one in item (c), namely

$$c_{\gamma,b} = \frac{1}{3} \frac{1}{c_Y} \frac{1}{\sqrt{c_1^*}} \frac{1}{\|D\Pi_0 Q_1 D^+\|_\infty \|D\mathcal{L}_1\|_\infty} \frac{1}{\sqrt{1 + K\|D\mathcal{L}_1\|_\infty^{-2}}},$$

$$c_{\gamma,c} = \frac{1}{12\sqrt{2}} \frac{1}{c_Y} \frac{1}{\sqrt{c_1^*}} \frac{1}{\|DQ_1\|_\infty \|D\mathcal{L}_1\|_\infty},$$

that have been derived in completely different ways. The first constant is obtained by a special, straightforward proof, the second one appears as a particular case of a much more complicated general proof. Nevertheless, they have main components in common.

Owing to the special form of D, $|D|= 1, |D^+|= 1$, and $DQ_1 = D\Pi_0 Q_1 D^+ D$, it holds that $\|DQ_1\|_\infty = \|D\Pi_0 Q_1 D^+\|_\infty$.

Note that $\Pi_1 = 0$ implies $K = 0$, thus $c_{\gamma,b} = 4\sqrt{2}\, c_{\gamma,c}$.

4.4. On possible stronger estimates if $1 \leq N < \mu - 1$.

The question if the stronger estimate (17) is valid matters for DAEs with index $\mu \geq 3$ only. We address this question in the context of Subsection 4.3, that is, the proof of Part (c) of Theorem 4.1, and we take the notation from Subsection 4.3.

For $K \geq 0$, let $\mathcal{P}_{\pi,K}^m$ denote the set of componentwise piecewise polynomial functions $[a,b] \to \mathbb{R}^m$ of degree less than or equal to K.

Let $K_{\mu-i}$ denote the minimal polynomial degree such that

$$q_{\mu-i} := \mathfrak{A}_{\pi,\mu-1}(Dp)' + \mathfrak{B}_{\pi,\mu-1} p \in \mathcal{P}_{\pi,K_{\mu-i}}^m \quad \text{for all } p \in X_\pi, \quad i = 1,\ldots,\mu-1.$$

Since $\mathfrak{A}_{\pi,\mu-1}$ and $\mathfrak{B}_{\pi,\mu-1}$ have degree at most $2(\mu-i)$, by construction, it holds that $1 \leq N \leq K_{\mu-i} \leq 2(\mu-i) + N \leq 2(\mu-1) + N$.

In the case of constant coefficients A and B, one has $K_{\mu-i} = N$ for $i = 1, \ldots, \mu-1$. The following theorem generalizes the respective result obtained in [3] for $N = 1$.

Theorem 4.7. *Let the bounded DA operator $T : H_D^1 \to L^2$ be associated with a constant-coefficient DAE with index $\mu \in \mathbb{N}$ and characteristic values* (12) *and let the boundary conditions be restricted by* (4). *Let the condition* (13) *be valid.*

Let X_π be given by (5) *as before, and $N \geq 1$.*

Then, there is a constant $c_\gamma > 0$ *such that*

$$\gamma_\pi \geq c_\gamma h_{\min}^{\min(N,\mu-1)} \geq c_\gamma \frac{1}{\rho^{\min(N,\mu-1)}} h^{\min(N,\mu-1)}$$

for all partitions π *with sufficiently small maximal stepsizes* h *and uniformly bounded ratios* $\frac{h}{h_{\min}} \leq \rho$.

Proof. Since Theorem 4.1 applies[9], it remains to show the stronger inequality for the case $1 \leq N \leq \mu - 2$, $\mu \geq 3$. Put $i_* := \mu - N$, $2 \leq i_* \leq \mu - 1$. We continue using the framework of Theorem 4.1 and its proof.

Let $p \in X_\pi$ and $q = T_\pi p$. Then, $D\mathcal{L}_{\mu-i}q_{\mu-i} = D\mathcal{L}_{\mu-i}q \in \mathcal{P}^k_{\pi,N}$ such that $\|D\mathcal{L}_{\mu-i}q_{\mu-i}\|_{L^2} \leq \|D\mathcal{L}_{\mu-i}\|_\infty \|q\|_{L^2}$, $i = 1, \ldots, \mu-1$. Regarding that the derivatives $(D\mathcal{L}_{\mu-i}q_{\mu-i})^{(s)}$, $s \geq N+1$, vanish, and applying Lemma 4.2, we derive

$$\begin{aligned}
&\sum_{i=1}^{\mu-1}\sum_{s=0}^{\mu-i} d_{i,s}\|(D\mathcal{L}_{\mu-i}q_{\mu-i})^{(s)}\|^2_{L^2} \\
&\quad = \sum_{i=1}^{i_*}\sum_{s=0}^{\mu-i_*} d_{i,s}\|(D\mathcal{L}_{\mu-i}q_{\mu-i})^{(s)}\|^2_{L^2} + \sum_{i=i_*+1}^{\mu-1}\sum_{s=0}^{\mu-i} d_{i,s}\|(D\mathcal{L}_{\mu-i}q_{\mu-i})^{(s)}\|^2_{L^2} \\
&\quad \leq \frac{1}{h_{\min}^{2N}}\Big[\sum_{i=1}^{i_*} \underbrace{d_{i,\mu-i_*}c_N^*\|D\mathcal{L}_{\mu-i_*}\|^2_\infty}_{=g_{\mu-i_*}} + O(h^2)\Big]\,\|q\|^2_{L^2}
\end{aligned}$$

and then

$$\begin{aligned}
&\|q\|^2_Y + |G_a p(a) + G_b p(b)|^2 \\
&\qquad \leq \|q\|^2_{L^2} + |G_a p(a) + G_b p(b)|^2 + \sum_{i=1}^{\mu-1}\sum_{s=0}^{\mu-i} d_{i,s}\|(D\mathcal{L}_{\mu-i}q_{\mu-i})^{(s)}\|^2_{L^2} \\
&\qquad \leq \|q\|^2_{L^2} + |G_a p(a) + G_b p(b)|^2 + \frac{1}{h_{\min}^{2N}}\Big[\sum_{i=1}^{i_*} g_{\mu-i_*} + O(h^2)\Big]\|q\|^2_{L^2} \\
&\qquad \leq \frac{1}{h_{\min}^{2N}}\sum_{i=1}^{i_*} 2g_{\mu-i_*}(\|q\|^2_{L^2} + |G_a p(a) + G_b p(b)|^2).
\end{aligned}$$

This leads to

$$\mathfrak{C} \geq h_{\min}^N \frac{1}{\sqrt{\sum_{i=1}^{i_*} 2g_{\mu-i_*}}},$$

and hence

$$\gamma_\pi \geq h_{\min}^N \frac{1}{\sqrt{8}c_Y\sqrt{\sum_{i=1}^{i_*} 2g_{\mu-i_*}}} = c_\gamma h_{\min}^N.$$

□

It may happen also for DAEs with time-varying coefficients A and B that $K_{\mu-i} < \mu - i$, and possibly, the order reduces.

[9] Note that for constant A and B the proof of Theorem 4.1 simplifies essentially regarding that then $T = T_\pi$.

Example 4.8. We inspect the index-3 DAE from Example 1.1 in more detail. The projector functions

$$Q_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, Q_1(t) = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & -t\eta & 0 \end{bmatrix}, Q_2(t) = \begin{bmatrix} 0 & t\eta & 1 \\ 0 & -t\eta & -1 \\ 0 & t\eta(1+t\eta) & 1+t\eta \end{bmatrix},$$

generate a fine decoupling and yield further

$$\mathcal{L}_1(t) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -t\eta & 0 \end{bmatrix}, U_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \mathcal{L}_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, U_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

and, on each interval $[t_{j-1}, t_j)$,

$$q_1(t) = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} (Dp)'(t) + \begin{bmatrix} 1+t-t_{j-1/2} & 0 & 0 \\ 0 & 0 & t\eta \\ 0 & 0 & 1+t-t_{j-1/2} \end{bmatrix} p(t), \quad p \in X_\pi,$$

$$q_2(t) = \begin{bmatrix} 1 & 0 \\ t\eta & 1 \\ 0 & 0 \end{bmatrix} (Dp)'(t) + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1+\eta & 0 \\ 0 & 0 & 0 \end{bmatrix} p(t), \quad p \in X_\pi.$$

We observe that $K_1 = N + 1$ and $K_2 = N$. Recall that Theorem 4.1 provides the estimate $\gamma_\pi \geq c_\gamma h^{-2}$ for all $N \geq 1$.

Set $N = 1$. Then the derivative q_2'' disappears in the treatment of the term $\mathfrak{C}$, and, therefore, the stronger estimate $\gamma_\pi \geq \bar{c}_\gamma h^{-1}$ is also valid. □

In general, working with low-degree ansatz functions, that is, $1 \leq N \leq \mu - 2$, stronger estimates might be valid. When estimating the expression $\mathfrak{C}$ in Subsection 4.3, Part (c), we can then replace certain inequalities $\|q^{(s)}_{\mu-i}\|_{L^2} \leq \sqrt{c^*_s} h^{-s}_{\min} \|q_{\mu-i}\|_{L^2}$ by $\|q^{(s)}_{\mu-i}\|_{L^2} = 0$ accordingly.

For instance, if $K_{\mu-1} \leq \mu - 2$, then $\gamma_\pi \geq \bar{c}_\gamma h^{\mu-2}$ is valid, and $K_{\mu-1} \leq \mu - 3$, $K_{\mu-1} \leq \mu - 3$ imply $\gamma_\pi \geq \bar{c}_\gamma h^{\mu-3}$, and so on.

5. Least-squares collocation via discretized norms

The present section is devoted to the case where the coefficients A and B have only polynomial entries. We show convergence of the least-squares collocation for sufficiently large M.

As described in Section 1, the least-squares collocation applied to a uniquely solvable BVP (1)–(2) means that we solve the overdetermined collocation scheme (7)–(8) comprising $Mnm+l$ equations in the least-squares sense, that is, we actually minimize the functional (9),

(35)

$$\phi_{\pi,M}(p) = \sum_{j=1}^{n} \frac{h_j}{M} \sum_{t \in S_j} |A(t)(Dp)'(t) + B(t)p(t) - y(t)|^2 + |G_a p(a) + G_b p(b) - r|^2,$$

subject to $p \in X_\pi$. Let the linear space of piecewise polynomial functions X_π be defined as before in (5), let the $M > N$ interpolation nodes $0 < \tau_1 < \tau_2 < \cdots < \tau_M < 1$ be fixed, and let S_j be the resulting sets of collocation points on the subinterval (t_{j-1}, t_j) as before.

Assuming y to be sufficiently smooth, so that for every $t \in S_j$ the function value $y(t)$ is well-defined and interpolation makes sense, we denote by y_π the interpolating piecewise polynomial defined by

$$y_\pi|_{[t_{j-1},t_j)} \in \mathcal{P}^m_{M-1}, \quad y_\pi(t) = y(t),\ t \in S_j,\ j = 1, \ldots, n. \tag{36}$$

Set $z_\pi = (y_\pi, r)$ such that

$$\delta_\pi := \|z - z_\pi\|_Z = \|y - y_\pi\|_{L^2} \le c_\delta h^M. \tag{37}$$

Following [4], it makes sense to turn to the perturbed equation $\mathcal{T}x = y_\pi$ and provide the further approximate solution

$$p_\pi^{\delta_\pi} \in \operatorname{argmin}\{\|Tp - y_\pi\|^2_{L^2} + |G_a p(a) + G_b p(b) - r|^2 :\ p \in X_\pi\}, \tag{38}$$

which satisfies the inequality

$$\|p_\pi^{\delta_\pi} - x_*\|_{H^1_D} \le \frac{\beta_\pi + \delta_\pi}{\gamma_\pi} + \alpha_\pi. \tag{39}$$

Replacing in (35) y by y_π does not at all change the value $\phi_{\pi,M}(p)$, which allows us to restrict the matter to such a piecewise polynomial right-hand side y_π, that is, instead of $\phi_\pi(p)$, we may minimize the modified functional

$$\tilde{\phi}_\pi(p) := \|Tp - y_\pi\|^2_{L^2} + |G_a p(a) + G_b p(b) - r|^2,\ p \in X_\pi. \tag{40}$$

We write for brevity

$$w := A(Dp)' + Bp - y_\pi = Tp - y_\pi,$$

and

$$W := \begin{bmatrix} W_1 \\ \vdots \\ W_n \end{bmatrix} \in \mathbb{R}^{mMn}, \quad W_j := \Big(\frac{h_j}{M}\Big)^{1/2} \begin{bmatrix} w(t_{j-1} + \tau_1 h_j) \\ \vdots \\ w(t_{j-1} + \tau_M h_j) \end{bmatrix} \in \mathbb{R}^{mM},$$

but we keep in mind that both the function w and the vector W depend on p. We further denote the Euclidean norm of $W \in \mathbb{R}^{mMn}$ by $|W|_2$ and arrive at the representation

$$\phi_{\pi,M}(p) = |W|_2^2 + |G_a p(a) + G_b p(b) - r|^2.$$

Next, let the entries of the coefficients A and B be polynomials of at most degree $N_{A,B}$ (at least on each subinterval of the coarsest partition π we start with). Then, for each $p \in X_\pi$, 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[10]

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

we ensure

$$\{Tp - y_\pi\}|_{[t_{j-1},t_j)} = \{A(Dp)' + Bp - y_\pi\}|_{[t_{j-1},t_j)} \in \mathcal{P}^m_{M-1}.$$

For such a special piecewise polynomial $w := A(Dp)' + Bp - y_\pi$ we derive along the lines of [3, Subsection 2.3] that

$$\|w\|^2_{L^2} = W^T \mathcal{L} W, \tag{41}$$

[10] Note that we have $N_{A,B} = 1$ and $M = 2N + 1$ in Example 1.1.

with a symmetric, positive definite matrix $\mathcal{L}$. Its entries do not depend on the partition π.[11] Further, there are positive constants c_L, $\bar{c}_L$ depending only on $\mathcal{L}$ such that

$$c_L|W|_2 \leq \|w\|_{L^2} \leq \bar{c}_L|W|_2. \tag{42}$$

Actually, the relation (42) indicates the equivalence of the L^2-norm and the norm defined by

$$\|w\|_2 := |W|_2$$

on the related finite-dimensional subspace in L^2; cf. [3, Proposition 2.7]. This gives

$$\phi_{\pi,M}(p) = \|w\|_2^2 + |G_a p(a) + G_b p(b) - r|^2.$$

Consequently, the least-squares collocation generates an approximate solution $p_\pi^{\delta_\pi}$, if instead of minimizing

$$\tilde{\phi}_\pi(p) = \|w\|_{L^2}^2 + |G_a p(a) + G_b p(b) - r|^2, \tag{43}$$

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 $\|w\|_2$.

At this place it should be emphasized that our previous experiments using both norms indicate no significant differences; see [3, Section 6].

As a consequence of the estimates (39) and (37) as well as the Theorems 4.1 and 4.7 we obtain the following sufficient convergence conditions.

Theorem 5.1. *Let the BVP* (1)–(2), *with index $\mu \geq 1$, satisfy the assumptions of Theorem* 2.4(a) *and have the unique, sufficiently smooth solution x_*. If the entries of the coefficients A and B are polynomials at most of degree $N_{A,B}$, X_π be given by* (5), *and M is chosen in such a way that $M \geq 1 + N + N_{A,B}$,*[12] *with $N \geq 1$, then the following statements are valid for all partitions π with sufficiently small stepsize h and uniformly bounded ratios $\frac{h}{h_{\min}} \leq \rho$:*

(a) *The least-squares collocation solutions $p_\pi^{\delta_\pi}$ of the overdetermined system* (7)–(8) *defined by* (38) *satisfy*

$$\|p_\pi^{\delta_\pi} - x_*\|_{H_D^1} \leq ch^{N-\mu+1}.$$

Hence, the choice of N such that $N \geq \mu$ ensures convergence in H_D^1, that is, $p_\pi^{\delta_\pi} \to x_*$ for $h \to 0$.

(b) *Moreover, if the coefficients A and B are constant* (*that is, $N_{A,B} = 0$*), *the solutions $p_\pi^{\delta_\pi}$ fulfill even*

$$\|p_\pi^{\delta_\pi} - x_*\|_{H_D^1} \leq ch^{\max(0,N-\mu+1)}$$

and the discrete solutions remain bounded in H_D^1 also if $N < \mu - 1$.

It should be mentioned that, under the conditions of Theorem 5.1, it holds always $M > N$ such that it does not apply to the standard collocation method.

Theorem 5.1 confirms merely sufficient convergence conditions for the restricted class of DAEs with polynomial coefficients. In contrast, the numerical experiments

[11] In [3] only equidistant partitions are considered. By marginal modifications the arguments remain valid also for general partitions.

[12] This can be generalized to the case of piecewise polynomial entries featuring a finite number of breakpoints. Then the breakpoints have to be incorporated into the partitions.

promise similar convergence behavior in much more general cases, e.g., Section 6, but so far there is no theoretical backup for this.

It should be emphasized that we are dealing with ill-posed problems. In practice, the choice of the X_π must be adapted to the problem at hand. In particular, some sort of discrepancy principle would be helpful in order to balance X_π, noise level, and instability; cf. [4]. In this context, quite a lot of questions remain to be answered.

6. Numerical experiments

The least-squares collocation depends on many choices for the individual components of the implementation including, but not restricted to,

- selection of the ansatz space X_π (5) defined by the grid π and the polynomial order $N \geq 1$;
- the representation of the ansatz functions $p \in X_\pi$;
- choice of the number $M > N$ of collocation points τ_i (6) and their placement.

The following examples shall give a first impression of the merits of the proposed method. In particular, we are interested in verifying the results about the order of convergence. Therefore, in all following experiments, equidistant grids π are used.

For a given space X_π, the representation of a basis does not matter in the absence of rounding errors. So we follow good practices and use a Runge-Kutta basis with interpolation points equidistantly distributed (cf. [3]). In our test we did not observe significant changes when using Gaussian points instead.

In the examples below we provide the errors and estimates of the order of convergence with respect to

- the polynomial degree N;
- different choices of the number M of collocation points and their placement;
- choice of the functionals ϕ_π (10) and $\phi_{\pi,M}$ (9), respectively.

We apply two different least-squares criteria: the functional $\phi_{\pi,M}$ (formula (9), also (35)) and the approximation of the basic functional ϕ_π which is given by formula (43). In the tables below the columns labeled $\mathbb{R}$ show the results for minimizing (35) whereas the columns labeled L^2 show the results for minimizing the expression (43).

In the tables below, the error is measured in the H^1_D-norm. The column labeled order contains an estimation k_{est} of the order,[13]

$$k_{\text{est}} = \log(\|p_n - x\|_{H^1_D} / \|p_{2n} - x\|_{H^1_D}) / \log 2.$$

The norm of an element $e \in H^1_D$ is approximated by interpolating both e and $(De)'$ on each subinterval of the partition at the M collocation points by polynomials. All computations have been done in Matlab.

6.1. **Continuation of the introductory Example 1.1.** We consider Example 1.1 in slightly more detail. The collocation points τ_i in (6) have been chosen to be $M = 2N + 1$ uniformly distributed points and $M = N + 1$ Gaussian points scaled to $(0, 1)$, respectively. Note that the latter one is equal to $N + N_{A,B}$ thus not belonging to the scope of Theorem 5.1. The results can be found in Table 3.

[13] In contrast, $\log(|p_n(b) - x(b)|/|p_{2n}(b) - x(b)|)/\log 2$ is applied, e.g., in [2].

TABLE 3. Example 1.1: Error of the collocation solution for $\eta = -2$ and $N = 3$ and M collocation points.

	$M = 2N+1$ uniform points				$M = N+1$ Gaussian points			
	L^2		$\mathbb{R}$		L^2		$\mathbb{R}$	
n	error	order	error	order	error	order	error	order
10	6.31e-4		6.51e-4		6.46e-4		6.51e-4	
20	1.44e-4	2.1	1.47e-4	2.1	1.45e-4	2.2	1.46e-4	2.2
40	3.47e-5	2.1	3.52e-5	2.1	3.47e-5	2.1	3.49e-5	2.1
80	8.53e-6	2.0	8.59e-6	2.0	8.53e-6	2.0	8.56e-6	2.0
160	2.12e-6	2.0	2.12e-6	2.0	2.12e-6	2.0	2.12e-6	2.0
320	5.27e-7	2.0	5.28e-7	2.0	5.27e-7	2.0	5.28e-7	2.0

TABLE 4. Example 1.1: Error of the collocation solution for $\eta = -2$ and $N = 1$.

	$M = 3$ uniform points				$M = 2$ Gaussian points			
	L^2		$\mathbb{R}$		L^2		$\mathbb{R}$	
n	error	order	error	order	error	order	error	order
10	5.65e-1		4.94e-1		5.65e-1		5.65e-1	
20	3.93e-1	0.5	3.14e-1	0.6	3.93e-1	0.5	3.95e-1	0.5
40	2.49e-1	0.6	2.14e-1	0.6	2.49e-1	0.7	2.50e-1	0.7
80	1.85e-1	0.4	1.62e-1	0.4	1.85e-1	0.4	1.85e-1	0.4
160	1.42e-1	0.4	1.26e-1	0.4	1.42e-1	0.4	1.42e-1	0.4
320	1.12e-1	0.3	1.00e-1	0.3	1.12e-1	0.3	1.12e-1	0.3

The performance of the collocation method is similar in both cases of the choice of M. So we will restrict ourselves to using only $M = N+1$ scaled Gaussian points, that is, the minimal number of collocation points in our method. This choice is also motivated by the fact that, in later examples, the coefficients A, B are no longer polynomials. Observe that the numerically estimated order of convergence is even higher than expected in view of the theory.

In order to test the boundedness of the error suggested by the results for constant coefficient DAEs in the case $N < \mu - 1$, we show also the results for $N = 1$ in Table 4. Theorem 3.1 provides a bound on the error of the order h^{-1} and Example 4.8 a sharper bound of order h^0. We do not only observe boundedness but a convergence of order 0.3 – 0.4. This is even sharper than the behavior suggested by Example 4.8.

6.2. **A modification of the introductory Example 1.1.** A slightly more involved example is obtained by applying the transformation

$$x(t) = K(t)\tilde{x}(t), \quad K(t) = \begin{bmatrix} 1 & k_{12}(t) & k_{13}(t) \\ 0 & 1 & k_{23}(t) \\ 0 & 0 & 1 \end{bmatrix},$$

as well as a corresponding refactorization of the leading term in Example 1.1. This does not change the index of the DAE; see [5]. In particular, the number of dynamical degrees of freedom remains $l = 0$. Since the index of the DAE is three, Theorem 4.1 provides the estimate $\gamma_\pi \geq c_\gamma h^{-2}$. The DAE for $\tilde{x}$ reads

$$\tilde{A}(\tilde{D}\tilde{x})' + \tilde{B}\tilde{x} = y, \tag{44}$$

TABLE 5. Errors and estimation of the convergence order for (44) with $\eta = -0.2$ using $M = N+1$. The criterion (43) has been minimized.

	$N=2$		$N=3$		$N=4$		$N=5$	
n	error	order	error	order	error	order	error	order
10	1.06e-1		6.44e-2		3.14e-3		1.02e-4	
20	6.16e-2	0.8	3.23e-2	1.0	5.49e-4	2.5	1.28e-5	3.0
40	3.99e-2	0.6	1.59e-2	1.0	9.63e-5	2.5	1.60e-6	3.0
80	2.70e-2	0.6	7.89e-3	1.0	1.70e-5	2.5	2.02e-7	3.0
160	1.87e-2	0.5	3.92e-3	1.0	2.99e-6	2.5	5.30e-8	2.0
320	1.31e-2	0.5	1.95e-3	1.0	5.73e-7	2.4	2.81e-7	−2.4

TABLE 6. Errors and estimation of the convergence order for $N = 3$ and $M = 4$ with $\eta = -0.2$.

	uniform points				Gaussian points			
	L^2		$\mathbb{R}$		L^2		$\mathbb{R}$	
n	error	order	error	order	error	order	error	order
10	6.43e-2		4.26e-2		6.44e-2		6.25e-2	
20	3.23e-2	1.0	2.07e-2	1.0	3.23e-2	1.0	3.15e-2	1.0
40	1.59e-2	1.0	9.79e-3	1.1	1.59e-2	1.0	1.56e-2	1.0
80	7.89e-3	1.0	4.66e-3	1.1	7.89e-3	1.0	7.72e-3	1.0
160	3.92e-3	1.0	2.25e-3	1.0	3.92e-3	1.0	3.83e-3	1.0
320	1.95e-3	1.0	1.11e-3	1.0	1.95e-3	1.0	1.91e-3	1.0

where

$$\tilde{A}(t) = \begin{bmatrix} 1 & k_{23} \\ t\eta & k_{23}t\eta + 1 \\ 0 & 0 \end{bmatrix}, \tilde{D} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tilde{B}(t) = \begin{bmatrix} 1 & k_{12} & k_{13} + k_{23}' \\ 0 & \eta + 1 & (\eta+1)k_{23} + t\eta k_{23}' \\ 0 & t\eta & t\eta k_{23} + 1 \end{bmatrix}.$$

In the experiments below $\eta = -0.2$ has been chosen. The transformation is given by

$$k_{12}(t) = \sin t, \quad k_{13}(t) = -\sin t, \quad k_{23}(t) = \cos t,$$

such that the DAE coefficients are no longer polynomial.

Table 5 shows the errors as well as an estimation of the order of convergence. It can be observed that the orders are as predicted by Theorem 3.1 for $N = 3$ and $N = 5$ while the order is by 0.5 higher for the even orders N. So far we do not have any explanation for this behavior. We can also see that there is a certain saturation of the accuracy for $N = 5$ and $n = 320$: We cannot reach an accuracy which is better than approximately 10^{-7}. This is a result of the interplay of rounding errors and other approximations during the solution process in connection with the ill-posedness of a higher-index DAE.

In order to get an impression about the behavior of the method with respect to collocation point placement we present results for $N = 3$ and $M = 4$ collocation points which are either chosen uniformly distributed or as Gaussian points scaled to $(0, 1)$. The results are presented in Table 6. We observe that both versions behave similarly and support the considerations in Section 5.

TABLE 7. Errors and estimation of the convergence order for $N = 1$ and $M = 2$ with $\eta = -0.2$.

	uniform points				Gaussian points			
	L^2		$\mathbb{R}$		L^2		$\mathbb{R}$	
n	error	order	error	order	error	order	error	order
10	2.97e-1		2.75e-1		2.97e-1		2.97e-1	
20	1.75e-1	0.8	1.56e-1	0.8	1.75e-1	0.8	1.75e-1	0.8
40	1.03e-1	0.8	9.38e-2	0.7	1.03e-1	0.8	1.03e-1	0.8
80	7.06e-2	0.5	6.71e-2	0.5	7.06e-2	0.5	7.06e-2	0.5
160	5.87e-2	0.3	5.74e-2	0.2	5.87e-2	0.3	5.87e-2	0.3
320	5.33e-2	0.1	5.24e-2	0.1	5.33e-2	0.1	5.33e-2	0.1

Finally, we consider the case of $N = 1$. The experiment is done using the settings as in Table 6 but with a different N. The results are listed in Table 7. Note that Theorem 4.7 on DAEs with constant coefficients does not apply here, and Theorem 3.1 guarantees a bound of the order h^{-1} only. Nevertheless, again the approximate solution does not only remain bounded but we observe even convergence although rather slow. As already mentioned earlier we do not have any theoretical backup for this behavior.

We observe that the error behavior is again independent of the chosen minimization criterion (43) or (35). Therefore, we will use the former in all the following experiments thus being closer to the theoretical situation in Theorem 3.1.

6.3. **An index-3 example with four dynamical degrees of freedom.** In the following experiment we consider the IVP for the DAE

$$A(Dx)'(t) + B(t)x(t) = y(t), \quad t \in [0,5] \tag{45}$$

with

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, D = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix},$$

the smooth matrix coefficient

$$B(t) = \begin{bmatrix} 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & \sin t & 0 & 1 & -\cos t & -2\rho\cos^2 t \\ 0 & 0 & -\cos t & -1 & 0 & -\sin t & -2\rho\sin t\cos t \\ 0 & 0 & 1 & 0 & 0 & 0 & 2\rho\sin t \\ 2\rho\cos^2 t & 2\rho\sin t\cos t & -2\rho\sin t & 0 & 0 & 0 & 0 \end{bmatrix}, \quad \rho = 5,$$

Table 8. Errors and estimation of the convergence order for (45) using $M = N + 1$. The criterion (43) has been minimized.

	$N = 1$		$N = 2$		$N = 3$		$N = 4$		$N = 5$	
n	error	order	error	order	error	order	error	order		
10	2.57e+0		3.37e-1		6.27e-2		6.24e-3		5.69e-4	
20	1.52e+0	0.8	1.98e-1	1.4	1.77e-2	1.8	9.36e-4	2.7	6.11e-5	3.2
40	8.80e-1	0.8	9.36e-2	1.0	6.43e-3	1.5	1.66e-4	2.5	7.31e-6	3.1
80	4.72e-1	0.9	4.63e-2	1.0	2.84e-3	1.2	3.41e-5	2.3	9.02e-7	3.0
160	3.01e-1	0.7	2.33e-2	1.0	1.36e-3	1.1	7.69e-6	2.2	1.12e-7	3.0
320	2.30e-1	0.4	1.18e-2	1.0	6.75e-4	1.0	1.82e-6	2.1	1.42e-8	3.0

and the initial condition

$$x_2(0) = 1,\ x_3(0) = 2,\ x_5(0) = 0,\ x_6(0) = 0.$$

This index-3 DAE is the linearized version of the test example from [2].[14] Its number of dynamical degrees of freedom equals four.

In the following numerical experiments we choose the exact solution

$$\begin{aligned} x_1 &= \sin t, & x_4 &= \cos t, \\ x_2 &= \cos t, & x_5 &= -\sin t, \\ x_3 &= 2\cos^2 t, & x_6 &= -2\sin 2t, \\ x_7 &= -\rho^{-1}\sin t. \end{aligned}$$

The experiments are done under the same parameter choices as in the previous one: If not indicated differently, for a given N, the collocation points are chosen to be $M = N + 1$ Gaussian points scaled to $(0, 1)$. The results are shown in Table 8. For $N > 2$ the results confirm the prediction of Theorem 3.1. The convergence order for $N = 2$ is one, which is much higher than the expected order zero. Moreover, the error behavior for $N = 1$ is again much more favorable than predicted by Theorem 3.1.

6.4. **The Campbell-Moore example.**

The hitherto existing theory does not apply to nonlinear DAE problems. Certain results concerning the approximation of ill-posed nonlinear problems by projection onto subspaces can be found in [4]; see also the references cited therein. However, the assumption used there is unrealistic in the case of DAEs. Despite this fact, for illustration purposes, we also treat the

[14] We put $B(t) = b(x_*(t), t)$ with the solution x_* given in [2]; cf. also Section 6.4. It has tractability index $\mu = 3$ and dynamical degree $l = 4$. The nonlinear DAE is treated in [2] by a multistep integrator which goes along with the use of a derivative array. The integration is carried out on the interval $[0, 5]$.

nonlinear test problem from [2]:[15]

$$\begin{aligned}
x_1' - x_4 &= 0,\\
x_2' - x_5 &= 0,\\
x_3' - x_6 &= 0,\\
x_4' - x_6\cos t + x_3 \sin t + x_5 - 2x_1(1 - r(x_1^2 + x_2^2)^{-\frac{1}{2}})x_7 &= 0,\\
x_5' - x_6\sin t - x_3 \cos t - x_4 - 2x_2(1 - r(x_1^2 + x_2^2)^{-\frac{1}{2}})x_7 &= 0,\\
x_6' + x_3 - 2x_3x_7 &= 0,\\
x_1^2 + x_2^2 + x_3^2 - 2r(x_1^2 + x_2^2)^{\frac{1}{2}} + r^2 - \rho^2 &= 0.
\end{aligned}$$

This DAE can be formulated as

$$A(Dx)'(t) + b(x(t), t), \tag{46}$$

with

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & 0 & 1\\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, D = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix},$$

and

$$b(x,t) = \begin{bmatrix} -x_4\\ -x_5\\ -x_6\\ -x_6\cos t + x_3\sin t + x_5 - 2x_1(1 - r(x_1^2 + x_2^2)^{-\frac{1}{2}})x_7\\ -x_6\sin t - x_3\cos t - x_4 - 2x_2(1 - r(x_1^2 + x_2^2)^{-\frac{1}{2}})x_7\\ x_3 - 2x_3x_7\\ x_1^2 + x_2^2 + x_3^2 - 2r(x_1^2 + x_2^2)^{\frac{1}{2}} + r^2 - \rho^2 \end{bmatrix}.$$

The function b and its partial derivative with respect to x are smooth everywhere on the domain $\operatorname{dom} b = \{(x,t) \in \mathbb{R}^7 \times \mathbb{R} : x_1^2 + x_2^2 > 0\}$.

In [2], the following solution is considered:

$$\begin{aligned}
x_{*1} &= (\rho\cos(2\pi - t) + r)\cos t = (\rho\cos t + r)\cos t,\\
x_{*2} &= (\rho\cos(2\pi - t) + r)\sin t = (\rho\cos t + r)\sin t,\\
x_{*3} &= \rho\sin(2\pi - t) = -\rho\sin t,
\end{aligned}$$

yielding

$$\begin{aligned}
x_{*4} &= -(\rho\cos(2\pi - t) + r)\sin t + \rho\sin(2\pi - t)\cos t,\\
x_{*5} &= (\rho\cos(2\pi - t) + r)\cos t + \rho\sin(2\pi - t)\sin t,\\
x_{*6} &= -\rho\cos(2\pi - t),\\
x_{*7} &= 0.
\end{aligned}$$

[15] In [2], a slightly different notation is used: x_4, x_5, x_6, x_7 are denoted by u_1, u_2, u_3, and λ, respectively.

Table 9. Errors and estimation of the convergence order for (46) using $M = N + 1$. The corresponding approximation of criterion (47) has been minimized.

	$N = 1$		$N = 2$		$N = 3$		$N = 4$		$N = 5$	
n	error	order	error	order	error	order	error	order		
10	3.32e+1		4.53e+0		3.82e-1		7.02e-2		1.47e-3	
20	3.32e+1	0.0	7.51e-1	2.6	1.02e-1	1.9	1.26e-2	2.5	1.24e-4	3.6
40	3.32e+1	0.0	3.03e-1	1.3	3.14e-2	1.7	2.52e-3	2.3	1.30e-5	3.3
80	3.32e+1	0.0	1.80e-1	0.7	1.22e-2	1.4	5.45e-4	2.2	1.54e-6	3.1
160	3.32e+1	0.0	1.17e-1	0.6	5.67e-3	1.1	1.25e-4	2.1	1.20e-6	0.6
320	3.32e+1	0.0	7.95e-2	0.6	2.73e-3	1.1	1.25e-4	0.0	1.20e-6	0.0

Table 10. $L^\infty(0,5)$-norm of the errors for the components x_i of (46). The expression (47) has been used.

	$n = 50$	$n = 100$	$n = 200$
$\|\|p_{\pi,1} - x_{*1}\|\|_\infty$	3.56e-10	5.70e-12	6.53e-12
$\|\|p_{\pi,2} - x_{*2}\|\|_\infty$	3.75e-10	7.34e-12	7.74e-12
$\|\|p_{\pi,3} - x_{*3}\|\|_\infty$	2.46e-10	6.33e-12	7.77e-12
$\|\|p_{\pi,4} - x_{*4}\|\|_\infty$	2.50e-08	1.43e-09	2.04e-09
$\|\|p_{\pi,5} - x_{*5}\|\|_\infty$	2.16e-08	1.24e-09	1.76e-09
$\|\|p_{\pi,6} - x_{*6}\|\|_\infty$	3.86e-08	2.17e-09	3.04e-09
$\|\|p_{\pi,7} - x_{*7}\|\|_\infty$	1.16e-06	1.36e-07	1.63e-07

In [2], the inequality $r > \rho$ is supposed and the numerical experiments are carried out for $\rho = 5$ and $r = 10$. We use the same parameters in the following experiment.

We solved the nonlinear problem (46) in a way analogous to the linear one (7)–(8). The expressions for the functionals ϕ_π and $\phi_{\pi,M}$ are replaced by their nonlinear counterparts

$$\sum_{j=1}^{n} \frac{h_j}{M} \sum_{t \in S_j} |A(t)(Dp)'(t) + b(p(t), t)|^2 + |G_a p(a) + G_b p(b) - r|^2$$

and

$$\|A(Dp)' + b(p, \cdot)\|_{L^2}^2 + |G_a p(a) + G_b p(b) - r|^2, \tag{47}$$

respectively. The resulting nonlinear least-squares problem was solved using Matlab's `nonlinsq` function. The computational results are shown in Table 9. The observed orders of convergence correspond to what would be expected in view of Theorem 3.1 for $N \geq 3$. The solutions for $N = 1$ and $N = 2$ remain bounded with even a positive order of convergence for $N = 2$ which is similar to the linear case. Moreover, similar to the linear case, there is also a maximal reachable accuracy which is explained by rounding errors and other approximations in connection with the ill-posedness of the problem at hand.

In the reference [2], a fifth order method has been used and errors in the norm of $L^\infty(0, 5)$ have been provided. Table 10 presents results of the least-squares method for $N = 5$ which can be compared to [2, Table 1]. As expected, the algebraic variable x_7 has the least accuracy being much lower than that for the differential components.

Let us emphasize once again that we do not have any theoretical backup for the least-squares collocation applied to nonlinear problems, however, it seems to be quite desirable and meaningful to develop this in the future.

7. Conclusions

We have consolidated the recently developed new least-squares collocation method for the numerical solution of initial and boundary value problems in linear higher-index DAEs. The motivation for this method originates from the fact that higher-index DAEs are essentially ill-posed problems in natural topologies. We provided 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 are necessary. In the numerical experiments, the method behaves in a robust 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.

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] S. L. Campbell and E. Moore, *Constraint preserving integrators for general nonlinear higher index DAEs*, Numer. Math. **69** (1995), no. 4, 383–399, DOI 10.1007/s002110050099. MR1314594

[3] M. Hanke, R. März, C. Tischendorf, E. Weinmüller, and S. Wurm, *Least-squares collocation for linear higher-index differential-algebraic equations*, J. Comput. Appl. Math. **317** (2017), 403–431, DOI 10.1016/j.cam.2016.12.017. MR3606087

[4] B. Kaltenbacher and J. Offtermatt, *A convergence analysis of regularization by discretization in preimage space*, Math. Comp. **81** (2012), no. 280, 2049–2069, DOI 10.1090/S0025-5718-2012-02596-8. MR2945147

[5] R. Lamour, R. März, and C. Tischendorf, *Differential-Algebraic Equations: A Projector Based Analysis*, Differential-Algebraic Equations Forum, Springer, Heidelberg, 2013. MR3024597

[6] R. Lamour, R. März, and E. Weinmüller, *Boundary-value problems for differential-algebraic equations: a survey*, Surveys in differential-algebraic equations. III, Differ.-Algebr. Equ. Forum, Springer, Cham, 2015, pp. 177–309. MR3411039

[7] R. März, *Numerical methods for differential algebraic equations*, Acta numerica, 1992, Acta Numer., Cambridge Univ. Press, Cambridge, 1992, pp. 141–198. MR1165725

[8] R. März, *Differential-algebraic equations from a functional-analytic viewpoint: a survey*, Surveys in differential-algebraic equations. II, Differ.-Algebr. Equ. Forum, Springer, Cham, 2015, pp. 163–285. MR3331417

Department of Mathematics, School of Engineering Sciences, KTH Royal Institute of Technology, S-100 44 Stockholm, Sweden
Email address: hanke@nada.kth.se

Institute of Mathematics, Humboldt University of Berlin, D-10099 Berlin, Germany
Email address: maerz@math.hu-berlin.de

Institute of Mathematics, Humboldt University of Berlin, D-10099 Berlin, Germany
Email address: caren@math.hu-berlin.de