Least-Squares Collocation for Higher-Index Linear Differential-Algebraic Equations: Estimating the Instability Threshold

Michael Hanke[1], Roswitha März[2] and Caren Tischendorf[3]

Abstract

Differential-algebraic equations with higher index give rise to essentially ill-posed problems. The least-squares collocation by discretizing the pre-image space is not much more computationally expensive than standard collocation methods used in the numerical solution of ordinary differential equations and index-1 differential-algebraic equations. This approach has displayed excellent 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 DAEs.

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

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, $[a,b] \subset \mathbb{R}$ denotes a finite interval, $q : [a,b] \to \mathbb{R}^m$ 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 and the leading term of the special form,

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

[1] KTH Royal Institute of Technology, School of Engineering Sciences, Department of Mathematics, S-100 44 Stockholm, Sweden, hanke@nada.kth.se

[2] Humboldt University of Berlin, Institute of Mathematics, D-10099 Berlin, Germany, maerz@math.hu-berlin.de

[3] Humboldt University of Berlin, Institute of Mathematics, D-10099 Berlin, Germany, caren@math.hu-berlin.de

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. We refer to [2, Subsection 5.1] for more general DAEs.

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., [5, Section 2],[4, 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 [2], we put the problem in a Hilbert space setting and consider generalized solutions $x \in H_D^1$,

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

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

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

Then $G_ax(a) + G_bx(b) = G_aD^+Dx(a) + G_bD^+Dx(b)$ 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.

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, 5] for the respective comprehensive surveys.

Here, we follow the ansatz from [2]. 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 or equal to K.

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

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

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 subintervall. In contrast, as first proposed in [2], we specify $M > N$ least-squares collocation points by choosing values

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

and setting

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

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

$$A(t)(Dp)'(t) + B(t)p(t) = y(t), \quad t \in S_j, \quad j = 1, \ldots, n \tag{6}$$

$$G_a p(a) + G_b p(b) = r. \tag{7}$$

in the least-squares sense. In this context, IVPs and BVPs are treated in the same way, more precisely, IVPs are treated as BVPs. Here the substance inheres in the DAE and it is a secondary matter whether we have initial conditions or boundary conditions.

To demonstrate the great potential of the overdetermined least-squares collocation we resume one of the experiments from [2]. The related DAE is known to cause serious difficulties and failures in the numerical integration depending on the movement of characteristic subspaces, see [6, p. 168], also [4, 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 exact solution and this determines q. In order to have a unique solution also for the classical collocation system, the conditions

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

Table 1: Collocation results for Example 1.1. The table shows the error $\|x_1 - p_1\|_\infty$.

n	Standard	Least-squares
20	3.74e+006	3.26e-4
40	9.84e+016	7.52e-5
80	3.51e+038	1.81e-5
160	2.04e+082	4.42e-6
320	2.98e+170	1.11e-6
640	3.06e+307	1.06e-6

were posed.
Table 1 displays the errors in the algebraic component 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 phenomenal!
The computations have been carried out in MATLAB.[1] □

In the present paper, we provide estimates of the instability threshold for arbitrary-index linear DAEs with variable coefficients, which considerably generalizes the results from [2] obtained for constant-coefficient differential-algebraic equations. This way, we obtain general convergence results for the least-squares method applied to systems (1) – (2). In the case of constant coefficient systems, a conjecture made in [2] 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. The least-squares method is introduced in an abstract setting in Section 3 leading to convergence results of the proposed method. The necessary estimates of the instability threshold are proven in Section 4. We consider these estimates as the main result of the present paper. The least-squares method as formulated in Hilbert spaces requires the evaluation of certain intergrals. We discuss a numerical intergration technique and its convergence properties in Section 5. Finally, we present some numerical examples in Section 6. Conclusions will be drawn in Section 7.

2 Differential-Algebraic Operators acting on H^1_D

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

$$Tx = y, \qquad \text{and} \qquad \mathcal{T}x = (y, r). \tag{8}$$

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

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

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

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

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

is a Hilbert space [7, Lemma 6.9] and the DA operator T is bounded.
Next, we resume the notion of the tractability index of the DA operator T from [2], see also [7, 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 (8) is so.
The tractability index is specified by means of certain sequences of continuous matrix functions $G_{i+1} := G_i + B_iQ_i$, 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 [7, Definition 4.1].

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

$$r_0 \leq \cdots \leq r_{\mu-1} < r_\mu = m, \quad l := m - \sum_{i=0}^{\mu-1}(m-r_i), \tag{9}$$

if there is an admissible matrix function sequence $G_0,\ldots,G_\mu$ with (9).
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 $\mathit{\Pi}_{\mathrm{can}}$ denote the canonical projector function of the associated fine DAE, see [4, Definition 2.37]. The projector $\mathit{\Pi}_{\mathrm{can}}(t)$ acts in $\mathbb{R}^m$ and its rank is l given in (9) 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, $\mathit{\Pi}_{can}$ represents the spectral projector of the pencil onto the eigenspace corresponding to the finite eigenvalues along the ones corresponding to the infinite eigenvalues [4, Theorem 1.33].
We quote [2, 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* (9). *Then the following statements hold:*

1. $\ker T$ *has finite dimension,* $\dim\ker T = l = \operatorname{rank} \mathit{\Pi}_{\mathrm{can}}$.
2. *T is surjective, thus Fredholm, exactly if $\mu = 1$.*
3. *If $\mu > 1$, then* $\operatorname{im} T$ *is a nonclosed, proper subset of L^2.*
4. *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,

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

is defined on $H_D^1 = \{x_2, x_3 \in H^1(0,1),\ x_1 \in L^2(0,1)\}$ with $m = 3,\ k = 2$. Its image becomes

$$\operatorname{im} T = \{y_1, y_2, y_3 \in L^2(0,1) : y_3 \in H^1(0,1),\ y_2 - y_3' \in H^1(0,1)\} \subset L^2.$$

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

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

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

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

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

is again a Hilbert space. Here, $<,>$ denotes the Euclidean scalar product of $\mathbb{R}^l$. We quote [2, Theorem 2.4] to provide properties of the operator $\mathcal{T}$:

Theorem 2.4. *Let the bounded DA operator $T : H_D^1 \to L^2$ be fine with index $\mu \in \mathbb{N}$ and characteristic values* (9) *and let the boundary conditions be restricted by* (4)*. Then the following statements hold:*

1. *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*

$$\ker(G_a X(a,a) + G_b X(b,a)) = \ker \varPi_{\text{can}}(a) \qquad (10)$$

holds. Here, $X(t,a)$ denotes the maximal fundamental solution matrix of the associated DAE, normalized at point a.[2]

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

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

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

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

[2] X is the unique solution of the IVP $A(t)(DX)'(t) + B(t)X(t) = 0,\ t \in (a,b),\ X(a) = \varPi_{\text{can}}(a)$.

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}. In favor of an easier reading we drop the extra integer s but we thoroughly assure that the indicated constants do not depend on the partitions and stepsizes in fact.
Following ideas developed in [3] (see also [2, Section 2.2]), the approximate solution

$$p_\pi = \operatorname{argmin}\{\|A(Dp)' + Bp - q\|_{L^2}^2 + |G_a p(a) + G_b p(b) - r|^2 : p \in X_\pi\} \quad (11)$$

satisfies the inequality

$$\|p_\pi - x_*\|_{H_D^1} \leq \frac{\beta_\pi}{\gamma_\pi} + \alpha_\pi, \quad (12)$$

with x_* denoting the unique solution of the BVP and

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

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

$$\alpha_\pi \leq \|p_{\text{int}} - x_*\|_{H_D^1} \leq c_\alpha h^N, \quad \beta_\pi \leq c_\beta h^N, \quad (13)$$

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 [2], for equidistant partitions π with sufficiently small stepsizes, the estimate

$$\gamma_\pi \geq c_\gamma h^{\min(N,\mu-1)} \quad (14)$$

with a constant $c_\gamma > 0$, is conjectured owing to numerous numerical experiments and a strong proof for the cases $N \geq \mu - 1$ and $N = 1$ for constant-coefficient DAEs.
Theorem 4.1 below in Section 4 verifies the inequality

$$\gamma_\pi \geq c_\gamma h^{\mu-1} \tag{15}$$

for general regular linear DAEs, and this can be seen as main result of the present paper. In addition, the stronger estimate (14) 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 (12) and (13) as well as the Theorems 4.1 and 4.7 we obtain

Theorem 3.1. *Let the BVP* (1) – (2)*, with index* $\mu \geq 1$*, satisfy the assumptions of Theorem* 2.4(1) *with the unique solution* x_* *as well as the coefficients* A, B *of the BVP being 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* $\frac{h}{h_{\min}} \leq \rho$*:*

1. *The least-squares collocation solutions* p_π *defined by* (11) *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* $h \to 0$*.*

2. *Moreover, if the coefficients* A *und* B *are constant, the solutions* p_π *fulfill even*

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

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

For providing the approximation p_π in practice, one needs to replace the integral by a discretized version. In Section 5 we will deal with one possible variant which traces the matter back to overdetermined least-squares collocation.

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 as 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. In Subsection 4.4 we address the case $1 \leq N < \mu-1$. In particular we verify the stronger inequality (14) for arbitrary regular DAEs with constant coefficients, which has been conjectured in [2] and proved for $N = 1$.

Theorem 4.1. *Let the bounded DA operator $T : H^1_D \to L^2$ be fine with index $\mu \in \mathbb{N}$ and characteristic values (9) and let the boundary conditions be restricted by (4). Let the condition (10) 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 $\frac{h}{h_{\min}} \leq \rho$:

1. *If $\mu = 1$ then there is a constant $c_\gamma > 0$ such that $\gamma_\pi \geq c_\gamma$.*
2. *If $\mu = 2$ then there is a constant $c_\gamma > 0$ such that $\gamma_\pi \geq c_\gamma h_{\min} \geq c_\gamma \frac{1}{\rho} h$.*
3. *If $\mu \geq 2$ and the coefficients A and B are sufficiently smooth*[3] *then there is a constant $c_\gamma > 0$ such that $\gamma_\pi \geq c_\gamma h_{\min}^{\mu-1} \geq c_\gamma \frac{1}{\rho^{\mu-1}} h^{\mu-1}$.*

Remark 4.2. More details concerning the constant c_γ will be shown in the proof later on. In the index-1 case one has simply $c_\gamma = c_Y^{-1}$ with c_Y from Proposition 4.5(2). In the higher-index case the constant c_γ provided in Theorem 4.1(3) is inversely proportional to the value c_Y from Proposition 4.5(2) and also to $\sqrt{g_{\mu-1}}$, with $g_{\mu-1} = d_{1,\mu-1} c^*_{\mu-1} \|D\mathcal{L}_{\mu-1}\|^2_\infty > 0$, see Lemma 4.6 for $d_{1,\mu-1}$ and Lemma 4.4 for $c^*_{\mu-1}$. Note that $c^*_{\mu-1}$ increases with the polynomial degree N.

Remark 4.3. For index-2 DAEs Theorem 4.1 offers one constant in item (2) and another one in item (3), namely

$$c_\gamma|_{\text{item}(2)} = \frac{1}{3}\frac{1}{c_Y}\frac{1}{\sqrt{c_1^*}}\frac{1}{\|D\mathit{\Pi}_0 Q_1 D^+\|_\infty \|D\mathcal{L}_{\mu-1}\|_\infty}\frac{1}{\sqrt{1+K\|D\mathcal{L}_{\mu-1}\|_\infty^{-2}}},$$

$$c_\gamma|_{\text{item}(3)} = \frac{1}{24\sqrt{2}}\frac{1}{c_Y}\frac{1}{\sqrt{c_1^*}}\frac{1}{\|DQ_1\|_\infty \|D\mathcal{L}_{\mu-1}\|_\infty},$$

by completely different proofs.
Owing to the special form of D, $|D|= 1, |D^+|= 1$, and $DQ_1 = D\mathit{\Pi}_0 Q_1 D^+ D$, it holds that $\|DQ_1\|_\infty = \|D\mathit{\Pi}_0 Q_1 D^+\|_\infty$.
Note that $K = 0$ if $\mathit{\Pi}_{\mu-1} = 0$.

4.1 An auxiliary estimation concerning piecewise polynomials

The following lemma is a straightforward consequence of [2, Lemma 3.3].

Lemma 4.4. *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*

$$\|q^{(i)}\|^2_{L^2} \leq c_i^* \frac{1}{h_{\min}^{2i}} \|q\|^2_{L^2}, \quad \|q^{(i)}\|^2_{H^1_D} \leq C_i^* \frac{1}{h_{\min}^{2i}} \|q\|^2_{H^1_D}, \quad , i = 1, \cdots, K,$$
$$\|q^{(K+1)}\|^2_{L^2} = 0, \quad \|q^{(K+1)}\|^2_{H^1_D} = 0$$

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 [2, Lemma 3.3].

[3] See Subsection 4.3 below for details.

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

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

Then owing to $q_i'|_{[t_{j-1},t_j)} \in \mathcal{P}_{K-1}$ we obtain $\|q'\|_{L^2}^2 \geq \frac{h_{\min}^2}{4\lambda_{K-1}} \|q''\|_{L^2}^2$ and further $\|q\|_{L^2}^2 \geq \frac{h_{\min}^2}{4\lambda_K} \frac{h_{\min}^2}{4\lambda_{K-1}} \|q''\|_{L^2}^2$, 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 [4, 7] for details. Let the DA operator $T : H_D^1 \to L^2$ corresponding to the DAE (1) be fine with tractability index $\mu \geq 2$ and the characteristic values (9). Then there are an admissible sequence of matrix valued function starting from $G_0 := AD$ and ending up with a nonsingular G_μ, see [4, Definition 2.6], as well as associated projector valued functions

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

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

$$\begin{aligned} Q_i &= I - P_i,\ i = 0, \ldots, \mu-1, \\ \varPi_0 := P_0,\ \varPi_i &:= \varPi_{i-1} P_i \in \mathcal{C}([a,b], L(\mathbb{R}^m)),\ i = 1, \ldots, \mu-1, \\ D\varPi_i D^+ &\in \mathcal{C}^1([a,b], L(\mathbb{R}^k)),\ i = 1, \ldots, \mu-1. \end{aligned}$$

By means of the projector functions we decompose the unknown x and decouple the DAE itself into their characteristic parts, see [4, Section 2.4].
The component $u = D\varPi_{\mu-1} x = D\varPi_{\mu-1} D^+ D x$ satisfies the explicit regular ODE residing in $\mathbb{R}^k$,

$$u' - (D\varPi_{\mu-1} D^+)' u + D\varPi_{\mu-1} G_\mu^{-1} B \varPi_{\mu-1} D^+ u = D\varPi_{\mu-1} G_\mu^{-1} y, \tag{16}$$

and the components $v_i = \varPi_{i-1} Q_i x = \varPi_{i-1} Q_i D^+ D x,\ i = 1, \ldots, \mu-1$ satisfy the triangular subsystem involving several differentiations,

$$\begin{bmatrix} 0 & \mathcal{N}_{12} & \cdots & \mathcal{N}_{1,\mu-1} \\ & 0 & \ddots & \vdots \\ & & \ddots & \mathcal{N}_{\mu-2,\mu-1} \\ & & & 0 \end{bmatrix} \begin{bmatrix} (Dv_1)' \\ \\ \vdots \\ (Dv_{\mu-1})' \end{bmatrix} \tag{17}$$
$$+ \begin{bmatrix} I & \mathcal{M}_{12} & \cdots & \mathcal{M}_{1,\mu-1} \\ & I & \ddots & \vdots \\ & & \ddots & \mathcal{M}_{\mu-2,\mu-1} \\ & & & I \end{bmatrix} \begin{bmatrix} v_1 \\ \\ \vdots \\ v_{\mu-1} \end{bmatrix} = \begin{bmatrix} \mathcal{L}_1 \\ \\ \vdots \\ \mathcal{L}_{\mu-1} \end{bmatrix} y.$$

Finally, one has for $v_0 = Q_0 x$ the representation

$$v_0 = \mathcal{L}_0 y - \mathcal{H}_0 D^+ u - \sum_{j=1}^{\mu-1} \mathcal{M}_{0\,j} v_j - \sum_{j=1}^{\mu-1} \mathcal{N}_{0\,j} (Dv_j)'. \tag{18}$$

The subspace $\operatorname{im} D\Pi_{\mu-1}$ is an invariant subspace for the ODE (16). The components $v_0, v_1, \dots, v_{\mu-1}$ remain within their subspaces $\operatorname{im} Q_0$, $\operatorname{im} \Pi_{\mu-2} Q_1, \dots,$ $\operatorname{im} \Pi_0 Q_{\mu-1}$, respectively. The structural decoupling is associated with the decomposition

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

All coefficients in (16) – (18) are continuous and explicitly given in terms of an admissible matrix function sequence as

$$\begin{aligned}
\mathcal{N}_{01} &:= -Q_0 Q_1 D^+ \\
\mathcal{N}_{0j} &:= -Q_0 P_1 \cdots P_{j-1} Q_j D^+, && j = 2, \dots, \mu-1, \\
\mathcal{N}_{i,i+1} &:= -\Pi_{i-1} Q_i Q_{i+1} D^+, \\
\mathcal{N}_{ij} &:= -\Pi_{i-1} Q_i P_{i+1} \cdots P_{j-1} Q_j D^+, && j = i+2, \dots, \mu-1,\ i = 1, \dots, \mu-2, \\
\mathcal{M}_{0j} &:= Q_0 P_1 \cdots P_{\mu-1} \mathcal{M}_j D\Pi_{j-1} Q_j, && j = 1, \dots, \mu-1, \\
\mathcal{M}_{ij} &:= \Pi_{i-1} Q_i P_{i+1} \cdots P_{\mu-1} \mathcal{M}_j D\Pi_{j-1} Q_j, && j = i+1, \dots, \mu-1,\ i = 1, \dots, \mu-2, \\
\mathcal{L}_0 &:= Q_0 P_1 \cdots P_{\mu-1} G_\mu^{-1}, \\
\mathcal{L}_i &:= \Pi_{i-1} Q_i P_{i+1} \cdots P_{\mu-1} G_\mu^{-1}, && i = 1, \dots, \mu-2, \\
\mathcal{L}_{\mu-1} &:= \Pi_{\mu-2} Q_{\mu-1} G_\mu^{-1}, \\
\mathcal{H}_0 &:= Q_0 P_1 \cdots P_{\mu-1} \mathcal{K} \Pi_{\mu-1},
\end{aligned}$$

in which

$$\mathcal{K} := (I - \Pi_{\mu-1}) G_\mu^{-1} B_{\mu-1} \Pi_{\mu-1} + \sum_{l=1}^{\mu-1} (I - \Pi_{l-1})(P_l - Q_l)(D\Pi_l D^+)' D\Pi_{\mu-1},$$

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

It should be added at this point, that the coefficients of the ODE (16) are uniquely determined in the scope of the fine decoupling. This justifies to speak about *the inherent explicit regular ODE* (IERODE) of the given DAE.
We introduce the factitious function space (cf, also [7])

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

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 special and strongly depend on the decoupling coefficients which in turn depend on the given data A, D, B.

Proposition 4.5. *Let the DA operator $T : H_D^1 \to L^2$ be fine with characteristic values* (9) *and index $\mu \geq 2$. Then the following results:*

1. $\operatorname{im} T = Y$.
2. *The space Y equipped with the norm $\|\cdot\|_Y$ is complete.*
3. *Let the operator $\mathcal{T}$ corresponding to the BVP satisfy the conditions of Theorem 2.4(i). Then there is a constant c_Y such that the inequality*

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

becomes valid.

Proof. (1) and (2) can be checked by a straightforward use of the above decoupling formulas analogously to the case of the Banach space setting in [7].
(3) 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,

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

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

$$\begin{aligned} Y_\pi &:= \{ y \in L^2 : D\mathcal{L}_{\mu-i} y \in \mathcal{C}_\pi^{\mu-i}, \quad i = 1, \ldots, \mu-1 \}, \\ Y_\pi^0 &:= \{ y \in \mathcal{C}_\pi^0 : D\mathcal{L}_{\mu-i} y \in \mathcal{C}_\pi^{\mu-i}, \quad i = 1, \ldots, \mu-1 \}. \end{aligned}$$

Lemma 4.6. *Let the DA operator T be fine with index $\mu > 1$ and let its coefficients A and B be sufficiently smooth such that*

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

1. *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

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

2. *If, additionally,*

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

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

Proof. (1): 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. (17)) $v_{*\mu-1} = \mathcal{L}_{\mu-1}y_*$ 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 (18). 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 $\mu \ge 3$.

Let $y \in Y_\pi$ be given. Regarding the definition of the function space Y which is closely related to the decoupled system (17) 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

$$\begin{aligned}(Dv_1)' = (D\mathcal{L}_1 y)' &+ \sum_{i=1}^{\mu-2} D\Pi_0 Q_1\cdots Q_{i+1}D^+(D\mathcal{L}_{i+1}y)^{(i+1)} \\ &+ \sum_{i=1}^{\mu-2}\sum_{s=0}^{i}\mathcal{E}_{\mu-1,i,s}(D\mathcal{L}_{i+1}y)^{(s)}.\end{aligned}$$

We estimate

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

and thus

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

We rearrange the last formula to the form

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

with the coefficients

$$\begin{aligned}\tilde{d}_{1,\mu-1} =&\|D\varPi_0Q_1\cdots Q_{\mu-1}D^+\|_\infty,\\ \tilde{d}_{1,\mu-2} =&\|D\varPi_1Q_2\cdots Q_{\mu-1}D^+\|_\infty+e_{\mu-1,\mu-2,\mu-2},\\ \tilde{d}_{1,\mu-3} =&\|D\varPi_2Q_3\cdots Q_{\mu-1}D^+\|_\infty+e_{\mu-2,\mu-3,\mu-3}+e_{\mu-1,\mu-2,\mu-3},\\ &\cdots\\ \tilde{d}_{1,2} =&\|D\varPi_{\mu-3}Q_{\mu-2}Q_{\mu-1}D^+\|_\infty+\sum_{j=2}^{\mu-1}e_{j,j-1,2},\\ \tilde{d}_{1,1} =& 1+\sum_{j=2}^{\mu-1}e_{j,j-1,1},\quad \tilde{d}_{1,0}=\sum_{j=2}^{\mu-1}e_{j,j-1,0},\\ \tilde{d}_{2,\mu-2} =&\|D\varPi_0Q_1\cdots Q_{\mu-2}D^+\|_\infty,\\ &\cdots\\ \tilde{d}_{2,1} =& 1+\sum_{j=3}^{\mu-1}e_{j,j-2,1},\quad \tilde{d}_{2,0}=\sum_{j=3}^{\mu-1}e_{j,j-2,0},\\ &\cdots\\ \tilde{d}_{\mu-1,1} =& 1,\quad \tilde{d}_{\mu-1,0}=0.\end{aligned}$$

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

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

where $d_{1,\mu-1} := 2\tilde{d}^2_{1,\mu-1}$ and $d_{i,s} := 2(S-1)\tilde{d}^2_{i,s}$, 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.
(2): For each arbitrary $p \in X_\pi$ and the corresponding $y = Tp = A(Dp)' + Bp$ it holds that $y \in C^0_\pi$ and $\mathcal{L}_{\mu-j}y = \mathcal{L}_{\mu-j}A(Dp)' + \mathcal{L}_{\mu-j}Bp \in \mathcal{C}^{\mu-j}_\pi$, thus $Tp \in Y^0_\pi$. □

4.3 Proof of Theorem 4.1

Part (1)

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

Part (2)

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

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

and Lemma 4.4 implies

$$\begin{aligned} \|(D\Pi_0 Q_1 G_2^{-1} q)'\|^2_{L^2} &\leq 2\Big(\|(D\Pi_0 Q_1 D^+)'\|^2_\infty + \|D\Pi_0 Q_1 D^+\|^2_\infty \, \frac{c_1^*}{h^2_{\min}}\Big)\|Dp\|^2_{L^2} \\ &\leq \frac{1}{h^2_{\min}}\big(2c_1^*\|D\Pi_0 Q_1 D^+\|^2_\infty + O(h^2)\big)\|Dp\|^2_{L^2} \\ &\leq \frac{1}{h^2_{\min}} 3c_1^* \|D\Pi_0 Q_1 D^+\|^2_\infty \|Dp\|^2_{L^2}, \end{aligned} \tag{19}$$

for sufficiently small $h > 0$ where $c_1^*\|D\Pi_0 Q_1 D^+\|^2_\infty > 0$. 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

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

with $d := \|D\Pi_0 Q_1 G_2^{-1}\|^2_\infty = \|\mathcal{L}_1\|^2_\infty > 0$. 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.[5]). Inserting into (19) we arrive at

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

with $\tilde{g}_1 := c_1^* \|D\varPi_0 Q_1 D^+\|_\infty^2 (K+d) > 0$. 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 4.5(3) 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 (3)

Let the coefficients A and B be smooth enough to ensure that (cf. (17))

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

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

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

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

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

$$\begin{aligned}
&\mathfrak{A}^{(s)}_{\pi,\mu-i}(t_{j-1/2}) = \mathfrak{A}^{(s)}_{\mu-i}(t_{j-1/2}) = (U_{\mu-i}A)^{(s)}(t_{j-1/2}),\\
&\mathfrak{B}^{(s)}_{\pi,\mu-i}(t_{j-1/2}) = \mathfrak{B}^{(s)}_{\mu-i}(t_{j-1/2}) = (U_{\mu-i}B)^{(s)}(t_{j-1/2}),\\
&\qquad\qquad s = 0,\dots,\mu-i,\quad i = 1,\dots,\mu-1,\quad j = 1,\dots,n, \qquad (20)
\end{aligned}$$

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

$$\begin{aligned}
&\frac{1}{h^{\mu-i}}\ \|\mathfrak{A}_{\mu-i} - \mathfrak{A}_{\pi,\,\mu-i}\|_\infty := \frac{1}{h^{\mu-i}} \max_{a\le t\le b} |\mathfrak{A}_{\mu-i}(t) - \mathfrak{A}_{\pi,\,\mu-i}(t)| \to 0,\\
&\frac{1}{h^{\mu-i}}\ \|\mathfrak{B}_{\mu-i} - \mathfrak{B}_{\pi,\,\mu-i}\|_\infty := \frac{1}{h^{\mu-i}} \max_{a\le t\le b} |\mathfrak{B}_{\mu-i}(t) - \mathfrak{B}_{\pi,\,\mu-i}(t)| \to 0,\\
&\qquad\qquad i = 1,\dots,\mu-1. \qquad (21)
\end{aligned}$$

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

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

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 (21) we know that there is a constant $C_{L^2} > 0$ such that

$$\|Tx - T_\pi x\|_{L^2} \le hC_{L^2}\|x\|_{H^1_D}, \quad x \in H^1_D. \qquad (22)$$

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 4.6 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

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

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

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

become arbitrarily small if h tends to zero. Deriving

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

and using Lemma 4.4 we estimate

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

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

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

In the consequence, it follows that the inequalities

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

and, finally,

$$\|(Tp - T_\pi p)\|_Y \leq \varepsilon_Y \|p\|_{H^1_D}, \quad p \in X_\pi, \tag{23}$$

are valid with values $\varepsilon_{i,s}$, ε_Y being arbitrarily small if h is sufficiently small. Owing to Proposition 4.5 it holds that

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

On the other hand, regarding (23) we can estimate

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

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

$$
\inf_{p\in X_\pi, p\neq 0} \frac{\|T_\pi p\|_Y^2 + |G_a p(a) + G_b p(b)|^2}{\|p\|_{H_D^1}^2} \geq \frac{1}{8c_Y^2}. \tag{24}
$$

Moreover, with the help of (22), we find

$$
\|Tp\|_{L^2} \geq \|T_\pi p\|_{L^2} - \|Tp - T_\pi p\|_{L^2} \geq \|Tp\|_{L^2} - hC_{L^2}\|p\|_{H_D^1}
$$

and thus

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

Summarize what we have obtained so far:

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

Next we provide an estimate of the expression $\mathfrak{E}$.
For each given nontrivial $p \in X_\pi$, the corresponding $q = T_\pi p$ belongs to Y_π^0, and the inequality (24) implies

$$
\|p\|_{H_D^1}^2 \leq 8c_Y^2(\|q\|_Y^2 + |G_a p(a) + G_b p(b)|^2). \tag{25}
$$

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

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

Owing to Lemma 4.6 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 4.4 the further inequalities

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

and

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

So far we have the relation

$$\|q\|_Y^2 + |G_a p(a) + G_b p(b)|^2 \leq \|q\|_{L^2}^2 + |G_a p(a) + G_b p(b)|^2 + \sum_{i=1}^{\mu-1}\frac{2}{h_{\min}^{2\mu-2i}} g_{\mu-i}\ \|q_{\mu-i}\|_{L^2}^2, \tag{26}$$

with

$$g_{\mu-1} = d_{1,\mu-1}c_{\mu-1}^*\|D\mathcal{L}_{\mu-1}\|_\infty^2 > 0, \tag{27}$$

see Lemma 4.6 for $d_{1,\mu-1}$ and Lemma 4.4 for $c_{\mu-1}^*$.

If the projector functions $U_{\mu-i}$ are constant ones, we know that $q_{\mu-i} = U_{\mu-i}q_{\mu-i} = U_{\mu-i}q$, and, therefore,

$$\begin{aligned}\|q\|_Y^2 + |G_a p(a) + G_b p(b)|^2 &\le \|q\|_{L^2}^2 + |G_a p(a) + p_b(b)|^2 \\ &\quad + \frac{1}{h_{\min}^{2\mu-2}}\big(2g_{\mu-1} + O(h^2)\big)\ \|q\|_{L^2}^2 \\ &\le \frac{1}{h_{\min}^{2\mu-2}} 3g_{\mu-1}\ (\|q\|_{L^2}^2 + |G_a p(a) + G_b p(b)|^2).\end{aligned}$$

and hence,

$$\mathfrak{E} \ge \frac{1}{\sqrt{3g_{\mu-1}}} h_{\min}^{\mu-1}.$$

If the projector functions $U_{\mu-i}$ vary with t, the situation is slightly more difficult. Then, regarding the properties (20) we express, for $t \in [t_{j-1}, t_j),\ j = 1, \dots, n$,

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

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

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

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

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

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

$$\|q_{\mu-i}\|_{L^2} \le 2\|q\|_{L^2} + O(h^{\mu-i+1})\|p\|_{H_D^1}.$$

Regarding also (25) we arrive at

$$\|q_{\mu-i}\|_{L^2}^2 \leq 4\|q\|_{L^2}^2 + O(h^{2(\mu-i+1)})(\|q\|_Y^2 + |G_a p(a) + G_b p(b)|^2).$$

Inserting this result into (26) we arrive at the inequality

$$\begin{aligned}\|q\|_Y^2 + |G_a p(a) + G_b p(b)|^2 \leq \|q\|_{L^2}^2 + |G_a p(a) + G_b p(b)|^2 + \sum_{i=1}^{\mu-1} \frac{2}{h_{\min}^{2\mu-2i}} 4g_{\mu-i}\, \|q\|_{L^2}^2 \\ + \sum_{i=1}^{\mu-1} \frac{2}{h_{\min}^{2\mu-2i}} O(h^{2(\mu-i+1)})(\|q\|_Y^2 + |G_a p(a) + G_b p(b)|^2)\end{aligned}$$

and hence,

$$\begin{aligned}\|q\|_Y^2 + |G_a p(a) + G_b p(b)|^2 \leq{} & \frac{1}{h_{\min}^{2\mu-2}} \big(8g_{\mu-1} + O(h^2)\big)(\|q\|_{L^2}^2 + |G_a p(a) + G_b p(b)|^2) \\ & + O(h^2)(\|q\|_Y^2 + |G_a p(a) + G_b p(b)|^2).\end{aligned}$$

that is, for all sufficiently fine partitions,

$$\|q\|_Y^2 + |G_a p(a) + G_b p(b)|^2 \leq \frac{2}{h_{\min}^{2\mu-2}} 9g_{\mu-1}(\|q\|_{L^2}^2 + |G_a p(a) + G_b p(b)|^2)$$

yielding

$$\mathfrak{E} \geq \frac{1}{3\sqrt{2g_{\mu-1}}} h_{\min}^{\mu-1}.$$

Summarizing all we know means

$$\gamma_\pi \geq \frac{1}{\sqrt{8}} \frac{1}{c_Y} \frac{1}{3\sqrt{2g_{\mu-1}}} h_{\min}^{\mu-1} - hC_{L^2} = \frac{1}{12c_Y\sqrt{g_{\mu-1}}} h_{\min}^{\mu-1} - hC_{L^2} \geq \frac{1}{24c_Y\sqrt{g_{\mu-1}}} h_{\min}^{\mu-1}.$$

□

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

The question if the stronger estimate (14) 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 (3) 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 or equal to K.

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

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

By construction, it holds that $1 \leq N \leq K_{\mu-i} \leq 2(\mu-1) + N$.

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 [2] 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* (9) *and let the boundary conditions be restricted by* (4). *Let the condition* (10) *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[4], 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 4.4, we derive

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

and then

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

This leads to

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

and hence

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

□

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

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

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

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

generate a fine decoupling and yield further

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

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

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

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

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 $\mathfrak{E}$, 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 $\mathfrak{E}$ in Subsection 4.3, Part (3), we are then allowed to replace certain applied there inequalities $\|q^{(s)}_{\mu-i}\|_{L^2} \leq \sqrt{c_s^*} h_{\min}^{-s} \|q_{\mu-i}\|_{L^2}$ 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 Collocation via Discrete Norms

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 (6) – (7) comprising $Mnm + l$ equations in the least-squares sense, that is, we minimize the expression

$$\sum_{j=1}^{n} h_j \sum_{t \in S_j} |A(t)(Dp)'(t) + B(t)p(t) - y(t)|^2 + |G_a p(a) + G_b p(b) - r|^2. \tag{28}$$

subject to $p \in X_\pi$. The linear space of piecewise polynomial functions X_π be defined as before in (5), the $M > N$ interpolation nodes $0 < \tau_1 < \tau_2 <$

$\dots < \tau_M < 1$ be fixed, and S_j be the resulting sets of collocation points on the subinterval $[t_{j-1}, t_j)$ as before.

On the other hand, 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

$$y_\pi|_{[t_{j-1},t_j)} \in \mathcal{P}^m_{M-1}, \quad y_\pi(t) = y(t),\ t \in S_j,\ j = 1, \dots, n. \tag{29}$$

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

$$\delta_n := \|z - z_\pi\|_Z = \|y - y_\pi\|_{L^2} \leq c_\delta h^M. \tag{30}$$

Following [3], it makes sense to turn to the perturbed equation $\mathcal{T}x = y_\pi$ and provide the further approximate solution

$$p_\pi^{\delta_\pi} \in \operatorname{argmin}\{\|Tp - y_\pi\|^2_{L^2} + |G_a p(a) + G_b p(b) - r|^2 :\ p \in X_\pi\}, \tag{31}$$

which satisfies the inequality

$$\|p_\pi^{\delta_\pi} - x_*\|_{H^1_D} \leq \frac{\beta_\pi + \delta_\pi}{\gamma_\pi} + \alpha_\pi. \tag{32}$$

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

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

we ensure

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

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

$$\begin{aligned} w &:= A(Dp)' + Bp - y_\pi, \\ W_j &:= h_j^{\frac{1}{2}} \begin{bmatrix} w(t_{j-1} + \tau_1 h_j) \\ \vdots \\ w(t_{j-1} + \tau_M h_j) \end{bmatrix} \in \mathbb{R}^{mM}, \quad W = \begin{bmatrix} W_1 \\ \vdots \\ W_n \end{bmatrix} \in \mathbb{R}^{mMn}, \end{aligned}$$

and derive along the lines of [2, Subsection 2.3] that

$$\|w\|^2_{L^2} = W^T \mathcal{L} W, \tag{33}$$

with a symmetric, positive definite matrix $\mathcal{L}$. Its entries do not at all depend on the partition π.[5] Further, there are positive constants c_L, $\bar{c}_L$ depending only on $\mathcal{L}$ such that

$$c_L |W|_2 \leq \|w\|_{L^2} \leq \bar{c}_L |W|_2. \tag{34}$$

[5] In [2] only equidistant partitions are considered. By marginal modifications the arguments remain valid also for general partitions.

Here, we denote the Euclidean norm of $W \in \mathbb{R}^{mMn}$ by $|W|_2$ and introduce $|w|_2 := |W|_2$. Actually, the relation 34 indicates a norm equivalence on the related finite-dimensional subspace in L^2, cf. [2, Proposition 2.7].
Regarding the interpolation condition (29), expression (28) exactly coincides with

$$|W|_2^2 + |G_a p(a) + G_b p(b) - r|^2 = |w|_2^2 + |G_a p(a) + G_b p(b) - r|^2$$

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

$$\|w\|_{L^2}^2 + |G_a p(a) + G_b p(b) - r|^2, \tag{35}$$

we use the equivalent norm $|w|_2$ for $\|w\|_{L_2}$. In this context, $\|w\|_{L_2}$ can be interpreted as a weighted form of $|w|_2$. Experiments using both norms indicate no significant differences, see [2, Section 6].
As a consequence of the estimates (32) and (30) as well as the Theorems 4.1 and 4.7 we obtain

Theorem 5.1. *Let the BVP* (1) – (2)*, with index $\mu \geq 1$, satisfy the assumptions of Theorem* 2.4(1) *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 $M \geq 1 + N + N_{A,B}$,*[6] *with $N \geq 1$, then the following statements are valid for all partitions π with sufficiently small stepsize h and uniformly bounded ratios $\frac{h}{h_{\min}} \leq \rho$:*

1. *The least-squares collocation solutions $p_\pi^{\delta_\pi}$ of the overdetermined system* (6) – (7) *defined by* (31) *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 $h \to 0$.

2. *Moreover, if the coefficients A und 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 $N < \mu - 1$.

6 Numerical Experiments

In the first instance we exhibit once again order results of the experiments carried out in [2] for Example 1.1. The numerical orders of convergence are calculated using the norm of $L^2(0,1)$, cf. (35) and $\mathbb{R}^{3Mn}$, cf. (28)), in the image space. All results can be found in Table 2. In this and the following tables, the error is measured in the $H_D^1(0,1)$-norm. The column labelled order contains an estimation k_{est} of the order,

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

Table 2: Example 1.1: Error of the collocation solution for $\eta = -2$ and $N = 3$. The $M = 2N + 1$ collocation points τ_i are uniformly distributed. The columns labelled L^2 show results for minimizing the expression (35) while the columns labelled $\mathbb{R}$ show results for minimizing the expression (28).

	L^2		$\mathbb{R}$	
n	error	order	error	order
10	6.31e-4		6.51e-4	
20	1.44e-4	2.1	1.47e-4	2.1
40	3.47e-5	2.1	3.52e-5	2.1
80	8.53e-6	2.0	8.59e-6	2.0
160	2.12e-6	2.0	2.12e-6	2.0
320	5.27e-7	2.0	5.28e-7	2.0
640	1.79e-7	1.6	1.39e-7	1.9

Table 3: Example 1.1: Error of the collocation solution for $\eta = -2$ and $N = 1$. The $M = 3$ collocation points τ_i are uniformly distributed. The columns labelled L^2 show results for minimizing the expression (35) while the columns labelled $\mathbb{R}$ show results for minimizing the expression (28).

	L^2		$\mathbb{R}$	
n	error	order	error	order
10	5.65e-1		4.94e-1	
20	3.93e-1	0.5	3.14e-1	0.6
40	2.49e-1	0.6	2.14e-1	0.6
80	1.85e-1	0.4	1.62e-1	0.4
160	1.42e-1	0.4	1.26e-1	0.4
320	1.12e-1	0.3	1.00e-1	0.3
640	9.01e-2	0.3	8.17e-2	0.3

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 the results for the example for $N = 1$ in Table 3. Note that Theorem 4.1 provides a bound on the error of the order h^{-1}. 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.

A slightly more involved example is obtaint by applying the transformation

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

as well as a corresponding refactorization of the leading term. This does not change the index of the DAE, see [4]. 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 y reads

$$\tilde{A}(\tilde{D}y)' + \tilde{B}y = q(t), \tag{36}$$

[6] 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.

Table 4: Errors and estimation of the convergence order for (36). The expression (35) has been used.

	$N=2$		$N=3$		$N=4$		$N=5$	
n	error	order	error	order	error	order	error	order
10	1.06e-1		6.45e-2		2.39e-3		1.02e-4	
20	6.15e-2	0.8	3.23e-2	1.0	4.25e-4	2.5	1.28e-5	3.0
40	3.99e-2	0.6	1.60e-2	1.0	7.53e-5	2.5	1.60e-6	3.0
80	2.70e-2	0.6	7.89e-3	1.0	1.33e-5	2.5	2.04e-7	3.0
160	1.87e-2	0.5	3.92e-3	1.0	2.37e-6	2.5	8.54e-8	1.3
320	1.31e-2	0.5	1.95e-3	1.0	5.27e-7	2.2	3.42e-7	−2.0
640	9.23e-3	0.5	9.75e-4	1.0	1.59e-6	−1.6	5.29e-1	−2.0

Table 5: Errors and estimation of the convergence order for $N=5$. The columns labelled L^2 show results for minimizing the expression (35) while the columns labelled $\mathbb{R}$ show results for minimizing the expression (28).

	L^2		$\mathbb{R}$	
n	error	order	error	order
10	1.02e-4		4.33e-5	
20	1.28e-5	3.0	5.13e-6	3.1
40	1.60e-6	3.0	6.30e-7	3.0
80	2.04e-7	3.0	7.93e-8	3.0
160	8.54e-8	1.3	3.26e-8	1.3
320	3.42e-7	−2.0	1.42e-7	−2.1
640	5.29e-1	−2.0	5.07e-2	−1.8

where

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

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

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

Table 4 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 4.1 for $N = 3$ and $N = 5$ while the order is by 0.5 higher for the even orders N. We do not have any explanation for this behavior.

In the next experiment, we replaced the norm in L^2 by a discrete version as indicated in (28). The collocation points have been chosen as Gaussian points scaled to $(0, 1)$ and one additonal point close to 0.5 but not coinciding with a Gaussian point.[7] Table 5 compares both versions (11) and (28). Both versions behave similar and support the considerations in Section 5.

Finally, we consider the case of $N = 1$. The experiment is done using the settings as in Table 5 but with a different N.[8] The results are listed in Table 6. Note that Theorem 4.7 does not apply here. Theorem 4.1 guarantees a bound

[7] The exact value is $1/2(1 + 1/42(245 - 14 \cdot 70^{1/2})^{1/2})$ for $N = 5$.

[8] The collocation points are 1/2 and 3/4.

Table 6: Errors and estimation of the convergence order for $N = 1$. The columns labelled L^2 show results for minimizing the expression (35) while the columns labelled $\mathbb{R}$ show results for minimizing the expression (28).

	L^2		$\mathbb{R}$	
n	error	order	error	order
10	2.97e-1		1.76e-1	
20	1.75e-1	0.8	1.01e-1	0.8
40	1.03e-1	0.8	6.85e-2	0.6
80	7.06e-2	0.5	5.68e-2	0.3
160	5.87e-2	0.3	5.16e-2	0.1
320	5.33e-2	0.1	4.72e-2	0.1
640	4.88e-2	0.1	4.24e-2	0.2

of the order h^{-1}, only. However, the approximate solution does not only remain bounded but we oberserve even convergence although rather slow. We do not have any theoretical backup for this behavior.

7 Conclusions

We have consolidated the recently developed new least-squares collocation method for the numerical solution of initial value 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 provide the corresponding functional analytic setting.

The basic idea of the proposed numerical method is the approximation of such a problem by a least-squares method where both, the image and the pre-image space are discretized. In the context of DAEs, this idea results in an extremely simple algorithm whose computational complexity is comparable to standard polynomial collocation methods for systems of ordinary differential equations. In particular, neither analytical preprocessing nor special structures of the DAE is necessary. In the numerical experiments, the method behaves in a robust and stable way, showing fast convergence. In our opinion, treating the DAEs as ill-posed problems is a fruitful approach and this idea deserves further research interest.

8 Acknowledgements

The work of Caren Tischendorf was supported by the German Research Foundation (DFG CRC/Transregio 154, subproject C02).

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] M. Hanke, R. März, C. Tischendorf, E. Weinmüller, and S. Wurm, *Least-squares collocation for linear higher-index differential-algebraic equations*, Tech. report, Humboldt-Universität Berlin, Istitut für Mathematik, 2016, http://dx.doi.org/10.1016/j.cam.2016.12.017.

[3] B. Kaltenbacher and J. Offtermatt, *A convergence analysis of regularization by discretization in preimage space*, Math. Comp. **81** (2012), no. 280, 2049–2069.

[4] R. Lamour, R. März, and C. Tischendorf, *Differential-algebraic equations: A projector based analysis*, Differential-Algebraic Equations Forum, Springer-Verlag Berlin Heidelberg New York Dordrecht London, 2013, Series Editors: A. Ilchman, T. Reis.

[5] R. Lamour, R. März, and E. Weinmüller, *Surveys in differential-algebraic equations iii*, Differential-Algebraic Equations Forum, ch. Boundary-Value Problems for Differential-Algebraic Equations: A Survey, pp. 177–309, Springer Heidelberg, 2015, ed. by A. Ilchmann and T. Reis.

[6] R. März, *Numerical methods for differential-algebraic equations*, Acta Numer. **1** (1992), 141–198.

[7] ______, *Surveys in differential-algebraic equations II*, Differential-Algebraic Equations Forum, ch. Differential-Algebraic Equations from a Functional-Analytic Viewpoint: A Survey, pp. 163–285, Springer Heidelberg, 2015, ed. by A. Ilchmann and T. Reis.