A least-squares collocation method for nonlinear higher-index differential-algebraic equations

Michael Hanke^a, Roswitha März^b

^aKTH Royal Institute of Technology, School of Engineering Sciences, 10044 Stockholm, Sweden ^bHumboldt University of Berlin, Institute of Mathematics, D-10099 Berlin, Germany

Abstract

We introduce a direct numerical treatment of nonlinear higher-index differential-algebraic equations by means of overdetermined polynomial least-squares collocation. The procedure is not much more computationally expensive than standard collocation methods for regular ordinary differential equations. The numerical experiments show impressive results. In contrast, the theoretical basic concept turns out to be considerably challenging. So far, quite recently convergence proofs for linear problems have been published. In the present paper we come up to a first convergence result for nonlinear problems.

Keywords: differential-algebraic equation, higher-index, essentially ill-posed problem, overdetermined collocation, polynomial collocation, nonlinear problem

1. Introduction

For regular ordinary differential equations and index-1 differential-algebraic equations *standard collocation methods* which rely on closed discretized systems¹ are known to work well. Moreover, Hessenberg form index-2 differential-algebraic equations can be treated successfully by so-called *projected collocation methods* that complement standard collocation with an additional updating of the differential solution component by a projection step. This goes along with the well-posedness of the related initial and boundary value problems in natural settings; we refer to [12] for a detailed survey. In contrast, higher-index differential-algebraic equations lead to ill-posed² initial and boundary value problems, and standard collocation methods necessarily fail unless an elaborate index-reducing preprocessing is incorporated, which utilizes derivative array systems.

Recently ([7, 8]) first promising experiments concerning an least-squares overdetermined polynomial collocation directly applied to the DAE without any preprocessing have been reported. The theoretical justification appears to be quite challenging. So far, only sufficient convergence conditions are obtained for linear problems [7, 6, 8]. In the present paper we provide a first proof for nonlinear problems.

Email addresses: hanke@nada.kth.se(Michael Hanke), maerz@math.hu-berlin.de(Roswitha März)

¹The number of unknowns equals the number of equations.

²More precisely: Essentially ill-posed in Tichonov's sense, that is, the related operators feature nonclosed ranges.

The paper is organized as follows: In Section 2 we state the problem in detail. Then we provide a Hilbert space setting in Section 3. This setting is more comfortable for the treatment of the given ill-posed problems. In Section 4, we introduce and investigate a kind of Newton-iteration related to a single partition, which uses bounded outer inverses as discussed in [15] and which serves in the end as background for the Gauss-Newton iteration applied to an overdetermined collocation system. Then, we consider nested multiple partitions to ensure convergence of the iteration-projection method in Section 6. The examples in Section 5 confirm the capability of the approach, but, having said that, they also indicate that our sufficient convergence conditions seem to be too unsubtle still. Finally, we provide some remarks and conclusions.

We use the symbol $\|\cdot\|$ for different function and operator norms. In general, in the given context things will be unambiguous. Only on certain places, to prevent maybe imminent confusions we indicate the special norms by the corresponding subscripts, e.g., $\|\cdot\|_{L^2}$.

Some notations and abbreviations

\mathbb{R}	set of real numbers
$\mathcal{L}(\mathbb{R}^s,\mathbb{R}^n)$	space of linear operators from \mathbb{R}^s to \mathbb{R}^n , also set of $n \times s$ - matrices with real entries
$C([a,b],\mathbb{R}^m)$	space of continuous functions mapping $[a, b]$ into \mathbb{R}^m
$C^s([a,b],\mathbb{R}^m)$	space of s-times continuously differentiable functions mapping $[a,b]$ into \mathbb{R}^m
	Lebesque space of functions mapping (a, b) into \mathbb{R}^m
* * * * * * * * * * * * * * * * * * * *	$:= W^{k,2}((a,b),\mathbb{R}^m)$, Sobolev space of functions mapping (a,b) into \mathbb{R}^m
$H_D^1 := H_D^1((a,b), \mathbb{R}^m)$	$:= \{x \in L^2 : Dx \in H^1\}$
K^{-}	generalized inverse of the operator K : $KK^-K = K$, $K^-KK^- = K^-$
K^+	Moore-Penrose inverse of <i>K</i>
ker K	nullspace (kernel) of K
im K	range (image) of K
$\langle \cdot, \cdot angle$	Euclidean inner product in \mathbb{R}^m
(\cdot,\cdot)	inner product in a function space
•	Euclidean vector norm and spectral norm of a matrix
$\ \cdot\ $	norm of function space element and operator norm
\oplus	topological direct sum
\mathfrak{P}_N	set of all polynomials of degree less than or equal to N
DAE, DAO	differential-algebraic equation, differential-algebraic operator
ODE	ordinary differential equation
IVP, BVP	initial value problem, boundary value problem

2. The issue and basic technicalities

We deal with IVPs and BVPs given in the form

$$f((Dx)'(t), x(t), t) = 0, \quad t \in [a, b], \tag{1}$$

$$g(x(a), x(b)) = 0, (2)$$

with [a, b] being a compact interval, $D = [I \ 0] \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^k)$, rank D = k, and data $f(y, x, t) \in \mathbb{R}^m$, $y \in \mathbb{R}^k$, $x \in \mathcal{D}_x \subseteq \mathbb{R}^m$, $t \in \mathcal{D}_t \subset \mathbb{R}$, $\mathcal{D}_t \supset [a, b]$, $g(u, v) \in \mathbb{R}^l$, $u, v \in \mathbb{R}^m$. The functions f and g are supposed to be at least continuous together with their partial derivatives f_v , f_x , g_u , g_v .

We assume that the BVP (1), (2) has the solution $x_* : [a, b] \to \mathbb{R}^m$ to be approximated. x_* is supposed to be continuous with continuously differentiable part Dx_* . Later on, among others for obtaining convergence orders, additional smoothness will be required.

Moreover, the DAE (1) is supposed to be regular with (tractability) index $\mu \in \mathbb{N}$ and characteristics $0 < r_0 \le \cdots \le r_{\mu-1} < r_m = m$ around x_* , that means, the graph $\{(x_*(t), t) : t \in [a, b]\}$ resides within a regularity region having these characteristics (e.g., [11, Definition 3.28]). Note that then the derivative (Dx)' is properly involved in the DAE (1) so that $f_{\nu}(y, x, t)$ has full column-rank k.

Furthermore, in condition (2), we apply $l = m - \sum_{i=0}^{\mu-1} (m - r_i) \ge 0$ which is the dynamical degree of freedom of the DAE. Recall that regular ODEs are indicated by l = k = m, regular index-1 DAEs by l = k < m, but higher-index DAEs by l < k < m. We are mainly interested in the last case. We further suppose the function g to satisfy the relation

$$g(u, v) = g(D^{+}Du, D^{+}Dv), \quad u, v \in \mathbb{R}^{m},$$
 (3)

so that the initial or boundary condition (2) actually applies to the differentiable component Dx only.

Together with the BVP (1),(2) we consider the linear BVP,

$$A_*(t)(Dz)'(t) + B_*(t)z(t) = q(t), \quad t \in [a, b], \tag{4}$$

$$G_{*a}z(a) + G_{*b}z(b) = d, (5)$$

with

$$A_*(t) := f_{y}((Dx_*)'(t), x_*(t), t), \quad B_*(t) := f_{x}((Dx_*)'(t), x_*(t), t), \quad t \in [a, b],$$

$$G_{*a} := g_{u}(x_*(a), x_*(b)), \quad G_{*b} := g_{v}(x_*(a), x_*(b)).$$

We assume the solution x_* and possibly the data f to be sufficiently smooth so that the linearized DAE (4) is fine in the sense of [11, Section 2.6]. Since the solution x_* resides in a regularity region of the DAE (1), the linear DAE (4) inherits the characteristic values and the index μ of the nonlinear DAE, see [14, Page 279]. Furthermore, owing to condition (3) it holds that

$$\ker D \subseteq \ker G_{*a}, \quad \ker D \subseteq \ker G_{*b}.$$
 (6)

Condition (2) is supposed to be stated in such a way that the linear BVP (4),(5) features accurately stated boundary condition in the sense of [12, Definition 2.3]), meaning that the problems

$$A_*(t)(Dz)'(t) + B_*(t)z(t) = 0, \ t \in [a,b], \quad G_{*a}z(a) + G_{*b}z(b) = d, \tag{7}$$

are uniquely solvable for each $d \in \mathbb{R}^l$, and the solutions satisfy the inequality

$$\max_{t\in[a,b]}|z(t)|\leq \kappa_{BC}|d|,$$

with a constant κ_{BC} . In particular, the homogeneous linear BVP, that is, the so-called variational problem, has then the trivial solution only.

Given the partition

$$\pi : a = t_0 < t_1 < \dots < t_n = b,$$
 (8)

with stepsizes $h_j = t_j - t_{j-1}$, maximal stepsize h_{π} , and minimal stepsize $h_{\pi, \min}$. Denote by $\mathcal{M}_{[r]}$ the set of all partitions π the ratio of the maximal stepsize by the minimal stepsize of which is uniformly bounded by the constant $r < \infty$.

Let $C_{\pi}([a,b],\mathbb{R}^m)$ denote the space of piecewise continuous functions having breakpoints merely at the mesh points.

Next we fix a number $N \ge 1$ and introduce the space X_{π} of ansatz functions to approximate the solution x_* by piecewise polynomial functions,

$$X_{\pi} = \{ x \in C_{\pi}([a, b], \mathbb{R}^{m}) : Dx \in C([a, b], \mathbb{R}^{m}),$$

$$x_{\kappa|_{[t_{i-1}, t_{i})}} \in \mathfrak{P}_{N}, \ \kappa = 1, \dots, k, \quad x_{\kappa|_{[t_{i-1}, t_{i})}} \in \mathfrak{P}_{N-1}, \ \kappa = k + 1, \dots, m, \ j = 1, \dots, n \}.$$
 (9)

This ansatz space has dimension nNm + k. Choosing values

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

we specify M collocation points per subinterval, i.e.,

$$t_{ji} = t_{j-1} + \tau_i h_j, \quad i = 1, \dots, M, \ j = 1, \dots, n,$$

and are then confronted with the collocation system of nMm + l equations for providing an approximation $x \in X_{\pi}$, namely,

$$f((Dx)'(t_{ii}), x(t_{ii}), t_{ii}) = 0, \quad i = 1, \dots, M, \ j = 1, \dots, n,$$
 (10)

$$g(x(t_0), x(t_n)) = 0,.$$
 (11)

The choice M = N corresponds to the standard polynomial collocation yielding nNm + l equations, which works well for regular ODEs and index-1 DAEs, with dynamical degree l = k = m and l = k < m, respectively (cf. [12]). In contrast, higher-index DAEs feature always a dynamical degree $0 \le l < k < m$. As it is well-known, completing the collocation system by additional k - l consistent boundary conditions does not result in a suitable method owing to the ill-posedness of the higher-index problem, e.g., [7, Example 1.1]. As a matter of course, the choice M > N goes along with an overdetermined system (10),(11) comprising more equations than unknowns.

Here we always set M > N and treat the overdetermined collocation system in a least-squares sense. More precisely, let $R_{\pi,M}: C_{\pi}([a,b],\mathbb{R}^m) \to C_{\pi}([a,b],\mathbb{R}^m)$ denote the restriction operator which assigns to $w \in C_{\pi}([a,b],\mathbb{R}^m)$ the piecewise polynomial $R_{\pi,M}w \in C_{\pi}([a,b],\mathbb{R}^m)$ of degree less than M such that the interpolation conditions,

$$(R_{\pi,M}w)(t_{ii}) = w(t_{ii}), \quad i = 1, \dots, M, \ j = 1, \dots, n,$$

are satisfied. We also assign to $w \in C_{\pi}([a, b], \mathbb{R}^m)$ the vector $W \in \mathbb{R}^{mMn}$,

$$W = \begin{bmatrix} W_1 \\ \vdots \\ W_n \end{bmatrix} \in \mathbb{R}^{mMn}, \quad W_j = \left(\frac{h_j}{M}\right)^{1/2} \begin{bmatrix} w(t_{j1}) \\ \vdots \\ w(t_{jM}) \end{bmatrix} \in \mathbb{R}^{mM},$$

which yields (cf. [8, Subsection 3])

$$\|R_{\pi,M}w\|_{L^2}^2 = W^T \mathcal{L} W, \quad w \in C_\pi([a,b],\mathbb{R}^m),$$

with a positive definite, symmetric matrix \mathcal{L} . The entries of \mathcal{L} do not at all depend on the partition π . They are fully determined by the corresponding M Lagrange basis polynomials.

Letting $w_f(t) = f((Dx)'(t), x(t), t), t \in [a, b]$, we introduce the functional

$$\psi_{\pi,M}(x) = W_f^T \mathcal{L} W_f + |g(x(a), x(b))|^2$$
(12)

$$= ||R_{\pi,M}w_f||^2 + |g(x(a), x(b))|^2, \quad x \in X_{\pi}.$$
(13)

The overdetermined least-squares collocation means now that we seek an element \tilde{x}_{π} making the value $\psi_{\pi,M}(\tilde{x}_{\pi})$ as small as possible. Note that there are positive constants $c_{\mathcal{L}}$, $C_{\mathcal{L}}$ such that

$$c_{\mathcal{L}}|W|^2 = c_{\mathcal{L}}W^TW \le W^T\mathcal{L}W \le C_{\mathcal{L}}W^TW = C_{\mathcal{L}}|W|^2, \quad W \in \mathbb{R}^{mMn},$$

which justifies the labeling *least squares collocation*. We refer to [8, 7] for a number of promising numerical experiments, see also Section 5. Expression (12) serves to indicate the basic numerical procedure, whereas formula (13) suggests that the mathematics behind is closely related to special properties of the restriction operator $R_{\pi,M}$ on the one hand, but on the other hand, to the problem to minimize the functional

$$\psi(x) = ||w_f||^2 + |g(x(a), x(b))|^2 \quad \text{subject to } x \in X_\pi,$$
(14)

for which (13) serves as approximation. We refer to [6] for properties of the restriction operator in this context. The objective of the present paper is to contribute to the background problem (14).

3. Hilbert space setting

Following the ideas of [8, 7] concerning linear problems, we investigate also the nonlinear problem (1),(2) described in Section 1 as operator equation $\mathcal{F}x = 0$ in a Hilbert space setting, which is most comfortable for treating ill-posed problems. Besides standard function spaces such as L^2 , H^1 , C, etc., equipped with usual inner products and norms, we use the space

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

equipped with the inner product

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

 H_D^1 is a Hilbert space, [14, Lemma 6.9]. Owing to the continuous embedding $H^1((a,b), \mathbb{R}^k) \hookrightarrow C([a,b], \mathbb{R}^k)$, e.g., [1, Theorem 0.4], $x \in H_D^1$ implies $Dx \in C([a,b], \mathbb{R}^k)$, and it holds

$$||Dx||_{\infty} \le \kappa ||Dx||_{H^{1}} \le \kappa ||x||_{H^{1}_{D}}, \quad x \in H^{1}_{D}.$$
(15)

We introduce the nonlinear operators F, F_{BC} , and \mathcal{F} ,

 $F:\operatorname{dom} F\subseteq H^1_D\to L^2,\quad F_{BC}:\operatorname{dom} F\subseteq H^1_D\to \mathbb{R}^l,\quad \mathcal{F}:=(F,F_{BC}):\operatorname{dom} F\subseteq H^1_D\to L^2\times \mathbb{R}^l,$

$$(Fx)(t) := f((Dx)'(t), x(t), t), \ t \in (a, b), \quad x \in \text{dom } F,$$
 (16)

$$F_{BC} x := g(x(a), x(b)), \quad x \in \text{dom } F,$$
(17)

$$\mathcal{F}x := (Fx, F_{BC}x). \quad x \in \text{dom } F, \tag{18}$$

as well as the linear operators T, T_{BC} , and \mathcal{T} ,

$$T: H_D^1 \to L_2, \quad T_{BC}: H_D^1 \to \mathbb{R}^l, \quad \mathcal{T}:= (T, T_{BC}): H_D^1 \to L^2 \times \mathbb{R}^l,$$

$$(Tx)(t) := A_*(Dx)' + B_*x, \ t \in (a,b), \quad x \in H_D^1, \tag{19}$$

$$T_{BC} x := G_{*a} x(a) + G_{*b} x(b), \quad x \in H_D^1,$$
 (20)

$$\mathcal{T}x := (Tx, T_{BC}x). \quad x \in H_D^1. \tag{21}$$

We are merely interested in the local behavior of F and \mathcal{F} and suppose

$$\operatorname{dom} \mathcal{F} = \operatorname{dom} F = \mathfrak{B}(x_*, \rho) \subset H_D^1.$$

Regarding condition (3) as well as (15), we find the operators F_{BC} and T_{BC} well defined. F_{BC} is Fréchet-differentiable, which can be checked by straightforward computation. In particular, $F_{BC}(x_*) = T_{BC}$. Moreover, supposing the partial derivatives g_u, g_v to be Lipschitz continuous, there is a constant L_{BC} such that

$$\begin{split} \|F_{BC}'(x) - F_{BC}'(\bar{x})\|_{H_D^1 \to \mathbb{R}^l} &\leq L_{BC} \|x - \bar{x}\|_{H_D^1}, \quad x, \bar{x} \in \text{dom } F, \\ \|F_{BC}'(x)\|_{H_D^1 \to \mathbb{R}^l} &\leq L_{BC} \; \rho + \|T_{BC}\|_{H_D^1 \to \mathbb{R}^l}, \quad x \in \text{dom } F. \end{split}$$

The linear operators T and \mathcal{T} are obviously bounded. The operator F is closely related to a certain Nemyckij operator as Proposition 3.1 below indicates. In the convergence proofs we will need that F and thus \mathcal{F} are Gâteaux-differentiable on their domain with uniformly bounded Gâteaux-derivatives,

$$||F'(x)||_{H_D^1 \to L^2} \le C_F, \quad x \in \text{dom } F,$$

$$||\mathcal{F}'(x)||_{H_D^1 \to L^2 \times \mathbb{R}^l} \le ||F'(x)||_{H_D^1 \to L^2} + ||F'_{BC}(x)||_{H_D^1 \to \mathbb{R}^l} \le C_{\mathcal{F}}, \quad x \in \text{dom } F,$$

$$C_{\mathcal{F}} := C_F + L_{BC} \rho + ||T_{BC}||_{H_D^1 \to \mathbb{R}^l}.$$
(22)

Proposition 3.1 provides sufficient conditions to justify these assumptions.

Moreover, we will need the inequality

$$||F'(x) - F'(\bar{x})||_{H_D^1 \to L^2} \le L_F h_{\pi}^{-1/2} ||x - \bar{x}||_{H_{\pi}^1}, \quad x, \bar{x} \in \text{dom } F \cap X_{\pi}, \ \pi \in \mathcal{M}_{[r]},$$
 (23)

to be valid with a constant L_F for the Gâteaux-derivative F' where X_{π} is given by (9). Proposition 3.1 provides conditions also for this property to hold. Having (23), we are provided with a constant L such that

$$\|\mathcal{F}'(x) - \mathcal{F}'(\bar{x})\|_{H_D^1 \to L^2 \times \mathbb{R}^l} \le (L_F h_{\pi}^{-1/2} + L_{BC}) \|x - \bar{x}\|_{H_D^1}$$

$$\le L h_{\pi}^{-1/2} \|x - \bar{x}\|_{H_D^1}, \quad x, \bar{x} \in \text{dom } F \cap X_{\pi}, \ \pi \in \mathcal{M}_{[r]}.$$
(24)

Note that L_F and L depend on the stepsize ratio r.

Now the BVP (1),(2) is represented by the operator equation $\mathcal{F}x = 0$ and the least-squares functional (14) we are mainly interested in reads now

$$\psi(x) = \|\mathcal{F}x\|^2, \quad x \in \text{dom } F. \tag{25}$$

By construction, one has $T = F'(x_*)$ and $\mathcal{T} = \mathcal{F}'(x_*)$. The equation $\mathcal{F}'(x_*)z = 0$ represents the homogeneous variational BVP (7), with d = 0, which has the trivial solution only. Therefore, the operator $\mathcal{F}'(x_*)$ is injective. At this place we emphasize again, that higher-index DAEs lead to ill-posed problems. In the context here this means that im $F'(x_*)$ and im $\mathcal{F}'(x_*)$ are non-closed subsets in L^2 and $L^2 \times \mathbb{R}^l$, respectively, see [7, Theorem 2.4], and the inverse $\mathcal{F}'(x_*)^{-1}$ is unbounded.

Proposition 3.1. Let f and D be as described in Section 2, with $\mathcal{D}_x = \mathbb{R}^m$, $\mathcal{D}_y = \mathbb{R}^k$ and bounded partial derivatives f_y and f_x .

(i) Then, $x \in H_D^1$ implies $Fx \in L^2$, and F is Gâteaux-differentiable, with the Gâteaux-derivative F'(x),

$$F'(x)z = A_{(x)}(Dz)' + B_{(x)}z, \quad z \in H_D^1,$$
(26)

$$A_{(x)}(t) := f_y((Dx)'(t), x(t), t), \quad B_{(x)}(t) := f_x((Dx)'(t), x(t), t), \quad \text{a.e. } t \in (a, b).$$

Moreover, F'(x) *is uniformly bounded.*

(ii) If, additionally, the partial derivatives f_x and f_y satisfy the inequalities

$$|f_y(y_1, x_1, t) - f_y(y_2, x_2,)|^2 \le \tilde{L}^2(|y_1 - y_2|^2 + |x_1 - x_2|^2),$$

$$|f_x(y_1, x_1, t) - f_x(y_2, x_2,)|^2 \le \tilde{L}^2(|y_1 - y_2|^2 + |x_1 - x_2|^2),$$

for all $x_1, x_2 \in \mathbb{R}^m$ and $y_1, y_2 \in \mathbb{R}^k$, then there is a constant $L_F = L_F(r)$ such that

$$||F'(x) - F'(\bar{x})||_{H_D^1 \to L^2} \le L_F h_{\pi}^{-1/2} ||x - \bar{x}||_{H_D^1}, \quad x, \bar{x} \in \text{dom } F \cap X_{\pi}, \ \pi \in \mathcal{M}_{[r]},$$

that is (23).

Proof. (i) Consider the operators $J: H_D^1 \to L^2((a,b),\mathbb{R}^k) \times L^2$ given by Jx = ((Dx)',x) and $\tilde{F}: L^2((a,b),\mathbb{R}^k) \times L^2 \to L^2$ defined as the Nemytskij operator

$$F(y, x)(t) = f(y(t), x(t), t).$$

Under the stated conditions on f, \tilde{F} is well-defined [1, Theorem 1.2.2]. Moreover, it is Gâteaux-differentiable and its Gâteaux-differential is given by [1, Theorem 1.2.7]

$$\tilde{F}'(y,x)(u,v) = f_y(y,x,\cdot)u + f_x(y,x,\cdot)v.$$

Now, $F = \tilde{F} \circ J$. Hence,

$$\lim_{h \to 0} \frac{1}{h} (F(x+tz) - F(x)) = \lim_{h \to 0} \frac{1}{h} (\tilde{F}(J(x+tz)) - \tilde{F}(J(x)))$$

$$= \lim_{h \to 0} \frac{1}{h} (\tilde{F}(J(x) + tJ(z)) - \tilde{F}(J(x)))$$

$$= \lim_{h \to 0} \frac{1}{h} (\tilde{F}(u+tv) - \tilde{F}(u))$$

$$= \tilde{F}'(u)v$$

$$= f_v((Dx)', x, \cdot)(Dz)' + f_v((Dx)', x, \cdot)z$$

where we used u = J(x) and v = J(z).

The norm of the derivative can be estimated by

$$\begin{aligned} \|A_{(x)}(Dz)' + B_{(x)}z\|^2 &\leq \|A_{(x)}\|_{L^{\infty}((a,b),\mathbb{R}^{m\times k})}^2 \|(Dz)'\|^2 + \|B_{(x)}\|_{L^{\infty}((a,b),\mathbb{R}^{m\times m})}^2 \|z\|^2 \\ &\leq C^2 \|z\|_{H_D^1}^2, \end{aligned}$$

where C denotes a bound on the partial derivatives f_y and f_x . Hence, the Gâteaux-derivative is uniformly bounded.

(ii) We will need an inverse inequality for functions from X_{π} . A consequence of [3, Theorem 3.2.6] is the estimate

$$||x||_{L^{\infty}((a,b),\mathbb{R}^m)} \le ch_{\pi}^{-1/2}||x||, \quad x \in X_{\pi}$$
 (28)

for a constant c independent of $\pi \in \mathcal{M}_{[r]}$.

Let $\pi \in \mathcal{M}_{[r]}$ and $x, \bar{x} \in X_{\pi}$. Then ist holds

$$\begin{split} ||(F'(x) - F'(\bar{x}))z||^2 &= \int_a^b |f_y((Dx)'(t), x(t), t) - f_y((D\bar{x})'(t), \bar{x}(t), t)|^2 |(Dz)'(t)|^2 dt \\ &+ \int_a^b |f_x((Dx)'(t), x(t), t) - f_x((D\bar{x})'(t), \bar{x}(t), t)|^2 |z(t)|^2 dt \\ &\leq \tilde{L}^2 \int_a^b \left(|(Dx)'(t) - (D\bar{x})'(t)|^2 + |x(t) - \bar{x}(t)|^2 \right) |(Dz)'(t)|^2 dt \\ &+ \tilde{L}^2 \int_a^b \left(|(Dx)'(t) - (D\bar{x})'(t)|^2 + |x(t) - \bar{x}(t)|^2 \right) |z(t)|^2 dt \\ &\leq \tilde{L}^2 \max_{a \leq t \leq b} \left(|(Dx)'(t) - (D\bar{x})'(t)|^2 + |x(t) - \bar{x}(t)|^2 \right) \int_a^b \left(|(Dz)'(t)|^2 + |z(t)|^2 \right) dt \\ &\leq \tilde{L}^2 \left(||(Dx)' - (D\bar{x})'||_\infty^2 + ||x - \bar{x}||_\infty^2 \right) ||z||_{H_D^1}^2. \end{split}$$

Applying (28), we arrive at

$$\|(F'(x) - F'(\bar{x}))z\|^2 \le \tilde{L}^2 c^2 h_{\pi}^{-1} \|x - \bar{x}\|_{H_D^1}^2 \|z\|_{H_D^1}^2$$

which proves the assertion.

Remark 3.2. According to Propsition 3.1, the Gâteaux-derivative F' is continuous on each X_{π} . Hence, it is Fréchet-differentiable there. However, F is in general not Fréchet-differentiable on H_D^1 unless it has a very special structure. A discussion of related question can be found in [1, Section 1.2].

Corollary 3.3. Let the partial derivatives f_y and f_x satisfy the Lipschitz condition in Proposition 3.1(ii) locally. Let $\tilde{x} \in \text{dom } F$ be a sufficiently smooth function, possibly not belonging to X_{π} , $\|\tilde{x} - x_*\| \le \rho/2$. Then it holds, for all $\tau \in [0, 1]$,

$$||F'(x) - F'(x + (1 - \tau)(\tilde{x} - x))|| \le L_F h_{\pi}^{-1/2} (1 - \tau) ||\tilde{x} - x||_{H_D^1} + \hat{L} h_{\pi}^{N - 1/2}, \ x \in X_{\pi} \cap \text{dom } F, \ \pi \in \mathcal{M}_{[r]}, \ (29)$$

with a constants \hat{L} . In particular, for $\tau = 0$, we obtain

$$||F'(x) - F'(\tilde{x})|| \le L_F h_{\pi}^{-1/2} ||\tilde{x} - x||_{H_D^1} + \hat{L} h_{\pi}^{N-1/2}, \quad x \in X_{\pi} \cap \text{dom } F, \ \pi \in \mathcal{M}_{[r]}$$

and

$$\|\mathcal{F}'(x) - \mathcal{F}'(\tilde{x})\| \le Lh_{\pi}^{-1/2}\|\tilde{x} - x\|_{H_D^1} + \hat{L}h_{\pi}^{N-1/2}, \quad x \in X_{\pi} \cap \text{dom } F, \ \pi \in \mathcal{M}_{[r]}.$$

Proof. Let $I_{\pi}: H_D^1 \cap C_{\pi}([a,b], \mathbb{R}^m) \to X_{\pi}$ be a piecewise polynomial interpolation operator. In order to be specific, consider node sequences

$$0 = \sigma_0^d < \sigma_1^d < \dots < \sigma_N^d = 1, 0 < \sigma_1^a < \sigma_2^a < \dots < \sigma_N^a < 1,$$

and define I_{π} componentwise. For a component $x_{\kappa} \in C[a,b], 1 \leq \kappa \leq k, I_{\pi,\kappa}x_{\kappa}$ is the piecewise polynomial interpolation using the nodes $\bar{t}_{ji} = t_{j-1} + \sigma_i^d h_j, i = 0, \dots, N, j = 1, \dots, n$. Analogously, for $x_{\kappa} \in C_{\pi}[a,b], I_{\pi,\kappa}, k < \kappa \leq m$ is the piecewise polynomial iterpolation using the nodes $\bar{t}_{ji} = t_{j-1} + \sigma_i^a h_j, i = 1, \dots, N, j = 1, \dots, n$. Then we set $I_{\pi} = [I_{\pi,1}, \dots, I_{\pi,m}]^T$.

Let $R_{\pi} = \tilde{x} - I_{\pi}\tilde{x}$ be the remainder. Standard interpolation results provide the estimate

$$||(DR_{\pi})'||_{\infty} \leq Ch_{\pi}^{N}, \quad ||R_{\pi}||_{\infty} \leq Ch_{\pi}^{N}, \quad ||R_{\pi}||_{H_{D}^{1}} \leq (2(b-a))^{1/2}C \ h_{\pi}^{N}.$$

For all sufficiently fine partitions $\pi \in \mathcal{M}_{[r]}$, $I_{\pi}\tilde{x}$ belongs also to dom F. Since I_{π} is the identity on X_{π} , we have, for each $x \in X_{\pi} \cap \text{dom } F$,

$$x + (1 - \tau)(\tilde{x} - x) - I_{\pi}(x + (1 - \tau)(\tilde{x} - x)) = x + (1 - \tau)(\tilde{x} - x) - (x + (1 - \tau)(I_{\pi}\tilde{x} - x))$$
$$= (1 - \tau)R_{\pi}.$$

Following the lines of the proof of Proposition 3.1(ii) we arrive at the estimate

$$\begin{split} \|(F'(I_{\pi}(x+(1-\tau)(\tilde{x}-x)))-F'(x+(1-\tau)(\tilde{x}-x)))z\|^2 &\leq \tilde{L}^2\big(\|(DR_{\pi})'\|_{\infty}^2+\|R_{\pi}\|_{\infty}^2\big)\|z\|_{H_D^1}^2 \\ &\leq 2\tilde{L}^2C^2h_{\pi}^{2N}\|z\|_{H_D^1}^2, \end{split}$$

hence

$$||F'(I_{\pi}(x+(1-\tau)(\tilde{x}-x)))-F'(x+(1-\tau)(\tilde{x}-x))|| \leq \tilde{\tilde{L}}h_{\pi}^{N}.$$

Then we obtain

$$\begin{split} \|F'(x) - F'(x + (1 - \tau)(\tilde{x} - x))\| &\leq \|F'(x) - F'(I_{\pi}(x + (1 - \tau)(\tilde{x} - x)))\| \\ &+ \|F'(I_{\pi}(x + (1 - \tau)(\tilde{x} - x))) - F'(x + (1 - \tau)(\tilde{x} - x))\| \\ &\leq L_{F}h_{\pi}^{-1/2}(1 - \tau)\|I_{\pi}\tilde{x} - x\|_{H_{D}^{1}} + \tilde{L}h_{\pi}^{N} \\ &\leq L_{F}h_{\pi}^{-1/2}(1 - \tau)(\|\tilde{x} - x\|_{H_{D}^{1}} + Ch_{\pi}^{N}) + \tilde{L}h_{\pi}^{N}. \end{split}$$

This proofs the assertion.

4. Properties related to individual sufficiently fine partitions π

This section is to provide an approximation of the solution x_* by means of an iteration residing in X_{π} for an arbitrary sufficiently fine individual partition $\pi \in \mathcal{M}_{[r]}$.

The space of ansatz functions X_{π} is defined by (9) as before. Below we frequently apply the topological decompositions

$$H_D^1 = X_\pi \oplus X_\pi^\perp, \quad L^2 = F'(x)X_\pi \oplus (F'(x)X_\pi)^\perp, \quad L^2 \times \mathbb{R}^l = \mathcal{F}'(x)X_\pi \oplus (\mathcal{F}'(x)X_\pi)^\perp. \tag{30}$$

and the associated orthoprojectors

$$U_{\pi}: H_D^1 \to H_D^1, \quad V_{\pi}(x): L^2 \to L^2, \quad \mathcal{V}_{\pi}(x): L^2 \times \mathbb{R}^l \to L^2 \times \mathbb{R}^l,$$
 (31)

$$\operatorname{im} U_{\pi} = X_{\pi}, \quad \operatorname{im} V_{\pi}(x) = F'(x)X_{\pi}, \quad \operatorname{im} \mathcal{V}_{\pi}(x) = \mathcal{F}'(x)X_{\pi}, \tag{32}$$

in which $x \in \text{dom } F$. $F'(x_*)$ is a fine DAO with index μ and $\mathcal{F}'(x_*)$ is injective, but its inverse is unbounded if $\mu > 1$.

Lemma 4.1. Let x_* be sufficiently smooth. Let $N > \mu - 1$. Choose $s \in \mathbb{R}$ with $s > \mu - 1/2 > 0$ and $\rho_{\pi} := c_{\rho}h_{\pi}^s$, with a constant $c_{\rho} > 0$. Then there is a constant $c_{\gamma} > 0$, such that the following relations become valid:

$$\| (\mathcal{F}'(x_*)U_{\pi})^+ \| \le \Gamma_{\pi} := \frac{1}{c_{\gamma}} h_{\pi}^{1-\mu},$$
 (33)

$$\ker \mathcal{F}'(x)U_{\pi} = \ker U_{\pi}, \quad x \in \bar{\mathfrak{B}}(x_{*}, \rho_{\pi}) \cap X_{\pi}, \tag{34}$$

$$(\mathcal{F}'(x)U_{\pi})^{+} = (\mathcal{V}_{\pi}(x_{*})\mathcal{F}'(x)U_{\pi})^{+}\mathcal{V}_{\pi}(x_{*})\mathcal{V}_{\pi}(x), \quad x \in \bar{\mathfrak{D}}(x_{*}, \rho_{\pi}) \cap X_{\pi}, \tag{35}$$

$$\| (\mathcal{F}'(x)U_{\pi})^{+} \| \le \| (\mathcal{V}_{\pi}(x_{*})\mathcal{F}'(x)U_{\pi})^{+} \| \le 2\Gamma_{\pi}, \quad x \in \bar{\mathfrak{B}}(x_{*}, \rho_{\pi}) \cap X_{\pi}, \tag{36}$$

for each arbitrary mesh $\pi \in \mathcal{M}_{[r]}$ with sufficiently small h_{π} .

Proof. The existence of $c_{\gamma} > 0$ as well as the inequality (33) are ensured by [7, Theorem 4.1] concerning the instability threshold. c_{γ} may depend on the ratio r. The injectivity of $\mathcal{F}'(x_*)$ immediately implies $\ker \mathcal{F}'(x_*)U_{\pi} = \ker U_{\pi}$.

For $x \in \overline{\mathfrak{B}}(x_*, \rho_\pi) \cap X_\pi$, $\rho_\pi < \rho$, we have

$$\ker U_{\pi} \subseteq \mathcal{F}'(x)U_{\pi} \subseteq \mathcal{V}(x_*)\mathcal{F}'(x)U_{\pi} \tag{37}$$

and

$$\mathcal{V}(x_*)\mathcal{F}'(x)U_{\pi} = \underbrace{\mathcal{F}'(x_*)U_{\pi}}_{=:\mathfrak{I}} + \underbrace{\mathcal{V}(x_*)(\mathcal{F}'(x) - \mathcal{F}'(x_*))U_{\pi}}_{=:\mathfrak{I}} = \mathfrak{U} + \mathfrak{E}..$$

Making the stepsize h_{π} small enough and regarding Corollary 3.3, (22), and (33) yields

$$\|\mathfrak{A}^{+}\|\|\mathfrak{E}\| \leq \Gamma_{\pi}(Lh_{\pi}^{-1/2}\rho_{\pi} + \hat{L}h_{\pi}^{N-1/2}) \leq \frac{1}{c_{\gamma}} \left(c_{\rho}Lh_{\pi}^{s-\mu+1/2} + \hat{L}h_{\pi}^{N-\mu+1/2}\right) \leq \frac{1}{2}.$$
 (38)

Applying Lemma A.2 of the appendix it results that

$$\dim \ker \mathcal{V}(x_*)\mathcal{F}'(x)U_{\pi} = \dim \ker \mathcal{F}'(x_*)U_{\pi}, \quad \text{thus} \quad \ker \mathcal{V}(x_*)\mathcal{F}'(x)U_{\pi} = \ker U_{\pi},$$

and further

$$\|(\mathcal{V}(x_*)\mathcal{F}'(x)U_{\pi})^+\| \leq \frac{\|\mathfrak{A}^+\|}{1-\|\mathfrak{A}^+\|\|\mathfrak{E}\|} \leq 2\Gamma_{\pi}.$$

Taking into account (37) we have

$$\ker U_{\pi} = \ker \mathcal{F}'(x)U_{\pi} = \ker \mathcal{V}(x_*)\mathcal{F}'(x)U_{\pi},$$

and, in particular, (34). It also follows that

$$U_{\pi} = (\mathcal{F}(x)U_{\pi})^{+}\mathcal{F}(x)U_{\pi}, \quad U_{\pi} = (\mathcal{V}_{\pi}(x_{*})\mathcal{F}(x)U_{\pi})^{+}\mathcal{V}_{\pi}(x_{*})\mathcal{F}(x)U_{\pi}.$$

Multiplying the last identity from the right by $(\mathcal{F}(x)U_{\pi})^{+}$ yields

$$(\mathcal{F}(x)U_{\pi})^{+} = (\mathcal{V}_{\pi}(x_{*})\mathcal{F}(x)U_{\pi})^{+} \mathcal{V}_{\pi}(x_{*})\mathcal{V}_{\pi}(x),$$

that means (35), and (36) follows immediately.

It should be noted that s in the previous lemma is not restricted to be an integer.

As previously agreed upon, there exists x_* such that $\mathcal{F}x_* = 0$, thus $\psi(x_*) = 0$, $F'(x_*)$ is a fine DAO, the varionational problem $\mathcal{F}'(x_*)z = 0$ features accurately stated boundary condition, and the composed operator $\mathcal{F}'(x_*)$ is injective. Assuming the solution x_* to be smooth enough we apply the estimates (cf. [7])

$$\alpha_{\pi} := ||U_{\pi} x_* - x_*|| \le c_{\alpha} h_{\pi}^N, \tag{39}$$

in which N is again the polynomial degree used for the ansatz space X_{π} .

Since the inverse $\mathcal{F}'(x_*)^{-1}$ is unbounded, standard Newton-like iterations cannot be expected to work well here. Instead we apply a kind of projected Newton iteration using the bounded Moore-Penrose inverse³ $(\mathcal{F}'(x)U_{\pi})^+$ against the background of Lemma 4.1.

More precisely, supposing that h_{π} is small enough, we take an initial guess $x_0 \in \bar{\mathfrak{B}}(x_*, \rho_{\pi}) \cap X_{\pi}$ and provide the correction z_1 by means of the least-squares problem

$$z_1 = \operatorname{argmin}\{\|\mathcal{F}'(x_0)z + \mathcal{F}x_0\|^2: z \in X_{\pi}\} = -(\mathcal{F}'(x_0)U_{\pi})^+ \mathcal{F}x_0, \tag{40}$$

and then put $x_1 = x_0 + z_1$, and so on. By construction, z_1 is well defined and belongs to X_{π} , and so does the new iteration x_1 . Notice that $z_1 = U_{\pi}z_1$ serves as descent direction of the functional ψ at x_0 , as long as $\mathcal{V}_{\pi}(x_0)\mathcal{F}(x_0) \neq 0$, because of

$$\psi'(x_0)z_1 = 2(\mathcal{F}'(x_0)z_1, \mathcal{F}x_0) = 2(\mathcal{F}'(x_0)U_{\pi}z_1, \mathcal{F}x_0) = -2(\mathcal{F}'(x_0)U_{\pi}(\mathcal{F}'(x_0)U_{\pi})^+\mathcal{F}x_0, \mathcal{F}x_0)$$
$$= -2(\mathcal{V}(x_0)\mathcal{F}x_0, \mathcal{F}x_0) = -2||\mathcal{V}(x_0)\mathcal{F}x_0||^2.$$

Next we ask if x_1 belongs to the ball $\bar{\mathfrak{B}}(x_*, \rho_\pi)$. For this aim we derive

$$x_{1} - x_{*} = x_{0} - x_{*} - (\mathcal{F}'(x_{0})U_{\pi})^{+} (\mathcal{F}x_{0} - \mathcal{F}x_{*})$$

$$= U_{\pi}(x_{0} - x_{*}) - (I - U_{\pi})x_{*} - (\mathcal{F}'(x_{0})U_{\pi})^{+} \int_{0}^{1} \mathcal{F}'(\tau x_{0} + (1 - \tau)x_{*})d\tau (x_{0} - x_{*})$$

$$= \mathfrak{B} - \mathfrak{D}.$$

Then

$$\mathfrak{B} = U_{\pi}(x_{0} - x_{*}) - (\mathcal{F}'(x_{0})U_{\pi})^{+} \int_{0}^{1} \mathcal{F}'(\tau x_{0} + (1 - \tau)x_{*}) d\tau U_{\pi}(x_{0} - x_{*})$$

$$= (\mathcal{F}'(x_{0})U_{\pi})^{+} \mathcal{F}'(x_{0})U_{\pi}(x_{0} - x_{*}) - (\mathcal{F}'(x_{0})U_{\pi})^{+} \int_{0}^{1} \mathcal{F}'(\tau x_{0} + (1 - \tau)x_{*}) d\tau U_{\pi}(x_{0} - x_{*})$$

$$= (\mathcal{F}'(x_{0})U_{\pi})^{+} \int_{0}^{1} (\mathcal{F}'(x_{0}) - \mathcal{F}'(\tau x_{0} + (1 - \tau)x_{*})) d\tau U_{\pi}(x_{0} - x_{*}),$$

hence, applying Corollary 3.3 for $\tilde{x} = \tau x_0 + (1 - \tau)x_*$, $||\tilde{x} - x_*|| \le \rho_\pi \le \frac{1}{2}\rho$, and supposing N > s,

$$\begin{split} \|\mathfrak{B}\| &\leq 2\Gamma_{\pi} \left(\frac{1}{2} L h_{\pi}^{-1/2} \rho_{\pi} + \hat{L} h^{N-1/2} \right) \|x_{0} - x_{*}\| \leq \frac{1}{c_{\gamma}} \left(L c_{\rho} h_{\pi}^{s-\mu+1/2} \rho_{\pi} + 2 \hat{L} h^{N-\mu+1/2} \right) \|x_{0} - x_{*}\| \\ &\leq \frac{1}{2} \|x_{0} - x_{*}\|, \end{split}$$

for sufficiently small h_{π} , cf. (38). Next, for

$$\mathfrak{D} = (I - U_{\pi})x_* + (\mathcal{F}'(x_0)U_{\pi})^+ \int_0^1 \mathcal{F}'(\tau x_0 + (1 - \tau)x_*)d\tau (I - U_{\pi})(-x_*)$$

$$= \left\{ I - (\mathcal{F}'(x_0)U_{\pi})^+ \int_0^1 \mathcal{F}'(\tau x_0 + (1 - \tau)x_*)d\tau \right\} (I - U_{\pi})x_*$$

³Note that $(\mathcal{F}'(x)U_{\pi})^+$ is a bounded outer inverse of $\mathcal{F}'(x_*)$.

we obtain a constant c_* such that

$$\|\mathfrak{D}\| \leq (1 + \Gamma_{\pi} C_{\mathcal{F}}) \|x_* - U_{\pi} x_*\| \leq c_{\alpha} \left(\frac{2}{c_{\gamma}} C_{\mathcal{F}} + h_{\pi}^{\mu-1}\right) h_{\pi}^{N-\mu+1} \leq c_* h_{\pi}^{N-\mu+1}.$$

Now, to ensure that x_1 belongs to the ball $\bar{\mathfrak{B}}(x_*, \rho_{\pi})$, we are confronted with the requirement

$$\|\mathfrak{D}\| \le c_* h_\pi^{N-\mu+1} \le \frac{1}{2} c_\rho h_\pi^s,$$

which becomes valid by choosing N so that

$$N - \mu + 1 > s,\tag{41}$$

for all sufficiently fine meshes $\pi \in \mathcal{M}_{[r]}$. Then we continue the iterations by providing

$$x_{k+1} = x_k + z_{k+1}, (42)$$

$$z_{k+1} = \operatorname{argmin}\{\|\mathcal{F}'(x_k)z + \mathcal{F}x_k\|^2 : z \in X_{\pi}\} = -(\mathcal{F}'(x_k)U_{\pi})^+ \mathcal{F}x_k, \quad k \ge 0.$$
 (43)

The sequence $\{x_k\}$ remains in $\bar{\mathfrak{B}}(x_*, \rho_{\pi})$. Furthermore we have

$$||x_{k+1} - x_*|| \le \frac{1}{2} ||x_k - x_*|| + c_* h_\pi^{N-\mu+1} \le \dots \le \frac{1}{2^{k+1}} ||x_0 - x_*|| + \sum_{i=0}^k \frac{1}{2^i} c_* h_\pi^{N-\mu+1}$$

$$\le \left(\frac{1}{2}\right)^{k+1} ||x_0 - x_*|| + 2c_* h_\pi^{N-\mu+1} \le \left(\frac{1}{2}\right)^{k+1} c_\rho h_\pi^s + 2c_* h_\pi^{N-\mu+1}, \quad k \ge 0.$$

There is a number $k_{\pi} \in \mathbb{N}$ so that one has $\left(\frac{1}{2}\right)^{k+1} \leq \frac{c_*}{c_{\rho}} h_{\pi}^{N-\mu+1-s}$ for all $k \geq k_{\pi}$, and hence

$$||x_{k+1} - x_*|| \le 3c_* h_\pi^{N-\mu+1}, \quad k \ge k_\pi.$$
 (44)

We summarize what we get:

Theorem 4.2. Let $\mathcal{F}x = 0$ denote the operator formulation from Section 3 associated with the BVP (1),(2), $\mathcal{F}x_* = 0$, $\ker \mathcal{F}'(x_*) = \{0\}$, and x_* be sufficiently smooth for (39) to hold. Let the radius ρ_{π} and the bound Γ_{π} be as introduced in Lemma 4.1, and

$$N - \mu + 1 > s > \mu - 1/2, \tag{45}$$

and the mesh $\pi \in \mathcal{M}_{[r]}$ be sufficiently fine. Then the iteration (42) starting from $x_0 \in \bar{\mathfrak{B}}(x_*, \rho_\pi) \cap X_\pi$ remains therein and there is a number $k_\pi \in \mathbb{N}$ such that the estimate (44) is valid and

$$\psi(x_{k+1}) \le (3c_*C_{\mathcal{F}})^2 h_{\pi}^{2(N-\mu+1)}, \quad k \ge k_{\pi}. \tag{46}$$

Proof. It only remains to verify (46) which is a simple consequence of (22) and (44):

$$\psi(x_{k+1}) = \|\mathcal{F}x_{k+1}\|^2 = \|\mathcal{F}x_{k+1} - \mathcal{F}x_*\|^2 \le C_{\mathcal{F}}^2 \|x_{k+1} - x_*\|^2 \le C_{\mathcal{F}}^2 (3c_*h_\pi^{N-\mu+1})^2, \quad k \ge k_\pi.$$

Let us emphasize that the constants c_{γ} , c_{ρ} , c_{α} , c_{*} , and M_{*} are global bounds for all partitions $\pi \in \mathcal{M}_{[r]}$.

Table 1: Errror in $L^2(0, 1)$ for N = 3 for the pendulum example. n equidistant grid points and M = N + 1 uniformly distributed collocation points have been used

\overline{n}	х	x'	У	<i>y</i> ′	λ
10	4.42e-02	1.17e-01	1.83e-02	1.01e-01	6.25e-01
20	6.01e-03	1.76e-02	2.48e - 03	1.98e-02	3.33e-01
40	8.28e-04	3.07e-03	3.41e-04	4.47e - 03	1.72e-01
80	1.11e-04	6.26e - 04	4.59e - 05	1.07e-03	8.67e - 02
160	1.42e-05	1.44e - 04	5.87e - 06	2.64e-04	4.34e-02
320	1.86e-06	3.50e - 05	7.65e - 07	6.58e - 05	2.17e - 02
640	2.32e-07	8.68e - 06	9.57e - 08	1.64e - 05	1.08e-02

5. Numerical experiments

In this section, we present the results of some experiments in order to illustrate the properties of the proposed method.

The nonlinear least-squares method (25) has been implemented in Matlab. Instead of (25), its approximation $\psi_{\pi,M}$ of (13) has been used. The finite-dimensional problems have been solved using a Matlab implementation of a Gauss-Newton method following the lines of [4, Section 4.3]. The iteration has been stopped if no further improvement in $\psi_{\pi,M}(x_k)$ could be observed. For the purposes of investigating the convergence of the method, an interpolation of the exact solution has been used as an initial guess.

5.1. The mathematical pendulum

This problem has been used in many publications for demonstrating properties of algorithms for the solution of differential algebraic systems. We use the formulation

$$x'' = -x\lambda,$$

$$y'' = -y\lambda - g,$$

$$0 = x^2 + y^2 - L^2.$$

The underlying interval is (0, 1). The parameters are chosen to be g = 16, $L = \sqrt{8}$. We consider the initial values y(0) = 2 and y'(0) = 0. This problem has index 3. Therefore, the results of Theorem 4.2 are only valid if $N \ge 5$. For N = 5, $s = \mu - 1/4$ can be chosen. However, the expected orders are observed in all cases $N \ge 2$. The case N = 1 is rather surprising since we observed bounded solutions instead of diverging ones.

In Tables 1 and 2 as well as Tables 3 and 4 results for N=3 and N=5, respectively, are presented. In both cases, uniform grids and M=N+1 uniformly distributed collocation points per subinterval have been used.

Table 2: Order estimate for N=3 for the pendulum example. n equidistant grid points and M=N+1 uniformly distributed collocation points have been used

n	Х	x'	у	y'	λ
10	2.9	2.7	2.9	2.3	0.9
20	2.9	2.5	2.9	2.1	1.0
40	2.9	2.3	2.9	2.1	1.0
80	3.0	2.1	3.0	2.0	1.0
160	2.9	2.0	2.9	2.0	1.0
320	3.0	2.0	3.0	2.0	1.0

Table 3: Errror in $L^2(0, 1)$ for N = 5 for the pendulum example. n equidistant grid points and M = N + 1 uniformly distributed collocation points have been used

n	Х	χ'	у	y'	λ
10	4.13e-04	1.28e-03	1.75e-04	1.99e-03	3.61 <i>e</i> -02
20	4.59e - 05	1.22e-04	1.88e - 05	1.38e-04	4.90e-03
40	1.45e - 06	4.44e - 06	5.94e-07	6.94e - 06	6.18e-04
80	3.43e - 08	1.82e-07	1.41e-08	3.97e-07	7.74e - 05
160	1.02e-09	1.04e-08	4.17e - 10	2.45e - 08	9.68e - 06
320	5.57e - 11	$6.41e{-10}$	2.28e-11	1.52e-09	1.21 <i>e</i> -06

Table 4: Order estimate for N = 5 for the pendulum example. n equidistant grid points and M = N + 1 uniformly distributed collocation points

n	Х	x'	у	y'	λ
10	3.2	3.4	3.2	3.8	2.9
20	5.0	4.8	5.0	4.3	3.0
40	5.4	4.6	5.4	4.1	3.0
80	5.1	4.1	5.1	4.0	3.0
160	4.2	4.0	4.2	4.0	3.0

5.2. An example proposed by S.L. Campbell and E. Moore

In [2], the following system is used as an example:

$$x'_{1} - x_{4} = 0,$$

$$x'_{2} - x_{5} = 0,$$

$$x'_{3} - x_{6} = 0,$$

$$x'_{4} - x_{6} \cos t + x_{3} \sin t + x_{5} - 2x_{1}(1 - r(x_{1}^{2} + x_{2}^{2})^{-\frac{1}{2}})x_{7} = 0,$$

$$x'_{5} - x_{6} \sin t - x_{3} \cos t - x_{4} - 2x_{2}(1 - r(x_{1}^{2} + x_{2}^{2})^{-\frac{1}{2}})x_{7} = 0,$$

$$x'_{6} + x_{3} - 2x_{3}x_{7} = 0,$$

$$x'_{1} + x'_{2} + x'_{3}^{2} - 2r(x'_{1}^{2} + x'_{2}^{2})^{\frac{1}{2}} + r^{2} - \rho^{2} = 0.$$

The solution considered in the reference is

$$x_{*1} = (\rho \cos(2\pi - t) + r)\cos t = (\rho \cos t + r)\cos t,$$

$$x_{*2} = (\rho \cos(2\pi - t) + r)\sin t = (\rho \cos t + r)\sin t,$$

$$x_{*3} = \rho \sin(2\pi - t) = -\rho \sin t,$$

yielding

$$x_{*4} = -(\rho \cos(2\pi - t) + r) \sin t + \rho \sin(2\pi - t) \cos t,$$

$$x_{*5} = (\rho \cos(2\pi - t) + r) \cos t + \rho \sin(2\pi - t) \sin t,$$

$$x_{*6} = -\rho \cos(2\pi - t),$$

$$x_{*7} = 0.$$

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. Under these conditions, the problem has index 3.

A thorough discussion as well as numerical experiments of the version linearized in the solution x_* is given in [7]. In order to stimulate discussions of the least-squares method for nonlinear problems, also results for the original nonlinear version have been provided in this reference. We cite the results in Tables 5 and 6. Theorem 4.2 is only valid for $N \ge 5$ in this example and, thus, the corresponding order is strictly proven. However, the expected orders are observed in allowed cases $N \ge 2$. The case N = 1 is rather surprising besause we observe bounded solutions even if we expecteddiverging ones.

6. Multilevel approach

We use N and s as previously agreed, that is $N - \mu + 1 > s > \mu - 1/2 > 0$. Given an additional constant q with 0 < q < 1 we now deal with a sequence of partitions $\pi_i \in \mathcal{M}_{[r]}$,

$$\pi_i: a = t_0^{[\pi_i]} < t_1^{[\pi_i]} < \dots < t_{n_\pi}^{[\pi_i]} = b,$$
 with maximal stepsize $qh_{\pi_i} = h_{\pi_{i+1}}, \ i \ge 0,$

Table 5: Errors in $H_D^1(0,5)$ for (5.2) using M=N+1 for the Campbell-Moore example. n equidistant grid points and M Gaussian collocation points have been used

\overline{n}	N = 1	N = 2	N = 3	N = 4	<i>N</i> = 5
10	3.32e+1	4.53e+0	3.82e-1	7.02e-2	1.47e-3
20	3.32e+1	7.51e-1	1.02e-1	1.26e-2	1.24e-4
40	3.32e+1	3.03e-1	3.14e-2	2.52e-3	1.30e-5
80	3.32e+1	1.80e-1	1.22e-2	5.45e-4	1.54e-6
160	3.32e+1	1.17e-1	5.67e-3	1.25e-4	1.20e-6
320	3.32e+1	7.95e-2	2.73e-3	1.25e-4	1.20e-6

Table 6: Order estimation for (5.2) using M = N + 1 for the Campbell-Moore example. n equidistant grid points and M Gaussian collocation points have been used. The row "theory" contains the expected orders. Note that Theorem 4.2 is only valid for $N \ge 5$

n	N = 1	N = 2	N = 3	N = 4	N = 5
20	0.0	2.6	1.9	2.5	3.6
40	0.0	1.3	1.7	2.3	3.3
80	0.0	0.7	1.4	2.2	3.1
160	0.0	0.6	1.1	2.1	0.6
320	0.0	0.6	1.1	0.0	0.0
theory		(0)	(1)	(2)	3

such that the associated ansatz spaces are nested,

$$X_{\pi_0} \subset X_{\pi_1} \subset \cdots \subset X_{\pi_i} \subset X_{\pi_{i+1}} \subset \cdots$$

and $h_{\pi_i} \to 0$ if $i \to \infty$. Let π_0 be fine enough for Lemma 4.1 and Theorem 4.2 to hold. This means that

$$\Gamma_{\pi_0}(\rho_{\pi_0}h_{\pi_0}^{-1/2}L + 2\hat{L}h_{\pi_0}^{N-1/2}) \le \frac{1}{2}, \quad \text{and} \quad h_{\pi_0}^{N-\mu+1-s} \le \frac{1}{2}\frac{c_\rho}{c_s},$$

to ensure the applicability of Lemma 4.1 and to make the iterations on the level π_0 to stay in $\bar{\mathfrak{B}}(x_*, \rho_{\pi_0}) \cap X_{\pi_0}$. Both conditions are satisfied correspondingly a fortiori on the further levels due to the smaller stepsizes h_{π_i} . In the consequence, Theorem 4.2 applies on each level, i.e., for $x_0^{[\pi_i]} \in \bar{\mathfrak{B}}(x_*, \rho_{\pi_i}) \cap X_{\pi}$ the sequence

$$x_{k+1}^{[\pi_i]} = x_k^{[\pi_i]} + z_{k+1}^{[\pi_i]},\tag{47}$$

$$z_{k+1}^{[\pi_i]} = \operatorname{argmin}\{\|\mathcal{F}'(x_k^{[\pi_i]})z + \mathcal{F}x_k^{[\pi_i]}\|^2: z \in X_{\pi}\} = -(\mathcal{F}'(x_k^{[\pi_i]})U_{\pi})^+ \mathcal{F}x_k^{[\pi_i]}, \quad k \ge 0.$$
 (48)

remains in $\bar{\mathfrak{B}}(x_*, \rho_{\pi_i}) \cap X_{\pi_i}$ and there exists a number $k_{\pi_i} \in \mathbb{N}$ such that $\left(\frac{1}{2}\right)^{k+1} \leq \frac{c_*}{c_\rho} h_{\pi_i}^{N-\mu+1-s}$ for all $k \geq k_{\pi_i}$, and hence

$$||x_{k+1}^{[\pi_i]} - x_*|| \le 3c_* h_{\pi_i}^{N-\mu+1}, \quad k \ge k_{\pi_i}.$$

$$\tag{49}$$

Since the ansatz spaces are nested, $x_{k_{\pi_{i+1}}}^{[\pi_i]}$ belongs to $X_{\pi_{i+1}}$. Replacing the condition $h_{\pi_i}^{N-\mu+1-s} \leq \frac{1}{2} \frac{c_{\rho}}{c_*}$ by the stronger one

$$h_{\pi_i}^{N-\mu+1-s} \le \frac{1}{3} q^s \frac{c_\rho}{c_s} \tag{50}$$

yields

$$3c_*h_{\pi_i}^{N-\mu+1} \leq 3c_*\frac{1}{3}q^s\frac{c_\rho}{c_*}h_{\pi_i}^s = c_\rho q^sh_{\pi_i}^s = c_\rho h_{\pi_{i+1}}^s = \rho_{\pi_{i+1}}.$$

Then $x_{k_{\pi_i}+1}^{[\pi_i]}$ belongs to $\bar{\mathfrak{B}}(x_*, \rho_{\pi_{i+1}}) \cap X_{\pi_{i+1}}$ and we are allowed to choose at the next level

$$x_0^{[\pi_{i+1}]} := x_{k_{\pi_{i+1}}}^{[\pi_i]}. (51)$$

We summarize our result:

Theorem 6.1. Let $\mathcal{F}x = 0$ denote the operator formulation from Section 3 associated with the BVP (1),(2), $\mathcal{F}x_* = 0$, $\ker \mathcal{F}'(x_*) = \{0\}$, and x_* be sufficiently smooth for (39). Let (45) be given and 0 < q < 1.

Let the sequence of partitions $\pi_i \in \mathcal{M}_{[r]}$, $i \geq 0$, be such that the ansatz spaces are nested and the maximal stepsizes are related by $qh_{\pi_i} = h_{\pi_{i+1}}$. Let the the mesh π_0 be sufficiently fine,

$$\Gamma_{\pi_0}(\rho_{\pi_0}h_{\pi_0}^{-1/2}L + 2\hat{L}h_{\pi_0}^{N-1/2}) \le \frac{1}{2}, \quad and \quad h_{\pi_0}^{N-\mu+1-s} \le \frac{1}{3}q^s \frac{c_\rho}{c_*}.$$

Then the iteration (47),(48),(51), with the initial guess $x_0^{[\pi_0]} \in \bar{\mathfrak{B}}(x_*, \rho_{\pi_0}) \cap X_{\pi_0}$ is well defined and yields

$$||x_{k_{\pi_{i+1}}}^{[\pi_{i}]} - x_{*}|| \le 3c_{*}h_{\pi_{i}}^{N-\mu+1} = 3c_{*}h_{\pi_{0}}^{N-\mu+1}(q^{N-\mu+1})^{i+1} \to 0 \quad (i \to \infty).$$
(52)

7. Remarks and conclusions

We have presented and investigated a nonlinear least-squares method for approximating higher index differential-algebraic equations. The idea consists of discretizing the preimage space H_D^1 by piecewise polynomials and to form an overdetermined collocation system to determine an approximating solution. The resulting overdetermined system is solved in a least-squares sense. In the numerical experiments, the method behaved very well despite its simplicity. In particular, the method is not much more expensive than the standard collocation method applied to explicit ordinary differential equations and index-1 differential-algebraic equations.

The main tool both for the convergence proof and for the numerical solution of the discretized problems is a variant of the Newton method. For a large class of nonlinear index- μ tractable equations, this method applied to the discretized system is shown to deliver appropriate approximations provided that the polynomial order is large enough. The numerical experiments indicate, however, that the strong condition on the polynomial order does not seem to be necessary. In particular, the order of convergence corresponds to that of linear index- μ tractable differential-algebraic equations. So the present result should be considered as a first step towards a theoretical foundation of the method.

Remark 7.1. (i) Under the conditions of Theorem 4.2 we could not show that the sequence $\{x_k\}$ converges.

- (ii) If there is a minimizer $x_{\pi,*}$ of (14) in $\mathfrak{B}(x_*,\rho_\pi) \cap X_\pi$, then it holds $(\mathcal{F}'(x_{\pi,*})\mathcal{U}_\pi)^+\mathcal{F}(x_{\pi,*}) = 0$. Since $\bar{\mathfrak{B}}(x_*,\rho_\pi) \cap X_\pi$ is compact, the sequence $\{x_k\}$ has a convergent subsequence. However, we were not able to show that, for an accumulation point \hat{x}_π , it holds $(\mathcal{F}'(\hat{x}_\pi)\mathcal{U}_\pi)^+\mathcal{F}(\hat{x}_\pi) = 0$.
- (iii) In the context of regularization methods for nonlinear illposed problems, the so-called Scherzer, or tangential cone, condition is often used [16, 9, 10]. However, in the context of differential-algebraic equations, this conditions requires very hard conditions on the structure of the system. Therefore, it is of minor use here.

Appendix A. An auxillary result

The convergence proof for the Gauss-Newton method requires an estimation of the norm and distance of Moore-Penrose inverses of derivatives of a nonlinear operator. In the case of finite dimensional spaces, such results are well-known and can be found, for example, in [13]. However, we need similar statements in the case of infinite dimensional spaces. This appendix provides the necessary lemmas.

Let X and Y be Hilbert spaces (not necessarily finite dimensional) and $A: X \to Y$ a linear and compact operator. Both operators A^*A and AA^* are selfadjoint compact operators. Their spectra consist only of nonnegative eigenvalues with finite multiplicity (with the possible exception of $\lambda = 0$). If the eigenvalues have an accumulation point, then it is 0. The nonzero eigenvalues

are identical (even with respect to their multiplicity) for both A^*A and AA^* . Let the nonzero eigenvalues be sorted according to

$$||A^*A|| = \lambda_1 \ge \lambda_2 \ge \cdots > 0.$$

Let then $\{u_i\} \subset X$ and $\{v_i\} \subset Y$ be a complete orthonormal system of eigenvalues⁴ for the operators A^*A and AA^* ,

$$\lambda_i u_i = A^* A u_i, \quad \lambda_i v_i = A A^* v_i.$$

We set $\sigma_i = \sqrt{\lambda_i} > 0$. This provides us with

$$\sigma_i u_i = A^* v_i, \quad \sigma_i v_i = A u_i.$$

The system $\{\sigma_i, u_i, v_i\}$ is called a singular system of A with the singular values σ_i . In particular, we have the representations

$$Ax = \sum_{i} (x, u_i)v_i, \ x \in X,$$
$$A^*y = \sum_{i} (y, v_i)u_i, \ y \in Y.$$

Here, (\cdot, \cdot) denotes the scalar products in X and Y, respectively. Note that these sums can be both finite and infinite.

We are interested in perturbation results for the singular values of an operator A. The following lemma is proven in [5, Corollary VI.1.6].

Lemma 1. Let X and Y be Hilbert spaces. Let $A, B: X \to Y$ be compact linear operators. Let $\sigma_i(A)$, $i = 1, ..., \nu(A)$ and $\sigma_i(B)$, $i = 1, ..., \nu(B)$ be the singular values of A and B, respectively.⁵ Assume without loss of generality $\nu(A) \le \nu(B)$. Then it holds

$$|\sigma_i(A) - \sigma_i(B)| \le ||A - B||, \quad i = 1, \dots, \nu(A),$$

 $\sigma_i(B) \le ||A - B||, \quad i = \nu(A) + 1, \dots, \nu(B).$

The next step consists of the establishement of bounds for the Moore-Penrose pseudoinverse. For compact operators as above with the singular system $\{\sigma_i, u_i, v_i\}$ it has the representation

$$x = A^+ y = \sum_i \sigma_i^{-1}(y, v_i) u_i$$

for all $y \in \text{dom}(A^+)$. An immediate consequence is:

(i)
$$||A|| = \sigma_1$$
.

⁴The systems are not necessarily complete in X and Y, respectively!

⁵Both $\nu(A)$ and $\nu(B)$ may be finite or infinite.

(ii) A^+ is bounded if and only if the number of singular values is finite. In that case it holds $||A^+|| = \sigma_{\nu(A)}^{-1}$.

The following lemma presents modifications of [13, Theorem (8.15)].

Lemma 2. Let $A, B: X \to Y$ be compact linear operators acting in the Hilbert spaces X, Y.

(i) Assume dim $X < \infty$ and $\nu(B) \le \nu(A) = r$, $r \ge 1$. Moreover, let

$$||A^+||||B - A|| < 1$$

to hold and set $\epsilon = ||B - A||$. Then it holds $\nu(B) = \nu(A)$ and

$$||B^+|| \le \frac{||A^+||}{1 - ||A^+||||B - A||} = \frac{1}{\sigma_{\nu(A)} - \epsilon}$$

where $\sigma_{\nu(A)}$ is the smallest singular value of A.

(ii) Assume that X decomposes in $X = X_f \oplus X_f^{\perp}$, dim $X_f < \infty$, and

$$X_f^\perp = \ker A \subseteq \ker B, \quad \operatorname{im} B \subseteq \operatorname{im} A, \quad ||A^+||||B - A|| < 1.$$

Then it follows that

$$||B^+|| \le \frac{||A^+||}{1 - ||A^+||||B - A||}.$$

Proof. It holds $||A^+|| = \sigma_{\nu(A)}^{-1}$ such that, by assumption, $\sigma_i - \epsilon > 0$, $i = 1, ..., \nu(A)$. Hence, B = A + (B - A) has at least $\nu(A)$ nonvanishing singular values because of Lemma A.1. Hence, $\nu(B) \ge \nu(A)$. Together with the assumption, this provides $\nu(B) = \nu(A)$. Consequently, $||B^+|| \le 1/(\sigma_{\nu(A)} - \epsilon)$. (ii) is a consequence of (i).

References

- [1] A. Ambrosetti and G. Prodi. *A Primer of Nonlinear Analysis*, volume 34 of *Cambridge studies in advanced mathematics*. Cambridge University Press, 1995.
- [2] S. L. Campbell and E. Moore. Constraint preserving integrators for general nonlinear higher index DAEs. *Num.Math.*, 69:383–399, 1995.
- [3] P.G. Ciarlet. *The Finite Element Method for Elliptic Problems*, volume 40 of *Classics in Applied Mathematics*. SIAM, 2002.
- [4] P. Deuflhard. Newton Methods for Nonlinear Problems. Affine Invariance and Adaptive Algorithms, volume 35 of Springer Series in Computational Mathematics. Springer Verlag, 2004.

- [5] Israel Gohberg, Seymour Goldberg, and Marinus A. Kaashoek. *Classes of linear operators, Vol. I*, volume 49 of *Operator Theory: Advances and Applications*. Birkhäuser Verlag, Basel, Boston, Berlin, 1990.
- [6] M. Hanke and R. März. Questions concerning differential-algebraic operators: Towards a direct numerical treatment of differential-algebraic equations. Technical report, 2018.
- [7] M. Hanke, R. März, and C. Tischendorf. Least-squares collocation for higher-index linear differential-algebaic equations: Estimating the stability threshold. *Math. Comp.*, page To appear, 2018.
- [8] 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.*, 317:403–431, 2017. http://dx.doi.org/10.1016/j.cam.2016.12.017.
- [9] M. Hanke, A. Neubauer, and O. Scherzer. A convergence analysis for the Landweber iteration for nonlinear ill-posed problems. *Numer. Math.*, 72:21–37, 1995.
- [10] B. Kaltenbacher and J. Offtermatt. A convergence analysis of regularization by discretization in preimage space. *Math. Comp.*, 81(280):2049–2069, 2012.
- [11] 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. Ilchmann, T. Reis.
- [12] R. Lamour, R. März, and E. Weinmüller. *Surveys in Differential-Algebraic Equations III*, chapter Boundary-Value Problems for Differential-Algebraic Equations: A Survey, pages 177–309. Differential-Algebraic Equations Forum. Springer Heidelberg, 2015. ed. by A. Ilchmann and T. Reis.
- [13] C.L. Lawson and R.J. Hanson. *Solving Least Squares Problems*. Prentice Hall, Englewood Cliffs, NY, 1974.
- [14] R. März. *Surveys in Differential-Algebraic Equations II*, chapter Differential-Algebraic Equations from a Functional-Analytic Viewpoint: A Survey, pages 163–285. Differential-Algebraic Equations Forum. Springer Heidelberg, 2015. ed. by A. Ilchmann and T. Reis.
- [15] M.Z. Nashed and X. Chen. Convergence of Newton-like methods for singular operator equations using outer inverses. *Numer Math*, 66:235–257, 1993.
- [16] O. Scherzer. Convergence criteria for iterative methods based on Landweber iteration for solving nonlinear problems. *J. Math. Anal. Appl.*, 194:911–933, 1995.