Numerical Algorithms (2022) 91:1721–1754
https://doi.org/10.1007/s11075-022-01320-z

ORIGINAL PAPER

# On the sensitivity of implementations of a least-squares collocation method for linear higher-index differential-algebraic equations

**Michael Hanke**[1]

Received: 9 February 2022 / Accepted: 21 April 2022 / Published online: 6 June 2022


**Abstract**
The present paper continues our investigation of an implementation of a least-squares collocation method for higher-index differential-algebraic equations. In earlier papers, we were able to substantiate the choice of basis functions and collocation points for a robust implementation as well as algorithms for the solution of the discrete system. The present paper is devoted to an analytic estimation of condition numbers for different components of an implementation. We present error estimations, which show the sources for the different errors.

**Keywords** Least-squares collocation · Higher index differential-algebraic equations · Ill-posed problem

**Mathematics Subject Classification (2010)** 65L80 · 65L08 · 65F20

## 1 Introduction

In a series of papers [1–4], we were developing a new method for solving higher-index differential algebraic equations (DAEs). In naturally given functional analytic settings, higher index DAEs give rise to ill-posed problems [5, Section 3.9; 6]. Motivated by the well-known method of least-squares, or discretization on preimage space, for the approximation of ill-posed problems [7], this approach has been adapted to the case of higher-index DAEs. In particular, the ansatz spaces for the discrete least-squares problem have been chosen to be piecewise polynomials. Additionally, the integrals have been replaced by discrete versions based on simplified integration rules, in the most simple approach by a version resembling well-known

✉ Michael Hanke
hanke@kth.se

[1] Department of Mathematics, KTH Royal Institute of Technology, S-100 44, Stockholm, Sweden

collocation methods for solving boundary value problems for systems of ordinary differential equations (ODEs). The latter, extremely simplified version of the approach proposed in [7], has been motivated by the success of collocations methods for ODEs. This connection led us to coin the notion *least-squares collocation method* and calling the integration nodes also as *collocation points*.

For our method, a number of convergence results for both linear and nonlinear DAEs have been proven. Even our first attempts showed surprisingly accurate results when applying the method to some linear examples [1]. More recently, we investigated the algorithmic ingredients of the method in more detail [8, 9]. Not surprisingly, the basis representation and the choice of the integration nodes showed an important influence on the accuracy of the method.

The present note is intended to further quantify the conditioning of the individual ingredients of the implementation of the proposed method and to better understand the (high) accuracy of the computational results obtained so far. Taking the ill-posedness of higher-index DAEs into account, we expect very sensitive discrete problems for sufficiently fine discretizations.

The practical implementation of a projection method consists of two steps for a given approximation space $X_\pi$: Choice of a basis and formulation and solution of the arising discrete system by a suitable method. This in turn gives rise to two different operators, the first being the *representation map* connecting the elements of $x \in X_\pi$ with their vector of coefficients with respect to the chosen basis. The other operator is the discrete version of the the least-squares collocation method that becomes a linearly equality constrained linear least-squares problem in our case. Both operators are investigated in detail both analytically and numerically.

In particular, qualitative and quantitative estimations for the condition numbers and norms of the representation map are proven for bases whose usefulness in the present applications has been established earlier [8, 9].

For the constrained linear least-squares problem, a number of perturbation results are well-known, e.g., [10–12]. However, in the present application, the constraints play a special role: In the usual choices of the basis functions, some coefficient vectors do not represent a function in the approximation space. A coefficient vector represents a function in the approximation space if and only if the constraints are fulfilled. Therefore, a new error estimation is derived, which takes care of the exceptional role of the constraints. The important ingredients in this estimate are the condition number of the constraints and a restricted condition number for the least-squares functional. For the former, a complete analytical characterization for the chosen bases is provided. In a number of numerical examples, values for the restricted condition number are presented.

In Section 2, the least-squares method for approximating linear DAEs is introduced and the representation map is constructed. Section 3 is devoted to an in-depth investigation of the representation map. Then we derive a perturbation result for constrained linear least-squares problems in Section 4. Numerical examples for the condition numbers of the different ingredients are given in Section 5. Section 6 contains some conclusions.

## 2 The problem setting

### 2.1 The discrete functional

In this section, we repeat the problem setting from [8] for the reader's convenience. Consider a linear boundary-value problem for a DAE with properly involved derivative,

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

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

with $[a,b] \subset \mathbb{R}$ being a compact interval, $D = [I\ 0] \in \mathbb{R}^{k\times m}$, $k < m$, with the identity matrix $I \in \mathbb{R}^{k\times k}$. Furthermore, $A(t) \in \mathbb{R}^{m\times k}$, $B(t) \in \mathbb{R}^{m\times m}$, and $q(t) \in \mathbb{R}^m$ are assumed to be sufficiently smooth with respect to $t \in [a,b]$. Moreover, $G_a, G_b \in \mathbb{R}^{l_{dyn}\times m}$. Thereby, $l_{dyn}$ is the dynamical degree of freedom of the DAE, that is, the number of free parameters that can be fixed by initial and boundary conditions. We assume further that $\ker D \subseteq \ker G_a$ and $\ker D \subseteq \ker G_b$.

Unlike regular ODEs where $l_{dyn} = k = m$, for DAEs, it holds that $0 \le l_{dyn} \le k < m$, in particular, $l_{dyn} = k$ for index-one DAEs, $l_{dyn} < k$ for higher-index DAEs, and $l_{dyn} = 0$ can certainly happen.

The appropriate space for looking for solutions of (1)–(2) is (cf [2])

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

Let $\mathfrak{P}_K$ denote the set of all polynomials of degree less than or equal to $K \ge 0$. Given the partition $\pi$,

$$\pi : \quad a = t_0 < t_1 < \cdots < t_n = b, \tag{3}$$

with the stepsizes $h_j = t_j - t_{j-1}$, $h = \max_{1\le j\le n} h_j$, and $h_{\min} = \min_{1\le j\le n} h_j$. Let $C_\pi([a,b],\mathbb{R}^m)$ denote the space of piecewise continuous functions having breakpoints merely at the meshpoints of the partition $\pi$. Let $N \ge 1$ be a fixed integer. We are looking for an approximate solution of our boundary value problem from the ansatz space $X_\pi \subset H_D^1(a,b)$,

$$\begin{aligned} X_\pi = \{&x \in C_\pi([a,b],\mathbb{R}^m) : Dx \in C([a,b],\mathbb{R}^k),\\ &x_\kappa|_{[t_{j-1},t_j)} \in \mathfrak{P}_N,\ \kappa = 1,\dots,k,\\ &x_\kappa|_{[t_{j-1},t_j)} \in \mathfrak{P}_{N-1},\ \kappa = k+1,\dots,m,\ j = 1,\dots,n\}. \end{aligned} \tag{4}$$

The continuous version of the least-squares method reads: Find an $x_\pi \in X_\pi$ that minimizes the functional

$$\Phi(x) = \int_a^b |A(t)(Dx)'(t) + B(t)x(t) - q(t)|^2 dt + |G_a x(a) + G_b x(b) - d|^2. \tag{5}$$

Here and in the following, $|\cdot|$ denotes the Euclidean norm in the corresponding spaces $\mathbb{R}^\alpha$ for the appropriate $\alpha$. Let $\langle\cdot,\cdot\rangle$ denote the scalar product in $\mathbb{R}^\alpha$.

The functional values $\Phi(x)$, which are needed when minimizing for $x \in X_\pi$, cannot be evaluated exactly and the integral must be discretized accordingly. Taking into account that the boundary-value problem is ill-posed in the higher index case, perturbations of the functional may have a serious influence on the error of the

approximate least-squares solution or even prevent convergence towards the exact solution. Therefore, careful approximations of the integral in $\Phi$ are required. We take over the options provided in [8], in which $M \geq N+1$ so-called collocation points

$$0 \leq \rho_1 < \cdots < \rho_M \leq 1. \tag{6}$$

are used, and further, on the subintervals of the partition $\pi$,

$$t_{ji} = t_{j-1} + \rho_i h_j, \quad i = 1, \ldots, M, \ j = 1, \ldots, n.$$

Introducing, for each $x \in X_\pi$ and $w(t) = A(t)(Dx)'(t) + B(t)x(t) - q(t)$, the corresponding vector $W \in \mathbb{R}^{mMn}$ by

$$W = \begin{bmatrix} W_1 \\ \vdots \\ W_n \end{bmatrix} \in \mathbb{R}^{mMn}, \quad W_j = h_j^{1/2} \begin{bmatrix} w(t_{j1}) \\ \vdots \\ w(t_{jM}) \end{bmatrix} \in \mathbb{R}^{mM}, \tag{7}$$

we turn to an approximate functional of the form

$$\Phi_{\pi,M}(x) = W^T \mathcal{L} W + |G_a x(a) + G_b x(b) - d|^2, \quad x \in X_\pi, \tag{8}$$

with a positive definite symmetric matrix[1]

$$\mathcal{L} = \operatorname{diag}(L \otimes I_m, \ldots, L \otimes I_m). \tag{9}$$

As detailed in [8], we have different options for the positive definite symmetric matrix $L \in \mathbb{R}^{M \times M}$, namely

$$L = L^C = M^{-1} I_M, \tag{10}$$

$$L = L^I = \operatorname{diag}(\gamma_1, \ldots, \gamma_M), \tag{11}$$

$$L = L^R = (V^{-1})^T V^{-1}, \tag{12}$$

see [8, Section 3] for details concerning the selection of the quadrature weights $\gamma_1, \ldots, \gamma_M$ and the construction of the mass matrix $V$. We emphasize that the matrices $L^C$, $L^I$, $L^R$ depend only on $M$, the node sequence (6), and the quadrature weights, but do not depend on the partition $\pi$ and its stepsizes at all.

In the context of the numerical experiments below, we denote each of the different versions of the functional by $\Phi^C_{\pi,M}$, $\Phi^I_{\pi,M}$, and $\Phi^R_{\pi,M}$, respectively. The following convergence result is known [8, Theorem 2]:

**Theorem 1** *Let the DAE (1) be regular with index $\mu \in \mathbb{N}$ and let the boundary condition (2) be accurately stated. Let $x_*$ be a solution of the boundary value problem (1)–(2), and let $A$, $B$, $q$ and also $x_*$ be sufficiently smooth.*

*Let all partitions $\pi$ be such that $h/h_{\min} \leq \rho$, with a global constant $\rho$. Then, with*

$$M \geq N + \mu,$$

*the following statements are true:*

[1] $\otimes$ denotes the Kronecker product.

(1) *For sufficient fine partitions $\pi$ and each sequence of arbitrarily placed nodes (6), there exists exactly one $x_\pi^R \in X_\pi$ minimizing the functional $\Phi_{\pi,M}^R$ on $X_\pi$, and*

$$\|x_\pi^R - x_*\|_{H_D^1(a,b)} \le C_R h^{N-\mu+1}.$$

(2) *For each integration rule related to the interval $[0,1]$, with $M$ nodes (6) and positive weights $\gamma_1, \dots, \gamma_M$, that is exact for polynomials with degree less than or equal to $2M-2$, and sufficient fine partitions $\pi$, there exists exactly one $x_\pi^I \in X_\pi$ minimizing the functional $\Phi_{\pi,M}^I$ on $X_\pi$, and $x_\pi^I = x_\pi^R$, thus*

$$\|x_\pi^I - x_*\|_{H_D^1(a,b)} \le C_R h^{N-\mu+1}.$$

A corresponding result for $\Phi_{\pi,M}^C$ is not known. Numerical tests showed excellent convergence results even for cases not covered by Theorem 1. This holds in particular for any $M \ge N+1$ tested in all three cases of the functional $\Phi_{\pi,M}$. Thus, $M = N+1$ seems to be the preferable choice.

### 2.2 A basis representation of $\Phi_{\pi,M}$

By choosing an appropriate basis for $X_\pi$, the minimization of the functional (8) will be reduced to a minimization problem for the coefficients of the elements $x \in X_\pi$. For the subsequent considerations, it is appropriate to introduce the space

$$\begin{aligned} \tilde{X}_\pi \;=\; & \{x \in C_\pi([a,b], \mathbb{R}^m) : \\ x_\kappa|_{[t_{j-1},t_j)} \in \mathfrak{P}_N,\ \kappa \;=\; & 1, \dots, k, \\ x_\kappa|_{[t_{j-1},t_j)} \in \mathfrak{P}_{N-1},\ \kappa \;=\; & k+1, \dots, m,\ \ j = 1, \dots, n\}. \end{aligned} \tag{13}$$

In particular, the elements $x$ of $\tilde{X}_\pi$ are no longer required to have continuous components $Dx$. Obviously, it holds $X_\pi \subseteq \tilde{X}_\pi$. In general, $\tilde{X}_\pi$ is not a subspace of $H_D^1(a,b)$. However, it holds

$$\begin{aligned} X_\pi \;&=\; \{x \in \tilde{X}_\pi : x_\kappa \in C[a,b], \quad \kappa = 1, \dots, k\} \\ &=\; \tilde{X}_\pi \cap H_D^1(a,b). \end{aligned}$$

Based on the analysis in [8, Section 4], we provide a basis of the ansatz space $\tilde{X}_\pi$ to begin with. Assume that $\{p_0, \dots, p_{N-1}\}$ is a basis of $\mathfrak{P}_{N-1}$ defined on the reference interval $[0,1]$. Then, $\{\bar{p}_0, \dots, \bar{p}_N\}$ given by

$$\bar{p}_i(\tau) = \begin{cases} 1, & i = 0, \\ \int_0^\tau p_{i-1}(\sigma)\mathrm{d}\sigma, & i = 1, \dots, N, \quad \tau \in [0,1], \end{cases} \tag{14}$$

form a basis of $\mathfrak{P}_N$. The transformation to the interval $(t_{j-1}, t_j)$ of the partition $\pi$ (3) yields

$$p_{ji}(t) = p_i((t - t_{j-1})/h_j), \quad \bar{p}_{ji}(t) = h_j \bar{p}_i((t - t_{j-1})/h_j). \tag{15}$$

and in particular

$$\bar{p}_{ji}(t_{j-1}) = h_j \bar{p}_i(0) = h_j \begin{cases} 1, & i = 0, \\ 0, & i = 1, \dots, N, \end{cases}$$

$$\bar{p}_{ji}(t_j) = h_j \bar{p}_i(1) = h_j \begin{cases} 1, & i = 0, \\ \int_0^1 p_{i-1}(\sigma)\mathrm{d}\sigma, & i = 1, \dots, N. \end{cases}$$

Next, we form the matrix functions

$$\bar{\mathcal{P}}_j = \begin{bmatrix} \bar{p}_{j0} & \dots & \bar{p}_{jN} \end{bmatrix} : [t_{j-1}, t_j] \to \mathbb{R}^{1\times(N+1)}, \quad \mathcal{P}_j = \begin{bmatrix} p_{j0} & \dots & p_{j,N-1} \end{bmatrix} : [t_{j-1}, t_j] \to \mathbb{R}^{1\times N},$$

such that

$$\bar{\mathcal{P}}_j(t_{j-1}) = h_j \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}, \quad j = 1, \dots, n, \tag{16}$$

$$\bar{\mathcal{P}}_j(t_j) = h_j \begin{bmatrix} 1 & \int_0^1 p_0(\sigma)\mathrm{d}\sigma & \dots & \int_0^1 p_{N-1}(\sigma)\mathrm{d}\sigma \end{bmatrix}, \quad j = 1, \dots, n. \tag{17}$$

Following the discussions in [8], the following bases are suitable in applications:

Legendre basis Let $P_i$ denote the Legendre polynomials. Then, $p_i$ is chosen to be the shifted Legendre polynomial, that is

$$p_i(\tau) = P_i(2\tau - 1), \quad i = 0, 1, \dots.$$

Modified Legendre basis In this case, we set

$$\bar{p}_0(\tau) = 1, \quad \bar{p}_i(\tau) = P_i(2\tau - 1) - (-1)^i, \quad i = 1, 2, \dots,$$

such that $p_i = \bar{p}'_{i+1}$, $i = 0, 1, \dots$. This basis has not been considered in [8], but later experiments indicated its usefulness. This is supported by considerations later below.

Chebyshev basis Let $T_i$ denote the Chebyshev polynomials of the first kind. Then we define

$$p_i(\tau) = T_i(2\tau - 1), \quad i = 0, 1, \dots.$$

Runge-Kutta basis Let $0 < \tau_1 < \dots < \tau_N < 1$ be interpolation nodes. Then we set

$$p_i(\tau) = \frac{\prod_{\kappa\neq i+1}(\tau - \tau_\kappa)}{\prod_{\kappa\neq i+1}(\tau_{i+1} - \tau_\kappa)}. \tag{18}$$

The latter are the usual Lagrange interpolation polynomials. In the implementation, it is advantageous to represent these polynomials in terms of Chebyshev polynomials [8]. Of particular use is the Runge-Kutta basis if the shifted Chebyshev nodes $\tau_\kappa = \frac{1}{2}\left(1 + \cos\left(\frac{2\kappa-1}{2N}\pi\right)\right)$ are chosen as interpolation nodes.

For $x \in \tilde{X}_\pi$ we use the denotations

$$x(t) = x_j(t) = \begin{bmatrix} x_{j1}(t) \\ \vdots \\ x_{jm}(t) \end{bmatrix} \in \mathbb{R}^m, \quad Dx_j(t) = \begin{bmatrix} x_{j1}(t) \\ \vdots \\ x_{jk}(t) \end{bmatrix} \in \mathbb{R}^k, \quad t \in [t_{j-1}, t_j).$$

Then, we develop each $x_j$ componentwise

$$\begin{aligned} x_{j\kappa}(t) &= \sum_{l=0}^{N} c_{j\kappa l}\bar{p}_{jl}(t) = \bar{\mathcal{P}}_j(t)c_{j\kappa}, \quad \kappa = 1,\dots,k, \\ x_{j\kappa}(t) &= \sum_{l=0}^{N-1} c_{j\kappa l}p_{jl}(t) = \mathcal{P}_j(t)c_{j\kappa}, \quad \kappa = k+1,\dots,m. \end{aligned} \tag{19}$$

with

$$c_{j\kappa} = \begin{bmatrix} c_{j\kappa 0} \\ \vdots \\ c_{j\kappa N} \end{bmatrix} \in \mathbb{R}^{N+1}, \quad \kappa = 1,\dots,k, \quad c_{j\kappa} = \begin{bmatrix} c_{j\kappa 0} \\ \vdots \\ c_{j\kappa,N-1} \end{bmatrix} \in \mathbb{R}^{N}, \quad \kappa = k+1,\dots,m.$$

Introducing still

$$\Omega_j(t) = \begin{bmatrix} I_k \otimes \bar{\mathcal{P}}_j(t) & \mathcal{O}_1 \\ \mathcal{O}_2 & I_{m-k} \otimes \mathcal{P}_j(t) \end{bmatrix} \in \mathbb{R}^{m\times(mN+k)}, \quad c_j = \begin{bmatrix} c_{j1} \\ \vdots \\ c_{jm} \end{bmatrix} \in \mathbb{R}^{mN+k},$$

with $\mathcal{O}_1 \in \mathbb{R}^{k\times kN}$ and $\mathcal{O}_2 \in \mathbb{R}^{(m-k)\times(m-k)(N+1)}$ being matrices having only zero entries we represent, for $t \in I_j$, $j = 1,\dots,n$,

$$x_j(t) = \Omega_j(t)c_j, \tag{20}$$

$$(Dx_j)'(t) = (D\Omega_j)'(t)c_j = \begin{bmatrix} I_k \otimes \bar{\mathcal{P}}_j'(t) & \mathcal{O}_1 \end{bmatrix} c_j \tag{21}$$

where $\bar{\mathcal{P}}_j'(t) = \begin{bmatrix} 0 & p_{j0} & \dots & p_{j,N-1} \end{bmatrix}$. Now we collect all coefficients $c_{j\kappa l}$ in the vector $c$,

$$c = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} \in \mathbb{R}^{n(mN+k)}.$$

**Definition 1** The mapping $\mathcal{R} : \mathbb{R}^{n(mN+k)} \to \tilde{X}_\pi$ given by (20) is called the *representation map* of $\tilde{X}_\pi$ with respect to the basis (15).

**Fact 1** We observe that each $x \in \tilde{X}_\pi$ has a representation of the kind (20) and each function of the form (20) is an element of $\tilde{X}_\pi$. Since $\dim \tilde{X}_\pi = n(mN+k)$, $\mathcal{R}$ is a bijective mapping.

Consider an element $x \in \tilde{X}_\pi$ with its representation (20). This element belongs to $X_\pi$ if and only if its first $k$ components are continuous. Using the representation (19) we see that $x \in X_\pi$ if and only if

$$\mathcal{C}c = 0. \tag{22}$$

where $\mathcal{C} \in \mathbb{R}^{k(n-1)\times n(mN+k)}$ and

$$\mathcal{C} = \begin{bmatrix} I_k \otimes \bar{\mathcal{P}}_1(t_1) & \mathcal{O}_1 & -I_k \otimes \bar{\mathcal{P}}_2(t_1) & \mathcal{O}_1 & & & \\ & & I_k \otimes \bar{\mathcal{P}}_2(t_2) & \mathcal{O}_1 & -I_k \otimes \bar{\mathcal{P}}_3(t_2) & \mathcal{O}_1 & \\ & & & \ddots & & \ddots & \\ & & & & & & \\ & & & & I_k \otimes \bar{\mathcal{P}}_{n-1}(t_{n-1}) & \mathcal{O}_1 & -I_k \otimes \bar{\mathcal{P}}_n(t_{n-1}) & \mathcal{O}_1 \end{bmatrix}.$$

Owing to the construction, $\mathcal{C}$ has full row rank, cf. (16), (17).

**Fact 2** Let $\tilde{\mathcal{R}} = \mathcal{R}|_{\ker \mathcal{C}}$ be the restriction of the representation map $\mathcal{R}$ onto the kernel $\ker \mathcal{C}$ of $\mathcal{C}$. Since $\mathcal{C}$ has full row rank, $\dim \ker \mathcal{C} = n(mN + k) - k(n - 1) = nmN + k = \dim X_\pi$, and $\mathcal{R}$ is injective, $\tilde{\mathcal{R}}$ is bijective. In particular, it holds also $\tilde{\mathcal{R}}^{-1} = \mathcal{R}^{-1}\big|_{\mathrm{im}\tilde{\mathcal{R}}}$.

The representations (20)–(21) can be inserted into the functional $\Phi_{\pi,M}$ (8). The result becomes a least-squares functional of the form

$$\varphi(c) = |\mathcal{A}c - r|^2_{\mathbb{R}^{nmM+l_{dyn}}} \to \min! \tag{23}$$

where $\mathcal{A}$ has the structure

$$\mathcal{A} = \begin{bmatrix} \mathcal{A}_1 & 0 & \cdots & & 0 \\ 0 & \ddots & & & \vdots \\ \vdots & & \ddots & & \\ & & & \ddots & 0 \\ 0 & & & & \mathcal{A}_n \\ G_a\Omega_1(t_0) & 0 & \cdots & 0 & G_b\Omega_n(t_n) \end{bmatrix}$$

where $\mathcal{A}_j \in \mathbb{R}^{mM\times(mN+k)}$ and $G_a\Omega_1(t_0), G_b\Omega_n(t_n) \in \mathbb{R}^{l_{dyn}\times(mN+k)}$.

So the discrete version of the least-squares method (8) becomes the linear least-squares problem (23) under the linear equality constraint (22).

Note that it holds $r \in \mathbb{R}^{nmM+l_{dyn}}$ and $\mathcal{A} \in \mathbb{R}^{(nmM+l_{dyn})\times n(mN+k)}$. The matrices $\mathcal{A}$ and $\mathcal{C}$ are very sparse. More details of the construction of $\mathcal{A}$ and $\mathcal{C}$ can be found in [9].

### 2.3 Conditioning of the implementation

The implementation for solving the least-squares problem (8) consists of the following steps:

1. Form $\mathcal{A}$, $\mathcal{C}$, and $r$.
2. Solve the constrained least-squares problem (23)–(22).
3. Form the approximation $x_\pi$.

What are the errors to be expected? Consider the individual steps:

1. The computation of $\mathcal{C}$ is not critical. Depending on the chosen basis, the entries of $\mathcal{C}$ may be available analytically. So we expect at most rounding errors for the

representation of the analytical data.[2] While the components of $\mathcal{A}$ corresponding to the boundary conditions are only subject to truncation errors when representing real numbers in floating point arithmetic, the DAE-related entries are subject to rounding errors as well as certain amplification factors stemming from the multiplication by the square root of the matrix $\mathcal{L}$ (9). The conditioning of the versions (10) and (11) is easy to infer while that of (12) has been discussed extensively in [8]. Under reasonable assumptions on the choice of collocation points, they are rather small.

Similar considerations apply to the computation of $r$.

2. This algorithmic step corresponds to the solution of a linearly constrained linear least-squares problem. A number of classical perturbation results are available, e.g., [11–13]. Further below, we represent a modified version that is taking into account the special role that the equality constraint $\mathcal{C}c = 0$ is playing in our application.
3. This step is described by the representation map $\mathcal{R}$, which assigns, to each solution $c$ of the previous step, the corresponding solution $x_\pi = \mathcal{R}c$. If $c \in \ker \mathcal{C}$, it holds $x_\pi \in X_\pi \subseteq H^1_D(a, b)$. However, due to the errors made in the previous step, the condition $c \in \ker \mathcal{C}$ cannot be guaranteed such that $\mathcal{R}c \in \tilde{X}_\pi$ but not necessarily $\mathcal{R}c \in H^1_D(a, b)$! In the next section, we will discuss the properties of $\mathcal{R}$.

## 3 Properties of the representation map $\mathcal{R}$

In the present section, we will investigate the properties of the representation map $\mathcal{R} : \mathbb{R}^{n(mN+k)} \to \tilde{X}_\pi$ in more detail. Previously, we have established a representation of $\mathcal{R}$ on each subinterval; see (20). We intend to derive a representation of $\mathcal{R}^{-1}$. The main tool will be interpolation.

Choose two sets of interpolation nodes

$$0 \le \bar{\sigma}_1 < \cdots < \bar{\sigma}_{N+1} \le 1 \text{ and } 0 \le \sigma_1 < \cdots < \sigma_N \le 1, \tag{24}$$

and shifted ones

$$\bar{\tau}_{ji} = t_{j-1} + \bar{\sigma}_i h_j, \quad \tau_{ji} = t_{j-1} + \sigma_i h_j$$

such that the integration formulae

$$\int_0^1 f(\sigma) d\sigma \approx \sum_{i=1}^{N+1} \bar{\gamma}_i f(\bar{\sigma}_i), \quad \text{and} \quad \int_0^1 f(\sigma) d\sigma \approx \sum_{i=1}^{N} \gamma_i f(\sigma_i)$$

[2] In the case of the Legendre and modified Legendre bases, all entries are integers weighted by the stepsizes.

have positive weights and so that they are exact for polynomials up to degree $2N$ and $2N-2$, respectively. With matrices

$$\bar{V}_j = \begin{bmatrix} \bar{p}_{j0}(\bar{\tau}_{j1}) & \cdots & \bar{p}_{jN}(\bar{\tau}_{j1}) \\ \vdots & & \vdots \\ \bar{p}_{j0}(\bar{\tau}_{j,N+1}) & \cdots & \bar{p}_{jN}(\bar{\tau}_{j,N+1}) \end{bmatrix} = h_j \begin{bmatrix} \bar{p}_0(\bar{\sigma}_1) & \cdots & \bar{p}_N(\bar{\sigma}_1) \\ \vdots & & \vdots \\ \bar{p}_0(\bar{\sigma}_{N+1}) & \cdots & \bar{p}_N(\bar{\sigma}_{N+1}) \end{bmatrix} =: h_j \bar{V}, \tag{25}$$

$$V_j = \begin{bmatrix} p_{j0}(\tau_{j1}) & \cdots & p_{j,N-1}(\tau_{j1}) \\ \vdots & & \vdots \\ p_{j0}(\tau_{jN}) & \cdots & p_{j,N-1}(\tau_{jN}) \end{bmatrix} = \begin{bmatrix} p_0(\sigma_1) & \cdots & p_{N-1}(\sigma_1) \\ \vdots & & \vdots \\ p_0(\sigma_N) & \cdots & p_{N-1}(\sigma_N) \end{bmatrix} =: V, \tag{26}$$

and

$$\bar{V}'_j = \begin{bmatrix} \bar{p}'_{j0}(\bar{\tau}_{j1}) & \cdots & \bar{p}'_{jN}(\bar{\tau}_{j1}) \\ \vdots & & \vdots \\ \bar{p}'_{j0}(\bar{\tau}_{j,N+1}) & \cdots & \bar{p}'_{jN}(\bar{\tau}_{j,N+1}) \end{bmatrix} = \begin{bmatrix} 0 & p_0(\bar{\sigma}_1) & \cdots & p_{N-1}(\bar{\sigma}_1) \\ \vdots & \vdots & & \vdots \\ 0 & p_0(\bar{\sigma}_{N+1}) & \cdots & p_{N-1}(\bar{\sigma}_{N+1}) \end{bmatrix} =: \mathring{V}, \tag{27}$$

we represent, for $\kappa = 1, \ldots, k$,

$$X_{j\kappa} := \begin{bmatrix} x_{j\kappa}(\bar{\tau}_{j1}) \\ \vdots \\ x_{j\kappa}(\bar{\tau}_{j,N+1}) \end{bmatrix} = \bar{V}_j c_{j\kappa} = h_j \bar{V} c_{j\kappa},$$

$$X'_{j\kappa} := \begin{bmatrix} x'_{j\kappa}(\bar{\tau}_{j1}) \\ \vdots \\ x'_{j\kappa}(\bar{\tau}_{j,N+1}) \end{bmatrix} = \bar{V}'_j c_{j\kappa} = \mathring{V} c_{j\kappa},$$

and, for $\kappa = k+1, \ldots, m$,

$$X_{j\kappa} := \begin{bmatrix} x_{j\kappa}(\tau_{j1}) \\ \vdots \\ x_{j\kappa}(\tau_{jN}) \end{bmatrix} = V_j c_{j\kappa} = V c_{j\kappa}.$$

The matrices $\bar{V}$ and $V$ are nonsingular. This amounts to the relation

$$c_j = \begin{bmatrix} c_{j1} \\ \vdots \\ c_{jk} \\ c_{j,k+1} \\ \vdots \\ c_{jm} \end{bmatrix} = \begin{bmatrix} I_k \otimes \bar{V}^{-1} & \\ & I_{m-k} \otimes V^{-1} \end{bmatrix} \begin{bmatrix} \frac{1}{h_j} X_{j1} \\ \vdots \\ \frac{1}{h_j} X_{jk} \\ X_{j,k+1} \\ \vdots \\ X_{jm} \end{bmatrix}, \quad j = 1, \ldots, n. \tag{28}$$

Owing to the fact, that polynomials of degree $N$ and $N-1$ are uniquely determined by their values at $N+1$ and $N$ different nodes, respectively, formula (28) provides $c = \mathcal{R}^{-1} x$ for each arbitrary given $x \in \tilde{X}_\pi$.

Next, we equip $\tilde{X}_\pi$ with the norms

$$\|x\|_{L^2}^2 = \sum_{j=1}^{n} \left\{ \sum_{\kappa=1}^{k} \int_{t_{j-1}}^{t_j} |x_{j\kappa}(t)|^2 dt + \sum_{\kappa=k+1}^{m} \int_{t_{j-1}}^{t_j} |x_{j\kappa}(t)|^2 dt \right\}, \tag{29}$$

$$\|x\|_{H^1_{D,\pi}}^2 = \sum_{j=1}^{n} \left\{ \sum_{\kappa=1}^{k} \int_{t_{j-1}}^{t_j} (|x_{j\kappa}(t)|^2 + |x'_{j\kappa}(t)|^2) dt + \sum_{\kappa=k+1}^{m} \int_{t_{j-1}}^{t_j} |x_{j\kappa}(t)|^2 dt \right\}. \tag{30}$$

The latter norm reduces, for $x \in X_\pi$, to $\|x\|_{H^1_{D,\pi}} = \|x\|_{H^1_D(a,b)}$. Moreover, $\|\cdot\|_{L^2} = \|\cdot\|_{L^2((a,b),\mathbb{R}^m)}$. On $\mathbb{R}^{n(mN+k)}$, we use the Euclidean norm. Then $\mathcal{R}$ becomes a homeomorphism in each case, and we are interested in the respective operator norms $\|\mathcal{R}\|_{\mathbb{R}^{n(mN+k)} \to L^2}$, $\|\mathcal{R}\|_{\mathbb{R}^{n(mN+k)} \to H^1_{D,\pi}}$, $\|\mathcal{R}^{-1}\|_{L^2 \to \mathbb{R}^{n(mN+k)}}$, and $\|\mathcal{R}^{-1}\|_{H^1_{D,\pi} \to \mathbb{R}^{n(mN+k)}}$. Regarding the properties of the related integration formulae and introducing the diagonal matrices

$$\bar{\Gamma} = \operatorname{diag}(\bar{\gamma}_1^{1/2}, \cdots, \bar{\gamma}_{N+1}^{1/2}), \ \Gamma = \operatorname{diag}(\gamma_1^{1/2}, \cdots, \gamma_N^{1/2}) \tag{31}$$

we compute for any $x = \mathcal{R}c$, and $\kappa = 1, \ldots, k$,

$$\begin{aligned} \int_{t_{j-1}}^{t_j} |x_{j\kappa}(t)|^2 dt &= h_j \sum_{i=1}^{N+1} \bar{\gamma}_i |x_{j\kappa}(\bar{\tau}_{ji})|^2 = h_j \sum_{i=1}^{N+1} |\bar{\gamma}_i^{1/2} x_{j\kappa}(\bar{\tau}_{ji})|^2 = h_j |\bar{\Gamma} X_{j\kappa}|^2 \\ &= h_j |\bar{\Gamma} \bar{V}_j c_{j\kappa}|^2 = h_j |\bar{\Gamma} h_j \bar{V} c_{j\kappa}|^2, \end{aligned}$$

$$\begin{aligned} \int_{t_{j-1}}^{t_j} (|x_{j\kappa}(t)|^2 + |x'_{j\kappa}(t)|^2) dt &= h_j \sum_{i=1}^{N+1} \bar{\gamma}_i (|x_{j\kappa}(\bar{\tau}_{ji})|^2 + |x'_{j\kappa}(\bar{\tau}_{ji})|^2) \\ &= h_j \sum_{i=1}^{N+1} (|\bar{\gamma}_i^{1/2} x_{j\kappa}(\bar{\tau}_{ji})|^2 + (|\bar{\gamma}_i^{1/2} x'_{j\kappa}(\bar{\tau}_{ji})|^2) = h_j |\bar{\Gamma} X_{j\kappa}|^2 + h_j |\bar{\Gamma} X'_{j\kappa}|^2 \\ &= h_j |\bar{\Gamma} \bar{V}_j c_{j\kappa}|^2 + h_j |\bar{\Gamma} \mathring{V} c_{j\kappa}|^2 = h_j |\bar{\Gamma} h_j \bar{V} c_{j\kappa}|^2 + h_j |\bar{\Gamma} \mathring{V} c_{j\kappa}|^2 \\ &= h_j \left| \begin{bmatrix} h_j \bar{\Gamma} \bar{V} \\ \bar{\Gamma} \mathring{V} \end{bmatrix} c_{j\kappa} \right|^2, \end{aligned}$$

and, in addition, for $\kappa = k+1, \ldots, m$,

$$\begin{aligned} \int_{t_{j-1}}^{t_j} |x_{j\kappa}(t)|^2 dt &= h_j \sum_{i=1}^{N} \gamma_i |x_{j\kappa}(\tau_{ji})|^2 = h_j \sum_{i=1}^{N} |\gamma_i^{1/2} x_{j\kappa}(\tau_{ji})|^2 = h_j |\Gamma X_{j\kappa}|^2 \\ &= h_j |\Gamma V c_{j\kappa}|^2. \end{aligned}$$

Summarizing, the following representations result:

$$\|x\|_{L^2}^2 = \sum_{j=1}^{n} \left\{ \sum_{\kappa=1}^{k} |h_j^{3/2} \bar{\Gamma} \bar{V} c_{j\kappa}|^2 + \sum_{\kappa=k+1}^{m} |h_j^{1/2} \Gamma V c_{j\kappa}|^2 \right\} = \sum_{j=1}^{n} |U_j c_j|^2 = |\mathcal{U} c|^2, \tag{32}$$

with matrices

$$\begin{aligned} \mathcal{U} &= \operatorname{diag}(U_1, \cdots, U_n) \in \mathbb{R}^{n(mN+k)\times n(mN+k)}, \\ U_j &= \begin{bmatrix} I_k \otimes h_j^{3/2}\bar{\Gamma}\bar{V} & \\ & I_{m-k} \otimes h_j^{1/2}\Gamma V \end{bmatrix} \in \mathbb{R}^{(mN+k)\times(mN+k)}, \end{aligned} \tag{33}$$

and

$$\|x\|_{H^1_{D,\pi}}^2 = \sum_{j=1}^{n} \left\{ \sum_{\kappa=1}^{k} \left| \begin{bmatrix} h_j^{3/2}\bar{\Gamma}\bar{V} \\ h_j^{1/2}\bar{\Gamma}\mathring{V} \end{bmatrix} c_{j\kappa} \right|^2 + \sum_{\kappa=k+1}^{m} |h_j^{1/2}\Gamma V c_{j\kappa}|^2 \right\} = \sum_{j=1}^{n} |\hat{U}_j c_j|^2 = |\hat{\mathcal{U}}c|^2, \tag{34}$$

with matrices

$$\begin{aligned} \hat{\mathcal{U}} &= \operatorname{diag}(\hat{U}_1, \cdots, \hat{U}_n) \in \mathbb{R}^{n(mN+k+k(N+1))\times n(mN+k)}, \\ \hat{U}_j &= \begin{bmatrix} I_k \otimes \begin{bmatrix} h_j^{3/2}\bar{\Gamma}\bar{V} \\ h_j^{1/2}\bar{\Gamma}\mathring{V} \end{bmatrix} & \\ & I_{m-k} \otimes h_j^{1/2}\Gamma V \end{bmatrix} \in \mathbb{R}^{(mN+k+k(N+1))\times(mN+k)}. \end{aligned} \tag{35}$$

**Proposition 1** *The singular values of $\mathcal{U}$ and $\hat{\mathcal{U}}$ are independent of the choice of the nodes $\sigma_i$ and $\bar{\sigma}_i$. Moreover, all singular values are positive.*

*Proof* $U_j$ and $\hat{U}_j$ have full column-rank. Consequently, $\mathcal{U}^T\mathcal{U}$ and $\hat{\mathcal{U}}^T\hat{\mathcal{U}}$ are symmetric and positive definite. Hence, their eigenvalues are all positive and, thus, also their singular values being the square root of the eigenvalues. The eigenvalues are independent of the choice of the nodes $\sigma_i$ and $\bar{\sigma}_i$ since, owing to the properties of the involved integration formulae, it holds that

$$\begin{aligned} (V^T\Gamma^2 V)_{\alpha\beta} &= \int_0^1 p_{\alpha-1}p_{\beta-1}(\sigma)d\sigma,\ \alpha,\beta = 1,\cdots,N, \\ (\bar{V}^T\bar{\Gamma}^2\bar{V})_{\alpha\beta} &= \int_0^1 \bar{p}_{\alpha-1}\bar{p}_{\beta-1}(\sigma)d\sigma,\ \alpha,\beta = 1,\cdots,N+1, \\ (\mathring{V}^T\bar{\Gamma}^2\mathring{V})_{\alpha\beta} &= \int_0^1 \bar{p}'_{\alpha-1}\bar{p}'_{\beta-1}(\sigma)d\sigma,\ \alpha,\beta = 1,\cdots,N+1, \end{aligned}$$

such that the entries of $\mathcal{U}^T\mathcal{U}$ and $\hat{\mathcal{U}}^T\hat{\mathcal{U}}$ are independent of the choice of the integration formulae. □

**Theorem 2** *Let $\sigma_{\min}(\mathcal{U})$ and $\sigma_{\max}(\mathcal{U})$ denote the maximal and minimal singular values of $\mathcal{U}$. Similarly, let $\sigma_{\min}(\hat{\mathcal{U}})$ and $\sigma_{\max}(\hat{\mathcal{U}})$ denote the maximal and minimal singular values of $\hat{\mathcal{U}}$. Then it holds*

$$\begin{aligned} \|\mathcal{R}\|_{\mathbb{R}^{n(mN+k)}\to L^2} &= \sigma_{\max}(\mathcal{U}), \quad \|\mathcal{R}^{-1}\|_{L^2\to\mathbb{R}^{n(mN+k)}} = \sigma_{\min}(\mathcal{U})^{-1}, \\ \|\mathcal{R}\|_{\mathbb{R}^{n(mN+k)}\to H^1_{D,\pi}} &= \sigma_{\max}(\hat{\mathcal{U}}), \quad \|\mathcal{R}^{-1}\|_{H^1_{D,\pi}\to\mathbb{R}^{n(mN+k)}} = \sigma_{\min}(\hat{\mathcal{U}})^{-1}. \end{aligned}$$

*Proof* It holds $\hat{\mathcal{U}} \in \mathbb{R}^{\nu\times\lambda}$ with $\nu = n(mN+k+k(N+1))$ and $\lambda = n(mN+k)$. Let $\hat{\mathcal{U}} = U\Sigma V^T$ be the singular value decomposition of $\mathcal{U}$. Here,

$$\Sigma = \begin{bmatrix} s_1 & & \\ & \ddots & \\ & & s_\nu \\ 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{\nu\times\lambda}$$

with $s_1 = \sigma_{\max}(\hat{\mathcal{U}})$ and $s_\nu = \sigma_{\min}(\hat{\mathcal{U}})$. According to Proposition 1, $\sigma_{\min}(\hat{\mathcal{U}}) > 0$. By (34), this leads to

$$\|\mathcal{R}\|_{\mathbb{R}^{n(mN+k)}\to H^1_{D,\pi}} = \sup_{c\neq 0}\frac{\|\mathcal{R}c\|_{H^1_{D,\pi}}}{|c|_{\mathbb{R}^\lambda}} = \sup_{c\neq 0}\frac{|\hat{\mathcal{U}}c|_{\mathbb{R}^\nu}}{|c|_{\mathbb{R}^\lambda}} = \sup_{\chi\neq 0}\frac{|\Sigma\chi|_{\mathbb{R}^\nu}}{|\chi|_{\mathbb{R}^\lambda}} = \sigma_{\max}(\hat{\mathcal{U}})$$

and

$$\|\mathcal{R}^{-1}\|_{H^1_{D,\pi}\to\mathbb{R}^{n(mN+k)}} = \sup_{x\neq 0}\frac{|\mathcal{R}^{-1}x|_{\mathbb{R}^\lambda}}{\|x\|_{H^1_{D,\pi}}} = \sup_{c\neq 0}\frac{|c|_{\mathbb{R}^\lambda}}{\|\mathcal{R}c\|_{H^1_{D,\pi}}} = \sup_{\chi\neq 0}\frac{|\chi|_{\mathbb{R}^\lambda}}{|\Sigma\chi|_{\mathbb{R}^\nu}} = \sigma_{\min}(\hat{\mathcal{U}})^{-1}.$$

The statements concerning $\|\mathcal{R}\|_{\mathbb{R}^{n(mN+k)}\to L^2}$ and $\|\mathcal{R}^{-1}\|_{H^1_{D,\pi}\to\mathbb{R}^{n(mN+k)}}$ follow similarly. □

Using the structure (33) of $\mathcal{U}$, we obtain

$$\begin{aligned}
\sigma_{\max}(\mathcal{U}) &= \max_{j=1,\dots,n} \max\{h_j^{3/2}\sigma_{\max}(\bar{\Gamma}\bar{V}), h_j^{1/2}\sigma_{\max}(\Gamma V)\} \\
&= \max_{j=1,\dots,n} h_j^{1/2}\max\{h_j\sigma_{\max}(\bar{\Gamma}\bar{V}), \sigma_{\max}(\Gamma V)\}, \\
\sigma_{\min}(\mathcal{U}) &= \min_{j=1,\dots,n} \min\{h_j^{3/2}\sigma_{\min}(\bar{\Gamma}\bar{V}), h_j^{1/2}\sigma_{\min}(\Gamma V)\} \\
&= \min_{j=1,\dots,n} h_j^{1/2}\min\{h_j\sigma_{\min}(\bar{\Gamma}\bar{V}), \sigma_{\min}(\Gamma V)\}.
\end{aligned}$$

The estimation of the singular values of $\hat{\mathcal{U}}$ leads to slightly more involved expressions. Let $U_{j,red} = \begin{bmatrix} h_j\bar{\Gamma}\bar{V} \\ \bar{\Gamma}\mathring{V} \end{bmatrix}$. Then, it holds

$$\begin{aligned}
\sigma_{\max}(\hat{\mathcal{U}}) &= \max_{j=1,\dots,n} h_j^{1/2}\max\{\sigma_{\max}(U_{j,red}), \sigma_{\max}(\Gamma V)\}, \\
\sigma_{\min}(\hat{\mathcal{U}}) &= \min_{j=1,\dots,n} h_j^{1/2}\min\{\sigma_{\min}(U_{j,red}), \sigma_{\min}(\Gamma V)\}.
\end{aligned}$$

We note that $\sigma_{\min}(\bar{\Gamma}\mathring{V}) = 0$ and $\sigma_{\max}(\Gamma V) = \sigma_{\max}(\bar{\Gamma}\mathring{V})$. This follows immediately from the construction of the basis for the differential components (14). The definition of singular values and Weyl's Theorem [14, Theorem III.2.1] provides us with

$$\begin{aligned}
\lambda_{\max}(\mathring{V}^T\bar{\Gamma}^2\mathring{V}) &\le \lambda_{\max}(h_j^2\bar{V}^T\bar{\Gamma}^2\bar{V} + \mathring{V}^T\bar{\Gamma}^2\mathring{V}) = \sigma_{\max}(U_{j,red})^2 \\
&\le h_j^2\lambda_{\max}(\bar{V}^T\bar{\Gamma}^2\bar{V}) + \lambda_{\max}(\mathring{V}^T\bar{\Gamma}^2\mathring{V}), \\
h_j^2\lambda_{\min}(\bar{V}^T\bar{\Gamma}^2\bar{V}) &\le \lambda_{\min}(h_j^2\bar{V}^T\bar{\Gamma}^2\bar{V} + \mathring{V}^T\bar{\Gamma}^2\mathring{V}) = \sigma_{\min}(U_{j,red})^2 \le h_j^2\lambda_{\max}(\bar{V}^T\bar{\Gamma}^2\bar{V})
\end{aligned}$$

since $\lambda_{\min}(\mathring{V}^T \bar{\Gamma}^2 \mathring{V}) = 0$. Then,

$$\begin{aligned}\sigma_{\max}(\Gamma V) &= \lambda_{\max}(V^T \Gamma^2 V)^{1/2} \\ &\le \max\{\sigma_{\max}(U_{j,red}), \sigma_{\max}(\Gamma V)\} \\ &\le \sigma_{\max}(\Gamma V) + O(h).\end{aligned}$$

Moreover,

$$\begin{aligned}\min\{h_j \sigma_{\min}(\bar{\Gamma}\bar{V}), \sigma_{\min}(\Gamma V)\} &\le \min\{\sigma_{\min}(U_{j,red}), \sigma_{\min}(\Gamma V)\} \\ &\le \min\{h_j \sigma_{\max}(\bar{\Gamma}\bar{V}), \sigma_{\min}(\Gamma V)\}\end{aligned}$$

Collecting all estimates ,Theorem 2 provides

**Theorem 3** *Let the grid* (3) *have the maximal stepsize $h$ and the minimal stepsize $h_{\min}$. Furthermore, let $\Gamma$ and $\bar{\Gamma}$ be given by* (31) *and let $V$, $\bar{V}$, and $\mathring{V}$ be given by* (26), (25), (27). *Then it holds, for sufficiently small $h$,*

$$\begin{aligned}\|\mathcal{R}\|_{\mathbb{R}^{n(mN+k)} \to L^2} &= h^{1/2} \sigma_{\max}(\Gamma V) = O(h^{1/2}), \\ \|\mathcal{R}^{-1}\|_{L^2 \to \mathbb{R}^{n(mN+k)}} &= h_{\min}^{-3/2} \sigma_{\min}(\bar{\Gamma}\bar{V})^{-1} = O(h_{\min}^{-3/2}), \\ \|\mathcal{R}\|_{\mathbb{R}^{n(mN+k)} \to H^1_{D,\pi}} &= h^{1/2} \sigma_{\max}(\Gamma V) + O(h^{3/2}) = O(h^{1/2}),\end{aligned}$$

*and*

$$h_{\min}^{-3/2} \sigma_{\max}(\bar{\Gamma}\bar{V})^{-1} \le \|\mathcal{R}^{-1}\|_{H^1_{D,\pi} \to \mathbb{R}^{n(mN+k)}} \le h_{\min}^{-3/2} \sigma_{\min}(\bar{\Gamma}\bar{V})^{-1}.$$

*In particular,* $\|\mathcal{R}\|_{\mathbb{R}^{n(mN+k)} \to H^1_{D,\pi}} = \|\mathcal{R}\|_{\mathbb{R}^{n(mN+k)} \to L^2} + O(h^{3/2})$.

In these estimates, we used the fact $\sigma_{\min}(\Gamma V) > 0$. Note that the constants hidden in the big-O notation in this theorem depend on both $N$ and the chosen basis. For the restriction $\tilde{\mathcal{R}}$ of $\mathcal{R}$ onto $\ker \mathcal{C}$ we obtain, obviously,

$$\|\tilde{\mathcal{R}}\| \le \|\mathcal{R}\|, \quad \|\tilde{\mathcal{R}}^{-1}\| \le \|\mathcal{R}^{-1}\|.$$

For some special cases, the singular values can be easily derived.

**Proposition 2** *Let $V$, $\bar{V}$, and $\mathring{V}$ be given by* (25)–(27) *and $\Gamma$, $\bar{\Gamma}$ by* (31). *Then it holds:*

(1) *Let $p_0, \dots, p_{N-1}$ be an orthogonal basis in $L^2(0,1)$. Then*

$$\begin{aligned}\sigma_{\min}(\Gamma V) &= \min\left\{\|p_\alpha\|_{L^2(0,1)} : \alpha = 0, \dots, N-1\right\}, \\ \sigma_{\max}(\Gamma V) &= \max\left\{\|p_\alpha\|_{L^2(0,1)} : \alpha = 0, \dots, N-1\right\}.\end{aligned}$$

*In particular, if $p_0, \dots, p_{N-1}$ is the Legendre basis, $\sigma_{\min}(\Gamma V) = (2N-1)^{-1/2}$ and $\sigma_{\max}(\Gamma V) = 1$.*

(2) *For an orthonormal basis $p_0, \dots, p_{N-1}$ in $L^2(0,1)$, $\sigma_{\min}(\Gamma V) = \sigma_{\max}(\Gamma V) = 1$.*

(3) *If $p_0, \dots, p_{N-1}$ is the modified Legendre basis, it holds $\sigma_{\min}(\bar{\Gamma}\bar{V}) \ge (2N+1)^{-1/2}$ and $\sigma_{\max}(\bar{\Gamma}\bar{V}) \le (N+2)^{1/2}$. Furthermore, the estimates*

$$\sigma_{\min}(\Gamma V) \ge \left(\frac{1}{2 - 2\cos\frac{N}{N+2}\pi}\right)^{1/2} \ge \frac{1}{2}, \quad \sigma_{\max}(\Gamma V) \le \left(\frac{2N-1}{2 - 2\cos\frac{1}{N+2}\pi}\right)^{1/2}$$

*hold true.*

*Proof* First, we observe that $(V^T\Gamma^2 V)_{\alpha\beta} = \int_0^1 p_{\alpha-1}(\rho)p_{\beta-1}(\rho)d\rho = \delta_{\alpha\beta}\|p_{\alpha-1}\|^2_{L^2(a,b)}$. This provides (1) and (2) as special cases.

Consider the modified Legendre basis now. It holds $\int_0^1 \bar{p}_0^2(\rho)d\rho = 1$ and $\int_0^1 \bar{p}_0(\rho)\bar{p}_\alpha(\rho)d\rho = \int_0^1 (P_\alpha(2\rho-1) - (-1)^\alpha)d\rho = (-1)^{\alpha+1}$ for $\alpha = 1, 2, \ldots$. Moreover, for $\alpha, \beta = 1, 2, \ldots$, we have

$$\begin{aligned}\int_0^1 \bar{p}_\alpha(\rho)\bar{p}_\beta(\rho)d\rho &= \int_0^1 (P_\alpha(2\rho-1) - (-1)^\alpha)(P_\beta(2\rho-1) - (-1)^\beta)d\rho \\ &= \int_0^1 P_\alpha(2\rho-1)P_\beta(2\rho-1)d\rho + (-1)^{\alpha+\beta} \\ &= (2\alpha+1)^{-1}\delta_{\alpha\beta} + (-1)^{\alpha+\beta}.\end{aligned}$$

Collecting these expressions, we obtain the compact representation

$$\bar{V}^T\bar{\Gamma}^2\bar{V} = \mathrm{diag}(1, \frac{1}{3}, \ldots, (2N+1)^{-1}) + ff^T$$

with $f^T = [1, 1, -1, +1, -1, \ldots, \pm 1] \in \mathbb{R}^{N+1}$. $ff^T$ is a rank-1 matrix having, therefore, the $N$-fold eigenvalue 0. Moreover, $f$ is an eigenvector to the eigenvalue $f^T f = N+1$. In particular, $ff^T$ is positive semidefinite. Invoking Weyl's theorem again, we obtain

$$\begin{aligned}(2N+1)^{-1} &= \lambda_{\min}(\mathrm{diag}(1, \frac{1}{3}, \ldots, (2N+1)^{-1})) \le \lambda_{\min}(\bar{V}^T\bar{\Gamma}^2\bar{V}) \\ \lambda_{\max}(\bar{V}^T\bar{\Gamma}^2\bar{V}) &\le \lambda_{\max}(\mathrm{diag}(1, \frac{1}{3}, \ldots, (2N+1)^{-1})) + \lambda_{\max}(ff^T) = N+2.\end{aligned}$$

This proves the first assertion of (3).

The relation $(\mathring{V}^T\bar{\Gamma}^2\mathring{V})_{\alpha\beta} = \int_0^1 \bar{p}'_{\alpha-1}\bar{p}'_{\beta-1}(\sigma)d\sigma$ shows that $K = \mathring{V}^T\bar{\Gamma}^2\mathring{V}$ is the stiffness matrix of the basis functions. For the modified Legendre basis, it has been investigated in [1, cp Eq. (31)]. According to the proof of Proposition A.2 of [1], the nonvanishing eigenvalues can be estimated by[3]

$$\lambda_{\min}(K) \ge \frac{1}{2 - 2\cos\frac{N}{N+2}\pi}, \quad \lambda_{\max}(K) \le \frac{2N-1}{2 - 2\cos\frac{1}{N+2}\pi}.$$

$K' = V^T\Gamma^2 V$ is the submatrix of $K$ obtained by omitting the first row and column of $K$, which consist entirely of zeros. This provides the final relations of assertion (3). □

An asymptotic analysis shows that $\sigma_{\max}(\Gamma V) \le \frac{2}{\sqrt{\pi}}N^{3/2} + O(N^{1/2})$ in the case of the modified Legendre basis.

[3] In [1], the stiffness matrix is scaled to the interval $(-1, 1)$ in contrast to the interval $(0, 1)$ used here. Therefore, an additional factor of $1/2$ appears in the present estimations.

*Remark 1* We are able to estimate the size of the jump of elements of $\tilde{X}_\pi$ at the grid points. For any $\tilde{x} \in \tilde{X}_\pi$ and $\tilde{c} = \mathcal{R}^{-1}\tilde{x}$, it holds

$$\begin{aligned}\|\tilde{x}_{j\kappa}\|_{C[t_{j-1},t_j)} &\le C_{h_j}\|\tilde{x}_{j\kappa}\|_{H^1(t_{j-1},t_j)} = C_{h_j}h_j^{1/2}|U_{j,red}\tilde{c}_{j\kappa}| \\ &\le C_{h_j}h_j^{1/2}\sigma_{\max}(U_{j,red})|\tilde{c}|\end{aligned}$$

with $C_{h_j} = \left(\max\{2/h_j, h_j\}\right)^{1/2}$. Here, we used [15, Lemma 3.2]. For sufficiently small $h_j$, this estimate reduces to

$$\|\tilde{x}_{j\kappa}\|_{C[t_{j-1},t_j)} \le \sqrt{2}\sigma_{\max}(\bar{\Gamma}\mathring{V})|\tilde{c}| = \sqrt{2}\sigma_{\max}(\Gamma V)|\tilde{c}|.$$

Let $x$ be any element of $X_\pi$ and $c = \mathcal{R}^{-1}x$. Replacing $\tilde{c}$ by $\Delta c = \tilde{c} - c$ in the last estimate, we obtain

$$\begin{aligned}|\tilde{x}_\kappa(t_{j-0}) - \tilde{x}_\kappa(t_{j+0})| &= |\tilde{x}_\kappa(t_{j-0}) - x_\kappa(t_{j-0}) + x_\kappa(t_{j+0}) - \tilde{x}_\kappa(t_{j+0})| \\ &\le \|\tilde{x}_{\kappa j} - x_{\kappa j}\|_{C[t_{j-1},t_j)} + \|\tilde{x}_{\kappa,j+1} - x_{\pi,\kappa,j+1}\|_{C[t_j,t_{j+1})} \\ &\le 2\sqrt{2}\sigma_{\max}(\Gamma V)|\Delta c|.\end{aligned}$$

Proposition 2 provides estimations for the factor $\sigma_{\max}(\Gamma V)$. In particular, for some bases, it does not depend on the polynomial degree $N$. □

## 4 Error estimation for the constrained minimization problem

The aim of this section is the derivation of bounds for perturbations of the solution $c$ for the problem (23)–(22), that is,

$$\begin{aligned}\varphi(z) = |\mathcal{A}z - r|^2 \to \min! \\ \text{subject to } \mathcal{C}z = 0,\end{aligned}$$

under perturbation of the data $\mathcal{A}, \mathcal{C}, r$. Such bounds are known for a long time, e.g., [11, 12]. However, we will provide different bounds in this section. The reason for this is that the constraint $\mathcal{C}c = 0$ has an exceptional meaning in the present context: It holds $\mathcal{C}c = 0$ if and only if $\mathcal{R}c \in H_D^1(a,b)$. If a perturbation $\Delta\mathcal{C}$ of $\mathcal{C}$ changes the kernel of $\mathcal{C}$, it does no longer hold $\mathcal{R}c \in H_D^1(a,b)$ in general! Therefore, we will consider the two cases $\ker(\mathcal{C}+\Delta\mathcal{C}) = \ker\mathcal{C}$ and $\ker(\mathcal{C}+\Delta\mathcal{C}) \ne \ker\mathcal{C}$ separately.

Let $\tilde{c}$ the solution of the perturbed problem

$$\min\{|(\mathcal{A}+\Delta\mathcal{A})z - (r+\Delta r)|^2 : (\mathcal{C}+\Delta\mathcal{C})z = 0\}. \tag{36}$$

Then, let $\Delta c = c - \tilde{c}$ denote the error. We are interested in deriving an error bound on $\Delta c$ in terms of the perturbations of the data.

Let us for a matrix $\mathcal{M}$, denote the Moore-Penrose inverse by $\mathcal{M}^+$. Moreover, let $\|\mathcal{M}\|$ be its spectral norm.

Let $\mathcal{D}$ be an orthonormal basis of $\ker\mathcal{C}$. Then, $P = I_{n(mN+k)} - \mathcal{C}^+\mathcal{C}$ is the orthogonal projector onto $\ker\mathcal{C}$ and $P\mathcal{D} = \mathcal{D}$. Some more properties are collected in the following proposition.

**Proposition 3** *It holds, for any matrix $\mathcal{M} \in \mathbb{R}^{\nu\times n(mN+k)}$, $\nu \in \mathbb{N}$,*

1. $\mathcal{D}^T\mathcal{D} = I_{nmN+k}$ *and* $\mathcal{D}\mathcal{D}^T = P$.
2. *If* $c = \mathcal{D}d$, *then* $|c| = |d|$.
3. $\|\mathcal{A}\mathcal{D}\| = \|\mathcal{A}P\|$.
4. $(\mathcal{A}P)^+ = \mathcal{D}(\mathcal{A}\mathcal{D})^+$.
5. $\|(\mathcal{A}P)^+\| = \|(\mathcal{A}\mathcal{D})^+\|$.

The proofs are obvious. For the following, we note that the matrix $\mathcal{A}\mathcal{D}$ has full column rank [9, Proposition 1].

### 4.1 $\ker(\mathcal{C} + \Delta\mathcal{C}) = \ker\mathcal{C}$

Each element $c$ of $\ker\mathcal{C}$ has a unique representation $c = \mathcal{D}d$ with $d \in \mathbb{R}^{nmN+k}$. Therefore, (23)–(22) is equivalent to the unconstrained minimization problem

$$\min_{d\in\mathbb{R}^{nmN+k}} \|\mathcal{A}\mathcal{D}d - r\| \tag{37}$$

while (36) becomes the unconstrained minimization problem

$$\min_{d\in\mathbb{R}^{nmN+k}} \|(\mathcal{A}+\Delta\mathcal{A})\mathcal{D}d - (r+\Delta r)\|. \tag{38}$$

Since $\mathcal{A}\mathcal{D}$ has full column rank, standard perturbation results for unconstrained least squares problems apply. As a consequence of [16, Satz 8.2.7] and Proposition 3, we obtain

**Theorem 4** *Let* $\omega = \|(\mathcal{A}P)^+\|\|\Delta\mathcal{A}P\| < 1$. *Then it holds*

$$|\Delta c| \le \frac{\|(\mathcal{A}P)^+\|}{1-\omega}\left\{\|\Delta\mathcal{A}P\|\left[|c| + \|(\mathcal{A}P)^+\||\mathfrak{r}| + |\Delta r|\right]\right\}$$

*and*

$$\frac{|\Delta c|}{|c|} \le \frac{1}{1-\omega}\left\{\left[\kappa_{\mathcal{C}}(\mathcal{A}) + \frac{|\mathfrak{r}|}{\|\mathcal{A}P\||c|}\kappa_{\mathcal{C}}(\mathcal{A})^2\right]\frac{\|\Delta\mathcal{A}P\|}{\|\mathcal{A}P\|} + \frac{\|(\mathcal{A}P)^+\||r|}{|c|}\cdot\frac{|\Delta r|}{|r|}\right\}.$$

*Here,* $\mathfrak{r} = r - \mathcal{A}c$ *and*

$$\kappa_{\mathcal{C}}(\mathcal{A}) = \|\mathcal{A}P\|\|(\mathcal{A}P)^+\|.$$

Theorem 4 corresponds to classical results for unconstrained minimization problems (e.g., [10], [13, Theorem 9.12] and is a small generalization of them. Let us emphasize that the estimation is independent of the perturbations of $\mathcal{C}$ as long as the null space of $\mathcal{C}$ is not changed by the perturbation.

*Remark 2* In the case of the Legendre basis, the elements of $\mathcal{C}$ consist only of three nonzero elements being equal to 1 and $-1$, respectively, possibly scaled by the step-sizes, cf. (16), (17). So we expect $\Delta\mathcal{C} = 0$ such that the estimates of this section apply. □

### 4.2 $\ker(\mathcal{C}+\Delta\mathcal{C}) \neq \ker \mathcal{C}$

The estimation of the error becomes much more involved than in the previous case. In a first step, we will construct a basis for the kernel of the perturbed constraint $(\mathcal{C}+\Delta\mathcal{C})z = 0$.

**Lemma 1** *Let* $\varkappa = \|\mathcal{C}^+\|\|\Delta\mathcal{C}\| < 1/2$. *Then* $\mathcal{C}+\Delta\mathcal{C}$ *has full rank and* $P_\Delta = I_{n(mN+k)} - (\mathcal{C}+\Delta\mathcal{C})^+(\mathcal{C}+\Delta\mathcal{C})$ *is a projector onto* $\ker(\mathcal{C}+\Delta\mathcal{C})$. *Furthermore,* $\mathcal{D}_\Delta = P_\Delta\mathcal{D}$ *is a basis of* $\ker(\mathcal{C}+\Delta\mathcal{C})$. *Moreover, the estimates*

$$\|(\mathcal{C}+\Delta\mathcal{C})^+\| \leq \frac{\|\mathcal{C}^+\|}{1-\varkappa}$$

*and*

$$\|(\mathcal{C}+\Delta\mathcal{C})^+ - \mathcal{C}^+\| \leq \frac{\sqrt{2}\|\mathcal{C}^+\|^2}{1-\varkappa}\|\Delta\mathcal{C}\|$$

*hold true.*

*Proof* The proposition of $\mathcal{C}+\Delta\mathcal{C}$ having full rank as well as the error estimates follow from [16, Satz 8.2.5].

For showing that $\mathcal{D}_\Delta$ is a basis of $\ker(\mathcal{C}+\Delta\mathcal{C})$ consider

$$\begin{aligned}(I-P_\Delta)P &= (\mathcal{C}+\Delta\mathcal{C})^+(\mathcal{C}+\Delta\mathcal{C})(I-\mathcal{C}^+\mathcal{C})\\ &= (\mathcal{C}+\Delta\mathcal{C})^+\Delta\mathcal{C}(I-\mathcal{C}^+\mathcal{C}).\end{aligned}$$

It holds

$$\|(I-P_\Delta)P\| \leq \|(\mathcal{C}+\Delta\mathcal{C})^+\|\|\Delta\mathcal{C}\| \leq \frac{\|\mathcal{C}^+\|}{1-\varkappa}\|\Delta\mathcal{C}\| \leq \frac{\varkappa}{1-\varkappa} < 1.$$

Therefore, the assumptions of [17, Theorem I-6.34] are fulfilled. Since $\dim\ker(\mathcal{C}+\Delta\mathcal{C}) = \dim\ker\mathcal{C}$,the first alternative of that theorem applies and $P_\Delta$ is a one-to-one mapping of $\ker\mathcal{C}$ onto $\ker(\mathcal{C}+\Delta\mathcal{C})$. Hence, $\mathcal{D}_\Delta$ is a basis of the latter space. □

By using the bases $\mathcal{D}$ and $\mathcal{D}_\Delta$, the unperturbed and the perturbed least squares problems become (37) and

$$\min_{d\in\mathbb{R}^{nmN+k}} \|(\mathcal{A}+\Delta\mathcal{A})\mathcal{D}_\Delta d - (r+\Delta r)\|. \tag{39}$$

In a first step, the deviations of the bases shall be estimated. It holds

$$\begin{aligned}P_\Delta - P &= \mathcal{C}^+\mathcal{C} - (\mathcal{C}+\Delta\mathcal{C})^+(\mathcal{C}+\Delta\mathcal{C})\\ &= \mathcal{C}^+\mathcal{C} - (\mathcal{C}+\Delta\mathcal{C})^+\mathcal{C} - (\mathcal{C}+\Delta\mathcal{C})^+\Delta\mathcal{C}\\ &= \left[\mathcal{C}^+ - (\mathcal{C}+\Delta\mathcal{C})^+\right]\mathcal{C} - (\mathcal{C}+\Delta\mathcal{C})^+\Delta\mathcal{C}.\end{aligned}$$

Invoking Lemma 1, we obtain[4]

$$\|P_\Delta - P\| \le \left[\frac{\sqrt{2}\|\mathcal{C}^+\|^2}{1-\varkappa}\|\mathcal{C}\| + \frac{\|\mathcal{C}^+\|}{1-\varkappa}\right]\|\Delta\mathcal{C}\| = \frac{\|\mathcal{C}^+\|}{1-\varkappa}\left[\sqrt{2}\kappa(\mathcal{C})+1\right]\|\Delta\mathcal{C}\|$$

with $\kappa(\mathcal{C}) = \|\mathcal{C}^+\|\|\mathcal{C}\|$. Consequently,

$$\|\mathcal{D}_\Delta - \mathcal{D}\| = \|(P_\Delta - P)\mathcal{D}\| \le \|P_\Delta - P\|\|\mathcal{D}\| \le \frac{\|\mathcal{C}^+\|}{1-\varkappa}\left[\sqrt{2}\kappa(\mathcal{C})+1\right]\|\Delta\mathcal{C}\|. \quad (40)$$

Let us transform (39) now. It holds

$$\begin{aligned}(\mathcal{A}+\Delta\mathcal{A})\mathcal{D}_\Delta &= (\mathcal{A}+\Delta\mathcal{A})\mathcal{D} + (\mathcal{A}+\Delta\mathcal{A})(\mathcal{D}_\Delta - \mathcal{D})\\ &= \mathcal{A}\mathcal{D} + \mathfrak{R}\end{aligned}$$

where $\mathfrak{R} = \Delta\mathcal{A}\mathcal{D} + (\mathcal{A}+\Delta\mathcal{A})(\mathcal{D}_\Delta - \mathcal{D})$. The representation of $\mathfrak{R}$ provides the estimate

$$\|\mathfrak{R}\| \le \|\Delta\mathcal{A}P\| + \|\mathcal{A}+\Delta\mathcal{A}\|\frac{\|\mathcal{C}^+\|}{1-\varkappa}\left[\sqrt{2}\kappa(\mathcal{C})+1\right]\|\Delta\mathcal{C}\|. \quad (41)$$

Denote $\omega_\Delta = \|(\mathcal{A}P)^+\|\|\mathfrak{R}\|$. The condition $\omega_\Delta < 1$ is obviously fulfilled if

$$\|(\mathcal{A}P)^+\|\left\{\|\Delta\mathcal{A}P\| + \|\mathcal{A}+\Delta\mathcal{A}\|\frac{\|\mathcal{C}^+\|}{1-\varkappa}\left[\sqrt{2}\kappa(\mathcal{C})+1\right]\|\Delta\mathcal{C}\|\right\} < 1. \quad (42)$$

Let $d + \Delta d$ be the solution of (39). Using the fact that $\mathcal{A}\mathcal{D}$ has full rank, Theorem 8.2.7 of [16] provides the estimates

$$|\Delta d| \le \frac{\|(\mathcal{A}P)^+\|}{1-\omega_\Delta}\left\{\|\mathfrak{R}\|\left[|d| + \|(\mathcal{A}P)^+\||\mathfrak{r}|\right] + |\Delta r|\right\} \quad (43)$$

and

$$\begin{aligned}\frac{|\Delta d|}{|d|} \le \frac{1}{1-\omega_\Delta}\Bigg\{&\left[\kappa_\mathcal{C}(\mathcal{A}) + \frac{|\mathfrak{r}|}{\|\mathcal{A}\mathcal{D}\||d|}\kappa_\mathcal{C}(\mathcal{A})^2\right]\frac{\|\mathfrak{R}\|}{\|\mathcal{A}\mathcal{D}\|}\\ &+ \frac{\|(\mathcal{A}\mathcal{D})^+\||r|}{|d|}\cdot\frac{|\Delta r|}{|r|}\Bigg\}.\end{aligned} \quad (44)$$

with $\mathfrak{r} = r - \mathcal{A}c$.

**Theorem 5** *Let $\|\Delta\mathcal{A}\|$ and $\|\Delta\mathcal{C}\|$ be sufficiently small such that (42) and $\varkappa = \|\mathcal{C}^+\|\|\Delta\mathcal{C}\| < 1/2$ hold true. Then it holds*

$$|\Delta c| \le \frac{\|(\mathcal{A}P)^+\|}{1-\omega_\Delta}\left\{\|\mathfrak{R}\|\left[|c| + \|(\mathcal{A}P)^+\||\mathfrak{r}|\right] + |\Delta r|\right\} + \frac{\|\mathcal{C}^+\|}{1-\varkappa}\left[\sqrt{2}\kappa(\mathcal{C})+1\right]\|\Delta\mathcal{C}\||c|$$

*and*

$$\begin{aligned}\frac{|\Delta c|}{|c|} \le \frac{1}{1-\omega_\Delta}\Bigg\{&\left[\kappa_\mathcal{C}(\mathcal{A}) + \frac{|\mathfrak{r}|}{\|\mathcal{A}P\||c|}\kappa_\mathcal{C}(\mathcal{A})^2\right]\frac{\|\mathfrak{R}\|}{\|\mathcal{A}P\|}\\ &+ \frac{\|(\mathcal{A}P)^+\||r|}{|c|}\cdot\frac{|\Delta r|}{|r|}\Bigg\} + \frac{\|\mathcal{C}^+\|}{1-\varkappa}\left[\sqrt{2}\kappa(\mathcal{C})+1\right]\|\Delta\mathcal{C}\|.\end{aligned}$$

[4] In case that $\ker(\mathcal{C}+\Delta\mathcal{C}) = \ker\mathcal{C}$ we obtain $\mathcal{P}_\Delta - \mathcal{P} = 0$ and $\mathcal{D}_\Delta = \mathcal{D}$ such that the present estimations coincide with those of the previous section.

*Proof* It holds $c = \mathcal{D}d$ and $\Delta c = \mathcal{D}_\Delta \Delta d + (\mathcal{D}_\Delta - \mathcal{D})d$ such that $|\Delta c| \le |\Delta d| + \|P_\Delta - P\||d|$. Inserting this estimate in (43) and (44) and using $|c| = |\mathcal{D}d| = |d|$ provides the claim. □

*Remark 3* $|\mathfrak{r}|$ is a measure for the accuracy of the discrete solution. Let $x_\pi \in X_\pi$ denote the discrete solution obtained by minimizing $\Phi_{\pi,M}$ (8). Its representation becomes $c = \mathcal{R}^{-1}x_\pi$. Then it holds $|\mathfrak{r}|^2 = |\mathcal{A}c - r|^2 = \Phi_{\pi,M}(x_\pi)$. Hence, $\Phi_{\pi,M}(x_\pi) \le 2(\Phi_{\pi,M}(x_*) + \Phi_{\pi,M}(x_\pi - x_*))$. Under the conditions of Theorem 1, it holds, therefore, $|\mathfrak{r}| \le ch^{N-\mu+1}$. □

The critical quantities to estimate the influence of perturbations are $\kappa_{\mathcal{C}}(\mathcal{A})$ and $\|\mathcal{C}^+\|$, $\kappa(\mathcal{C})$ as well as $\|(\mathcal{A}P)^+\|$. The norms of $\mathcal{C}$ and its pseudoinverse depend only on the choice of $X_\pi$ and the basis chosen for it, but not on the DAE. It holds $\|\mathcal{C}\| = \sigma_{\max}(\mathcal{C})$ and $\|\mathcal{C}^+\| = \sigma_{\min}(\mathcal{C})^{-1}$ with $\sigma_{\min}(\mathcal{C})$ being the smallest nonvanishing singular value of $\mathcal{C}$. Since $\mathcal{C}$ has full row rank, $\sigma_{\min}(\mathcal{C}) = \left(\lambda_{\min}(\mathcal{C}\mathcal{C}^T)\right)^{1/2}$ and $\sigma_{\max}(\mathcal{C}) = \left(\lambda_{\max}(\mathcal{C}\mathcal{C}^T)\right)^{1/2}$.

With $\mathcal{C}$ from (22) we observe that

$$\mathcal{C} = \Pi_1 \left[I_k \otimes \mathcal{C}_{\mathrm{s}} | \mathcal{O}_{\mathrm{s}}\right] \Pi_2$$

with

$$\mathcal{C}_{\mathrm{s}} = \begin{bmatrix} \bar{\mathcal{P}}_1(t_1) & -\bar{\mathcal{P}}_2(t_1) & & & & \\ & \bar{\mathcal{P}}_2(t_2) & -\bar{\mathcal{P}}_3(t_2) & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & \ddots & \\ & & & & \bar{\mathcal{P}}_{n-1}(t_{n-1}) & -\bar{\mathcal{P}}_n(t_{n-1}) \end{bmatrix} \in \mathbb{R}^{(n-1)\times n(N+1)}$$

and $\mathcal{O}_{\mathrm{s}} \in \mathbb{R}^{k(n-1)\times nN(m-k)}$ consists entirely of zero elements. The permutation matrices $\Pi_1$ and $\Pi_2$ are constructed as follows: Let $x = [x_1, x_2, \ldots, x_m]^T \in \tilde{X}_\pi$. First, the equations in $\mathcal{C}c = 0$ are reordered such that first all equations related to the first component $x_1$, then those of $x_2$, and so on until $x_k$ are available. This reordering is expressed via $\Pi_1$. The column permutation $\Pi_2$ reorders the coefficients such that the ones describing the differential components are taken first, and then the ones belonging to the algebraic components. In particular, the coefficients $c^\kappa$ describing $x_\kappa$ are given by $c^\kappa = [c_{1\kappa 0}, c_{1,\kappa 1}, \ldots, c_{1\kappa N}, c_{2\kappa 0}, \ldots, c_{n\kappa N}]^T$. Then we have

$$\mathcal{C}\mathcal{C}^T = \Pi_1 \left[I_k \otimes \mathcal{C}_{\mathrm{s}} | \mathcal{O}_{\mathrm{s}}\right] \Pi_2 \Pi_2^T \begin{bmatrix} I_k \otimes \mathcal{C}_{\mathrm{s}}^T \\ \mathcal{O}_{\mathrm{s}}^T \end{bmatrix} \Pi_1^T = \Pi_1 (I_k \otimes \mathcal{C}_{\mathrm{s}}\mathcal{C}_{\mathrm{s}}^T) \Pi_1^T, \quad (45)$$

Using (16) and (17), it holds

$$\mathcal{C}_{\mathrm{s}} = C_{\mathrm{s}} \mathrm{diag}(h_1, \ldots, h_n)$$

with

$$C_{\mathrm{s}} = \begin{bmatrix} f & -e_1^T & & \\ & f & -e_1^T & \\ & & \ddots & \ddots \\ & & & f \quad -e_1^T \end{bmatrix} \tag{46}$$

where $e_1$ is the first unit vector and $f = [1, \int_0^1 p_0(\sigma)d\sigma, \ldots, \int_0^1 p_{N-1}(\sigma)d\sigma]$. This leads to

$$\mathcal{C}_{\mathrm{s}}\mathcal{C}_{\mathrm{s}}^T = \begin{bmatrix} h_1^2|f|^2 + h_2^2 & -h_2^2 & & \\ -h_2^2 & h_2^2|f|^2 + h_3^2 & -h_3^2 & \\ & \ddots & \ddots & \ddots \\ & & -h_{n-1}^2 & h_{n-1}^2|f|^2 + h_n^2 \end{bmatrix}. \tag{47}$$

The eigenvalues of $\mathcal{C}\mathcal{C}^T$ are those of (47).

**Proposition 4** *Let the grid* (3) *have the maximal stepsize h and the minimal stepsize* $h_{\min}$. *Then it holds*

*(1)* $|f| > 1$.
*(2)* $0 < h_{\min}^2(|f|^2 - 1) \le \lambda_{\min}(\mathcal{C}_s\mathcal{C}_s^T)$ *and* $\lambda_{\max}(\mathcal{C}_s\mathcal{C}_s^T) \le h^2(|f|^2 + 3)$.

*Proof* Since the first component of $f$ is equal to 1, we have $|f| \ge 1$ and $|f| = 1$ if and only if $\int_0^1 p_0(\sigma)d\sigma = \cdots = \int_0^1 p_{N-1}(\sigma)d\sigma = 0$. Assume that the latter condition holds true. This means in particular that $p_0, \ldots, p_{N-1}$ are orthogonal to the polynomial $p(\tau) \equiv 1 \in \mathfrak{P}_{N-1}$. The latter space has dimension $N$. Since $p_0, \ldots, p_{N-1} \in \mathfrak{P}_{N-1}$ are $N$ polynomials being orthogonal to $p$, they must be linearly dependent in contradiction to the assumption that they form a basis. This proves (1).

In order to prove (2), we observe that $\mathcal{C}_s\mathcal{C}_s^T$ is symmetric such that all eigenvalues are real. Invoking Gershgorin's circle theorem [16, Theorem 1.2.10], the eigenvalues $\lambda$ of $\mathcal{C}_s\mathcal{C}_s^T$ fulfill

$$\min_{j=1,\ldots,n-1} h_j^2(|f|^2 - 1) \le \lambda \le \max_{j=1,\ldots,n-1} h_j^2(|f|^2 + 1) + 2h_{j+1}^2.$$

This proves (2). □

We obtain immediately the following corollary. Note that $f$ depends only on $N$ and the chosen basis, but not on the grid.

**Corollary 1** *Let the grids* (3) *be quasiuniform, that is* $h/h_{\min} \le \rho < \infty$ *with* $\rho$ *independent of* $\pi$. *Then it holds* $\kappa(\mathcal{C}) \le \rho\left(\frac{|f|^2+3}{|f|^2-1}\right)^{1/2}$ *and* $\|\mathcal{C}^+\| \le h_{\min}(|f|^2 - 1)^{-1/2}$.

For constant stepsize $h$, we have $\mathcal{C}_s\mathcal{C}_s^T = h^2 C_s C_s^T$, which is a Toeplitz tridiagonal matrix. In this case, the eigenvalues of $C_s C_s^T$ are given by [18, Theorem 2.2]

$$\lambda_j = 1 + |f|^2 - 2\cos\left(\frac{j\pi}{n}\right), \quad j = 1, \ldots, n-1. \tag{48}$$

**Proposition 5** *Let the grid* (3) *be equidistant with stepsize h, and* $C_s$ *be given by* (46)*. Then it holds*

- *For the Legendre basis* $1 \le \lambda_{\min}(C_s C_s^T) \le \lambda_{\max}(C_s C_s^T) \le 5$*;*
- *For the modified Legendre basis* $2N \le \lambda_{\min}(C_s C_s^T) \le \lambda_{\max}(C_s C_s^T) \le 2N + 6$*;*
- *For the Chebyshev basis* $1 \le \lambda_{\min}(C_s C_s^T) \le \lambda_{\max}(C_s C_s^T) \le 4 + 2\ln 2$.
- *For the Runge-Kutta basis assume additionally that* $\int_0^1 p_i(\sigma)d\sigma \ge 0$*,* $i = 0, 1, \ldots, N-1$*. Then* $N^{-1} \le \lambda_{\min}(C_s C_s^T) \le \lambda_{\max}(C_s C_s^T) \le 5$.

*Proof* In the case of the Legendre basis, it holds $f = [1, 1, 0, \ldots, 0]$. Hence, $|f|^2 = 2$ such that the statement follows.

For the modified Legendre basis, we have $f = [1, 2, 0, 2, 0, \ldots]$ such that

$$|f|^2 = \begin{cases} 2N+1, & N \text{ even}, \\ 2N+3, & N \text{ odd}. \end{cases}$$

For the Chebyshev basis, we observe

$$\int_0^1 p_i(\sigma)d\sigma = \begin{cases} \frac{1}{2}\frac{1+(-1)^i}{1-i^2}, & i \neq 1, \\ 0, & i = 1. \end{cases}$$

This leads to $f = [1, 1, 0, -\frac{1}{3}, 0, -\frac{1}{8}, 0, \ldots]$. Hence,

$$2 \le |f|^2 \le 2 + \sum_{i=1}^{\infty}\left(\frac{1}{1-(2i)^2}\right)^2 \le 2 + \sum_{i=1}^{\infty}\frac{1}{i(4i^2-1)} = 2 + 2\ln 2 - 1.$$

For the sum of the series, cf. [19, p. 269, series 110.d]. This provides the estimate for the Chebyshev basis.

In case of the Runge-Kutta basis, it holds $\sum_{i=0}^{N-1} p_i(\sigma) \equiv 1$. With $f = [1, f_2, \ldots, f_{N+1}]$ it holds then $f_i \ge 0$ and $\sum_{i=2}^{N+1} f_i = 1$. Hence,

$$\frac{1}{N} = \frac{1}{N}\left(\sum_{i=2}^{N+1} f_i\right)^2 \le \sum_{i=2}^{N+1} f_i^2 \le \sum_{i=2}^{N+1} f_i = 1.$$

This yields $1 + N^{-1} \le |f|^2 \le 2$ and the claim follows. □

*Remark 4* For the Runge-Kutta basis, the values $f_i = \int_0^1 p_{i-1}(\sigma)d\sigma$ are just the weights of the interpolatory quadrature rule corresponding to the nodes $\tau_1, \ldots, \tau_N$ of (18). For a number of common choices of nodes, these weights are known to be positive. Examples are the Gauss-Legendre nodes, Radau nodes, and Lobatto nodes [20, Section 2.7]. It holds also true for Chebyshev nodes and many others; see, e.g., [20, pp. 85f]. □

Note that the claims of Proposition 5 could also be shown using Gershgorin's theorem. This indicates that the estimates of Proposition 4 are rather tight.

**Corollary 2** *For equidistant grids* (3)*, it holds*

- *For the Legendre basis* $\kappa(\mathcal{C}) \leq \sqrt{5}$ *and* $\|\mathcal{C}^+\| \leq h^{-1}$*;*
- *For the modified Legendre basis* $\kappa(\mathcal{C}) \leq \left(\frac{2N+6}{2N}\right)^{1/2}$ *and* $\|\mathcal{C}^+\| \leq (2N)^{-1/2}h^{-1}$*;*
- *For the Chebyshev basis* $\kappa(\mathcal{C}) \leq (4 + 2\ln 2)^{1/2} \approx 2.32$ *and* $\|\mathcal{C}^+\| \leq h^{-1}$.
- *For the Runge-Kutta basis* $\kappa(\mathcal{C}) \leq (5N)^{1/2}$ *and* $\|\mathcal{C}^+\| \leq N^{1/2}h^{-1}$ *provided that* $\int_0^1 p_i(\sigma)d\sigma \geq 0$*,* $i = 0, 1, \ldots, N-1$.

It should be emphasized again that, if $\ker(\mathcal{C}+\Delta\mathcal{C}) \neq \ker\mathcal{C}$, it cannot be guaranteed that the solution of the perturbed problem $\mathcal{R}(c + \Delta c)$ belongs to $X_\pi$. Instead, it belongs to $\tilde{X}_\pi$, only. Simple projection algorithms of elements of $\tilde{X}_\pi$ onto $X_\pi$ can be derived. In our experiments so far, these projections did not lead to a better accuracy than the unprojected numerical solutions.

## 5 Some examples

### 5.1 Conditioning of the representation map $\mathcal{R}$

For each selection $\{p_0, \ldots, p_{N-1}\}$ of basis polynomials, the conditioning of the representation map depends both on the grid and on $N$. For simplicity, we assume here that an equidistant grid with stepsize $h$ is used for defining $X_\pi$. Besides the bases introduced before, we will additionally consider the Runge-Kutta basis with uniform interpolation points as used in our very first paper on the subject [1].

The norms of the representation map and its inverse have been computed for both settings (mapping into $L^2((a,b),\mathbb{R}^m)$ and $H_D^1(a,b)$) and for polynomial degrees $N = 3, 5, 10, 20$ and $h = n^{-1}$ where $N = 10, 20, 40, 80, 160, 320$. These are the first observations:

- $\sigma_{\min}(\hat{\mathcal{U}})$ is independent of the chosen basis and independent of $N$ for $h \leq 0.1$. However, this is not true for larger stepsizes, cf. Table 2.
- For every basis, $\sigma_{\max}(\mathcal{U}) \approx \sigma_{\max}(\hat{\mathcal{U}})$ up to a relative error below $10^{-3}$. This coincides with the findings of Theorem 3.

In Tables 1, 2, 3, 4, 5, and 6, we present more detailed results. From these tables, we can draw the following conclusions:

- The asymptotic behavior with respect to the stepsize $h$ as indicated in Theorem 3 is clearly visible.
- For both the Legendre and the Chebyshev bases, $\sigma_{\max}(\mathcal{U})$ and $\sigma_{\max}(\hat{\mathcal{U}})$ do not depend on $N$. This is reasonable for the Legendre basis if Proposition 2 is taken into account.
- The asymptotics of $\sigma_{\min}(\mathcal{U})$ coincides with the results of Theorem 3 and Proposition 2 for the modified Legendre basis.

**Table 1** $\sigma_{\min}(\hat{\mathcal{U}})$

| $n = 1/h$ | $\sigma_{\min}(\hat{\mathcal{U}})$ |
|---|---|
| 10 | 3.16e-2 |
| 20 | 1.12e-2 |
| 40 | 3.95e-3 |
| 80 | 1.40e-3 |
| 160 | 4.94e-4 |
| 320 | 1.75e-4 |

- The norm of the representation map behaves similarly for all considered bases. Not unexpectedly, an exception is the Runge-Kutta basis for uniform nodes, which has a much larger norm than that for other bases. When comparing $\sigma_{\min}(\mathcal{U})$ and $\sigma_{\max}(\mathcal{U})$ for different bases, we observe that the difference between the Legendre basis and the Chebyshev basis on one hand and the modified Legendre basis on the other hand it seems that they have different scaling only, but their conditioning (being the product of the norms of the representation map and its inverse) is similar. A similar property holds for $\hat{\mathcal{U}}$.
- The Runge-Kutta basis has surprisingly good properties. However, this property depends on the representation with respect to an orthogonal polynomial basis (in

**Table 2** $\sigma_{\min}(\hat{\mathcal{U}})$. The column headings denote the Legendre basis (*L*), the modified Legendre basis (*mL*), the Chebyshev basis (*Ch*), the Runge-Kutta basis (*RK*), and the Runge-Kutta basis with uniform nodes (*RKu*)

| $n = 1/h$ | L | mL | Ch | RK | RKu |
|---|---|---|---|---|---|
| $N = 3$ | | | | | |
| 1 | 4.47e-1 | 8.56e-1 | 5.52e-1 | 4.05e-1 | 5.57e-1 |
| 3 | 1.88e-1 | 1.89e-1 | 1.88e-1 | 1.87e-1 | 1.88e-1 |
| 5 | 8.88e-2 | 8.88e-2 | 8.88e-2 | 8.88e-2 | 8.88e-2 |
| $N = 5$ | | | | | |
| 1 | 3.33e-1 | 8.56e-1 | 4.31e-1 | 2.54e-1 | 4.12e-1 |
| 3 | 1.88e-1 | 1.89e-1 | 1.88e-1 | 1.46e-1 | 1.85e-1 |
| 5 | 8.88e-2 | 8.88e-2 | 8.88e-2 | 8.86e-2 | 8.87e-2 |
| $N = 10$ | | | | | |
| 1 | 2.29e-1 | 8.56e-1 | 2.93e-1 | 1.31e-1 | 2.33e-1 |
| 3 | 1.32e-1 | 1.89e-1 | 1.69e-1 | 7.57e-2 | 1.35e-1 |
| 5 | 8.88e-2 | 8.88e-2 | 8.88e-2 | 5.86e-2 | 8.84e-2 |
| $N = 20$ | | | | | |
| 1 | 1.60e-1 | 8.56e-1 | 2.10e-1 | 6.65e-2 | 1.50e-1 |
| 3 | 9.24e-2 | 1.89e-1 | 1.21e-1 | 3.84e-2 | 8.68e-2 |
| 5 | 7.16e-2 | 8.88e-2 | 1.21e-1 | 3.84e-2 | 8.68e-2 |

**Table 3** $\sigma_{\min}(\mathcal{U})$. The column headings denote the Legendre basis (*L*), the modified Legendre basis (*mL*), the Chebyshev basis (*Ch*), the Runge-Kutta basis (*RK*), and the Runge-Kutta basis with uniform nodes (*RKu*)

| $n = 1/h$ | L | mL | Ch | RK | RKu |
|---|---|---|---|---|---|
| $N = 3$ | | | | | |
| 10 | 1.17e-3 | 6.83e-3 | 1.35e-3 | 9.81e-4 | 1.93e-3 |
| 20 | 4.12e-4 | 2.42e-3 | 4.78e-4 | 3.47e-4 | 6.84e-4 |
| 40 | 1.46e-4 | 8.54e-4 | 1.69e-4 | 1.23e-4 | 2.42e-4 |
| 80 | 5.15e-5 | 3.02e-4 | 5.98e-5 | 4.33e-5 | 8.55e-5 |
| 160 | 1.82e-5 | 1.07e-4 | 2.11e-5 | 1.53e-5 | 3.02e-5 |
| 320 | 6.44e-6 | 3.78e-5 | 7.47e-6 | 5.42e-6 | 1.07e-5 |
| $N = 5$ | | | | | |
| 10 | 4.51e-4 | 4.76e-3 | 5.06e-4 | 2.96e-4 | 1.00e-3 |
| 20 | 1.59e-4 | 1.68e-3 | 1.79e-4 | 1.04e-4 | 3.54e-4 |
| 40 | 5.63e-5 | 5.95e-4 | 6.32e-5 | 3.69e-5 | 1.25e-4 |
| 80 | 1.99e-5 | 2.10e-4 | 2.23e-5 | 1.31e-5 | 4.43e-5 |
| 160 | 7.04e-6 | 7.44e-5 | 7.90e-6 | 4.62e-6 | 1.56e-5 |
| 320 | 2.49e-6 | 2.63e-5 | 2.79e-6 | 1.63e-6 | 5.53e-6 |
| $N = 10$ | | | | | |
| 10 | 1.08e-4 | 2.73e.3 | 1.10e-4 | 4.94e-5 | 2.35e-4 |
| 20 | 3.83e-5 | 9.59e-4 | 3.90e-5 | 1.75e-5 | 8.29e-5 |
| 40 | 1.36e-5 | 3.39e-4 | 1.38e-5 | 6.17e-6 | 2.93e-5 |
| 80 | 4.79e-6 | 1.20e-4 | 4.88e-6 | 2.18e-6 | 1.04e-5 |
| 160 | 1.69e-6 | 4.24e-5 | 1.72e-6 | 7.71e-7 | 3.66e-6 |
| 320 | 5.99e-7 | 1.50e-5 | 6.10e-7 | 2.73e-7 | 1.30e-6 |
| $N = 20$ | | | | | |
| 10 | 2.28e-5 | 1.46e-3 | 2.30e-5 | 7.26e-6 | 5.73e-5 |
| 20 | 8.06e-6 | 5.16e-4 | 8.12e-6 | 2.57e-6 | 2.03e-5 |
| 40 | 2.85e-6 | 1.82e-4 | 2.87e-6 | 9.08e-7 | 7.17e-6 |
| 80 | 1.01e-6 | 5.45e-5 | 1.01e-6 | 3.21e-7 | 2.53e-6 |
| 160 | 3.56e-7 | 2.28e-5 | 3.59e-7 | 1.13e-7 | 8.96e-7 |
| 320 | 1.26e-7 | 8.06e-6 | 1.27e-7 | 4.01e-8 | 3.17e-7 |

the present example, Chebyshev polynomials). Thus, it is much more expensive to work with it compared to using Legendre or Chebyshev bases directly. □

### 5.2 Conditioning of the constrained minimization problems

In order to provide a first insight into the conditioning of the constrained minimization problem (23)–(22), we computed the condition numbers $\kappa_{\mathcal{C}}(\mathcal{A})$ which have crucial importance for the behavior of the computational error. Discussions of $\kappa(\mathcal{C})$

**Table 4** $\sigma_{\max}(\hat{\mathcal{U}}) = \sigma_{\max}(\mathcal{U})$. The column headings denote the Legendre basis (*L*), the modified Legendre basis (*mL*), the Chebyshev basis (*Ch*), the Runge-Kutta basis (*RK*), and the Runge-Kutta basis with uniform nodes (*RKu*)

| $n = 1/h$ | L | mL | Ch | RK | RKu |
|---|---|---|---|---|---|
| $N = 3$ | | | | | |
| 10 | 3.16e-1 | 1.57e+0 | 3.41e-1 | 2.19e-1 | 2.64e-1 |
| 20 | 2.24e-1 | 1.11e+0 | 2.41e-1 | 1.55e-1 | 1.86e-1 |
| 40 | 1.58e-1 | 7.87e-1 | 1.70e-1 | 1.10e-1 | 1.32e-1 |
| 80 | 1.12e-1 | 5.56e-1 | 1.20e-1 | 7.74e-2 | 9.32e-2 |
| 160 | 7.91e-2 | 3.93e-1 | 8.52e-2 | 5.48e-2 | 6.59e-2 |
| 320 | 5.59e-2 | 2.78e-1 | 6.02e-2 | 3.87e-2 | 4.66e-2 |
| $N = 5$ | | | | | |
| 10 | 3.16e-1 | 2.69e+0 | 3.41e-1 | 1.74e-1 | 3.96e-1 |
| 20 | 2.24e-1 | 1.90e+1 | 2.41e-1 | 1.23e-1 | 2.80e-1 |
| 40 | 1.58e-1 | 1.35e+1 | 1.70e-1 | 8.70e-2 | 1.98e-1 |
| 80 | 1.12e-1 | 9.51e-1 | 1.20e-1 | 6.15e-2 | 1.40e-1 |
| 160 | 7.91e-2 | 6.73e-1 | 8.52e-2 | 4.35e-2 | 9.90e-2 |
| 320 | 5.59e-2 | 4.76e-1 | 6.02e-2 | 3.08e-2 | 7.00e-2 |
| $N = 10$ | | | | | |
| 10 | 3.16e-1 | 6.26e+0 | 3.41e-1 | 1.25e-1 | 3.86e+0 |
| 20 | 2.24e-1 | 4.43e+0 | 2.41e-1 | 8.81e-2 | 2.73e+0 |
| 40 | 1.58e-1 | 3.13e+0 | 1.70e-1 | 6.23e-2 | 1.93e+0 |
| 80 | 1.12e-1 | 2.21e+0 | 1.20e-1 | 4.40e-2 | 1.36e+0 |
| 160 | 7.91e-2 | 1.57e+0 | 8.52e-2 | 3.11e-2 | 9.64e-1 |
| 320 | 5.59e-2 | 1.11e+0 | 6.02e-2 | 2.20e-2 | 6.82e-1 |
| $N = 20$ | | | | | |
| 10 | 3.16e-1 | 1.60e+1 | 3.41e-1 | 8.85e-2 | 1.47e+3 |
| 20 | 2.24e-1 | 1.13e+1 | 2.41e-1 | 6.26e-2 | 1.04e+3 |
| 40 | 1.58e-1 | 7.98e+0 | 1.70e-1 | 4.43e-2 | 7.37e+2 |
| 80 | 1.12e-1 | 5.64e+0 | 1.20e-1 | 3.13e-2 | 5.21e+2 |
| 160 | 7.91e-2 | 3.99e+0 | 8.52e-2 | 2.21e-2 | 3.68e+2 |
| 320 | 5.59e-2 | 2.82e+0 | 6.02e-2 | 1.56e-2 | 2.61e+2 |

and $\|\mathcal{C}^+\|$ have been provided earlier (Proposition 5 and Corollary 2). The examples below are chosen from our earlier investigations that led to surprisingly accurate results.

As done before, we use the bases as introduced in Section 5.1. We abandon the use the Runge-Kutta basis with uniform nodes since this basis has a bad conditioning. We choose $M = N + 1$ and the Gauss-Legendre nodes as collocation points (6). For this choice, $\Phi^R_{\pi,M} = \Phi^I_{\pi,M}$ (see (12), (11)) and $\kappa_{\mathcal{C}}(\mathcal{A})$ is identical for both choices.

**Table 5** $\kappa(\hat{\mathcal{U}}) = \sigma_{\max}(\hat{\mathcal{U}})/\sigma_{\min}(\hat{\mathcal{U}})$. The column headings denote the Legendre basis (*L*), the modified Legendre basis (*mL*), the Chebyshev basis (*Ch*), the Runge-Kutta basis (*RK*), and the Runge-Kutta basis with uniform nodes (*RKu*)

| $n = 1/h$ | L | mL | Ch | RK | RKu |
|---|---|---|---|---|---|
| $N = 3$ | | | | | |
| 10 | 1.03e+1 | 4.98e+1 | 1.08e+1 | 6.95e+0 | 8.35e+0 |
| 20 | 2.00e+1 | 9.95e+1 | 2.16e+1 | 1.39e+1 | 1.67e+1 |
| 40 | 4.00e+1 | 1.99e+2 | 4.31e+1 | 2.77e+1 | 3.34e+1 |
| 80 | 8.00e+1 | 3.98e+2 | 8.62e+1 | 5.54e+1 | 6.67e+1 |
| 160 | 1.60e+2 | 7.96e+2 | 1.72e+2 | 1.11e+2 | 1.33e+2 |
| 320 | 3.20e+2 | 1.59e+3 | 3.45e+2 | 2.22e+2 | 2.67e+2 |
| $N = 5$ | | | | | |
| 10 | 1.00e+1 | 8.53e+1 | 1.08e+1 | 5.52e+0 | 1.25e+1 |
| 20 | 2.00e+1 | 1.70e+2 | 2.16e+1 | 1.10e+1 | 2.51e+1 |
| 40 | 4.00e+1 | 3.40e+2 | 4.31e+1 | 2.20e+1 | 5.01e+1 |
| 80 | 8.00e+1 | 6.81e+2 | 8.62e+1 | 4.40e+1 | 1.00e+2 |
| 160 | 1.60e+2 | 1.36e+3 | 1.72e+2 | 8.80e+1 | 2.00e+2 |
| 320 | 3.20e+2 | 2.72e+3 | 3.45e+2 | 1.76e+2 | 4.01e+2 |
| $N = 10$ | | | | | |
| 10 | 1.00e+1 | 1.98e+2 | 1.08e+1 | 3.95e+0 | 1.22e+2 |
| 20 | 2.00e+1 | 3.96e+2 | 2.16e+1 | 7.88e+0 | 2.44e+2 |
| 40 | 4.00e+1 | 7.92e+2 | 4.31e+1 | 1.58e+1 | 4.88e+2 |
| 80 | 8.00e+1 | 1.58e+3 | 8.62e+1 | 3.15e+1 | 9.76e+2 |
| 160 | 1.60e+2 | 3.17e+3 | 1.72e+2 | 6.30e+1 | 1.95e+3 |
| 320 | 3.20e+2 | 6.34e+3 | 3.45e+2 | 1.26e+2 | 3.90e+3 |
| $N = 20$ | | | | | |
| 10 | 1.00e+1 | 5.05e+2 | 1.08e+1 | 4.21e+0 | 4.67e+4 |
| 20 | 2.00e+1 | 1.01e+3 | 2.16e+1 | 5.60e+0 | 9.32e+4 |
| 40 | 4.00e+1 | 2.02e+3 | 4.31e+1 | 1.12e+1 | 1.86e+5 |
| 80 | 8.00e+1 | 4.04e+3 | 8.62e+1 | 2.24e+1 | 3.73e+5 |
| 160 | 1.60e+2 | 8.07e+3 | 1.72e+2 | 4.48e+1 | 7.46e+5 |
| 320 | 3.20e+2 | 1.61e+4 | 3.45e+2 | 9.00e+1 | 1.49e+6 |

*Example 1* The first example is an index-3 DAE without dynamic degrees of freedom. It has been used before in numerous papers, e.g., [1, 2, 8]. The problem is given by

$$\begin{aligned} x_2'(t) + x_1(t) &= q_1(t), \\ t\eta x_2'(t) + x_3'(t) + (\eta+1)x_2(t) &= q_2(t), \\ t\eta x_2(t) + x_3(t) &= 7q_3(t), \quad t \in [0,1]. \end{aligned}$$

**Table 6** $\kappa(\mathcal{U}) = \sigma_{\max}(\mathcal{U})/\sigma_{\min}(\mathcal{U})$. The column headings denote the Legendre basis (*L*), the modified Legendre basis (*mL*), the Chebyshev basis (*Ch*), the Runge-Kutta basis (*RK*), and the Runge-Kutta basis with uniform nodes (*RKu*)

| $n = 1/h$ | L | mL | Ch | RK | RKu |
|---|---|---|---|---|---|
| $N = 3$ | | | | | |
| 10 | 2.71e+2 | 2.30e+2 | 2.52e+2 | 2.23e+2 | 1.36e+2 |
| 20 | 5.43e+2 | 4.61e+2 | 5.04e+2 | 4.47e+2 | 2.73e+2 |
| 40 | 1.09e+3 | 9.21e+2 | 1.01e+3 | 8.93e+2 | 5.45e+2 |
| 80 | 2.17e+3 | 1.84e+3 | 2.01e+3 | 1.79e+3 | 1.09e+3 |
| 160 | 4.34e+3 | 3.68e+3 | 4.03e+3 | 3.57e+3 | 2.18e+3 |
| 320 | 8.68e+3 | 7.37e+3 | 8.06e+3 | 7.15e+3 | 4.36e+3 |
| $N = 5$ | | | | | |
| 10 | 7.03e+2 | 5.65e+2 | 6.74e+2 | 5.89e+2 | 3.96e+2 |
| 20 | 1.40e+3 | 1.13e+3 | 1.35e+3 | 1.18e+3 | 7.91e+2 |
| 40 | 2.81e+3 | 2.26e+3 | 2.70e+3 | 2.36e+3 | 1.58e+3 |
| 80 | 5.61e+3 | 4.52e+3 | 5.39e+3 | 4.71e+3 | 3.16e+3 |
| 160 | 1.12e+4 | 9.05e+3 | 1.08e+4 | 9.42e+3 | 6.33e+3 |
| 320 | 2.25e+4 | 1.81e+4 | 2.16e+4 | 1.88e+4 | 1.27e+4 |
| $N = 10$ | | | | | |
| 10 | 2.92e+3 | 2.31e+3 | 3.09e+3 | 2.53e+3 | 1.65e+4 |
| 20 | 5.83e+3 | 4.62e+3 | 6.18e+3 | 5.05e+3 | 3.29e+4 |
| 40 | 1.17e+4 | 9.23e+3 | 1.24e+4 | 1.01e+4 | 6.58e+4 |
| 80 | 2.33e+4 | 1.85e+4 | 2.47e+4 | 2.02e+4 | 1.32e+5 |
| 160 | 4.67e+4 | 3.69e+4 | 4.94e+4 | 4.04e+4 | 2.63e+5 |
| 320 | 9.33e+4 | 7.39e+4 | 9.88e+4 | 8.07e+4 | 5.26e+5 |
| $N = 20$ | | | | | |
| 10 | 1.39e+4 | 1.09e+4 | 1.48e+4 | 1.22e+4 | 2.57e+7 |
| 20 | 2.78e+4 | 2.19e+4 | 2.97e+4 | 2.44e+4 | 5.14e+7 |
| 40 | 5.55e+4 | 4.37e+4 | 5.94e+4 | 4.87e+4 | 1.03e+8 |
| 80 | 1.11e+5 | 8.74e+4 | 1.19e+5 | 9.75e+4 | 2.06e+8 |
| 160 | 2.22e+5 | 1.75e+5 | 2.37e+5 | 1.95e+5 | 4.11e+8 |
| 320 | 4.44e+5 | 3.50e+5 | 4.75e+5 | 3.90e+5 | 8.22e+8 |

For unique solvability, no boundary or initial conditions are necessary. We choose the exact solution

$$\begin{aligned} x_{*,1}(t) &= e^{-t}\sin t, \\ x_{*,2}(t) &= e^{-2t}\sin t, \\ x_{*,3}(t) &= e^{-t}\cos t \end{aligned}$$

and adapt the right-hand side $q$ accordingly. In Table 7, the values of $\kappa_{\mathcal{C}}(\mathcal{A})$ for $\Phi^R_{\pi,M}$ and $\Phi^C_{\pi,M}$ are provided. It turns out that the behavior for different functionals

is comparable. Therefore, in the following examples, we present only the values for $\Phi^R_{\pi,M}$. □

*Example 2* We continue with an example of a Hessenberg index-2 system used previously in [1]. Consider the DAE system

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

with the right-hand side $q$ chosen in such a way that

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

**Table 7** $\kappa_C(\mathcal{A})$ for $\Phi^R_{\pi,M}$ and $\Phi^C_{\pi,M}$. Here, *L* denotes the Legendre basis, *mL* the modified Legendre basis, *Ch* the Chebyshev basis, and *RK* the Runge-Kutta basis. The smallest values are set in boldface

| $n = 1/h$ | $\kappa_C(\mathcal{A})$ for $\Phi^R_{\pi,M}$ | | | | $\kappa(\mathcal{A})$ for $\Phi^C_{\pi,M}$ | | | |
|---|---|---|---|---|---|---|---|---|
| | L | mL | Ch | RK | L | mL | Ch | RK |
| $N = 3$ | | | | | | | | |
| 10 | 5.77e+4 | 5.76e+4 | 6.22e+4 | **4.96e+4** | 6.01e+4 | 7.04e+4 | 6.04e+4 | **4.53e+4** |
| 20 | 2.37e+5 | 2.40e+5 | 2.55e+5 | **2.03e+5** | 2.47e+5 | 2.93e+5 | 2.48e+5 | **1.85e+5** |
| 40 | 9.62e+5 | 9.79e+5 | 1.04e+6 | **8.25e+5** | 1.00e+6 | 1.20e+6 | 1.01e+6 | **7.52e+5** |
| 80 | 3.88e+6 | 3.96e+6 | 4.18e+6 | **3.32e+6** | 4.05e+6 | 4.86e+6 | 4.06e+6 | **3.03e+6** |
| $N = 5$ | | | | | | | | |
| 10 | 4.41e+5 | **2.79e+5** | 4.69e+5 | 3.84e+5 | 4.58e+5 | 4.06e+5 | 4.43e+5 | **3.33e+5** |
| 20 | 1.80e+6 | **1.23e+6** | 1.91e+6 | 1.56e+6 | 1.86e+6 | 1.68e+6 | 1.80e+6 | **1.34e+6** |
| 40 | 7.25e+6 | **5.02e+6** | 7.71e+6 | 6.32e+6 | 7.52e+6 | 6.85e+6 | 7.27e+6 | **5.39e+6** |
| 80 | 2.92e+7 | **2.03e+7** | 3.10e+7 | 2.54e+7 | 3.02e+7 | 2.77e+7 | 2.92e+7 | **2.16e+7** |
| $N = 10$ | | | | | | | | |
| 10 | 7.16e+6 | **3.92e+6** | 7.71e+6 | 6.50e+6 | 7.02e+6 | 6.09e+6 | 6.73e+6 | **5.11e+6** |
| 20 | 2.89e+7 | **1.59e+7** | 3.11e+7 | 2.64e+7 | 2.84e+7 | 2.46e+7 | 2.71e+7 | **2.04e+7** |
| 40 | 1.16e+8 | **6.39e+7** | 1.25e+8 | 1.06e+8 | 1.14e+8 | 9.92e+7 | 1.09e+8 | **8.17e+7** |
| 80 | 4.67e+8 | **2.57e+8** | 5.02e+8 | 4.27e+8 | 4.58e+8 | 3.98e+8 | 4.37e+8 | **3.27e+8** |
| $N = 20$ | | | | | | | | |
| 10 | 1.34e+8 | **6.79e+7** | 1.49e+8 | 1.23e+8 | 1.17e+8 | 1.21e+8 | 1.13e+8 | **8.40e+7** |
| 20 | 5.39e+8 | **2.25e+8** | 5.99e+8 | 4.98e+8 | 4.71e+8 | 4.89e+8 | 4.56e+8 | **3.34e+8** |
| 40 | 2.16e+9 | **1.10e+9** | 2.41e+9 | 2.01e+9 | 1.89e+9 | 1.97e+9 | 1.83e+9 | **1.34e+9** |
| 80 | 8.67e+9 | **4.42e+9** | 9.65e+9 | 8.06e+9 | 7.59e+9 | 7.89e+9 | 7.34e+9 | **5.34e+9** |

is a solution. It has one dynamical degree of freedom. We choose the special condition

$$x_1(0) = 0.$$

The results for $\eta = -25$ and $\lambda = -1$ are provided in Table 8.

*Example 3* Our next example is a linearized problem proposed by Campbell and More [21]. It has been used previously in the experiments in [2, 8, 9] and others. Let

$$A(Dx)'(t) + B(t)x(t) = q(t), \quad t \in [0, 5],$$

where

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

**Table 8** $\kappa_C(\mathcal{A})$ for $\Phi^R_{\pi,M}$. Here, *L* denotes the Legendre basis, *mL* the modified Legendre basis, *Ch* the Chebyshev basis, and *RK* the Runge-Kutta basis. The smallest values are set in boldface

| $n = 1/h$ | $\kappa_C(\mathcal{A})$ | | | |
|---|---|---|---|---|
| | L | mL | Ch | RK |
| $N = 3$ | | | | |
| 10 | 1.95e+5 | 3.42e+5 | 2.00e+5 | **1.96e+5** |
| 20 | 2.56e+5 | 6.90e+5 | 2.65e+5 | **2.48e+5** |
| 40 | 4.01e+5 | 1.58e+6 | 4.22e+5 | **3.50e+5** |
| 80 | 8.17e+5 | 3.82e+6 | 8.73e+5 | **6.06e+5** |
| $N = 5$ | | | | |
| 10 | 6.23e+5 | **5.25e+5** | 6.13e+5 | 7.05e+5 |
| 20 | 8.54e+5 | 1.19e+6 | **8.48e+5** | 9.32e+5 |
| 40 | 1.31e+6 | 2.75e+6 | 1.32e+6 | **1.26e+6** |
| 80 | 2.36e+6 | 6.61e+6 | 2.41e+6 | **2.03e+6** |
| $N = 10$ | | | | |
| 10 | 3.06e+6 | **1.33e+6** | 3.02e+6 | 4.45e+6 |
| 20 | 4.28e+6 | **3.02e+6** | 4.19e+6 | 5.98e+6 |
| 40 | 6.63e+6 | 6.85e+6 | **6.55e+6** | 7.78e+6 |
| 80 | 1.19e+7 | 1.61e+7 | 1.19e+7 | **1.14e+7** |
| $N = 20$ | | | | |
| 10 | 1.68e+7 | **4.73e+6** | 1.71e+7 | 3.05e+7 |
| 20 | 2.12e+7 | **1.03e+7** | 2.18e+7 | 3.77e+7 |
| 40 | 3.23e+7 | **2.23e+7** | 3.35e+7 | 4.76e+7 |
| 80 | 6.12e+7 | **4.93e+7** | 6.37e+7 | 6.74e+7 |

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

subject to the initial conditions

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

This problem has index 3 and dynamical degree of freedom $l_{dyn} = 4$. The right-hand side $q$ has been chosen in such a way that the exact solution becomes

$$\begin{aligned} x_{*,1} &= \sin t, & x_{*,4} &= \cos t, \\ x_{*,2} &= \cos t, & x_{*,5} &= -\sin t, \\ x_{*,3} &= 2\cos^2 t, & x_{*,6} &= -2\sin 2t, \\ x_{*,7} &= -\rho^{-1}\sin t. \end{aligned}$$

**Table 9** $\kappa_C(\mathcal{A})$ for $\Phi^R_{\pi,M}$. Here, *L* denotes the Legendre basis, *mL* the modified Legendre basis, *Ch* the Chebyshev basis, and *RK* the Runge-Kutta basis. The smallest values are set in boldface

| $n = 5/h$ | $\kappa_C(\mathcal{A})$ | | | |
|---|---|---|---|---|
| | L | mL | Ch | RK |
| $N = 3$ | | | | |
| 10 | 4.64e+2 | 1.97e+3 | 5.00e+2 | **4.12e+2** |
| 20 | 1.43e+3 | 5.06e+3 | 1.54e+3 | **1.21e+3** |
| 40 | 5.543e+3 | 1.38e+4 | 5.96e+3 | **4.70e+3** |
| 80 | 2.19e+4 | 4.12e+4 | 2.36e+4 | **1.86e+4** |
| $N = 5$ | | | | |
| 10 | 1.63e+3 | 3.74e+3 | 1.70e+3 | **1.46e+3** |
| 20 | 6.26e+3 | 1.21e+4 | 6.55e+3 | **5.60e+3** |
| 40 | 2.48e+4 | 4.63e+4 | 2.60e+4 | **2.22e+4** |
| 80 | 9.88e+4 | 1.83e+5 | 1.04e+5 | **8.85e+4** |
| $N = 10$ | | | | |
| 10 | 3.27e+4 | 4.69e+4 | 3.58e+4 | **3.20e+4** |
| 20 | 1.29e+5 | 1.89e+5 | 1.42e+5 | **1.18e+5** |
| 40 | 5.14e+5 | 7.58e+5 | 5.65e+5 | **4.69e+5** |
| 80 | 2.06e+6 | 3.03e+6 | 2.26e+6 | **1.88e+6** |
| $N = 20$ | | | | |
| 10 | **7.30e+5** | 8.64e+5 | 8.16e+5 | 9.69e+5 |
| 20 | 2.91e+6 | 3.63e+6 | 3.26e+6 | **2.71e+6** |
| 40 | 1.17e+7 | 1.50e+7 | 1.31e+7 | **1.09e+7** |
| 80 | 4.69e+7 | 6.10e+7 | 5.25e+7 | **4.36e+7** |

The results are shown in Table 9. Note that, in the present example, $h = 5/n$ in contrast to all previous computations where $h = 1/n$. □

The numerical experiments give rise to the following observations:

- The condition numbers of the discrete problem have almost the same size for given polynomial degree $N$ and stepsize $h$.
- The experiments indicate that the Runge-Kutta basis seems to provide the lowest condition number for smaller stepsizes. In the case of higher order ansatz functions and larger stepsizes, the modified Legendre basis seems to provide the smallest condition numbers.
- In order to obtain a complete picture of the relative merits of the different bases, in the case discussed in Theorem 5, not only the condition number $\kappa(\mathcal{C})$ of $\mathcal{C}$ but the term $\|\mathcal{C}^+\|\kappa(\mathcal{C})$ has to be taken into account. Corollary 2 shows that the modified Legendre basis is well-suited for higher orders $N$.
- If the perturbed solution $\tilde{c}$ of (36) is projected back onto the nullspace $\ker \mathcal{C}$, we can assume that the conditions of Theorem 4 are fulfilled. In this case, $\mathcal{C}$ does not have any influence on the error estimation.

## 6 Conclusions

In this paper, we investigated the conditioning of the discrete problems arising in the least-squares collocation method for DAEs. In particular, the solution algorithm has been split into a representation mapping that connects the coefficients of the basis representation to the function to be represented, and a linearly equality constrained linear least-squares problem. A careful investigation of the representation map allowed for a characterization of errors in the function spaces by those made in the solution of the discrete problem.

The perturbation estimates for the constrained least-squares problem have been derived with the application in mind: the approximation of a DAE. The constraints play an exceptional role. If they are satisfied, the resulting numerical solution belongs to the solution space $H_D^1(a, b)$. If this cannot be guaranteed, the convergence theory for the least-squares method does not apply. Some of the characterizing quantities could be estimated analytically for reasonable choices of bases while others have been estimated numerically in certain examples. We believe that these considerations contribute to a robust and efficient implementation of the proposed method, which seems to provide surprisingly accurate numerical solutions to higher-index DAEs.

**Acknowledgements** The author wants to thank Roswitha März for many discussions that led to a great enhancement of the presentation. In particular, her contributions simplified the derivations leading to Theorem 2 considerably.

**Funding** Open access funding provided by Royal Institute of Technology.

**Data availability** The code used for generating all datasets is available on request.

## Declarations

**Conflict of interest** The author declares no competing interests.



## References

1. Hanke, M., März, R., Tischendorf, C., Weinmüller, E., Wurm, S.: Least-squares collocation for linear higher-index differential-algebraic equations. J. Comput. Appl. Math. **317**, 403–431 (2017). https://doi.org/10.1016/j.cam.2016.12.017
2. Hanke, M., März, R., Tischendorf, C.: Least-squares collocation for higher-index linear differential-algebaic equations: Estimating the stability threshold. Math. Comp. **88**(318), 1647–1683 (2019). https://doi.org/10.1090/mcom/3393
3. Hanke, M., März, R.: A reliable direct numerical treatment of differential-algebraic equations by overdetermined collocation: An operator approach. J. Comput. Appl. Math. **387**, 112520 (2021)
4. Hanke, M., März, R.: Convergence analysis of least-squares collocation metods for nonlinear higher-index differential-algebraic equations. J. Comput. Appl. Math. **387**(11), 2021 (2514)
5. Hanke, M.: Linear differential-algebraic equations in spaces of integrable functions. J. Differential Equations **79**(1), 14–30 (1989)
6. Lamour, R., März, R., Tischendorf, C.: Differential-Algebraic Equations: A Projector Based Analysis. Differential-Algebraic Equations Forum. Springer, Berlin (2013). Series Editors: A. Ilchmann, T. Reis
7. Kaltenbacher, B., Offtermatt, J.A.: Convergence analysis of regularization by discretization in preimage space. Math. Comp. **81**(280), 2049–2069 (2012)
8. Hanke, M., März, R.: Towards a reliable implementation of least-squares collocation for higher-index linear differential-algebaic equations. Part 1: Basics and ansatz choices. Numerical Algorithms **89**, 931–963 (2022)
9. Hanke, M., März, R.: Towards a reliable implementation of least-squares collocation for higher-index linear differential-algebaic equations. Part 2: The discrete least-squares problem. Numerical Algorithms **89**, 965–986 (2022)
10. Wedin, P.-A.: Perturbation theory for pseudo-inverses. BIT **13**, 217–232 (1973)
11. Eldén, L.: Perturbation theory for the least squares problem with linear equality constraints. SIAM J. Numer Anal. **17**(3), 338–350 (1980)
12. Cox, A.J., Higham, N.J.: Accuracy and stability of the nullspace method for solving the equality constraind least squares problem. BIT **39**(1), 34–50 (1999)
13. Lawson, C.L., Hanson, R.J.: Solving least squares problems. Prentice Hall, Englewood Cliffs NY (1974)
14. Bhatia, R.: Matrix Analysis. Graduate Texts in Mathematics. Springer, New York (1997)
15. Hanke, M., März, R.: Least-Squares Collocation for Higher-Index Daes: Global Approach and Attempts Towards a Time-Stepping Version. In: Reis, T., et al. (eds.) Progress in Differential-Algebraic Equations II, pp. 91–136. Springer, Cham (2020)
16. Kiełbasiński, A., Schwetlick, H.: Numerische Lineare Algebra. Deutscher Verlag der Wissenschaften, Berlin (1988)
17. Kato, T.: Perturbation Theory for Linear Operators, 2nd edn. Classics in Mathematics. Springer, Berlin (1995)
18. Kulkarni, D., Schmidt, D., Tsui, S.-K.: Eigenvalues of tridiagonal pseudo-Toeplitz matrices. Lin. Alg. Appl. **297**, 63–80 (1999)

19. Knopp, K.: Application of Infinite Series Theory. Blackie&Son Glasgow, London (1954)
20. Davis, P.J., Rabinowithz, P.: Methods of Numerical Intergration, 2nd edn. Academic Press, San Diego, London (1984)
21. Campbell, S.L., Moore, E.: Constraint preserving integrators for general nonlinear higher index DAEs. Num. Math. **69**, 383–399 (1995)