Least-Squares Collocation for Higher-Index DAEs: Global Approach and Attempts Toward a Time-Stepping Version

Michael Hanke and Roswitha März

Abstract Overdetermined polynomial least-squares collocation for two-point boundary value problems for higher index differential-algebraic equations shows excellent convergence properties while at the same time being only slightly more expensive than the widely used collocation method for ordinary differential equations by piecewise polynomials. In the present paper, basic properties of this method when applied to initial value problems by a windowing technique are proven. Some examples are provided in order to show the potential of time-stepping approach.

Keywords Differential-algebraic equation · Higher index · Initial-value problem · Essentially ill-posed problem · Least-squares problem · Polynomial collocation

Mathematics Subject Classification (2010) 65L80, 65L08

1 Introduction

In a number of recent papers [7–10] convergence results for an overdetermined polynomial least-squares collocation for two-point boundary value problems for higher index differential-algebraic equations (DAEs) have been established. This method is comparable in computational efficiency with the widely used collocation

M. Hanke (✉)
School of Engineering Sciences, Department of Mathematics, KTH Royal Institute of Technology, Stockholm, Sweden
e-mail: hanke@nada.kth.se

R. März
Institute of Mathematics, Humboldt-University of Berlin, Berlin, Germany
e-mail: maerz@math.hu-berlin.de

T. Reis et al. (eds.), *Progress in Differential-Algebraic Equations II*, Differential-Algebraic Equations Forum,
https://doi.org/10.1007/978-3-030-53905-4_4

method for ordinary differential equations using piecewise polynomials. For initial value problems (IVPs), a considerable increase in numerical efficiency of the overdetermined polynomial least-squares collocation method is expected if one can construct time-stepping or windowing techniques. Below, we consider some key issues in this respect. Our ultimate goal is that overdetermined collocation is used on succeeding individual time-windows, though we emphasize that the present note deals with the very first attempts in this context only.

We are interested in general initial-value problems (IVPs),

$$f((Dx)'(t), x(t), t) = 0, \quad t \in [a, b], \quad G_a x(a) = r. \tag{1.1}$$

$x : [a, b] \to \mathbb{R}^m$ is the unknown vector-valued function defined on the finite interval $[a, b] \subset \mathbb{R}$. We assume an explicit partitioning of the unknowns into differentiated and nondifferentiated (also called algebraic) components by selecting

$$D \in \mathbb{R}^{k\times m}, \quad D = [I_k\ 0]$$

with the identity matrix $I_k \in \mathbb{R}^{k\times k}$. The function $f : \mathbb{R}^k \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^m$ is assumed to be sufficiently smooth, at least continuous and with continuous partial derivatives with respect to the first and second arguments.

The initial values deserve some special attention. For a solution to exist they must be consistent. We will ensure this by requiring special properties on the matrix G_a. It is reasonable to assume that at most the differentiated components $x_1, \ldots, x_k$ are fixed by initial conditions, which leads to the requirement

$$G_a \in \mathbb{R}^{l\times m}, \quad \ker G_a \supseteq \ker D,$$

such that $G_a x(a) = G_a D^+ Dx(a)$. Moreover, we will assume that the initial conditions are independent of each other, that is $\operatorname{rank} G_a = l$, where l denotes the actual dynamical degree of freedom. Later on, more detailed requirements, depending on the DAE will be posed.

Let the interval $[a, b]$ be decomposed into L subintervals,

$$a = w_0 < w_1 < \cdots < w_L = b,$$

with lengths $H_\lambda = w_\lambda - w_{\lambda-1}$, $\lambda = 1, \ldots, L$. First, for $\lambda = 1$, we provide an approximating segment $\tilde{x}^{[1]} : [w_0, w_1] \to \mathbb{R}^m$ by applying overdetermined collocation to the IVP

$$f((D\tilde{x}^{[1]})'(t), \tilde{x}^{[1]}(t), t) = 0, \quad t \in [w_0, w_1], \quad G_a\tilde{x}^{[1]}(a) = r. \tag{1.2}$$

For $\lambda > 1$, having already the segment $\tilde{x}^{[\lambda-1]} : [w_{\lambda-2}, w_{\lambda-1}] \to \mathbb{R}^m$, we intend to determine the next segment $\tilde{x}^{[\lambda]} : [w_{\lambda-1}, w_\lambda] \to \mathbb{R}^m$ by solving the DAE

$$f((D\tilde{x}^{[\lambda]})'(t), \tilde{x}^{[\lambda]}(t), t) = 0, \quad t \in [w_{\lambda-1}, w_\lambda]. \tag{1.3}$$

In order to obtain an appropriate approximation to the solution of (1.1), we need to compensate the now unavailable initial conditions by certain transfer conditions using $\tilde{x}^{[\lambda-1]}$. Below we investigate two different approaches, namely,

$$G(w_{\lambda-1})\tilde{x}^{[\lambda]}(w_{\lambda-1}) = G(w_{\lambda-1})\tilde{x}^{[\lambda-1]}(w_{\lambda-1}), \tag{1.4}$$

with a suitably prescribed matrix function $G : [a, b] \to \mathbb{R}^{l\times m}$, and

$$D\tilde{x}^{[\lambda]}(w_{\lambda-1}) = D\tilde{x}^{[\lambda-1]}(w_{\lambda-1}). \tag{1.5}$$

The construction of appropriate transfer conditions is crucial for the success of the method.[1]

In the present note we merely deal with the linear version of the IVP,

$$A(t)(Dx)'(t) + B(t)x(t) - q(t) = 0, \quad t \in [a, b], \tag{1.6}$$

$$G_a x(a) = r, \tag{1.7}$$

in which the right-hand side $q : [a, b] \to \mathbb{R}^m$ and the matrix coefficients $A : [a, b] \to \mathbb{R}^{m\times k}$ and $B : [a, b] \to \mathbb{R}^{m\times m}$ are assumed to be sufficiently smooth, however at least continuous, thus uniformly bounded.

As it is well-known,[2] conventional time-stepping methods such as the BDF in the famous DAE solver DASSL work well only when applied to index-1 DAEs and special form index-2 DAEs. The so far available time-stepping solvers for more general higher-index DAEs are definitely bound to the construction and evaluation of so-called derivative array systems,[3] e.g., [3, 4, 12, 16, 17], which accounts for a serious limitation in view of applications. The recently discussed ansatz of overdetermined least-squares collocation [7–10] fully avoids the use of derivative arrays and no reduction procedures are incorporated, which is highly beneficial. However, this is a global ansatz over the entire interval, not a time-stepping method and large ill-conditioned discrete systems may arise. For this reason, eventually, a time-stepping version would be much more advantageous. Recall that we come up with very first related ideas here.

The paper is organized as follows: We describe the global overdetermined collocation procedure in Sect. 2 and collect there the relevant convergence results. In Sect. 3 we derive basic error estimates for overdetermined collocation on arbitrary individual subintervals corresponding to both procedures (1.2)–(1.3) and (1.4). A corresponding result for the approach (1.2)–(1.3) and (1.5) is provided in Sect. 4. We study the simpler time-stepping version with uniform window-size H and the

[1] It should be noted that also an appropriate continuous functional of $\tilde{x}^{[\lambda-1]}$ can be considered as a suitable candidate for defining a transfer condition.

[2] We refer to [1, 6] for an early discussion and to [2, 13] for a topical one.

[3] Also called prolongation. The necessary differentiations have to be provided analytically or via automatic differentiation.

same uniform stepsize on all subintervals in Sect. 5. Convergence of the method using the transfer condition (1.4) is shown in Sect. 5.1. However, our estimates in Sect. 4 are not sufficient to show convergence for the case (1.3), (1.5). Therefore, an investigation of a very special system in Sect. 5.3 provides some hints on what could be expected in that case. In order to demonstrate the behavior of the method, we provide a more complex example in Sect. 6 using both approaches, (1.4) as well as (1.5).

2 Global Overdetermined Collocation

2.1 The Global Procedure

Let us consider first the case of global overdetermined collocation, that is $L = 1$ and $H = b - a$. Let, for a given $n \in \mathbb{N}$, a grid π on the interval $[a, b]$ be defined:

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

where $t_j = a + jh$ and $h = (b - a)/n$.[4]

In order to be able to introduce collocation conditions we will need a space of piecewise continuous functions. Let $C_\pi([a, b], \mathbb{R}^m)$ denote the space of all functions $x : [a, b] \to \mathbb{R}^m$ which are continuous on each subinterval (t_{j-1}, t_j) and feature continuous extensions onto $[t_{j-1}, t_j]$, $j = 1, \ldots, n$. Furthermore, let $\mathscr{P}_N$ denote the set of polynomials of degree less than or equal to N, $N \geq 1$. We define the ansatz space

$$\begin{aligned} X_\pi = \{p \in C_\pi([a, b], \mathbb{R}^m) | Dp \in C([a, b], \mathbb{R}^k), \\ p_\kappa|_{(t_{j-1}, t_j)} \in \mathscr{P}_N, \kappa = 1, \ldots, k, \\ p_\kappa|_{(t_{j-1}, t_j)} \in \mathscr{P}_{N-1}, \ \kappa = k + 1, \ldots, m, \\ j = 1, \ldots, n\}. \end{aligned}$$

Let now M points τ_i be given such that $0 < \tau_1 < \cdots < \tau_M < 1$. The set of collocation points is given by

$$S_{\pi,M} = \{t_{ji} = t_{j-1} + \tau_i h | \ j = 1, \ldots, n, \ i = 1, \ldots, M\}. \tag{2.1}$$

[4] A generalization to quasi-uniform grids is easily possible.

Using this set $S_{\pi,M}$, an interpolation operator $R_{\pi,M} : C_\pi([a,b],\mathbb{R}^m) \to C_\pi([a,b],\mathbb{R}^m)$ is given by assigning, to each $w \in C_\pi([a,b],\mathbb{R}^m)$, the piecewise polynomial $R_{\pi,M}w$ with

$$R_{\pi,M}w|_{(t_{j-1},t_j)} \in \mathscr{P}_{M-1},\ j = 1,\dots,n, \quad R_{\pi,M}w(t) = w(t),\ t \in S_{\pi,M}.$$

The functional

$$\Phi_{\pi,M}(x) = \|R_{\pi,M}(f((Dx)'(\cdot),x(\cdot),\cdot))\|_{L^2}^2 + |G_a x(a) - r|^2, \quad x \in X_\pi,$$

can be represented as (cf. [10, Subsection 2.3], also [8, 9])

$$\Phi_{\pi,M}(x) = W^T \mathscr{L} W + |G_a x(a) - r|^2, \quad x \in X_\pi,$$

with the vector $W \in \mathbb{R}^{mMn}$,

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

with the matrix $\mathscr{L}$ being positive definite, symmetric and independent[5] of h. Moreover there are constants $\kappa_l, \kappa_u > 0$ such that

$$\kappa_l\ |V|^2 \le V^T \mathscr{L} V \le \kappa_u\ |V|^2, \quad V \in \mathbb{R}^{mMn}. \tag{2.2}$$

If the DAE in (1.1) is regular with index one, $l = k$, and $M = N$, then there is an element $\tilde{x}_\pi \in X_\pi$ such that $\Phi_{\pi,M}(\tilde{x}_\pi) = 0$, which corresponds to the classical collocation method resulting in a system of $nMm + l$ equations for $nNm + k = nMm + l$ unknowns. Though classical collocation works well for regular index-1 DAEs (e.g., [14]), it is known to be useless for higher-index DAEs.

Reasonably, one applies l initial conditions in compliance with the dynamical degree of freedom of the DAE. In the case of higher-index DAEs, the dynamical degree of freedom is always less than k. For $0 \le l \le k$ and $M \ge N+1$, necessarily an overdetermined collocation system results since $nMm + l > nNm + k$. *Overdetermined least-squares collocation* consists of choosing $M \ge N+1$ and then determining an element $\tilde{x}_\pi \in X_\pi$ which minimizes the functional $\Phi_{\pi,M}$, i.e.,

$$\tilde{x}_\pi \in \operatorname{argmin}\{\Phi_{\pi,M}(x) | x \in X_\pi\}.$$

This runs astonishingly well [9, 10], see also Sect. 6.

[5] The entries of $\mathscr{L}$ are fully determined by the corresponding M Lagrangian basis polynomials, thus, by M and $\tau_1,\dots,\tau_M$.

2.2 Convergence Results for the Global Overdetermined Collocation Applied to Linear IVPs

We now specify results obtained for boundary value problems in [8–10] for a customized application to IVPs. Even though we always assume a sufficiently smooth classical solution $x_* : [a, b] \to \mathbb{R}^m$ of the IVP (1.6), (1.7) to exist, for the following, an operator setting in Hilbert spaces will be convenient. The spaces to be used are:

$$L^2 = L^2((a, b), \mathbb{R}^m\}, \quad H_D^1 = \{x \in L^2 | Dx \in H^1((a, b), \mathbb{R}^k\}, \quad Y = L^2 \times \mathbb{R}^l.$$

The operator $T : H_D^1 \to L^2$ given by

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

is bounded. Since, for $x \in H_D^1$, the values $Dx(a)$ and thus $G_a x(a) = G_a D^+ Dx(a)$ are well-defined, the composed operator $\mathscr{T} : X \to Y$ given by

$$\mathscr{T}x = \begin{bmatrix} Tx \\ G_a x(a) \end{bmatrix}, \quad x \in H_D^1,$$

is well-defined and also bounded.

Let $U_\pi : H_D^1 \to H_D^1$ denote the orthogonal projector of the Hilbert space H_D^1 onto X_π.

For a more concise notation later on, we introduce the composed interpolation operator $\mathscr{R}_{\pi,M} : C_\pi([a, b], \mathbb{R}^m) \times \mathbb{R}^l \to Y$,

$$\mathscr{R}_{\pi,M} \begin{bmatrix} w \\ r \end{bmatrix} = \begin{bmatrix} R_{\pi,M} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} w \\ r \end{bmatrix}.$$

With these settings, overdetermined least-squares collocation reduces to the minimization of

$$\Phi_{\pi,M}(x) = \|R_{\pi,M}(Tx - q)\|_{L^2}^2 + |G_a x(a) - r|^2 = \|\mathscr{R}_{\pi,M}(\mathscr{T}x - y)\|_Y^2, \ x \in X_\pi,$$

that is, to find

$$\tilde{x}_\pi \in \operatorname{argmin}\{\Phi_{\pi,M}(x) | x \in X_\pi\}.$$

Later on, we will provide conditions which ensure that $\ker \mathscr{R}_{\pi,M} \mathscr{T} U_\pi = X_\pi^\perp$ such that $\tilde{x}_\pi$ is uniquely defined. Therefore,

$$\tilde{x}_\pi = (\mathscr{R}_{\pi,M} \mathscr{T} U_\pi)^+ \mathscr{R}_{\pi,M}\, y.$$

We consider also the related functional

$$\Phi(x) = \|Tx - q\|_{L^2}^2 + |G_a x(a) - r|^2 = \|\mathscr{T}x - y\|_Y^2, \quad x \in H_D^1,$$

and the corresponding method for approximating the solution x_* by determining

$$x_\pi \in \operatorname{argmin}\{\Phi(x) | x \in X_\pi\}.$$

As before, the conditions assumed below will guarantee that the minimizer x_π is unique such that

$$x_\pi = (\mathscr{T}U_\pi)^+ y.$$

Below, the operator $\mathscr{T}$ is ensured to be injective. Since $\mathscr{T}$ is associated with a higher-index DAE, the inverse $\mathscr{T}^{-1}$ is unbounded and the IVP is essentially ill-posed in the sense of Tikhonov. Following ideas to treat ill-posed problems, e.g., [11], the proofs in [8–10] are based on estimates of the type

$$\|x_\pi - x_*\|_{H_D^1} \le \frac{\beta_\pi}{\gamma_\pi} + \alpha_\pi,$$

$$\|\tilde{x}_\pi - x_*\|_{H_D^1} \le \frac{\tilde{\beta}_\pi}{\tilde{\gamma}_\pi} + \alpha_\pi,$$

in which

$$\alpha_\pi = \|(I - U_\pi)x_*\|_{H_D^1},$$

$$\beta_\pi = \|\mathscr{T}(I - U_\pi)x_*)\|_Y,$$

$$\tilde{\beta}_\pi = \|\mathscr{R}_{\pi,M}\mathscr{T}(I - U_\pi)x_*\|_Y,$$

$$\gamma_\pi = \inf_{p \in X_\pi, p \neq 0} \frac{\|\mathscr{T}p\|_Y}{\|p\|_{H_D^1}} = \inf_{p \in X_\pi, p \neq 0} \left(\frac{\|Tp\|_{L^2}^2 + |G_a p(a)|^2}{\|p\|_{H_D^1}} \right)^{1/2},$$

$$\tilde{\gamma}_\pi = \inf_{p \in X_\pi, p \neq 0} \frac{\|\mathscr{R}_{\pi,M}\mathscr{T}p\|_Y}{\|p\|_{H_D^1}} = \inf_{p \in X_\pi, p \neq 0} \left(\frac{\|R_{\pi,M}Tp\|_{L^2}^2 + |G_a p(a)|^2}{\|p\|_{H_D^1}} \right)^{1/2}.$$

The most challenging task in this context is to provide suitable positive lower bounds of the *instability thresholds* γ_π and $\tilde{\gamma}_\pi$, [8–10] and, what is the same, upper bounds for the Moore-Penrose inverses

$$\|(\mathscr{T}U_\pi)^+\| = \frac{1}{\gamma_\pi}, \quad \|(\mathscr{R}_{\pi,M}\mathscr{T}U_\pi)^+\| = \frac{1}{\tilde{\gamma}_\pi}.$$

It should be noted that $\mathscr{T}$ and $\mathscr{R}_{\pi,M}\mathscr{T}$ are of very different nature: While $\mathscr{T}$ is bounded, $\mathscr{R}_{\pi,M}\mathscr{T}$ is unbounded owing to the fact that $R_{\pi,M}$ is an unbounded operator in L^2, see [8].

We now briefly summarize the relevant estimations resulting from [8, 9] for IVPs. For details we refer to [8, 9].

The general assumptions with respect to the DAE and the initial conditions are:[6]

1. The operator T is fine with tractability index $\mu \geq 2$ and characteristic values $0 < r_0 \leq \cdots \leq r_{\mu-1} < r_\mu = m$.
2. The initial conditions are accurately stated such that $l = m - \sum_{i=0}^{\mu-1}(m - r_i)$ and $G_a = G_a \Pi_{can}(a)$, with the canonical projector Π_{can}. This implies $\operatorname{im}\mathscr{T} = \operatorname{im} T \times \mathbb{R}^l$, see [14, Theorem 2.1].
3. The coefficients A, B, the right-hand side $q \in \operatorname{im} T$, and the solution x_* are sufficiently smooth.

Result (a), see [9]: Assume $M \geq N+1$. Then there are positive constants c_α, c_β, c_γ and c such that, for all sufficiently small stepsizes $h > 0$,

$$\gamma_\pi \geq c_\gamma h^{\mu-1}, \quad \alpha_\pi \leq c_\alpha h^N, \quad \beta_\pi \leq c_\beta h^N,$$

and eventually

$$\|x_\pi - x_*\|_{H_D^1} \leq c\, h^{N-\mu+1}.$$

Result (b), see [8]: Assume $M \geq N+\mu$. Then there are positive constants c_α, $\tilde{c}_\beta$, $\tilde{c}_\gamma$, and $\tilde{c}$ such that, for all sufficiently small stepsizes $h > 0$,

$$\tilde{\gamma}_\pi \geq \tilde{c}_\gamma h^{\mu-1}, \quad \alpha_\pi \leq c_\alpha h^N, \quad \tilde{\beta}_\pi \leq \tilde{c}_\beta h^N,$$

and eventually

$$\|\tilde{x}_\pi - x_*\|_{H_D^1} \leq \tilde{c}\, h^{N-\mu+1}.$$

By [8], one can do with $\tilde{c}_\gamma = c_\gamma/2$. We refer to [9, 10] for a series of tests which confirm these estimations or perform even better. Recall that so far, IVPs for higher-index DAEs are integrated by techniques which evaluate derivative arrays, e.g., [5]. Comparing with those methods even the global overdetermined collocation method features beneficial properties. However, a time-stepping version could be much more advantageous.

[6] The following results are also valid for index-1 DAEs. However, we do not recommend this approach for $\mu = 1$ since standard collocation methods work well, see [14].

3 Overdetermined Collocation on an Arbitrary Subinterval $[\bar{t}, \bar{t}+H] \subset [a,b]$

3.1 Preliminaries

We continue to consider the IVP (1.6), (1.7) as described above, but instead of the global approach immediately capturing the entire interval $[a,b]$ we now aim at stepping forward by means of consecutive time-windows applying overdetermined least-squares collocation on each window. As special cases, we have in mind the two windowing procedures outlined by (1.2), (1.3), and (1.4), and by (1.2), (1.3), and (1.5). At the outset we ask how overdetermined collocation works on an arbitrary subinterval,

$$[\bar{t}, \bar{t}+H] \subseteq [a,b].$$

It will become important to relate global quantities (valid for overdetermined least-squares collocation on $[a,b]$) to their local counterparts (appropriate on subintervals of length H). We introduce the function spaces related to this subinterval,

$$L^2_{sub} = L^2((\bar{t}, \bar{t}+H), \mathbb{R}^m\}, \quad H^1_{sub} = H^1((\bar{t}, \bar{t}+H), \mathbb{R}^k),$$
$$H^1_{D,sub} = \{x \in L^2_{sub} | Dx \in H^1_{sub}\}, \quad Y_{sub} = L^2_{sub} \times \mathbb{R}^l, \qquad \hat{Y}_{sub} = L^2_{sub} \times \mathbb{R}^k,$$

equipped with natural norms, in particular,

$$\|x\|_{H^1_{D,sub}} = (\|x\|^2_{L^2_{sub}} + \|(Dx)'\|^2_{L^2_{sub}})^{1/2}, \quad x \in H^1_{D,sub}.$$

Note that we indicate quantities associated to the subinterval by the extra subscript *sub* only if necessary and otherwise misunderstandings could arise.

Now we assume that the grid π is related to the subinterval only,

$$\pi : \quad \bar{t} = t_0 < \cdots < t_n = \bar{t} + H,$$

where $t_j = \bar{t} + jh$ and $h = H/n$. The ansatz space reads now

$$X_\pi = \{p \in C_\pi([\bar{t}, \bar{t}+H], \mathbb{R}^m) | \ Dp \in C([\bar{t}, \bar{t}+H], \mathbb{R}^k),$$
$$p_\kappa|_{(t_{j-1},t_j)} \in \mathscr{P}_N, \ \kappa = 1, \ldots, k, \ p_\kappa|_{(t_{j-1},t_j)} \in \mathscr{P}_{N-1}, \ \kappa = k+1, \ldots, n,$$
$$j = 1, \ldots, n\}.$$

With $0 < \tau_1 < \cdots < \tau_M < 1$, the set of collocation points

$$S_{\pi,M} = \{t_{ji} = t_{j-1} + \tau_i h | \ j = 1, \ldots, n, \ i = 1, \ldots, M\} \tag{3.1}$$

belongs to the subinterval $[\bar{t}, \bar{t}+H]$. Correspondingly, the interpolation operator $R_{\pi,M}$ acts on $C_\pi([\bar{t}, \bar{t}+H], \mathbb{R}^m)$. We introduce the operator $T_{sub} : H^1_{D,sub} \to L^2_{sub}$,

$$(T_{sub}x)(t) = A(t)(Dx)'(t) + B(t)x(t), \ a.e.\ t \in (\bar{t}, \bar{t}+H), \ x \in H^1_{D,sub},$$

and the composed operators $\mathscr{T}_{sub} : H^1_{D,sub} \to Y_{sub}$ and $\hat{\mathscr{T}}_{sub} : H^1_{D,sub} \to \hat{Y}_{sub}$,

$$\mathscr{T}_{sub}x = \begin{bmatrix} T_{sub}x \\ G(\bar{t})x(\bar{t}) \end{bmatrix}, \quad \hat{\mathscr{T}}_{sub}x = \begin{bmatrix} T_{sub}x \\ Dx(\bar{t}) \end{bmatrix}, \quad x \in H^1_{D,sub}.$$

Occasionally, we also use the operators $T_{IC,sub} : H^1_{D,sub} \to \mathbb{R}^l$ and $T_{ICD,sub} : H^1_{D,sub} \to \mathbb{R}^k$ given by

$$T_{IC,sub}x = G(\bar{t})x(\bar{t}), \quad T_{ICD,sub}x = Dx(\bar{t}), \quad x \in H^1_{D,sub},$$

which are associated with the initial condition posed at $\bar{t}$. Here, aiming for injective composed operators, we suppose a function $G : [a, b] \to \mathbb{R}^l$ such that

$$\ker G(t) = \ker \Pi_{can}(t), \ \operatorname{im} G(t) = \mathbb{R}^l, \ |G(t)| \le c_G, \ t \in [a, b]. \tag{3.2}$$

Since T_{sub} inherits the tractability index, the characteristic values of T, and also the canonical projector (restricted to the subinterval, see [13, Section 2.6]), the local initial condition at $\bar{t}$, $G(\bar{t})x(\bar{t}) = r$, is accurately stated. Then $\operatorname{im} \mathscr{T}_{sub} = \operatorname{im} T_{sub} \times \mathbb{R}^l$ and $\ker \mathscr{T}_{sub} = \{0\}$, so that the overdetermined least-squares collocation on $[\bar{t}, \bar{t}+H]$ works analogously to the global one described in Sect. 2.

The composed interpolation operators $\mathscr{R}_{\pi,M}$ and $\hat{\mathscr{R}}_{\pi,M}$ act now on $C_\pi([\bar{t}, \bar{t}+H], \mathbb{R}^m) \times \mathbb{R}^l$ and $C_\pi([\bar{t}, \bar{t}+H], \mathbb{R}^m) \times \mathbb{R}^k$,

$$\mathscr{R}_{\pi,M}\begin{bmatrix} w \\ r \end{bmatrix} = \begin{bmatrix} R_{\pi,M} & 0 \\ 0 & I_l \end{bmatrix}\begin{bmatrix} w \\ r \end{bmatrix}, \quad \hat{\mathscr{R}}_{\pi,M}\begin{bmatrix} w \\ \hat{r} \end{bmatrix} = \begin{bmatrix} R_{\pi,M} & 0 \\ 0 & I_k \end{bmatrix}\begin{bmatrix} w \\ \hat{r} \end{bmatrix}.$$

Let $U_{\pi,sub} : H^1_{D,sub} \to H^1_{D,sub}$ be the orthogonal projector of $H^1_{D,sub}$ onto $X_\pi \subset H^1_{D,sub}$.

Accordingly, we define $\alpha_{\pi,sub}$ and, furthermore, $\beta_{\pi,sub}$, $\gamma_{\pi,sub}$, $\tilde{\beta}_{\pi,sub}$, $\tilde{\gamma}_{\pi,sub}$, associated with the operator $\mathscr{T}_{sub}$ and, similarly, $\hat{\beta}_{\pi,sub}$, $\hat{\gamma}_{\pi,sub}$, $\hat{\tilde{\beta}}_{\pi,sub}$, $\hat{\tilde{\gamma}}_{\pi,sub}$ associated with $\hat{\mathscr{T}}_{sub}$.

The following lemma provides conditions for the existence of a function $G : [a, b] \to \mathbb{R}$ having the properties (3.2). The latter is a necessary prerequisite for the transition condition (1.4).

Lemma 3.1 *Let the operator T be fine with tractability index $\mu \geq 2$, characteristic values $0 < r_0 \leq \cdots \leq r_{\mu-1} < r_\mu = m$, $l = m - \sum_{i=0}^{\mu-1}(m - r_i)$, and the canonical projector function Π_{can}.*
Then there are continuously differentiable functions $G : [a, b] \to \mathbb{R}^{l\times m}$ and $K : [a, b] \to \mathbb{R}^{k\times k}$ such that

$$\operatorname{im} G(t) = \mathbb{R}^l, \quad \ker G(t) = \ker \Pi_{can}(t), \quad [I_l\ 0]K(t)D = G(t), \quad t \in [a, b],$$

$K(t)$ remains nonsingular on $[a, b]$, and, with $\kappa = (\max_{a\leq t\leq b}|K(t)|)^{-1}$,

$$|Dz| = |K(t)^{-1}K(t)Dz| \geq \kappa|K(t)Dz| \geq \kappa|G(t)z|, \quad z \in \mathbb{R}^k,\ t \in [a, b].$$

Proof We choose an admissible matrix function sequence with admissible projector functions $Q_0, \ldots, Q_{\mu-1}$, see [13, Section 2.2]. Denote $P_i = I - Q_i, \Pi_i = P_0 \cdots P_i$. Then, $\Pi_{\mu-1}$ and $D\Pi_{\mu-1}D^+$ are also projector functions, both with constant rank l. Since $D\Pi_{\mu-1}D^+$ is continuously differentiable, we find a continuously differentiable matrix function $\Gamma_{dyn} : [a, b] \to \mathbb{R}^{l\times k}$ so that

$$\operatorname{im} \Gamma_{dyn}(t) = \mathbb{R}^l, \quad \ker \Gamma_{dyn}(t) = \ker(D\Pi_{\mu-1}D^+)(t), \quad t \in [a, b].$$

Furthermore, there is a pointwise reflexive generalized inverse $\Gamma_{dyn}^- : [a, b] \to \mathbb{R}^{k\times l}$, also continuously differentiable, such that $\Gamma_{dyn}\Gamma_{dyn}^- = I$ and $\Gamma_{dyn}^-\Gamma_{dyn} = D\Pi_{\mu-1}D^+$. Similarly, we find constant-rank continuously differentiable matrix functions $\Gamma_{nil,i} : [a, b] \to \mathbb{R}^{(m-r_i)\times k}$ and pointwise generalized inverses $\Gamma_{nil,i}^- : [a, b] \to \mathbb{R}^{k\times(m-r_i)}$ such that

$$\Gamma_{nil,i}\Gamma_{nil,i}^- = I, \quad \Gamma_{nil,i}^-\Gamma_{nil,i} = D\Pi_{i-1}Q_iD^+, \quad i = 1, \ldots, \mu - 1.$$

The resulting $k \times k$ matrix function

$$K = \begin{bmatrix} \Gamma_{dyn} \\ \Gamma_{nil,1} \\ \vdots \\ \Gamma_{nil,\mu-1} \end{bmatrix} = \begin{bmatrix} \Gamma_{dyn} \\ \Gamma_{nil} \end{bmatrix}$$

remains nonsingular on $[a, b]$ owing to the decomposition $I_k = DD^+ = D\Pi_0Q_1D^+ + \cdots + D\Pi_{\mu-2}Q_{\mu-1}D^+ + D\Pi_{\mu-1}D^+$.

Set $G = \Gamma_{dyn}D = [I_l\ 0]KD$. This implies $\ker G(t) = \ker \Pi_{\mu-1}$. Taking into account the fact that $\ker \Pi_{\mu-1} = \ker \Pi_{can}$, see [13, Theorem 2.8], one has actually $\ker G(t) = \ker \Pi_{can}$.

Finally, we derive for $z \in \mathbb{R}^k,\ t \in [a,b]$,

$$|Dz|^2 = |K(t)^{-1}K(t)Dz|^2 \geq \kappa^2 |K(t)Dz|^2 = \kappa^2(|G(t)z|^2 + |\Gamma_{nil}(t)Dz|^2)$$
$$\geq \kappa^2 |G(t)z|^2,$$

which completes the proof. □

Lemma 3.2 *For $\bar{t} \in [a,b]$, $0 < H \leq b - \bar{t}$, and*

$$C_H = \left(\max\left(\frac{2}{H}, 2H\right)\right)^{1/2}$$

it holds that

$$|Dx(t)| \leq C_H \|Dx\|_{H^1_{sub}} \leq C_H \|x\|_{H^1_{D,sub}}, \quad t \in [\bar{t}, \bar{t}+H], \quad x \in H^1_{D,sub}.$$

Proof By definition, $x \in H^1_{D,sub}$ implies $u = Dx \in H^1_{sub}$. Since H^1_{sub} is continuously embedded in C_{sub}, it follows that

$$u(t) = u(s) + \int_s^t u'(\tau)\mathrm{d}\tau, \quad t, s \in [\bar{t}, \bar{t}+H],$$

which gives

$$|u(t)|^2 \leq 2|u(s)|^2 + 2\left(\int_s^t |u'(\tau)|\mathrm{d}\tau\right)^2 \leq 2|u(s)|^2 + 2H\int_{\bar{t}}^{\bar{t}+H} |u'(\tau)|^2\mathrm{d}\tau.$$

Integrating this inequality with respect to s leads to

$$H|u(t)|^2 \leq 2\int_{\bar{t}}^{\bar{t}+H} |u(s)|^2\mathrm{d}s + 2H^2\int_{\bar{t}}^{\bar{t}+H} |u'(\tau)|^2\mathrm{d}\tau.$$

Finally, with C_H as defined in the assertion, it holds that

$$\|u\|^2_{C_{sub}} \leq C_H^2 \|u\|^2_{H^1_{sub}} \leq C_H^2 \|x\|^2_{H^1_{D,sub}}$$

and the assertion follows. □

Lemma 3.3 *Let the function G fulfilling (3.2) with the bound c_G be given, and denote $c_T = (2\max\{\|A\|_\infty^2, \|B\|_\infty^2\})^{1/2}$.*

(1) *Then, for each subinterval, the inequalities*

$$\|T_{sub}x\|_{L^2_{sub}} \le c_T\|x\|_{H^1_{D,sub}}, \quad x \in H^1_{D,sub},$$
$$|T_{IC,sub}x| \le c_G C_H\|x\|_{H^1_{D,sub}}, \quad |T_{ICD,sub}x| \le C_H\|x\|_{H^1_{D,sub}}, \quad x \in H^1_{D,sub}, \tag{3.3}$$

are valid.

(2) *If $M \ge N+1$ and A, B are of class C^M, then there are constants C_{AB1}, C_{AB2}, both independent of the size H of the subinterval, such that*

$$\|R_{\pi,M}T_{sub}U_\pi x\|_{L^2_{sub}} \le C_{AB1}\|x\|_{H^1_{D,sub}}, \quad x \in H^1_{D,sub},$$
$$\|R_{\pi,M}T_{sub}U_\pi x - T_{sub}U_\pi x\|_{L^2_{sub}} \le C_{AB1}h^{M-N-1/2}\|x\|_{H^1_{D,sub}}, \quad x \in H^1_{D,sub}.$$

Proof

(1) Regarding that A, B are given on $[a, b]$, by straightforward computation we obtain

$$\|T_{sub}x\|^2_{L^2_{sub}} \le 2\max\{\|A\|^2_{\infty,sub}, \|B\|^2_{\infty,sub}\}\|x\|^2_{H^1_{D,sub}} \le c_T\|x\|^2_{H^1_{D,sub}}.$$

Applying Lemma 3.2 we find the inequalities (3.3).

(2) These inequalities can be verified analogously to the first two items of [8, Proposition 4.2]. □

We are now prepared to estimate the values $\alpha_{\pi,sub}$, $\beta_{\pi,sub}$, $\tilde\beta_{\pi,sub}$, $\hat\beta_{\pi,sub}$, and $\hat{\tilde\beta}_{\pi,sub}$.

Theorem 3.4 *Let the operator T described in Sect. 2 be fine with tractability index $\mu \ge 2$ and characteristic values $0 < r_0 \le \cdots \le r_{\mu-1} < r_\mu = m$, $l = m - \sum_{i=0}^{\mu-1}(m - r_i)$. Let the coefficients A, B, as well as the solution x_* of the IVP (1.6), (1.7) be sufficiently smooth. Let the function G with (3.2) be given and $[\bar t, \bar t + H] \subset [a, b]$.*

Then there are positive constants $\alpha_{\pi,sub}$, C_β, $\tilde C_\beta$, $\hat C_\beta$, $\hat{\tilde C}_\beta$ such that

$$\alpha_{\pi,sub} \le C_\alpha H^{1/2}h^N,$$
$$\beta_{\pi,sub} \le C_\beta h^N, \ \tilde\beta_{\pi,sub} \le \tilde C_\beta h^N,$$
$$\hat\beta_{\pi,sub} \le \hat C_\beta h^N, \ \hat{\tilde\beta}_{\pi,sub} \le \hat{\tilde C}_\beta h^N.$$

uniformly for all individual subintervals $[\bar t, \bar t + H]$ and all sufficient fine grids X_π.

Proof First we choose N nodes $0 < \tau_{*,1} < \cdots < \tau_{*,N} < 1$ and construct the interpolating function $p_{*,int} \in X_\pi$ so that

$$Dp_{*,int}(\bar{t}) = Dx_*(\bar{t}), \quad p_{*,int}(t_j + \tau_{*,i}h) = x_*(t_j + \tau_{*,i}h),\ i = 1, \ldots, N,\ j = 1, \ldots, n,$$

yielding

$$\|x_* - p_{*,int}\|_{\infty,sub} + \|(Dx_*)' - (Dp_{*,int})'\|_{\infty,sub} \le C_* h^N,$$

with a uniform constant C_* for all subintervals. C_* is determined by x_* and its derivatives given on $[a, b]$. Now we have also

$$\|x_* - p_{*,int}\|_{H^1_{D,sub}} \le C_* \sqrt{2H} h^N,$$

and therefore, with $C_\alpha = C_* \sqrt{2}$,

$$\alpha_{\pi,sub} = \|(I - U_{\pi,sub})x_*\|_{H^1_{D,sub}} = \|(I - U_{\pi,sub})(x_* - p_{*,int})\|_{H^1_{D,sub}} \le C_\alpha \sqrt{H} h^N.$$

Set $C_D = \sqrt{2} \max\{1, b - a\} C_\alpha$ such that $C_H \sqrt{H} C_\alpha \le C_D$ for all H. Using Lemma 3.2 we derive

$$|D((I - U_{\pi,sub})x_*)(\bar{t})| \le C_H \alpha_{\pi,sub} \le C_D h^N.$$

We derive further

$$\begin{aligned}
\beta^2_{\pi,sub} &= \|\mathscr{T}_{sub}(I - U_{\pi,sub})x_*\|^2_{Y_{sub}} \\
&= \|T_{sub}(I - U_{\pi,sub})x_*\|^2_{L^2_{sub}} + |G(\bar{t})D^+ D((I - U_{\pi,sub})x_*)(\bar{t})|^2 \\
&\le \|T_{sub}\|^2 \alpha^2_{\pi,sub} + c_G^2 C_D^2 h^{2N} \le (c_T^2 C_\alpha^2 (b - a) + c_G^2 C_D^2) h^{2N} = C^2_{\beta} h^{2N},
\end{aligned}$$

$$\begin{aligned}
\hat{\beta}^2_{\pi,sub} &= \|\hat{\mathscr{T}}_{sub}(I - U_{\pi,sub})x_*\|^2_{Y_{sub}} \\
&= \|T_{sub}(I - U_{\pi,sub})x_*\|^2_{L^2_{sub}} + |D((I - U_{\pi,sub})x_*)(\bar{t})|^2 \\
&\le \|T_{sub}\|^2 \alpha^2_{\pi,sub} + c_G^2 C_D^2 h^{2N} \le (c_T^2 C_\alpha^2 (b - a) + C_D^2) h^{2N} = \hat{C}^2_{\beta} h^{2N}.
\end{aligned}$$

Following [8, Section 2.3], we investigate also $w_* = T_{sub}(x_* - p_{*,int}) \in C_\pi([\bar{t}, \bar{t} + H], \mathbb{R}^m)$ and use the estimate (cf. [8, Section 2.3])

$$H^{-1/2} \|R_{\pi,M} w_*\|_{L^2,sub} \le \|R_{\pi,M} w_*\|_{\infty,sub} \le C_L \|w_*\|_{\infty,sub} \le \max\{\|A\|_\infty, \|B\|_\infty\} C_L h^N.$$

Here, C_L denotes a constant that depends only on the choice of the interpolation nodes $\tau_{*,1}, \ldots, \tau_{*,N}$. Then we derive

$$\begin{aligned}\|R_{\pi,M}T_{sub}(I-U_{\pi,sub})x_*\|_{L^2,sub} &\le \|R_{\pi,M}T_{sub}(I-U_{\pi,sub})(x_*-p_{*,int})\|_{L^2,sub}\\ &\le \|R_{\pi,M}T_{sub}(x_*-p_{*,int})\|_{L^2,sub}\\ &\qquad + \|R_{\pi,M}T_{sub}U_{\pi,sub}(x_*-p_{*,int})\|_{L^2,sub}\\ &\le \|R_{\pi,M}w_*\|_{L^2,sub} + C_{AB1}\|x_*-p_{*,int}\|_{H^1_{D,sub}}\\ &\le C_{RT}\sqrt{H}h^N,\end{aligned}$$

where $C_{RT} = C_L \max\{\|A\|_\infty, \|B\|_\infty\} + \sqrt{2}C_*C_{AB1}$. Therefore,

$$\begin{aligned}\tilde{\beta}^2_{\pi,sub} &= \|\mathscr{R}_{\pi,m}\mathscr{T}_{sub}(I-U_{\pi,sub})x_*\|^2_{\hat{Y}_{sub}}\\ &= \|R_{\pi,M}T_{sub}(I-U_{\pi,sub})x_*\|^2_{L^2_{sub}} + |G(\bar{t})D^+D((I-U_{\pi,sub})x_*)(\bar{t})|^2\\ &\le C^2_{RT}Hh^{2N} + c^2_Gc^2_Dh^{2N} \le C^2_{RT}(b-a)h^{2N} + c^2_Gc^2_Dh^{2N} = \tilde{C}^2_{\beta}h^{2N},\end{aligned}$$

$$\begin{aligned}\tilde{\hat{\beta}}^2_{\pi,sub} &= \|\mathscr{R}_{\pi,m}\hat{\mathscr{T}}_{sub}(I-U_{\pi,sub})x_*\|^2_{\hat{Y}_{sub}}\\ &= \|R_{\pi,M}T_{sub}(I-U_{\pi,sub})x_*\|^2_{L^2_{sub}} + |D((I-U_{\pi,sub})x_*)(\bar{t})|^2\\ &\le C^2_{RT}Hh^{2N} + C^2_Dh^{2N} \le C^2_{RT}(b-a)h^{2N} + C^2_Dh^{2N} = \tilde{\hat{C}}^2_{\beta}h^{2N}.\end{aligned}$$

□

3.2 *Overdetermined Collocation on $[\bar{t}, \bar{t}+H] \subset [a,b]$, with Accurately Stated Initial Condition at $\bar{t}$*

We ask if there are positive constants c_γ and $\tilde{c}_\gamma$ serving as lower bounds for all the individual constants characterizing the instability thresholds associated to each arbitrary subinterval $[\bar{t}, \bar{t}+H] \subset [a,b]$.

Theorem 3.5 *Let the operator T described in Sect. 2 be fine with tractability index $\mu \ge 2$ and characteristic values $0 < r_0 \le \cdots \le r_{\mu-1} < r_\mu = m$, $l = m - \sum_{i=0}^{\mu-1}(m-r_i)$. Let the coefficients A, B, the right-hand side $q \in \operatorname{im} T$, as well as the solution x_* of the IVP (1.6), (1.7) be sufficiently smooth. Let q_{sub} denote the restriction of q onto the subinterval $[\bar{t}, \bar{t}+H] \subset [a,b]$.*

Let a function G with (3.2) be given.

(1) *Then, for each arbitrary $r \in \mathbb{R}^l$, there is exactly one solution $x_{[r]}$ of the equation $\mathscr{T}_{sub} x = (q_{sub}, r)$ and*

$$\|x_{[r]} - x_*\|_{H^1_{D,sub}} \le c_{sub}\, |r - G(\bar{t})x_*(\bar{t})|.$$

$x_{[r]}$ coincides on the subinterval with x_*, if and only if $r = G(\bar{t})x_*(\bar{t})$. Furthermore, there is a bound C_p such that $c_{sub} \le C_p$ is valid for all subintervals.

(2) *If $M \ge N + 1$, there is a constant $C_\gamma > 0$ such that,*

$$\gamma_{\pi,sub} \ge C_\gamma h^{\mu-1}, \quad \|(\mathscr{T}_{sub} U_{\pi,sub})^+\|_{Y_{sub} \to H^1_{D,sub}} = \frac{1}{\gamma_{\pi,sub}} \le \frac{1}{C_\gamma h^{\mu-1}}$$

uniformly for all subintervals and sufficiently small stepsizes $h > 0$.

(3) *If $M \ge N + \mu$, there is a positive constant $\tilde{C}_\gamma = \frac{C_\gamma}{2}$ such that*

$$\|(\mathscr{R}_{\pi,M} \mathscr{T}_{sub} U_{\pi,sub})^+\|_{Y_{sub} \to H^1_{D,sub}} = \frac{1}{\tilde{\gamma}_{\pi,sub}} \le \frac{1}{\tilde{C}_\gamma h^{\mu-1}}$$

uniformly for all subintervals and sufficiently small stepsizes $h > 0$.

Proof

(1) This is a consequence of Proposition A.1 in the Appendix.

(2) The constant C_γ can be obtained by a careful inspection and adequate modification of the proof of [9, Theorem 4.1] on the basis of Proposition A.1 below instead of [9, Proposition 4.3]. Similarly to [9, Lemma 4.4], we provide the inequality

$$\|q\|^2_{Z_{sub}} \le \|q\|^2_\pi := \|q\|^2_{L^2_{sub}} + \sum_{i=1}^{\mu-1} \sum_{s=0}^{\mu-i} d_{i,s} \|(D\mathscr{L}_{\mu-i} q)^{(s)}\|^2_{L^2_{sub}}, \quad q \in Z_\pi,$$

with $Z_\pi = \{q \in L^2_{sub} | D\mathscr{L}_{\mu-i} q \in C^{\mu-i}_\pi([\bar{t}, \bar{t} + H], \mathbb{R}^k), i = 1, \ldots, \mu - 1\} \subset T_{sub} X_\pi$, with coefficients $d_{i,s}$ being independent of the subinterval.

(3) This statement proves by a slight modification of [8, Proposition 4.2]. □

Theorem 3.5 allows to apply homogeneous error estimations on all subintervals. Note that the involved constants C_α etc. may depend on N and M. For providing the function G with (3.2), the canonical nullspace $N_{can} = \ker \Pi_{can}$ must be available, not necessarily the canonical projector itself. Owing to [13, Theorem 2.8], it holds that $N_{can} = \ker \Pi_{\mu-1}$ for any admissible matrix function sequence, which makes N_{can} easier accessible. Nevertheless, though the function G is very useful in theory it is hardly available in practice.

For problems with dynamical degree $l = 0$ the canonical projector Π_{can} vanishes identically, that is, the initial condition is absent, and T_{sub} itself is injective. This happens, for example, for Jordan systems, see also Sect. 5.3. In those cases, with no initial conditions and no transfer the window-wise forward stepping works well.

Let $\tilde{x}_{\pi,old}$ be already computed as approximation of the solution x_* on an certain *old* subinterval of length H_{old} straight preceding the current one $[\bar{t}, \bar{t} + H]$. Motivated by Theorems 3.4 and 3.5 assume

$$\|\tilde{x}_{\pi,old} - x_*\|_{H^1_{sub,old}} \le C h_{old}^{N-\mu+1}$$

for sufficiently small stepsize h_{old}. Applying Lemma 3.2 we obtain

$$|D\tilde{x}_{\pi,old}(\bar{t}) - Dx_*(\bar{t})| \le C_{H_{old}} C h_{old}^{N-\mu+1}.$$

Next we apply overdetermined least-squares collocation on the current subinterval $[\bar{t}, \bar{t} + H]$. We use the transfer condition $r = G(\bar{t})\tilde{x}_{\pi,old}(\bar{t})$ to state the initial condition for the current subinterval. The overdetermined collocation generates the new segment $\tilde{x}_\pi$,

$$\tilde{x}_\pi = \operatorname{argmin}\{\|R_{\pi,M}(T_{sub}x - q)\|^2_{H^1_{D,sub}} + |G(\bar{t})x(\bar{t}) - G(\bar{t})\tilde{x}_{\pi,old}(\bar{t})|^2 | x \in X_\pi\},$$

which is actually an approximation of $x_{[r]}$ being neighboring to x_*, such that

$$\|\tilde{x}_\pi - x_{[r]}\|_{H^1_{D,sub}} \le \tilde{c} h^{N-\mu+1}.$$

Owing to Theorem 3.5 we have also

$$\begin{aligned}\|x_{[r]} - x_*\|_{H^1_{D,sub}} &\le c_{sub}|r - G(\bar{t})x_*(\bar{t})| = c_{sub}|G(\bar{t})\tilde{x}_{\pi,old}(\bar{t}) - G(\bar{t})x_*(\bar{t})| \\ &\le c_{sub}c_G C_{H_{old}} C h_{old}^{N-\mu+1}.\end{aligned}$$

If $h = h_{old}$, it follows that

$$\|\tilde{x}_\pi - x_*\|_{H^1_{D,sub}} \le C_{sub} h^{N-\mu+1}$$

with $C_{sub} = c_{sub}c_G C_{H_{old}} C + \tilde{c}$. This is the background which ensures the windowing procedure (1.2), (1.3), (1.4) to work.

4 Overdetermined Collocation on a Subinterval $[\bar{t}, \bar{t}+H] \subset [a,b]$, with Initial Conditions Related to $Dx(\bar{t})$

Here we proceed as in the previous section, but now we use the initial condition $Dx(\bar{t}) = \hat{r}$ instead of $G(\bar{t})x(\bar{t}) = r$, to avoid the use of the function G. Obviously, this formulation is easier to use in practice since D is given. However, in contrast to the situation in Theorem 3.5, the equation $\hat{\mathscr{T}}_{sub}x = (q_{sub}, \hat{r})$ is no longer solvable for arbitrary $\hat{r} \in \mathbb{R}^k$. For solvability, $\hat{r}$ must be consistent.

Theorem 4.1 *Let the operator T described in Sect. 2.2 be fine with tractability index $\mu \geq 2$ and characteristic values $0 < r_0 \leq \cdots \leq r_{\mu-1} < r_\mu = m$, $l = m - \sum_{i=0}^{\mu-1}(m - r_i)$. Let the coefficients A, B, the right-hand side $q \in \operatorname{im} T$, as well as the solution x_* of the IVP (1.6),(1.7) be sufficiently smooth. Then the following holds:*

(1) *$\hat{\mathscr{T}}_{sub}$ is injective.*
(2) *If $M \geq N+1$, there is a constant $\hat{C}_\gamma$ uniformly for all possible subintervals and sufficiently small stepsizes $h > 0$ such that*

$$\hat{\gamma}_{\pi,sub} \geq \hat{C}_\gamma h^{\mu-1}.$$

and hence

$$\|(\hat{\mathscr{T}}_{sub}U_{\pi,sub})^+\|_{\hat{Y}_{sub}} = \frac{1}{\hat{\gamma}_{\pi,sub}} \leq \frac{1}{\hat{C}_\gamma h^{\mu-1}}.$$

(3) *If $M \geq N+\mu$, there is a constants $\tilde{\hat{C}}_\gamma > 0$ uniformly for all possible subintervals and sufficiently small stepsizes $h > 0$, such that*

$$\|(\hat{\mathscr{R}}_{\pi,M}\hat{\mathscr{T}}_{sub}U_{\pi,sub})^+\|_{\hat{Y}_{sub}} = \frac{1}{\tilde{\hat{\gamma}}_{\pi,sub}} \leq \frac{1}{\tilde{\hat{C}}_\gamma h^{\mu-1}}.$$

Proof The assertions are straightforward consequences of Theorem 3.5 and Lemma 3.1.

$\hat{\mathscr{T}}x = 0$ means $Tx = 0$ and $Dx(\bar{t}) = 0$, thus also $G(\bar{t})x(\bar{t}) = [I_l\, 0]K(\bar{t})Dx(\bar{t}) = 0$, finally $\mathscr{T}x = 0$. Since $\mathscr{T}$ is injective it follows that $x = 0$. For $p \in X_\pi$,

$$\begin{aligned}\|\hat{\mathscr{T}}_{sub}p\|^2_{\hat{Y}_{sub}} &= \|T_{sub}p\|^2_{L^2_{sub}} + |Dp(\bar{t})|^2 \geq \|T_{sub}p\|^2_{L^2_{sub}} + \kappa^2|G(\bar{t})p(\bar{t})|^2 \\ &\geq \min\{1,\kappa^2\}\|\mathscr{T}_{sub}p\|^2_{Y_{sub}} \geq \min\{1,\kappa^2\}\left(C_\gamma h^{\mu-1}\|p\|_{H^1_{D,sub}}\right)^2,\end{aligned}$$

and

$$\|\hat{\mathscr{R}}_{\pi,M}\hat{\mathscr{T}}_{sub}p\|^2_{\hat{Y}_{sub}} = \|R_{\pi,M}T_{sub}p\|^2_{L^2_{sub}} + |Dp(\bar{t})|^2 \geq \|R_{\pi,M}T_{sub}p\|^2_{L^2_{sub}} + \kappa^2|G(\bar{t})p(\bar{t})|^2$$
$$\geq \min\{1,\kappa^2\}\|\mathscr{R}_{\pi,M}\mathscr{T}_{sub}p\|^2_{\hat{Y}_{sub}} \geq \min\{1,\kappa^2\}\left(\tilde{C}_\gamma h^{\mu-1}\|p\|_{H^1_{D,sub}}\right)^2. \square$$

In contrast to the situation in Sect. 3.2 the equation $\hat{\mathscr{T}}_{sub}x = (q_{sub},\hat{r})$ is no longer solvable for all $\hat{r}\in\mathbb{R}^k$. Recall that q_{sub} is the restriction of $q = Tx_*$ so that $q_{sub}\in \operatorname{im} T_{sub}$. Denote

$$\hat{y} = \begin{bmatrix} q_{sub} \\ Dx_*(\bar{t}) \end{bmatrix},\ \hat{y}^{[\delta]} = \begin{bmatrix} q_{sub} \\ \hat{r} \end{bmatrix},\ \delta = \|\hat{y}-\hat{y}^{[\delta]}\| = |Dx_*(\bar{t})-\hat{r}|,$$

and, following [11], we take $\hat{y}^{[\delta]}$ as noisy data and compute

$$\tilde{\hat{x}}^{[\delta]}_\pi = \operatorname{argmin}\{\|\hat{\mathscr{R}}_{\pi,M}(\hat{\mathscr{T}}_{sub}x - y^{[\delta]})\|^2_{L^2_{sub}\times\mathbb{R}^k} | x\in X_\pi\}$$
$$= \operatorname{argmin}\{\|R_{\pi,M}(T_{sub}x - q_{sub})\|^2_{L^2_{sub}} + |Dx(\bar{t})-\hat{r}|^2 | x\in X_\pi\}$$

and similarly,

$$\hat{x}^{[\delta]}_\pi = \operatorname{argmin}\{\|\hat{\mathscr{T}}_{sub}x - y^{[\delta]}\|^2_{L^2_{sub}\times\mathbb{R}^k} | x\in X_\pi\}$$
$$= \operatorname{argmin}\{\|T_{sub}x - q_{sub}\|^2_{L^2_{sub}} + |Dx(\bar{t})-\hat{r}|^2 | x\in X_\pi\}.$$

Applying the error representation [11, Equation (2.9)] we arrive at

$$\hat{\tilde{x}}^{[\delta]}_\pi - x_* = (\hat{\mathscr{R}}_{\pi,M}\hat{\mathscr{T}}U_\pi)^+(\hat{y}^{[\delta]}-\hat{y})$$
$$+ (\hat{\mathscr{R}}_{\pi,M}\hat{\mathscr{T}}U_\pi)^+\hat{\mathscr{R}}_{\pi,M}\hat{\mathscr{T}}_{sub}(I-U_\pi)x_* - (I-U_\pi)x_*$$

and, correspondingly,

$$\hat{x}^{[\delta]}_\pi - x_* = (\hat{\mathscr{T}}U_\pi)^+(\hat{y}^{[\delta]}-\hat{y}) + (\hat{\mathscr{T}}U_\pi)^+\hat{\mathscr{T}}_{sub}(I-U_\pi)x_* - (I-U_\pi)x_*.$$

Thus,

$$\|\hat{x}^{[\delta]}_\pi - x_*\|_{H^1_{D,sub}} \leq \frac{1}{\tilde{C}_\gamma h^{\mu-1}}\{\|\hat{y}^{[\delta]}-\hat{y}\| + \hat{\beta}_{\pi,sub}\} + \alpha_\pi = \frac{1}{\tilde{C}_\gamma h^{\mu-1}}\{\delta + \hat{\beta}_{\pi,sub}\} + \alpha_\pi,$$

$$\|\tilde{\hat{x}}^{[\delta]}_\pi - x_*\|_{H^1_{D,sub}} \leq \frac{1}{\tilde{\hat{C}}_\gamma h^{\mu-1}}\{\|\hat{y}^{[\delta]}-\hat{y}\| + \tilde{\hat{\beta}}_{\pi,sub}\} + \alpha_\pi = \frac{1}{\tilde{\hat{C}}_\gamma h^{\mu-1}}\{\delta + \tilde{\hat{\beta}}_{\pi,sub}\} + \alpha_\pi.$$

All these estimations can be put together in order to arrive at a recursive error estimation for the application of (1.3), (1.5). Unfortunately, this estimate is not sufficient for proving convergence of the windowing technique in contrast to the approach using accurately stated initial conditions of Sect. 3.2!

5 Time-Stepping with $b - a = LH$ and $H = nh$

We set now $H = (b-a)/L$, $w_\lambda = a + \lambda H,\ \lambda = 0, \ldots, L,$ and $h = H/n$, and study the somehow uniform time-stepping procedures.

5.1 Time-Stepping with Accurate Transfer Conditions

In the time-stepping approach corresponding to (1.3)–(1.4), the transfer conditions are given so that G is chosen according to (3.2). Let $\tilde{x}^{[\lambda]}$ be the approximation provided by the overdetermined least-squares collocation for the subinterval $[a + (\lambda-1)H, a + \lambda H]$ corresponding to the initial and transfer conditions

$$\begin{aligned} G_a \tilde{x}_\pi^{[1]}(a) &= r, \\ G(w_\lambda)\tilde{x}_\pi^{[\lambda]}(a + (\lambda-1)H) &= G(w_\lambda)\tilde{x}_\pi^{\lambda-1}(a + (\lambda-1)H), \quad \lambda > 1. \end{aligned}$$

Then we obtain from Theorem 3.5 and Lemma 3.2, for $\lambda = 1$,

$$\|\tilde{x}_\pi^{[1]} - x_*\|_{H^1_{D,sub}} \le \tilde{C} h^{N-\mu+1} =: d_1.$$

For $\lambda > 1$, let $r = G_\lambda \tilde{x}_\pi^{[\lambda-1]}(a + (\lambda-1)H)$. Then it holds

$$\begin{aligned} \|\tilde{x}_\pi^{[\lambda]} - x_*\|_{H^1_{D,sub}} &\le \|\tilde{x}_\pi^{[\lambda]} - x_{[r]}\|_{H^1_{D,sub}} + \|x_{[r]} - x_*\|_{H^1_{D,sub}} \\ &\le \tilde{C} h^{N-\mu+1} + C_p |r - G_\lambda x_*(a + (\lambda-1)H)| \\ &\le \tilde{C} h^{N-\mu+1} + C_p c_G C_H \|\tilde{x}_\pi^{[\lambda-1]} - x_*\|_{H^1_{D,sub}} \\ &\le \bar{C}(h^{N-\mu+1} + C_H \|\tilde{x}_\pi^{[\lambda-1]} - x_*\|_{H^1_{D,sub}}) =: d_\lambda \end{aligned}$$

where $\bar{C} = \max\{C_p c_G, \tilde{C}\}$. Hence,

$$d_1 \le \bar{C} h^{N-\mu+1}, \quad d_\lambda \le \bar{C}(C_H d_{\lambda-1} + h^{N-\mu+1}).$$

A solution of this recursion provides us with

$$d_\lambda \le \sum_{\iota=0}^{\lambda-1} \bar{C}(\bar{C}C_H)^\iota h^{N-\mu+1} = \bar{C}\frac{1-(\bar{C}C_H)^\lambda}{1-\bar{C}C_H}h^{N-\mu+1}.$$

A similar estimation can be derived for the least-squares approximations using the operator $(\mathscr{T}_{sub}U_{\pi,sub})^+$.

Example 5.1 The index-2 DAE with $k = 2,\ m = 3,\ l = 1$,

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} x \right)'(t) + \begin{bmatrix} \theta & -1 & -1 \\ \eta t(1-\eta t) - \eta & \theta & -\eta t \\ 1-\eta t & 1 & 0 \end{bmatrix} x(t) = q(t), \tag{5.1}$$

is taken from [10, Example 1.1]. One has $N_{can}(t) = \{z \in \mathbb{R}^3 |\ \eta t z_1 - z_2 = 0\}$ so that

$$G(t) = \begin{bmatrix} \eta t & -1 & 0 \end{bmatrix}$$

will do. We consider the DAE on the interval (0,1). The right-hand side q is chosen in such a way that

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

is a solution. This solution becomes unique if an appropriate initial condition is added. With $G_a = G(0)$, the initial condition becomes

$$G_a x(0) = G_a \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T = 0.$$

In the following experiments, $\eta = -25$ and $\theta = -1$ have been chosen. This allows for a comparison with the experiments in [10].

This problem is solved on equidistant grids using, for each polynomial degree N, $M = N + 1$ Gaussian collocation points scaled to (0, 1). The tables show the errors of the approximate solutions in $H_D^1(0,1)$. The columns labeled order contain an estimation k_{est} of the order

$$k_{\text{est}} = \log(\|x_\pi - x_*\|_{H_D^1(0,1)} / \|x_{\pi'} - x_*\|_{H_D^1(0,1)}) / \log 2.$$

Here, π' is obtained from π by stepsize halving. It should be noted that the norm is taken for the complete interval (0, 1) even in the windowing approach. In order

Table 1 Errors and estimation of the convergence order for (5.1) and $\bar{t} = 0$, $H = 1$ using $M = N + 1$

	$N = 1$		$N = 2$		$N = 3$		$N = 4$		$N = 5$	
n	Error	Order	Error	Order	Error	Order	Error	Order	Error	Order
10	1.21e+0		1.65e−1		2.84e−3		7.55e−6		2.82e−7	
20	1.12e+0	0.1	3.74e−2	2.1	5.04e−4	2.5	9.66e−7	3.0	1.51e−8	4.2
40	1.29e−0	−0.2	1.55e−2	1.3	9.59e−5	2.4	1.25e−7	2.9	7.74e−10	4.3
80	1.16e−0	0.2	6.65e−3	1.2	1.83e−5	2.4	1.31e−8	3.3	1.32e−10	2.6
160	9.80e−1	0.2	3.21e−3	1.0	3.05e−6	2.6	1.31e−9	3.3	1.75e−10	−0.4
320	8.63e−1	0.2	1.60e−3	1.0	4.94e−7	2.6	2.00e−10	2.7	3.62e−10	−1.1

Table 2 Errors and estimation of the convergence order for (5.1) and $n = 1$ using $H = 1/L$

	$N = 1$		$N = 2$		$N = 3$		$N = 4$		$N = 5$	
L	Error	Order	Error	Order	Error	Order	Error	Order	Error	Order
10	3.76e+0		2.19e−1		2.82e−3		9.34e−6		2.84e−7	
20	2.67e+0	0.5	7.62e−2	1.5	5.06e−4	2.5	1.29e−6	2.9	1.53e−8	4.2
40	1.77e+0	0.6	3.30e−2	1.2	9.72e−5	2.4	1.92e−7	2.7	7.90e−10	4.3
80	1.62e+0	0.1	1.39e−2	1.2	1.89e−5	2.4	2.38e−8	3.0	4.67e−11	4.1
160	1.65e+0	−0.0	5.06e−3	1.5	3.20e−6	2.6	2.26e−9	3.4	1.13e−10	−1.3
320	1.66e+0	−0.0	1.91e−3	1.4	5.26e−7	2.6	2.21e−10	3.4	1.46e−10	−0.4

to enable a comparison, we provide the results for solving the problem without windowing in Table 1. This corresponds to $\bar{t} = 0$ and $H = 1$.

In the next experiment, the time-stepping approach using accurately stated transfer conditions has been tested with $n = 1$. The results are shown in Table 2. □

A more complex example is presented in Sect. 6.

5.2 *Time-Stepping with Transfer Conditions Based on D*

In our experiments in fact, the situation is much better than indicated by the estimates in Sect. 4. The latter are not sufficient to show convergence of the present time-stepping approach when the transfer conditions are based on D, see (1.5).

Example 5.2 (Continuation of Example 5.1) We apply the time-stepping procedure under the same conditions as in Example 5.1, however, this time the transfer conditions are chosen as

$$\tilde{x}_i^{[\lambda]}(\bar{t}) = \tilde{x}_i^{[\lambda-1]}(\bar{t}), \quad i = 1, 2.$$

The results are presented in Table 3. The errors are slightly worse than those of Table 2 where accurately stated transfer conditions are used. However, the observed orders of convergence are similar, at least for $N \geq 2 = \mu - 1$. The values for $n = 2$ and $n = 3$ have also been checked. The orders are identical to those of Table 3

Table 3 Errors and estimation of the convergence order for (5.1) and $n = 1$ using $H = 1/L$

	$N = 1$		$N = 2$		$N = 3$		$N = 4$		$N = 5$	
L	Error	Order	Error	Order	Error	Order	Error	Order	Error	Order
10	1.80e+0		1.46e−1		3.27e−3		9.85e−6		3.16e−7	
20	2.36e+0	−0.4	4.65e−2	1.6	5.84e−4	2.5	1.35e−6	2.9	1.71e−8	4.2
40	2.77e+1	−3.5	1.66e−2	1.5	1.09e−4	2.4	1.75e−7	2.9	8.78e−10	4.3
80	5.07e+2	−4.2	6.64e−3	1.3	2.03e−5	2.4	1.76e−8	3.3	6.65e−11	3.7
160	1.11e+3	−1.1	3.19e−3	1.1	3.51e−6	2.5	1.60e−9	3.5	1.50e−10	−1.2
320	7.46e+2	0.6	1.59e−3	1.0	6.44e−7	2.4	1.85e−10	3.1	3.07e−10	−1.0

even if the errors are smaller due to the smaller stepsize h. For $N = 1$, divergent approximations are obtained. However, this is beyond the scope of our theoretical results even in the case of accurate transfer conditions. □

5.3 Studying the Damping of Inconsistent Transition Values

The results of the previous sections show that the windowing method converges if the transfer conditions used refer to the dynamic components, only. The latter are, in general, not easily available unless a detailed analysis of the DAE is available. However, so far we do not know any conditions for convergence if the practically accessible values of the differentiated components Dx are used in the transfer conditions.[7] Example 5.2 indicates, however, that the use of (1.5) may be possible. In order to gain some more insight into what could be expected in the setting of Sect. 5.2, we will consider a simple special case in this section.

The model problem in question here is a simple system featuring only one Jordan block,

$$\begin{aligned} J(Dx)' + x &= 0, \\ Dx(\bar{t}) &= r. \end{aligned}$$

Here, $J \in \mathbb{R}^{\mu\times(\mu-1)}$, $D \in \mathbb{R}^{(\mu-1)\times\mu}$ where

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

[7] In the index-1 case, Dx describes just the dynamic components such that convergence is assured for using all differentiated components. However, for index-1 DAEs, much more efficient collocation methods are available.

This system has index μ and no dynamic components, $l = 0$. The system is solvable for $r = 0$, only, leading to the unique solution $x_*(t) \equiv 0$. When trying to solve the system using the proposed windowing technique, the only information transferred from the subinterval $[\bar{t}, \bar{t}+H]$ to the next one is the value of the approximate solution x_π at the end of the interval, $Dx_\pi(\bar{t}+H)$. The latter is an approximation to the exact solution $Dx_*(\bar{t}+H) \equiv 0$ that cannot be guaranteed to be consistent with the DAE. Therefore, we ask the question of how $Dx_\pi(\bar{t}+H)$ depends on r.

Let

$$\begin{aligned} x_{[r],\pi} &= \operatorname{argmin}\{\|\hat{\mathscr{T}}_{sub}x\|^2_{L^2_{sub}\times\mathbb{R}^k} | x \in X_\pi\} \\ &= \operatorname{argmin}\{\|T_{sub}x\|^2_{L^2_{sub}} + |Dx(\bar{t}) - r|^2_{\mathbb{R}^k} | x \in X_\pi\} \end{aligned}$$

where $Tx = J(Dx)' + x$. Obviously, $Dx_{[r],\pi}(\bar{t}+H)$ depends linearly on r. There exists a matrix $S = S(N, H, n)$ such that $Dx_{[r],\pi}(\bar{t}+H) = Sr$ which we will denote as the transfer matrix. For convergence of the method, it is necessary that the spectral radius $\rho(S)$ of the transfer matrix is bounded by 1.

The analytical computation of S is rather tedious. After some lengthy calculations, we found that, for $\mu = 2$, it holds, with $\eta = (N+1)^{-1}$,

$$\begin{aligned} \rho(S(N, H, n) &= \eta^n \left| \frac{2}{\left(-1+\sqrt{1-\eta^2}\right)^n + \left(-1-\sqrt{1-\eta^2}\right)^n} \right| \\ &\approx \eta^n 2^{1-n}. \end{aligned}$$

In particular, $\rho(S)$ is independent of H and n can be chosen arbitrarily. Moreover, the damping of the inconsistent value r is the better the larger n is. This result can be compared to the experiments in Example 5.2 (an index-2 problem) where we cannot identify any influence of an inaccuracy due to inconsistent transfer conditions.

For larger values of μ, we determined $\rho(S)$ by numerical means. Results are shown in Tables 4, 5 and 6. We observe that, for an index $\mu > 2$, n must be chosen

Table 4 Spectral radius of the transfer matrix $S(N, H, n)$ for $n = 1$ and $H = 0.1$ (left panel) and $H = 0.01$ (right panel). The column headings show the index μ

N	2	3	4	5
2	3.3e−1	2.1e+0	1.3e+0	1.1e+0
3	2.5e−1	1.8e+0	5.9e+0	2.9e+0
4	2.0e−1	1.5e+0	7.1e+0	1.4e+1
5	1.7e−1	1.3e+0	7.0e+0	2.3e+1
6	1.5e−1	1.1e+0	6.5e+0	2.7e+1
7	1.2e−1	9.7e−1	6.1e+0	2.9e+1
8	1.1e−1	8.7e−1	5.6e+0	2.9e+1

N	2	3	4	5
2	3.3e−1	2.1e+0	1.1e+0	1.0e+0
3	2.5e−1	1.8e+0	5.9e+0	1.5e+0
4	2.0e−1	1.5e+0	7.1e+0	1.4e+1
5	1.7e−1	1.3e+0	7.0e+0	2.3e+1
6	1.5e−1	1.1e+0	6.6e+0	2.8e+1
7	1.2e−1	9.7e−1	6.1e+0	2.9e+1
8	1.1e−1	8.7e−1	5.6e+0	2.9e+1

Table 5 Spectral radius of the transfer matrix $S(N, H, n)$ for $n = 2$ and $H = 0.1$ (left panel) and $H = 0.01$ (right panel). The column headings show the index μ

N	2	3	4	5
2	5.9e−2	1.4e+0	1.5e+0	1.2e+0
3	3.2e−2	6.4e−1	8.0e+0	9.9e+0
4	2.0e−2	3.7e−1	5.0e+0	2.0e+1
5	1.4e−2	2.5e−1	3.1e+0	3.0e+1
6	1.0e−2	1.8e−1	2.1e+0	2.2e+1
7	7.9e−3	1.3e−1	1.5e+0	1.6e+1
8	6.2e−3	1.0e−1	1.2e+0	1.2e+1

N	2	3	4	5
2	5.9e−2	1.4e+0	1.2e+0	1.0e+0
3	3.2e−2	6.4e−1	8.2e+0	2.5e+0
4	2.0e−2	3.7e−1	5.0e+0	3.6e+2
5	1.4e−2	2.5e−1	3.1e+0	3.2e+2
6	1.0e−2	1.8e−1	2.1e+0	8.1e−1
7	7.9e−3	1.3e−1	1.5e+0	2.1e+0
8	6.2e−3	1.0e−1	1.2e+0	1.2e+1

Table 6 Spectral radius of the transfer matrix $S(N, H, n)$ for $n = 3$ and $H = 0.1$ (left panel) and $H = 0.01$ (right panel). The column headings show the index μ

N	2	3	4	5
2	1.0e−2	6.8e−1	1.8e+0	1.4e+0
3	4.1e−3	1.8e−2	6.1e+0	2.5e+0
4	2.1e−3	8.1e−2	2.1e+0	1.7e+1
5	1.2e−3	4.3e−2	9.2e−1	1.8e+1
6	7.4e−4	2.6e−2	5.1e−1	8.5e+0
7	4.9e−4	1.7e−2	3.1e−1	4.8e+0
8	3.5e−4	1.2e−2	2.1e−1	3.0e+0

N	2	3	4	5
2	1.0e−2	6.8e−1	1.3e+0	1.0e+0
3	4.1e−3	1.8e−1	6.3e+0	5.5e+0
4	2.1e−3	8.1e−2	2.1e+0	4.2e+1
5	1.2e−3	4.3e−2	9.2e−1	7.8e−1
6	7.4e−4	2.6e−2	5.1e−1	7.5e−1
7	4.9e−4	1.7e−2	3.1e−1	2.8e−1
8	3.5e−4	1.2e−2	2.1e−1	2.3e−1

larger than 1 in order to ensure $\rho(S) < 1$. Moreover, $\rho(S)$ depends on H only marginally for the investigated cases.

Details of the derivations are collected in the appendix.

6 A More Complex Example

In order to show the merits of the windowing technique, we will continue to use the example considered in [9]. This example is the linearized version of a test example from [5]. We consider an initial value problem for the DAE

$$A(Dx)'(t) + B(t)x(t) = y(t), \quad t \in [0, 5] \tag{6.1}$$

with

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

the smooth matrix coefficient

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

This DAE is obtained if the test example from [5] is linearized in the solution $x_*(t)$ considered there.[8] It has tractability index $\mu = 3$ and dynamical degree of freedom $l = 4$. In order to use the windowing technique with accurately stated initial conditions, we will need a function $G : [0, 5] \to \mathbb{R}^{4\times 7}$ fulfilling the assumptions of Theorem 3.5. The nullspace of the projector Π_2 has the representation

$$\ker \Pi_2 = \ker \begin{bmatrix} I - \Omega & 0 & 0 \\ \Omega'\Omega & I - \Omega & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \Omega = b(t)b(t)^T, \quad b(t) = \begin{bmatrix} -\cos^2 t \\ -\cos t \sin t \\ \sin t \end{bmatrix}.$$

Based on this representation, we can use

$$G(t) = \begin{bmatrix} \sin t & -\cos t & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & \cos t & 0 & 0 & 0 & 0 \\ -\cos^3 t & -\sin t\cos^2 t & \sin t\cos t & \sin t & -\cos t & 0 & 0 \\ -(\sin t\cos t)^2 & -\sin^3 t\cos t & \sin^3 t & 0 & 1 & \cos t & 0 \end{bmatrix}. \tag{6.2}$$

In the following numerical experiments we choose the exact solution

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

[8] Compare also [9, Sections 6.3 and 6.4].

Table 7 Errors and estimation of the convergence order for (6.1) and $\bar{t} = 0$, $H = 5$ using $M = N + 3$

	$N = 1$		$N = 2$		$N = 3$		$N = 4$		$N = 5$	
n	Error	Order	Error	Order	Error	Order	Error	Order	Error	Order
10	2.64e+0		5.24e−1		6.29e−2		6.33e−3		5.73e−4	
20	1.54e+0	0.8	1.99e−1	1.4	1.77e−2	1.8	9.39e−4	2.8	6.12e−5	3.2
40	8.79e−1	0.8	9.36e−2	1.1	6.44e−3	1.5	1.66e−4	2.5	7.31e−6	3.1
80	4.69e−1	0.9	4.63e−2	1.0	2.84e−3	1.2	3.42e−5	2.3	9.02e−7	3.0
160	3.00e−1	0.6	2.33e−2	1.0	1.37e−3	1.1	7.69e−6	2.2	1.12e−7	3.0
320	2.30e−1	0.4	1.18e−2	1.0	6.75e−4	1.0	1.82e−6	2.1	1.40e−8	3.0

which is also the one used in [9]. Setting $G_a = G(0)$, this provides us with the initial condition[9]

$$G_a x(0) = \begin{bmatrix} -1 \\ 3 \\ 0 \\ 0 \end{bmatrix}.$$

The problem is solved on equidistant grids using, for each polynomial degree N, $M = N+3$ Gaussian collocation points scaled to $(0, 1)$. This number of collocation points has been chosen such that the assumptions of Theorem 3.5(3) are fulfilled. The tables show the errors of the approximate solutions in $H_D^1(0, 5)$. Similarly as in previous examples, the columns labeled order contain an estimation k_{est} of the order

$$k_{\text{est}} = \log(\|x_\pi - x_*\|_{H_D^1(0,5)} / \|x_{\pi'} - x_*\|_{H_D^1(0,5)}) / \log 2.$$

Here, π' is obtained from π by stepsize halving.

In order to enable a comparison, we provide the results for solving the problem without windowing in Table 7. This corresponds to $\bar{t} = 0$ and $H = 5$. Note that the results are almost identical to those obtained in [9] using a slightly different formulation of the initial condition and a different number of collocation points.

In Tables 8, 9 and 10 the results using the windowing technique with transfer conditions (1.5) for different numbers of subdivisions n of the individual windows $[\bar{t}, \bar{t} + H]$ are shown. Since the transfer condition is based on all of the differentiated components Dx, they are expected to be inconsistent away from the initial point $t = 0$. For $n = 1$ and $N \leq 3$, the method delivers exponentially divergent approximations.

[9] This initial condition is slightly different from the one used in [9]. However, both conditions are equivalent.

Table 8 Errors and estimation of the convergence order for (6.1) and $n = 1$, $H = 5/L$ using $M = N + 3$

	$N = 4$		$N = 5$	
L	Error	Order	Error	Order
10	1.21e−2		7.18e−4	
20	2.28e−3	2.4	7.65e−5	3.2
40	5.16e−4	2.1	9.36e−6	3.0
80	1.25e−4	2.0	1.18e−6	3.0
160	3.10e−5	2.0	1.48e−7	3.0
320	7.74e−6	2.0	1.93e−8	2.9

Table 9 Errors and estimation of the convergence order for (6.1) and $n = 2$, $H = 5/L$ using $M = N + 3$

	$N = 1$		$N = 2$		$N = 3$		$N = 4$		$N = 5$	
L	Error	Order	Error	Order	Error	Order	Error	Order	Error	Order
10	2.30e+0		2.66e−1		2.99e−2		1.99e−3		7.64e−5	
20	1.64e+0	0.5	2.98e−1	−0.2	1.25e−2	1.3	4.89e−4	2.0	9.24e−6	3.0
40	1.49e+0	0.1	2.41e+1	−6.3	5.99e−3	1.1	1.22e−4	2.0	1.16e−6	3.0
80	1.45e+0	0.0	4.16e+5	−14.1	3.03e−3	1.0	3.06e−5	2.0	1.46e−7	3.0
160	1.44e+0	0.0	1.15e+14	−28.0	1.54e−3	1.0	7.65e−6	2.0	1.84e−8	3.0
320	1.44e+0	0.0	1.48e+31	−56.8	7.77e−4	1.0	1.91e−6	2.0	1.09e−8	0.8

Table 10 Errors and estimation of the convergence order for (6.1) and $n = 3$, $H = 5/L$ using $M = N + 3$

	$N = 1$		$N = 2$		$N = 3$		$N = 4$		$N = 5$	
L	Error	Order	Error	Order	Error	Order	Error	Order	Error	Order
10	1.74e+0		1.38e−1		1.38e−2		7.31e−4		2.05e−5	
20	1.64e+0	0.0	6.92e−2	1.0	6.20e−3	1.2	1.83e−4	2.0	2.53e−6	3.0
40	1.66e+0	0.0	3.95e−2	0.8	3.07e−3	1.0	4.61e−5	2.0	3.18e−7	3.0
80	1.67e+0	0.0	2.75e−2	0.5	1.55e−3	1.0	1.15e−5	2.0	3.99e−8	3.0
160	1.68e+0	0.0	2.35e−2	0.2	7.81e−4	1.0	2.89e−6	2.0	6.43e−9	2.6
320	1.68e+0	0.0	2.23e−2	0.1	3.93e−4	1.0	7.22e−7	2.0	2.41e−8	−1.9

Finally, we consider the case of using accurately stated initial conditions as transfer conditions. So they correspond to choosing $G(\bar{t})$ according to (6.2). The results are collected in Table 11. The latter can be compared to the behavior of the global method as shown in Table 7. The results are rather close to each other.

Table 11 Errors and estimation of the convergence order for (6.1) and accurately posed transfer conditions with $n = 1$, $H = 5/L$ and $M = N + 3$

	$N = 1$		$N = 2$		$N = 3$		$N = 4$		$N = 5$	
L	Error	Order	Error	Order	Error	Order	Error	Order	Error	Order
10	5.32e+0		5.12e−1		8.46e−2		1.20e−2		1.03e−3	
20	2.56e+0	1.1	2.67e−1	0.9	2.64e−2	1.7	2.47e−3	2.3	8.85e−5	3.5
40	2.20e+0	0.2	2.03e−1	0.4	1.09e−2	1.3	5.85e−4	2.1	9.51e−6	3.2
80	2.17e+0	0.0	1.88e−1	0.1	5.14e−3	1.1	1.44e−4	2.0	1.14e−6	3.1
160	2.17e+0	0.0	1.84e−1	0.0	2.53e−3	1.0	3.59e−5	2.0	1.40e−7	3.0
320	2.17e+0	0.0	1.83e−1	0.0	1.26e−3	1.0	8.97e−6	2.0	1.76e−8	3.0

7 Conclusions

We continued the investigation of overdetermined least-squares collocation using piecewise polynomial ansatz functions. This method is known to efficiently produce accurate numerical approximations of solutions for two-point boundary value problems for higher-index DAEs including IVPs as a special case. Since a further increase in computational efficiency is expected if modified for a customized application to IVPs, we considered time-stepping techniques for IVPs in this paper. It turned out that the success of such techniques depends strongly on the transfer conditions used. In the case that the intrinsic structure is available, meaning in particular that the dynamic solution components are known, the time-stepping method has convergence properties similar to the boundary value approach. However, if only the information about the differentiated components of the DAE is used, so far our estimates do not secure convergence of the time-stepping approach. Investigations of a model problem indicate that even in this case convergence can be obtained provided that the method parameters are chosen appropriately.

The overdetermined least-squares collocation method shows impressive convergence results in our experiments. On one hand, the accuracy is impressive, on the other hand, the computational efficiency is comparable to widely used collocation methods for ordinary differential equations. Opposed to that, there are severe difficulties to theoretically justify these methods. The underlying reason is the ill-posedness of higher-index DAEs. To the best of our knowledge, available convergence results are rather sparse and important questions of practical relevance for constructing efficient algorithms are completely open, e.g., *a-posteriori* error estimations, the choice of grids, polynomial orders, collocation points etc. However, the results so far are encouraging.

A Proof of Theorem 3.5

The Proposition A.1 below plays its role when verifying the statements of Theorem 3.5. We collect the necessary ingredients of the projector based DAE analysis to prove Proposition A.1. We refer to [13, 15] for more details. Let the DAE (1.6) be fine with tractability index $\mu \geq 2$ and characteristic values

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

Recall that this property is determined by the given coefficients $A : [a, b] \to \mathbb{R}^{m\times k}$, $D = [I\ 0] \in \mathbb{R}^{k\times m}$, and $B : [a, b] \to \mathbb{R}^{m\times m}$. A and B are sufficiently smooth, at least continuous. Then there are an admissible sequence of matrix valued functions starting from $G_0 := AD$ and ending up with a nonsingular G_μ, see [13, Definition 2.6], as well as associated projector valued functions

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

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

$$\begin{aligned} Q_i &= I - P_i,\ i = 0, \ldots, \mu - 1, \\ \Pi_0 := P_0,\ \Pi_i := \Pi_{i-1}P_i &\in C([a, b], \mathbb{R}^{m\times m}),\ i = 1, \ldots, \mu - 1, \\ D\Pi_i D^+ &\in C^1([a, b], \mathbb{R}^{k\times k}),\ i = 1, \ldots, \mu - 1. \end{aligned}$$

By means of the projector functions we decompose the unknown x and decouple the DAE itself into their characteristic parts, see [13, Section 2.4].

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

$$u' - (D\Pi_{\mu-1}D^+)'u + D\Pi_{\mu-1}G_\mu^{-1}B\Pi_{\mu-1}D^+u = D\Pi_{\mu-1}G_\mu^{-1}q. \tag{A.2}$$

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

$$\begin{bmatrix} 0 & \mathscr{N}_{12} & \cdots & \mathscr{N}_{1,\mu-1} \\ & 0 & \ddots & \vdots \\ & & \ddots & \mathscr{N}_{\mu-2,\mu-1} \\ & & & 0 \end{bmatrix} \begin{bmatrix} (Dv_1)' \\ \\ \vdots \\ (Dv_{\mu-1})' \end{bmatrix} \tag{A.3}$$

$$+ \begin{bmatrix} I & \mathscr{M}_{12} & \cdots & \mathscr{M}_{1,\mu-1} \\ & I & \ddots & \vdots \\ & & \ddots & \mathscr{M}_{\mu-2,\mu-1} \\ & & & I \end{bmatrix} \begin{bmatrix} v_1 \\ \\ \vdots \\ v_{\mu-1} \end{bmatrix} = \begin{bmatrix} \mathscr{L}_1 \\ \\ \vdots \\ \mathscr{L}_{\mu-1} \end{bmatrix} q.$$

The coefficients $\mathscr{N}_{i,j}$, $\mathscr{M}_{i\,j}$, and $\mathscr{L}_i$ are subsequently given. Finally, one has for $v_0 = Q_0 x$ the representation

$$v_0 = \mathscr{L}_0 y - \mathscr{H}_0 D^+ u - \sum_{j=1}^{\mu-1} \mathscr{M}_{0\,j} v_j - \sum_{j=1}^{\mu-1} \mathscr{N}_{0\,j} (Dv_j)'. \tag{A.4}$$

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

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

All coefficients in (A.2)–(A.4) are continuous on $[a, b]$ and explicitly given in terms of the used admissible matrix function sequence as

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

in which

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

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

$$j = 1, \ldots, \mu - 1.$$

Consider an arbitrary subinterval $[\bar{t}, \bar{t} + H] \subseteq [a, b]$ and use the function spaces

$$L^2_{sub} = L^2((\bar{t}, \bar{t} + H), \mathbb{R}^m), \quad H^1_{sub} = H^1((\bar{t}, \bar{t} + H), \mathbb{R}^k), \quad H^1_{D,sub} = \{x \in L^2_{sub} | Dx \in H^1_{sub}\},$$

equipped with their natural norms. Additionally, we introduce the function space (cf., [9, 15])

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

and its norm

$$\|q\|_{Z_{sub}} := \left(\|q\|^2_{L^2_{sub}} + \sum_{i=1}^{\mu-1} \|(Dv_i)'\|^2_{L^2_{sub}} \right)^{1/2}, \quad q \in Z_{sub}.$$

The latter function space is very special, it strongly depends on the decoupling coefficients which in turn are determined by the given data A, D, B.

We also assume a function $G : [a, b] \to \mathbb{R}^l$ with $G(t) = G(t)D^+D$ for all $t \in [a, b]$ to be given, and introduce the operator related to the subinterval $T_{sub} : H^1_{D,sub} \to L^2_{sub}$ and the composed operator $\mathscr{T}_{sub} : H^1_{D,sub} \to L^2_{sub} \times \mathbb{R}^l$, by

$$T_{sub}x = A(Dx)' + Bx, \quad \mathscr{T}_{sub}x = \begin{bmatrix} T_{sub}x \\ G(\bar{t})x(\bar{t}) \end{bmatrix}, \quad x \in H^1_{D,sub}.$$

Here, trivially, the restrictions of A and B to the subinterval are meant. The operators T_{sub} and $\mathscr{T}_{sub}$ are well-defined and bounded. Regarding

$$\begin{aligned}\|T_{sub}x\|^2_{L^2_{sub}} &= \int_{\bar t}^{\bar t+H} |A(t)(Dx)'(t) + B(t)x(t)|^2 \mathrm{dt}\\ &\le 2\max\{\max_{t\in[\bar t,\bar t+H]} |A(t)|^2, \max_{t\in[\bar t,\bar t+H]} |B(t)|^2\}\|x\|^2_{H^1_{D,sub}}\\ &\le 2\max\{\max_{t\in[a,b]} |A(t)|^2, \max_{t\in[a,b]} |B(t)|^2\}\|x\|^2_{H^1_{D,sub}}\end{aligned}$$

we see that there is an upper bound on the operator norm of T_{sub} uniformly for all subintervals. Similarly, supposing G to be bounded on $[a, b]$, there is a uniform upper bound for the norm of $\mathscr{T}_{sub}$, too.

Proposition A.1 *Let the DAE be fine on* $[a, b]$ *with characteristic values* (A.1) *and index* $\mu \ge 2$.

Let the function $G : [a, b] \to \mathbb{R}^l$ *be such that*

$$\ker G(t) = \ker \Pi_{\mu-1}(t),\ |G(t)| \le c_G,\ |G(t)^-| \le c_{G^-},\quad t \in [a, b],$$

in which c_G *and* c_{G^-} *denote constants and* $G(t)^-$ *is a reflexive generalized inverse of* $G(t)$. *Then it holds:*

(1) $\mathrm{im}\, T_{sub} = Z_{sub}$, $\mathrm{im}\, \mathscr{T}_{sub} = Z_{sub} \times \mathbb{R}^l$, $\ker \mathscr{T}_{sub} = \{0\}$.
(2) *The function space* Z_{sub} *equipped with the norm* $\|\cdot\|_{Z_{sub}}$ *is complete.*
(3) *There is a constant* c_Z, *uniformly for all subintervals* $[\bar t, \bar t + H] \subseteq [a, b]$, *such that*

$$\|x\|_{H^1_{D,sub}} \le c_Z\ (\|q\|^2_{Z_{sub}} + |r|^2)^{1/2} \ \textit{for all}\ \ q \in Z_{sub}, r \in \mathbb{R}^l, x = \mathscr{T}^{-1}_{sub}(q, r).$$

Note that such a functions G exists always. For instance, applying Lemma 3.1 one can set $G(t) = [I_l\ 0]K(t)D$ and supplement it by $G(t)^- = D^+K(t)^{-1}[I_l\ 0]^+$.

Proof

(1) The first assertions can be verified by means of the above decoupling formulas, which are given on $[a, b]$, and which are valid in the same way on each arbitrary subinterval, too. In particular, examining the equation $\mathscr{T}_{sub}x = 0$, we know from (A.3) that $q \in L^2_{sub}$, $q = 0$ implies $v_j = 0$ on the subinterval successively for $j = \mu - 1, \ldots, 1$. On the other hand, $G(\bar t)x(\bar t) = 0$ leads to $u(\bar t) = D\Pi_{\mu-1}(\bar t)x(\bar t) = D\Pi_{\mu-1}(\bar t)G(\bar t)^-G(\bar t)x(\bar t) = 0$. Since $u \in H^1_{sub}$ solves the homogeneous ODE (A.2) on the subinterval, u vanishes there identically. Finally, from (A.4) it follows that $v_0 = 0$, and hence, $x = 0$.
(2) Let $q_n \in Z_{sub}$ be a fundamental sequence with respect to the $\|\cdot\|Z_{sub}$-norm, and $v_{n,i} \in H^1_{D,sub}$, $i = 1, \ldots, \mu-1$, correspondingly defined by (A.3), further $w_{n,i} = (Dv_{n,i})'$, $i = 1, \ldots, \mu - 1$. Then there exists an elements $q_* \in L^2_{sub}$

such that $q_n \xrightarrow{L^2} q_*$ and there are further elements $w_{*,i} \in L^2((\bar{t}, \bar{t}+H), \mathbb{R}^k)$ so that $w_{n,i} \xrightarrow{L^2} w_{*,i}, i = 1, \ldots, \mu-1$. The first line of the associated relations (A.3) leads to $v_{n,\mu-1} = \mathscr{L}_{\mu-1} q_n \xrightarrow{L^2} \mathscr{L}_{\mu-1} q_* =: v_{*,\mu-1}$, $Dv_{n,\mu-1} = D\mathscr{L}_{\mu-1} q_n \xrightarrow{L^2} Dv_{*,\mu-1}$, thus $Dv_{*,\mu-1} \in H^1_{sub}$, $(Dv_{*,\mu-1})' = w_{*,\mu-1}$. The next lines of (A.3) successively for $j = 2, \ldots, \mu-1$ provide

$$
\begin{aligned}
v_{n,\mu-j} &= \mathscr{L}_{\mu-j} q_n - \sum_{i=1}^{j-1} \mathscr{N}_{\mu-j,\mu-j+i} (Dv_{n,\mu-j+i})' - \sum_{i=1}^{j-1} \mathscr{M}_{\mu-j,\mu-j+i} v_{n,\mu-j+i} \\
&\xrightarrow{L^2} \mathscr{L}_{\mu-j} q_* - \sum_{i=1}^{j-1} \mathscr{N}_{\mu-j,\mu-j+i} (Dv_{*,\mu-j+i})' - \sum_{i=1}^{j-1} \mathscr{M}_{\mu-j,\mu-j+i} v_{*,\mu-j+i} =: v_{*,\mu-j}, \\
Dv_{*,\mu-j} &\in H^1_{sub}, \quad (Dv_{*,\mu-j})' = w_{*,\mu-j},
\end{aligned}
$$

and eventually we arrive at $q_* \in Z_{sub}$.

(3) The operator T_{sub} is bounded also with respect to the new image space Z_{sub} equipped with the norm $\|\cdot\|_{Z_{sub}}$. Namely, for each $x \in H^1_{D,sub}$ owing to the decoupling it holds that

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

This leads to $\|T_{sub} x\|_{Z_{sub}} \le c^Z_{T_{sub}} \|x\|_{H^1_{D,sub}}$, with a uniform constant $c^Z_{T_{sub}}$ for all subintervals. In the new setting, the associated operator $\mathscr{T}_{sub} : H^1_{D,sub} \to Z_{sub} \times \mathbb{R}^l$ is a homeomorphism, and hence, its inverse is bounded. It remains to verify the existence of a uniform upper bound c_Z of the norm of $\mathscr{T}^{-1}_{sub}$. □

Let an arbitrary pair $(q, r) \in Z_{sub} \times \mathbb{R}^l$ be given and the solution $x \in H^1_{D,sub}$ of $\mathscr{T}_{sub} x = (q, r)$, i.e., $T_{sub} x = q$, $G(\bar{t}) x(\bar{t}) = r$. We apply again the decomposition of the solution $x = D^+ u + v_0 + v_1 + \cdots + v_{\mu-1}$ and the decoupling (A.2), (A.3), (A.4). Owing to the properties of the function G it holds that $u(\bar{t}) = D\Pi_{\mu-1}(\bar{t}) x(\bar{t}) = D\Pi_{\mu-1}(\bar{t}) G(\bar{t})^- G(\bar{t}) x(\bar{t}) = D\Pi_{\mu-1}(\bar{t}) G(\bar{t})^- r$ and thus

$$
|u(\bar{t})| \le k_1 |r|,
$$

with a constant k_1 being independent of the subinterval. Below, all the further constants k_i are also uniform ones for all subintervals.

Let $U(t, \bar{t})$ denote the fundamental solution matrix normalized at $\bar{t}$ of the ODE (A.2). U is defined on the original interval $[a, b]$, there continuously differentiable and nonsingular. $U(t, \bar{t})$ and $U(t, \bar{t})^{-1} = U(\bar{t}, t)$ are uniformly bounded on

$[a, b]$. Turning back to the subinterval we apply the standard solution representation

$$u(t) = U(t, \bar{t})u(\bar{t}) + \int_{\bar{t}}^{t} U(t, s)D\Pi_{\mu-1}(s)G_{\mu}^{-1}(s)q(s)\mathrm{ds}$$
$$= U(t, \bar{t})D\Pi_{\mu-1}(\bar{t})G(\bar{t})^{-}r + \int_{\bar{t}}^{t} U(t, s)D\Pi_{\mu-1}(s)G_{\mu}^{-1}(s)q(s)\mathrm{ds},\ t \in [\bar{t}, \bar{t} + H].$$

Taking into account that the involved coefficients are defined on $[a, b]$ and continuous there we may derive an inequality

$$\|u\|_{H^1_{sub}}^2 \le k_2|r|^2 + k_3\|q\|_{sub}^2.$$

Next we rearrange system (A.3) to

$$\begin{bmatrix} v_1 \\ \vdots \\ v_{\mu-1} \end{bmatrix} = \mathfrak{M}^{-1} \begin{bmatrix} \mathscr{L}_1 \\ \vdots \\ \mathscr{L}_{\mu-1} \end{bmatrix} q - \mathfrak{M}^{-1} \begin{bmatrix} 0 & \mathscr{N}_{12} & \cdots & \mathscr{N}_{1,\mu-1} \\ & 0 & \ddots & \vdots \\ & & \ddots & \mathscr{N}_{\mu-2,\mu-1} \\ & & & 0 \end{bmatrix} \begin{bmatrix} (Dv_1)' \\ \\ \vdots \\ (Dv_{\mu-1})' \end{bmatrix},$$

in which the inverse of the matrix function

$$\mathfrak{M} = \begin{bmatrix} I & \mathscr{M}_{12} & \cdots & \mathscr{M}_{1,\mu-1} \\ & I & \ddots & \vdots \\ & & \ddots & \mathscr{M}_{\mu-2,\mu-1} \\ & & & I \end{bmatrix}$$

is again continuous on $[a, b]$ and upper triangular. This allows to derive the inequalities

$$\|v_j\|_{L^2_{sub}}^2 \le k_4\|q\|_{L^2_{sub}}^2 + k_5 \sum_{i=1}^{\mu-1} \|(Dv_i)'\|_{L^2_{sub}}^2, \quad j = 1, \ldots, \mu - 1.$$

Considering also (A.4) we obtain

$$\|x\|_{L^2_{sub}}^2 \le k_6\|q\|_{L^2_{sub}}^2 + k_7 \sum_{i=1}^{\mu-1} \|(Dv_i)'\|_{L^2_{sub}}^2 + k_8|r|^2.$$

Since $(Dx)' = u' + \sum_{i=1}^{\mu-1}(Dv_i)'$ we have further

$$\|x\|^2_{H^1_{D,sub}} \le k_9 \left\{ \|q\|^2_{L^2_{sub}} + \sum_{i=1}^{\mu-1} \|(Dv_i)'\|^2_{L^2_{sub}} + |r|^2 \right\} = k_9(\|q\|^2_{Z_{sub}} + |r|^2).$$

B On the Derivation of the Transfer Matrix $S(N, H, n)$

Consider an interval $(0, h)$. For the representation of polynomials we will use the Legendre polynomials P_k [18]. They have the properties

1. $\int_{-1}^{1} P_k(t)P_l(t)dt = \frac{2}{2k+1}\delta_{kl}$, $k, l = 0, 1, \ldots$.
2. $P_k(1) = 1$, $P_k(-1) = (-1)^k$, $k = 0, 1, \ldots$.
3. $P'_{k+1} - P'_{k-1} = (2k+1)P_k$, $k = 1, 2, \ldots$.

Let

$$p_k(t) = a_k P_k(1 - \frac{2}{h}t), \quad a_k = \left(\frac{2k+1}{h}\right)^{1/2}.$$

Then it holds

$$\int_0^h p_k(t)p_l(t)dt = \delta_{kl}, \quad p_k(0) = a_k, \quad p_k(h) = (-1)^k a_k.$$

From the representation for the derivatives, we obtain

$$\frac{h}{2}(c_k p'_{k-1} - d_k p'_{k+1}) = (2k+1)p_k$$

where

$$c_k = \frac{a_k}{a_{k-1}} = \left(\frac{2k+1}{2k-1}\right)^{1/2}, \quad d_k = \frac{a_k}{a_{k+1}} = \left(\frac{2k+1}{2k+3}\right)^{1/2}.$$

Since $p_0(t) \equiv a_0$ and $p'_1(t) = -2a_1/h$, we have the representation

$$\frac{h}{2}\bar{\Gamma}\begin{bmatrix} p'_1 \\ \vdots \\ p'_N \end{bmatrix} = D_- \begin{bmatrix} p_0 \\ \vdots \\ p_{N-1} \end{bmatrix},$$

with

$$\bar{\Gamma} = \begin{bmatrix} -d_0 & & & & \\ 0 & -d_1 & & & \\ c_2 & 0 & -d_2 & & \\ & \ddots & \ddots & \ddots & \\ & & c_{N-1} & 0 & -d_{N-1} \end{bmatrix}, \quad D_- = \begin{bmatrix} 1 & & & \\ & 3 & & \\ & & \ddots & \\ & & & 2N-1 \end{bmatrix}.$$

This provides

$$\begin{bmatrix} p_0' \\ \vdots \\ p_N' \end{bmatrix} = \frac{2}{h} \Gamma \begin{bmatrix} p_0 \\ \vdots \\ p_N \end{bmatrix}, \quad \Gamma = \begin{bmatrix} 0 & 0 \\ \bar{\Gamma}^{-1} D_- & 0 \end{bmatrix}.$$

A representation of Γ being more suitable for the subsequent derivations can be obtained by observing that

$$\bar{\Gamma} = D_-^{1/2} \begin{bmatrix} -1 & & & & \\ 0 & -1 & & & \\ 1 & 0 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 0 & -1 \end{bmatrix} D_+^{-1/2}, \quad D_+ = \begin{bmatrix} 3 & & & \\ & 5 & & \\ & & \ddots & \\ & & & 2N+1 \end{bmatrix}.$$

Let Z denote the tridiagonal matrix in this decomposition. Then it holds

$$(Z^{-1})_{ij} = \begin{cases} -1, & i \geq j, i-j \text{ even}, \\ 0, & \text{else}. \end{cases}$$

Hence,

$$\Gamma = \begin{bmatrix} 0 \\ D_+^{1/2} Z^{-1} D_-^{1/2} \; 0 \end{bmatrix} = D^{1/2} Y D^{1/2}$$

where

$$D = \begin{bmatrix} 1 & & & \\ & 3 & & \\ & & \ddots & \\ & & & 2N+1 \end{bmatrix}, \quad Y = -\begin{bmatrix} 0 & & & & & & & \\ 1 & 0 & & & & & & \\ 0 & 1 & 0 & & & & & \\ 1 & 0 & 1 & 0 & & & & \\ 0 & 1 & 0 & 1 & 0 & & & \\ 1 & 0 & 1 & 0 & 1 & 0 & & \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ & \cdots & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}.$$

Assume now

$$x_i = \sum_{k=0}^{N} \alpha_{ik} p_k, \quad \mathbf{a} = (a_0, \ldots, a_N)^T.$$

Then

$$x_i(0) = \sum_{n=0}^{N} \alpha_{in} p_n(0) = \sum_{n=0}^{N} \alpha_{in} a_n = \mathbf{a}^T \alpha_i \quad i = 1, \ldots, \mu - 1.$$

We collect the coefficients $\alpha = (\alpha_1, \ldots, \alpha_{\mu-1})^T$ and set

$$A = \begin{bmatrix} I & & & \\ \frac{2}{h} \Gamma^T & I & & \\ & \ddots & \ddots & \\ & & \frac{2}{h} \Gamma^T & I \end{bmatrix}, \quad C = \begin{bmatrix} \mathbf{a}^T & & \\ & \ddots & \\ & & \mathbf{a}^T \end{bmatrix}.$$

Let now $H > 0$ be fixed and $h = H/n$ for a given positive integer n. The functional to be minimized is

$$\Phi_{sub}(x) = \frac{1}{2} \|x_1\|^2_{L^2_{sub}} + \sum_{i=2}^{\mu} \frac{1}{2} \|x'_{i-1} - x_i\|^2_{L^2_{sub}}, \quad x = (x_1, \ldots, x_\mu)^T$$

on X_π under the condition $x_i = r_i$, $i = 1, \ldots, \mu - 1$. The term for $i = \mu$ in this sum can be omitted since, for given $x_{\mu-1} \in \mathscr{P}_N$, $x_\mu \in \mathscr{P}_{N-1}$ can always be set to $x'_{\mu-1}$ such that the last term amounts to 0.

For a shorthand notation, define $x_i^\nu = x_i|_{((\nu-1)h,\nu h)}$. Assuming the representation

$$x_i^\nu = \sum_{k=0}^{N} \alpha_{ik}^\nu p_k^\nu$$

on $((\nu-1)h, \nu h)$ with p_k^ν being the polynomials p_k transformed onto $((\nu-1)h, \nu h)$, we obtain

$$\Phi_{sub}(x) = \sum_{\nu=1}^{n} \left(\frac{1}{2} \left|\alpha_1^\nu\right|^2 + \frac{1}{2} \sum_{i=2}^{\mu-1} \left| \frac{2}{h} \Gamma^T \alpha_{i-1}^\nu + \alpha_i^\nu \right|^2 \right)$$

where $\alpha_i^\nu = (\alpha_{i0}^\nu, \ldots, \alpha_{iN}^\nu)^T$. Furthermore,

$$x_i^{\nu-1}(\nu h) = \sum_{k=0}^{N} \alpha_{ik}^{\nu-1} p_k(h) = \sum_{k=0}^{N} \alpha_{ik}^{\nu-1} a_k (-1)^k$$

for $i = 1, \ldots, \mu - 1$. Define $\mathbf{b} = (a_0, \ldots, (-1)^N a_N)^T$.

All these equations can be conveniently written down in a matrix fashion. The initial condition becomes

$$C\alpha^1 = r$$

while the transfer conditions read

$$B\alpha^{\nu-1} = C\alpha^\nu, \quad \nu = 2, \ldots, n$$

with

$$B = \begin{bmatrix} \mathbf{b}^T & & \\ & \ddots & \\ & & \mathbf{b}^T \end{bmatrix}.$$

Let $\boldsymbol{\alpha} = (\alpha^1, \ldots, \alpha^n)^T$ and

$$\mathscr{A} = \begin{bmatrix} A & & & \\ & A & & \\ & & \ddots & \\ & & & A \end{bmatrix}, \quad \mathscr{C} = \begin{bmatrix} C & & & \\ -B & C & & \\ & \ddots & \ddots & \\ & & -B & C \end{bmatrix}.$$

Note that $\mathscr{A}$ is nonsingular since A is so. Similarly, $\mathscr{C}$ has full row rank since C has the same property.

Finally, we obtain

$$\Phi_{sub}(x) = \Phi_{sub}(\boldsymbol{\alpha}) = \frac{1}{2}|\mathscr{A}\boldsymbol{\alpha}|^2 \to \min \text{ such that } \mathscr{C}\boldsymbol{\alpha} = (r, 0, \ldots, 0)^T.$$

The transfer matrix is then given by

$$S(N, H, n)r = B\alpha^n(r) \text{ for all } r \in \mathbb{R}^{\mu-1}.$$

In the case $\mu = 2$, a simple analytical solution is feasible.

B.1 The Case $\mu = 2$

In order to simplify the notation, the index i will be omitted. The transfer matrix reduces to a scalar

$$\rho_n = \left| \frac{x^n(H)}{r} \right|.$$

The Lagrange functional belonging to the present optimization problem reads

$$\varphi(\alpha, \lambda) = \sum_{\nu=1}^{n} \sum_{k=0}^{N} (\alpha_k^1)^2 + \lambda_1 \left(\sum_{k=0}^{N} a_k \alpha_k^1 - r \right) + \sum_{\nu=2}^{n} \lambda_\nu \left(\sum_{k=0}^{N} a_k \alpha_k^\nu - \sum_{k=0}^{N} (-1)^k a_k \alpha_k^{\nu-1} \right).$$

In the following, we will use the notations

$$a = \sum_{k=0}^{N} a_k^2 = \frac{1}{h}(N+1)^2, \quad b = \sum_{k=0}^{N} (-1)^k a_k^2 = \frac{(-1)^N}{h}(N+1), \quad c = \left| \frac{b}{a} \right|.$$

The derivatives of the Lagrange functional are

$$\begin{aligned}
\frac{\partial \varphi}{\partial \lambda_1} &= \sum_{k=0}^{N} a_k \alpha_k^1 - r, \\
\frac{\partial \varphi}{\partial \lambda_\nu} &= \sum_{k=0}^{N} a_k \alpha_k^\nu - \sum_{k=0}^{N} a_k (-1)^k \alpha_k^{\nu-1}, \quad \nu = 2, \ldots, n \\
\frac{\partial \varphi}{\partial \alpha_k^n} &= \alpha_k^n + \lambda_n a_k, \\
\frac{\partial \varphi}{\partial \alpha_k^\nu} &= \alpha_n^\nu + \lambda_\nu a_k - \lambda_{\nu+1} a_k (-1)^k, \quad \nu = 1, \ldots, n-1.
\end{aligned}$$

Hence, for $\nu = 1$,

$$\begin{aligned}
r &= \sum_{k=0}^{N} a_k \alpha_k^1 = \sum_{k=0}^{N} a_k \left(\lambda_2 a_k (-1)^k - \lambda_1 a_k \right) \\
&= b\lambda_2 - a\lambda_1.
\end{aligned}$$

Similarly, for $\nu = n$,

$$\begin{aligned}
0 &= \sum_{k=0}^{N} a_k \alpha_k^n - \sum_{k=0}^{N} a_k (-1)^k \alpha_k^{\nu-1} \\
&= -\sum_{k=0}^{N} a_k^2 \lambda_n + \sum_{k=0}^{N} a_k (-1)^k \left(\lambda_{n-1} a_k - \lambda_n a_k (-1)^k \right) \\
&= b\lambda_{n-1} - 2a\lambda_n.
\end{aligned}$$

And finally, for $1 < \nu < n$,

$$\begin{aligned}
0 &= \sum_{k=0}^{N} a_k \alpha_k^\nu - \sum_{k=0}^{N} a_k (-1)^k \alpha_k^{\nu-1} \\
&= \sum_{k=0}^{N} a_k \left(\lambda_{\nu+1} a_k (-1)^k - \lambda_\nu a_k \right) - \sum_{k=0}^{N} a_k (-1)^k \left(\lambda_\nu a_k (-1)^k - \lambda_{\nu-1} a_k \right) \\
&= b\lambda_{\nu+1} - 2a\lambda_\nu + b\lambda_{\nu-1}.
\end{aligned}$$

This provides us with the linear system of equations

$$\begin{bmatrix} -a & b & & & \\ b & -2a & b & & \\ & \ddots & \ddots & \ddots & \\ & & b & -2a & b \\ & & & b & -2a \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_{n-1} \\ \lambda_n \end{bmatrix} = \begin{bmatrix} r \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}.$$

Since $x^n(H) = \sum_{k=0}^{N} a_k(-1)^k \alpha_k^n = -\sum_{k=0}^{N} a_k(-1)^k \lambda_n a_k = -b\lambda_n$ it is sufficient to compute the last component λ_n of the solution to this system. Let A_n denote the system matrix and $\tilde{A}_n$ the matrix obtained from A_n by replacing the last column of A_n by the right-hand side. According to Cramer's rule it holds

$$\lambda_n = \frac{\det A_n}{\det \tilde{A}_n}.$$

Let $u_n = \det A_n$ and $v_n = \det \tilde{A}_n$. Then we obtain the recursion

$$\begin{aligned} v_1 &= r, \\ v_\nu &= -bv_{\nu-1}. \end{aligned}$$

Its solution is given by

$$v_\nu = (-b)^{\nu-1} r.$$

Analogously, we have

$$\begin{aligned} u_1 &= -a, \\ u_2 &= 2a^2 - b^2, \\ u_\nu &= -2au_{\nu-1} - b^2 u_{\nu-2}. \end{aligned}$$

This recursion is a simple difference equation with the general solution

$$u_\nu = c_1 z_1^\nu + c_2 z_2^\nu,$$

where

$$z_{1,2} = a\left(-1 \pm \sqrt{1 - c^2}\right).$$

Application of the initial condition leads to $c_1 = c_2 = 1/2$. Inserting these expressions we obtain

$$\begin{aligned}\rho_n &= \left|\frac{x^n(H)}{r}\right| \\ &= \left|\frac{-b\lambda_n}{r}\right| \\ &= \left|\frac{2b^n}{a^n\left[\left(-1+\sqrt{1-c^2}\right)^n + \left(-1-\sqrt{1-c^2}\right)^n\right]}\right| \\ &= c^n\left|\frac{2}{\left(-1+\sqrt{1-c^2}\right)^n + \left(-1-\sqrt{1-c^2}\right)^n}\right|.\end{aligned}$$

From the definition of c we obtain $c = (N+1)^{-1}$. Hence, $\sqrt{1-c^2} \approx 1$ such that

$$\rho_L \approx c^L 2^{1-L}.$$

B.2 An Approach for $\mu > 2$

In the case $\mu > 2$, the steps taken in the case $\mu = 2$ can be repeated. The Lagrangian system for the constraint optimization problem reads

$$\begin{bmatrix} \mathscr{A}^T\mathscr{A} & \mathscr{C}^T \\ \mathscr{C} & 0 \end{bmatrix}\begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{r} \end{bmatrix}$$

where

$$\mathbf{r} = (r, 0, \ldots, 0)^T.$$

The computation steps are then

(i) $\boldsymbol{\alpha} = -(\mathscr{A}^T\mathscr{A})^{-1}\mathscr{C}^T\boldsymbol{\lambda}$
(ii) $\boldsymbol{\lambda} = -[\mathscr{C}(\mathscr{A}^T\mathscr{A})^{-1}\mathscr{C}^T]^{-1}\mathbf{r}$
(iii) $\boldsymbol{\alpha} = (\mathscr{A}^T\mathscr{A})^{-1}\mathscr{C}^T[\mathscr{C}(\mathscr{A}^T\mathscr{A})^{-1}\mathscr{C}^T]^{-1}\mathbf{r}$
(iv) $x(H) = B\boldsymbol{\alpha} = B(\mathscr{A}^T\mathscr{A})^{-1}\mathscr{C}^T[\mathscr{C}(\mathscr{A}^T\mathscr{A})^{-1}\mathscr{C}^T]^{-1}\mathbf{r}$.

In the end, this yields

$$S(N, H, n) = B(\mathscr{A}^T\mathscr{A})^{-1}\mathscr{C}^T[\mathscr{C}(\mathscr{A}^T\mathscr{A})^{-1}\mathscr{C}^T]^{-1}.$$

This representation can easily be evaluated using symbolic computations. It should be mentioned that most terms in $S(N, H, n)$ lead to simple rational expressions in N. However, the results presented in Sect. 5.3 have been computed numerically.

References

1. Brenan, K.E., Campbell, S.L., Petzold, L.R.: Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations. North-Holland, Elsevier Science Publishing, Amsterdam (1989)
2. Burger, M., Gerdts, M.: In: Ilchmann, A., Reis, T. (eds.) Surveys in Differential-Algebraic Equations IV, chap. A Survey on Numerical Methods for the Simulation of Initial Value Problems with sDAEs, pp. 221–300. Differential-Algebraic Equations Forum. Springer, Heidelberg (2017)
3. Campbell, S.L.: The numerical solution of higher index linear time varying singular systems of differential equations. SIAM J. Sci. Stat. Comput. **6**(2), 334–348 (1985)
4. Campbell, S.L.: A computational method for general nonlinear higher index singular systems of differential equations. IMACS Trans. Sci. Comput. **1**(2), 555–560 (1989)
5. Campbell, S.L., Moore, E.: Constraint preserving integrators for general nonlinear higher index DAEs. Num. Math. **69**, 383–399 (1995)
6. Griepentrog, E., März, R.: Differential-Algebraic Equations and Their Numerical Treatment. Teubner Texte zur Mathematik 88. BSB Teubner Leipzig (1986)
7. Hanke, M., März, R.: Convergence analysis of least-squares collocation metods for nonlinear higher-index differential-algebraic equations. J. Comput. Appl. Math. (2019). doi:10.1016/j.cam.2019.112514
8. Hanke, M., März, R.: A reliable direct numerical treatment of differential-algebraic equations by overdetermined collocation: An operator approach. J. Comput. Appl. Math. (2019). doi:10.1016/j.cam.2019.112510
9. Hanke, M., März, R., Tischendorf, C.: Least-squares collocation for higher-index linear differential-algebaic equations: Estimating the stability threshold. Math. Comput. **88**(318), 1647–1683 (2019). https://doi.org/10.1090/mcom/3393
10. 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). http://dx.doi.org/10.1016/j.cam.2016.12.017
11. Kaltenbacher, B., Offtermatt, J.: A convergence analysis of regularization by discretization in preimage space. Math. Comput. **81**(280), 2049–2069 (2012)
12. Kunkel, P., Mehrmann, V.: Differential-Algebraic Equations. Analysis and Numerical Solution. Textbooks in Mathematics. European Mathematical Society, Zürich (2006)
13. Lamour, R., März, R., Tischendorf, C.: In: Ilchmann, A., Reis, T. (eds.) Differential-Algebraic Equations: A Projector Based Analysis. Differential-Algebraic Equations Forum. Springer, Berlin, Heidelberg, New York, Dordrecht, London (2013)
14. Lamour, R., März, R., Weinmüller, E.: In: Ilchmann, A., Reis, T. (eds.) Surveys in Differential-Algebraic Equations III, chap. Boundary-Value Problems for Differential-Algebraic Equations: A Survey, pp. 177–309. Differential-Algebraic Equations Forum. Springer, Heidelberg (2015)
15. März, R.: In: Ilchmann, A., Reis, T. (eds.) Surveys in Differential-Algebraic Equations II, chap. Differential-Algebraic Equations from a Functional-Analytic Viewpoint: A Survey, pp. 163–285. Differential-Algebraic Equations Forum. Springer, Heidelberg (2015)

16. Pryce, J.D.: Solving high-index DAEs by Taylor series. Numer. Algorithms **19**(1–4), 195–211 (1998)
17. Schwarz, D.E., Lamour, R.: A new approach for computing consistent initial values and Taylor coefficients for DAEs using projector-based constrained optimization. Numer. Algorithms **78**(2), 355–377 (2018)
18. Suetin, P.K.: Classical Orthogonal Polynomials (in Russian), 2nd edn. Nauka, Moskva (1979)