Symmetric Collocation for Unstructured Nonlinear Differential-Algebraic Equations of Arbitrary Index

Peter Kunkel[*] Volker Mehrmann[†] Ronald Stöver[‡]

November 15, 2003

Abstract

We examine a class of symmetric collocation schemes for the solution of nonlinear boundary value problems for unstructured nonlinear systems of differential-algebraic equations with arbitrary index. We show that these schemes converge with the same orders as one would expect for ordinary differential equations. In particular, we show superconvergence for a special choice of the collocation points. We demonstrate the efficiency of the new approach with some numerical examples.

1 Introduction

In this paper we discuss the numerical solution of nonlinear boundary value problems (BVPs) for systems of differential-algebraic equations of arbitrary index. There are many possibilities to design numerical methods for the solution of BVPs. We concentrate here on symmetric collocation methods. For shooting methods, see [14] and references therein. Collocation methods are well studied for ordinary differential equations, see [1], and also for special classes of systems of differential-algebraic equations (DAEs), see [2, 3]. A well-known software for differential-algebraic BVPs is `COLDAE` [3], but it is restricted to semi-explicit problems of index at most two.

[*] Mathematisches Institut, Universität Leipzig, Augustusplatz 10-11, D-04109 Leipzig, Fed. Rep. Germany. Supported by DFG research grant Ku964/4.

[†] Institut für Mathematik, MA 4-5, Technische Universität Berlin, Straße des 17. Juni 136, D-10623 Berlin, Fed. Rep. Germany. Supported by DFG research grant Me790/11.

[‡] Zentrum für Technomathematik, Fachbereich 3, Universität Bremen, Postfach 330 440, D-28334 Bremen, Fed. Rep. Germany.

In this paper we study general nonlinear differential-algebraic BVPs

$$\begin{array}{ll} \text{(a)} & F(t,x,\dot{x}) = 0, \\ \text{(b)} & r(x(\underline{t}), x(\overline{t})) = 0, \end{array} \tag{1.1}$$

where $F : [\underline{t}, \overline{t}] \times \mathbb{D}_x \times \mathbb{D}_{\dot{x}} \to \mathbb{R}^n$, $r : \mathbb{D}_x \times \mathbb{D}_x \to \mathbb{R}^d$ with $\mathbb{D}_x, \mathbb{D}_{\dot{x}} \subseteq \mathbb{R}^n$ open and d the number of differential components of x (we give a precise definition below).

Currently no collocation methods for such general differential-algebraic BVPs are available, but in the linear case (i. e., linear F and linear r), a new class of symmetric collocation methods was recently presented in [15] that exhibit the same convergence behavior as collocation methods for ordinary differential equations, including superconvergence. The main idea in [15] is based on the fact that by index reduction techniques one can distinguish between differential and algebraic equations. It combines two sets of collocation schemes, a Gauß-like scheme for the differential part and a Lobatto-type scheme for the algebraic part.

Here we generalize the results of [15] to the general nonlinear case. The paper is organized as follows. In Section 2, we recall some preliminaries on the theory of DAEs including the index definition that we are using. We then formulate the collocation equations and show solvability for sufficiently fine meshes in Section 3. Section 4 discusses how to realize the collocation method and exhibits the results of a number of numerical experiments. We then give some conclusions in Section 5. In the appendix we analyze a generalized simplified Newton method that presents the basis for our convergence analysis.

2 Preliminaries

For differential-algebraic equations, it is well-known that the solution may depend on derivatives of (1.1a). In particular, differentiation of (1.1a) may lead to hidden algebraic constraints on the possible states of the solution. It is then clear that for a theoretical and numerical treatment of (1.1) we must know the number of differentiations that must be performed to obtain all algebraic constraints that are present in the system. Assuming in the following that F and r are sufficiently smooth, we first introduce the so-called derivative array functions (see [4, 5])

$$F_\ell(t,x,\dot{x},\ldots,x^{(\ell+1)}) = \begin{bmatrix} F(t,x,\dot{x}) \\ \frac{d}{dt}F(t,x,\dot{x}) \\ \vdots \\ (\frac{d}{dt})^\ell F(t,x,\dot{x}) \end{bmatrix}, \tag{2.1}$$

obtained from (1.1a) by successive differentiation with respect to t. Note that F_ℓ is treated here as a function from some subset of $\mathbb{R}^{(\ell+2)n+1}$ into $\mathbb{R}^{(\ell+1)n}$, where the independent variables are denoted by $(t,x,\dot{x},\ldots,x^{(\ell+1)})$. In addition, we need partial derivatives of F_ℓ and other functions. We will denote these by subscripts as in

$$F_{\ell;x} = \frac{\partial}{\partial x}F_\ell, \quad F_{\ell;\dot{x},\ldots,x^{(\ell+1)}} = \begin{bmatrix} \frac{\partial}{\partial \dot{x}}F_\ell & \cdots & \frac{\partial}{\partial x^{(\ell+1)}}F_\ell \end{bmatrix}.$$

The following hypothesis will play a central role in the design and investigation of the collocation method that we present, see [9].

Hypothesis 2.1 *There exist integers μ, a and d such that for all values $(t,x,\dot{x},\ldots,x^{(\mu+1)}) \in \mathbb{L}_\mu$, with*

$$\mathbb{L}_\mu = \{(t,x,\dot{x},\ldots,x^{(\mu+1)}) \in \mathbb{R}^{(\mu+2)n+1} \mid F_\mu(t,x,\dot{x},\ldots,x^{(\mu+1)}) = 0\} \neq \emptyset \tag{2.2}$$

associated with F the following properties hold:

1. *We have*

$$\operatorname{rank} F_{\mu;\dot{x},\ldots,x^{(\mu+1)}}(t,x,\dot{x},\ldots,x^{(\mu+1)}) = (\mu+1)n - a,$$

such that there exists a smooth matrix function $\hat{Z}_2$ on $\mathbb{L}_\mu$ with orthonormal columns and size $((\mu+1)n, a)$ satisfying

$$\hat{Z}_2^T F_{\mu;\dot{x},\ldots,x^{(\mu+1)}} = 0 \quad \text{on } \mathbb{L}_\mu.$$

2. *We have*

$$\operatorname{rank} \hat{Z}_2^T F_{\mu;x}(t,x,\dot{x},\ldots,x^{(\mu+1)}) = a,$$

such that there exists a smooth matrix function $\hat{T}_2$ on $\mathbb{L}_\mu$ with orthonormal columns and size (n,d), where $d = n - a$, satisfying

$$\hat{Z}_2^T F_{\mu;x}\hat{T}_2 = 0 \quad \text{on } \mathbb{L}_\mu.$$

3. *We have*

$$\operatorname{rank} F_{\dot{x}}\hat{T}_2(t, x, \dot{x}, \dots, x^{(\mu+1)}) = d,$$

such that there exists a smooth matrix function $\hat{Z}_1$ on $\mathbb{L}_\mu$ with orthonormal columns and size (n, d) satisfying

$$\operatorname{rank} \hat{Z}_1^T F_{\dot{x}}\hat{T}_2 = d \quad \text{on } \mathbb{L}_\mu.$$

The minimal number μ (if it exists) such that Hypothesis 2.1 is fulfilled is called the *strangeness index* of F. The numbers a and d denote the size of the *algebraic and differential part* of (1.1a).

Remark 2.2 *The above definition of the strangeness index follows a different philosophy than the notion of the differentiation index. The differentiation index as defined in [5] aims for an ODE such that the solutions of the given DAE also solve the ODE. But in general this ODE has more solutions than the original DAE due to the loss of the information on the algebraic constraints. Hypothesis 2.1 is the weakest form of assumptions on a given DAE to guarantee that we can derive (theoretically) another DAE that satisfies Hypothesis 2.1 with $\mu = 0$ and has the same solutions as the original problem. In particular, one can see that the derived DAE has a differentiation index of at most one. Compared with the definition of the differentiation index in [5], one can show that Hypothesis 2.1 is invariant under a larger class of equivalence transformations. Moreover, it can be generalized to over- and underdetermined DAEs, cp. [10]. For a more detailed discussion of Hypothesis 2.1 and the strangeness index see [9].*

As usual in the investigation of computational methods for boundary value problems, we assume that there exists a sufficiently smooth solution of the given problem. In the context of (1.1) we therefore assume that there exists a sufficiently smooth $x^* \in C^1([\underline{t}, \overline{t}], \mathbb{R}^n)$ with

$$\begin{array}{lll} \text{(a)} & F(t, x^*(t), \dot{x}^*(t)) = 0 & \text{for all } t \in [\underline{t}, \overline{t}], \\ \text{(b)} & F_\mu(t, x^*(t), P(t)) = 0 & \text{for all } t \in [\underline{t}, \overline{t}], \\ \text{(c)} & r(x^*(\underline{t}), x^*(\overline{t})) = 0, & \end{array} \tag{2.3}$$

where $P : [\underline{t}, \overline{t}] \to \mathbb{R}^{(\mu+1)n}$ is some smooth function that coincides with $\dot{x}^*$ in the first n components. Sufficient conditions for the existence of such a function $P(t)$ can be found in [10, Theorem 3].

Since $(t, x^*(t), P(t)) \in \mathbb{L}_\mu$ for all $t \in [\underline{t}, \overline{t}]$, Hypothesis 2.1 implies the existence of matrix functions

$$Z_1 : [\underline{t}, \overline{t}] \to \mathbb{R}^{n,d}, \quad Z_2 : [\underline{t}, \overline{t}] \to \mathbb{R}^{(\mu+1)n,a}, \quad T_2 : [\underline{t}, \overline{t}] \to \mathbb{R}^{n,d} \tag{2.4}$$

as restrictions of $\hat{Z}_1$, $\hat{Z}_2$ and $\hat{T}_2$ to the path $(t, x^*(t), P(t))$. These satisfy

$$\begin{array}{lll} \text{(a)} & Z_2(t)^T F_{\mu;\dot{x},\ldots,x^{(\mu+1)}}(t, x^*(t), P(t)) = 0 & \text{for all } t \in [\underline{t}, \overline{t}], \\ \text{(b)} & Z_2(t)^T F_{\mu;x}(t, x^*(t), P(t)) T_2(t) = 0 & \text{for all } t \in [\underline{t}, \overline{t}], \\ \text{(c)} & \operatorname{rank} Z_1(t)^T F_{\dot{x}}(t, x^*(t), \dot{x}^*(t)) T_2(t) = d & \text{for all } t \in [\underline{t}, \overline{t}]. \end{array} \tag{2.5}$$

In addition, there exist smooth functions

$$\begin{array}{ll} Z_2' : [\underline{t}, \overline{t}] \to \mathbb{R}^{(\mu+1)n,(\mu+1)n-a}, & T_1 : [\underline{t}, \overline{t}] \to \mathbb{R}^{(\mu+1)n,a}, \\ T_2' : [\underline{t}, \overline{t}] \to \mathbb{R}^{n,a}, & T_1' : [\underline{t}, \overline{t}] \to \mathbb{R}^{(\mu+1)n,(\mu+1)n-a}, \end{array} \tag{2.6}$$

such that the matrix valued functions $[Z_2', Z_2]$, $[T_1', T_1]$ and $[T_2', T_2]$ are square and pointwise orthogonal and, furthermore,

$$Z_2'(t)^T F_{\mu;\dot{x},\ldots,x^{(\mu+1)}}(t, x^*(t), P(t)) T_1(t) = 0 \quad \text{for all } t \in [\underline{t}, \overline{t}]. \tag{2.7}$$

Using these functions we now consider the nonlinear system of equations $H(t, x, y) = 0$ given by

$$\begin{array}{ll} \text{(a)} & Z_2'(t)^T F_\mu(t, x, y) = 0, \\ \text{(b)} & T_1(t)^T (y - P(t)) = 0. \end{array} \tag{2.8}$$

We then have that $H(t, x^*(t), P(t)) = 0$ and

$$\operatorname{rank} H_y(t, x^*(t), P(t)) = \operatorname{rank} \left[\begin{array}{c} Z_2'(t)^T F_{\mu;\dot{x},\ldots,x^{(\mu+1)}}(t, x^*(t), P(t)) \\ T_1(t)^T \end{array} \right].$$

In particular, it follows from (2.7) that $H_y(t, x^*(t), P(t))$ is nonsingular. Thus, (2.8) locally defines a function K according to

$$y = K(t, x).$$

Introducing the functions

$$\begin{array}{ll} \text{(a)} & \hat{F}_1(t, x, \dot{x}) = Z_1(t)^T F(t, x, \dot{x}), \\ \text{(b)} & \hat{F}_2(t, x) = Z_2(t)^T F_\mu(t, x, K(t, x)), \end{array} \tag{2.9}$$

we have that the given solution x^* of (1.1) also solves the DAE

$$\begin{array}{ll} \text{(a)} & \hat{F}_1(t, x, \dot{x}) = 0, \\ \text{(b)} & \hat{F}_2(t, x) = 0. \end{array} \tag{2.10}$$

Conversely, using the definition of K, it follows that every $x \in \mathbb{R}^n$ in a neighborhood of $x^*(t)$ with $\hat{F}_2(t,x) = 0$ not only satisfies the relation $Z_2(t)^T F_\mu(t,x,K(t,x)) = 0$ but also $Z_2'(t)^T F_\mu(t,x,K(t,x)) = 0$. Hence, $F_\mu(t,x,K(t,x)) = 0$ and x satisfies all algebraic constraints at point t imposed by the DAE (1.1a). Equation (2.10b) therefore represents (by the implicit function theorem) all these algebraic constraints.

Remark 2.3 *If one can obtain (2.10) analytically then one should do it this way. Analogously, if one can evaluate $\hat{F}_2$ of (2.10) by a simpler system (see, e. g., [11]) than via $F_\mu = 0$ such as, e. g., for Hessenberg systems, one should do it this way. As formulated in the title, the approach taken here is discussed for systems without assuming any structure. If the problem is known to have a special structure, of course it should be utilized if possible. In the course of (automatic) modeling, however, one may rather be interested in a general purpose numerical procedure to investigate the generated model and to integrate it.*

If we linearize (2.10) along the given solution x^*, then we obtain a linear DAE

$$\begin{bmatrix} E_1(t) \\ 0 \end{bmatrix} \dot{x} = \begin{bmatrix} A_1(t) \\ A_2(t) \end{bmatrix} x, \tag{2.11}$$

where

$$\begin{array}{ll} \text{(a)} & E_1(t) = Z_1(t)^T F_{\dot{x}}(t, x^*(t), \dot{x}^*(t)), \\ \text{(b)} & A_1(t) = -Z_1(t)^T F_x(t, x^*(t), \dot{x}^*(t)), \\ \text{(c)} & A_2(t) = -Z_2(t)^T F_{\mu;x}(t, x^*(t), P(t)). \end{array} \tag{2.12}$$

Here (2.12c) follows from (2.5a). By (2.5c) and the definition of T_2', we have that (omitting arguments)

$$\operatorname{rank} \begin{bmatrix} E_1 \\ A_2 \end{bmatrix} = \operatorname{rank} \begin{bmatrix} * & Z_1^T F_{\dot{x}} T_2 \\ -Z_2^T F_{\mu;x} T_2' & 0 \end{bmatrix} = n. \tag{2.13}$$

It follows that the DAE (2.11) has differentiation index at most one. In particular, it satisfies the assumptions of [15].

In this section we have given a brief introduction to the basic theory of nonlinear systems of differential-algebraic equations as it was developed in [9, 10]. In the next section we generalize the collocation scheme of [15] to the case of nonlinear DAEs. Note that the transformations introduced above are only used in the investigation of the numerical method that we are going to present. But none of these transformations need to be carried out in the actual numerical computations.

3 Collocation discretization

In the previous section we have derived in (2.10) a new representation of the differential-algebraic system (1.1a). This representation has the advantage that differential and algebraic parts are well separated. This allows, as in the linear case of [15], to treat the differential and the algebraic part in a different way.

For the development and analysis of collocation methods, it is convenient to write the given boundary value problem as an operator equation. For the choice of the correct spaces, we must not only observe that (2.10a) and (2.10b) have different smoothness properties but also that the collocation solution is supposed to be piecewise polynomial but globally only to be continuous.

Consider a mesh

$$\begin{aligned} \pi \ :\ \underline{t} = t_0 < t_1 < \cdots < t_{N-1} < t_N = \overline{t}, \quad N \in \mathbb{N}, \\ h_i = t_{i+1} - t_i, \quad h = \max_{i=0,\ldots,N-1} h_i, \quad h \leq M \min_{i=0,\ldots,N-1} h_i, \end{aligned} \tag{3.1}$$

where $M > 0$ is some fixed constant when we consider $h \to 0$. We define the spaces

$$\begin{aligned} &\text{(a)} \quad \mathbb{X} = C^1_\pi([\underline{t},\overline{t}],\mathbb{R}^n) \cap C^0([\underline{t},\overline{t}],\mathbb{R}^n), \\ &\text{(b)} \quad \mathbb{Y} = C^0_\pi([\underline{t},\overline{t}],\mathbb{R}^d) \times C^1_\pi([\underline{t},\overline{t}],\mathbb{R}^a) \cap C^0([\underline{t},\overline{t}],\mathbb{R}^a) \times \mathbb{R}^d, \end{aligned} \tag{3.2}$$

where the subscript π denotes that we have the stated smoothness only piecewise with respect to the mesh with one-sided limits. This leads to an ambiguity of the corresponding function values at the mesh points, which, however, is not crucial in the following analysis.

If we equip the spaces in (3.2) with the norms

$$\begin{aligned} &\text{(a)} \quad \|x\|_{\mathbb{X}} = \max_{t\in[\underline{t},\overline{t}]} \|x(t)\|_\infty + \max_{i=0,\ldots,N-1} \{ \max_{t\in[t_i,t_{i+1}]} \|\dot{x}(t)\|_\infty \}, \\ &\text{(b)} \quad \|(f_1, f_2, v)\|_{\mathbb{Y}} = \max_{i=0,\ldots,N-1} \{ \max_{t\in[t_i,t_{i+1}]} \|f_1(t)\|_\infty \} + \\ &\qquad + \max_{t\in[\underline{t},\overline{t}]} \|f_2(t)\|_\infty + \max_{i=0,\ldots,N-1} \{ \max_{t\in[t_i,t_{i+1}]} \|\dot{f}_2(t)\|_\infty \} + \|v\|_\infty, \end{aligned} \tag{3.3}$$

where $\dot{x}(t)$ with $t \in [t_i, t_{i+1}]$ and similar quantities denote one-sided limits for $t = t_i, t_{i+1}$ taken within $[t_i, t_{i+1}]$, then the spaces $\mathbb{X}$ and $\mathbb{Y}$ become Banach spaces.

The BVP (1.1) takes the form of the operator equation

$$L(x) = 0, \tag{3.4}$$

with

$$\begin{array}{ll} \text{(a)} & L : \mathbb{X} \to \mathbb{Y}, \\ \text{(b)} & x \mapsto \begin{pmatrix} \hat{F}_1(t, x(t), \dot{x}(t)) \\ \hat{F}_2(t, x(t)) \\ r(x(\underline{t}), x(\overline{t})) \end{pmatrix}. \end{array} \tag{3.5}$$

Since x^* solves (1.1) according to (2.3), we have $L(x^*) = 0$.

For the construction of a Newton-like method we will later need the Fréchet derivative $DL[u]$ of L at $u \in \mathbb{X}$, which is given by

$$\begin{array}{ll} \text{(a)} & DL[u] : \mathbb{X} \to \mathbb{Y}, \\ \text{(b)} & x \mapsto \begin{pmatrix} \hat{F}_{1;x}(t, u(t), \dot{u}(t))x(t) + \hat{F}_{1;\dot{x}}(t, u(t), \dot{u}(t))\dot{x}(t) \\ \hat{F}_{2;x}(t, u(t))x(t) \\ r_{x_a}(u(\underline{t}), u(\overline{t}))x(\underline{t}) + r_{x_b}(u(\underline{t}), u(\overline{t}))x(\overline{t}) \end{pmatrix}. \end{array} \tag{3.6}$$

For linear DAEs, it has been suggested in [15] to use two different types of collocation schemes for the differential and algebraic parts of the DAE. In a similar fashion we introduce a Gauß-type scheme for the differential equations and a Lobatto-type scheme for the algebraic equations. These schemes are given by nodes

$$\begin{array}{ll} \text{(a)} & 0 < \varrho_1 < \cdots < \varrho_k < 1, \\ \text{(b)} & 0 = \sigma_0 < \cdots < \sigma_k = 1, \quad k \in \mathbb{N}, \end{array} \tag{3.7}$$

respectively, and define the collocation points

$$\begin{array}{ll} \text{(a)} & t_{ij} = t_i + h_i \varrho_j, \quad j = 1, \ldots, k, \\ \text{(b)} & s_{ij} = t_i + h_i \sigma_j, \quad j = 0, \ldots, k. \end{array} \tag{3.8}$$

Let $\mathbb{P}_{k+1,\pi}$ denote the space of piecewise polynomials of maximal degree k (order $k + 1$) and introduce the finite dimensional spaces

$$\begin{array}{ll} \text{(a)} & \mathbb{X}_\pi = \mathbb{P}_{k+1,\pi} \cap C^0([\underline{t}, \overline{t}], \mathbb{R}^n), \\ \text{(b)} & \mathbb{Y}_\pi = \mathbb{R}^{kNd} \times \mathbb{R}^{(kN+1)a} \times \mathbb{R}^d. \end{array} \tag{3.9}$$

Observe that $\dim \mathbb{X}_\pi = (k + 1)Nn - (N - 1)n = (kN + 1)n = \dim \mathbb{Y}_\pi$.

Then we apply the collocation discretization given by

$$L_\pi(x_\pi) = 0 \tag{3.10}$$

with

$$\begin{array}{ll} \text{(a)} & L_\pi : \mathbb{X} \to \mathbb{Y}_\pi, \\ \text{(b)} & x \mapsto \begin{pmatrix} \hat{F}_1(t_{ij}, x(t_{ij}), \dot{x}(t_{ij})) \\ \hat{F}_2(s_{ij}, x(s_{ij})) \\ r(x(\underline{t}), x(\overline{t})) \end{pmatrix} \end{array} \tag{3.11}$$

and we seek a solution $x_\pi \in \mathbb{X}_\pi$. For ease of notation in (3.11b) we have omitted that the indices i and j must run over the values $i = 0, \dots, N-1$, $j = 1, \dots, k$ in the first component and $i = 0, \dots, N-1$, $j = 1, \dots, k$ together with $i = 0$, $j = 0$ in the second component. Note that in the second component the indices $i = 1, \dots, N-1$, $j = 0$ must be omitted since the space $\mathbb{X}_\pi$ includes continuity of the solution. We will use this kind of abbreviation in the remainder of the paper. We will also need the Fréchet derivative $DL_\pi[u]$ of the discretized operator L_π at $u \in \mathbb{X}$, which is given by

$$\begin{array}{ll} \text{(a)} & DL_\pi[u] : \mathbb{X} \to \mathbb{Y}_\pi, \\ \text{(b)} & x \mapsto \begin{pmatrix} \hat{F}_{1;x}(t_{ij}, u(t_{ij}), \dot{u}(t_{ij}))x(t_{ij}) + \hat{F}_{1;\dot{x}}(t_{ij}, u(t_{ij}), \dot{u}(t_{ij}))\dot{x}(t_{ij}) \\ \hat{F}_{2;x}(s_{ij}, u(s_{ij}))x(s_{ij}) \\ r_{x_a}(u(\underline{t}), u(\overline{t}))x(\underline{t}) + r_{x_b}(u(\underline{t}), u(\overline{t}))x(\overline{t}) \end{pmatrix}. \end{array} \tag{3.12}$$

Note that we have defined L_π on the larger space $\mathbb{X}$ and not only on $\mathbb{X}_\pi \subseteq \mathbb{X}$. Because of this inclusion, we use the norm of $\mathbb{X}$ also for $\mathbb{X}_\pi$. For $\mathbb{Y}_\pi$, we take the ℓ_∞-norm. Finally, we need the restriction operator

$$\begin{array}{ll} \text{(a)} & R_\pi : \mathbb{Y} \to \mathbb{Y}_\pi, \\ \text{(b)} & \begin{pmatrix} f_1 \\ f_2 \\ v \end{pmatrix} \mapsto \begin{pmatrix} f_1(t_{ij}) \\ f_2(s_{ij}) \\ v \end{pmatrix}. \end{array} \tag{3.13}$$

Observe that $L_\pi = R_\pi L$ and $DL_\pi[u] = R_\pi DL[u]$.

The aim of the following discussion is to show that if x^* satisfies some regularity condition that guarantees that x^* is locally unique, then equation (3.10) is solvable in $\mathbb{X}_\pi$ at least for sufficiently small h. We will also show that we obtain the same orders of convergence with $h \to 0$ as in the linear case [15].

To do so, we follow the lines of [1, pp. 222-226] where a corresponding result is shown in the case of ordinary differential equations. In particular, we consider the iterative process

$$x_\pi^{m+1} = x_\pi^m - DL_\pi[x^*]^{-1} L_\pi(x_\pi^m) \tag{3.14}$$

and prove that under suitable assumptions it generates a sequence $\{x_\pi^m\}$ in $\mathbb{X}_\pi$ that converges to a solution of (3.10). Note that the iteration (3.14) is only a tool for the theoretical analysis. It cannot be used as a numerical method, since the value of the Fréchet derivative at the exact solution x^* is not available.

A typical convergence result for an iteration of the form (3.14) is given by Theorem A.1 of the appendix. In the present context, however, we are interested in properties of (3.14) for $h \to 0$. Thus, we must consider families of iterations (3.14) with the maximum mesh sizes tending to zero. For these we must show that certain constants are independent of h. Unfortunately, in the standard formulation of Theorem A.1 and its proof this does not hold for the constants β and γ. The main task of the following considerations is to replace the standard definition of β and γ by more appropriate quantities and to show that then the crucial estimates in the proof of Theorem A.1 still hold. Since the modified quantities will play the same roles as the original constants β and γ, we will keep the same notation.

We start our analysis by investigating $DL_\pi[x^*]$. For this we introduce the linearization (cp. (2.12))

$$\begin{array}{ll} E_1(t) = \hat{F}_{1;\dot{x}}(t, x^*(t), \dot{x}^*(t)), & C = r_{x_a}(x^*(\underline{t}), x^*(\overline{t})), \\ A_1(t) = -\hat{F}_{1;x}(t, x^*(t), \dot{x}^*(t)), & D = r_{x_b}(x^*(\underline{t}), x^*(\overline{t})), \\ A_2(t) = -\hat{F}_{2;x}(t, x^*(t)), & \end{array} \tag{3.15}$$

and hence we have that

$$\begin{array}{ll} \text{(a)} & DL[x^*] : \mathbb{X} \to \mathbb{Y}, \\ \text{(b)} & x \mapsto \left(\begin{array}{c} E_1(t)\dot{x}(t) - A_1(t)x(t) \\ -A_2(t)x(t) \\ Cx(\underline{t}) + Dx(\overline{t}) \end{array} \right) \end{array} \tag{3.16}$$

and $DL_\pi[x^*] = R_\pi DL[x^*]$. Since we have assumed sufficient smoothness of F, r and x^*, all assumptions of [15] are satisfied for linear boundary value problems that involve $DL[x^*]$. To make use of the results developed there, we need the following regularity property of x^*.

Definition 3.1 *A solution $x^* \in \mathbb{X}$ of (1.1) is called* regular *if the boundary value problem*

$$DL[x^*]x = 0 \tag{3.17}$$

only possesses the trivial solution.

Throughout the remainder of the paper, we assume x^* to be regular in the sense of Definition 3.1. For linear systems in [15, Prop. 2.2] a characterization of regularity in terms of the data E_1, A_1, A_2, C and D has been given.

The first step of our analysis of the iteration (3.14) is to construct a suitable initial function $x_\pi^0 \in \mathbb{X}_\pi$ to start the iteration. Note that due to [15] the operator $DL_\pi[x^*]$ is invertible for sufficiently small h when we restrict it to $\mathbb{X}_\pi$. Thus, there is a well-defined inverse

$$DL_\pi[x^*]^{-1} : \mathbb{Y}_\pi \to \mathbb{X}_\pi \subseteq \mathbb{X} \tag{3.18}$$

that satisfies

$$DL_\pi[x^*]^{-1} DL_\pi[x^*] = \mathrm{id}_{\mathbb{X}_\pi} . \tag{3.19}$$

Lemma 3.2 *For sufficiently small h, the linear collocation problem*

$$DL_\pi[x^*]x_\pi = DL_\pi[x^*]x^* \tag{3.20}$$

for x_π has a unique solution $x_\pi^0 \in \mathbb{X}_\pi$ with

$$\|x_\pi^0 - x^*\|_{\mathbb{X}} \leq Ch^k, \tag{3.21}$$

where C is independent of h.

Proof. The operator equation (3.20) is the collocation discretization of the linear boundary value problem

$$DL[x^*]x = DL[x^*]x^*$$

which possesses the solution x^*. Applying the results of [15] shows that (3.20) has a unique solution $x_\pi^0 \in \mathbb{X}_\pi$ with

$$\max_{t \in [\underline{t}, \overline{t}]} \|x^*(t) - x_\pi^0(t)\|_\infty \leq Ch^k$$

for sufficiently small h with C independent of h.

Due to the definition of $\|\cdot\|_{\mathbb{X}}$ we also need an estimate for the derivative $\dot{x}^*(t) - \dot{x}_\pi^0(t)$ on $[t_i, t_{i+1}]$ in order to prove (3.21). Observing that

$$\begin{aligned} E_1(t_{ij})(\dot{x}^*(t_{ij}) - \dot{x}_\pi^0(t_{ij})) &= A_1(t_{ij})(x^*(t_{ij}) - x_\pi^0(t_{ij})), \quad & j = 1, \ldots, k, \\ 0 &= A_2(s_{ij})(x^*(s_{ij}) - x_\pi^0(s_{ij})), \quad & j = 0, \ldots, k, \end{aligned}$$

and applying Lagrange interpolation of $x^* - x_\pi^0$ at the points $s_{ij} = t_i + h_i\sigma_j$ gives

$$x^*(t) - x_\pi^0(t) = \sum_{l=0}^{k} \Big(x^*(s_{il}) - x_\pi^0(s_{il})\Big) L_l(\tfrac{t-t_i}{h_i}) + O(h^{k+1}),$$

where

$$L_l(\tau) = \prod_{\substack{m=0 \\ m \neq l}}^{k} \frac{\tau - \sigma_m}{\sigma_l - \sigma_m}.$$

Since $(\frac{d}{dt})^{k+1}x_\pi^0 = 0$ on $[t_i,t_{i+1}]$ it follows that the constant involved in $O(h^{k+1})$ does not depend on h. Thus, with $L_l'(\tau) = \frac{d}{d\tau}\,L_l(\tau)$, we have

$$\begin{aligned}
&A_2(t_{ij})(\dot{x}^*(t_{ij}) - \dot{x}_\pi^0(t_{ij})) = \\
&\quad = A_2(t_{ij})\textstyle\sum_{l=0}^k \Big(x^*(s_{il}) - x_\pi^0(s_{il})\Big) L_l'(\frac{t_{ij}-t_i}{h_i})\frac{1}{h_i} + O(h^k) = \\
&\quad = \textstyle\sum_{l=0}^k \Big(A_2(s_{il}) + O(h)\Big)\Big(x^*(s_{il}) - x_\pi^0(s_{il})\Big) L_l'(\rho_j)\frac{1}{h_i} + O(h^k) = \\
&\quad = \textstyle\sum_{l=0}^k \Big(0 + O(h)\cdot O(h^k)\Big)\cdot O(1)\cdot O(h^{-1}) + O(h^k) = \\
&\quad = O(h^k),
\end{aligned}$$

where all constants do not depend on h. Combining the estimates for x and $\dot{x}$, we have

$$\left[\begin{array}{c} E_1(t_{ij}) \\ A_2(t_{ij}) \end{array}\right](\dot{x}^*(t_{ij}) - \dot{x}_\pi^0(t_{ij})) = O(h^k).$$

Property (2.13) says that the leading matrix is invertible and has a bounded inverse. Hence,

$$\|\dot{x}^*(t_{ij}) - \dot{x}_\pi^0(t_{ij})\|_\infty \le Ch^k,$$

possibly increasing the constant C. Lagrange interpolation of $\dot{x}^* - \dot{x}_\pi^0$ at the points t_{ij} gives

$$\dot{x}^*(t) - \dot{x}_\pi^0(t) = \sum_{l=1}^k \Big(\dot{x}^*(t_{il}) - \dot{x}_\pi^0(t_{il})\Big)\tilde{L}_l(\tfrac{t-t_i}{h_i}) + O(h^k),$$

with

$$\tilde{L}_l(\tau) = \prod_{\substack{m=1\\ m\neq l}}^k \frac{\tau - \varrho_m}{\varrho_l - \varrho_m}.$$

Again, since $(\frac{d}{dt})^k\dot{x}_\pi^0 = 0$ on $[t_i,t_{i+1}]$, it follows that the constant involved in $O(h^k)$ does not depend on h. Thus, we finally have

$$\max_{t\in[t_i,t_{i+1}]}\|\dot{x}^*(t) - \dot{x}_\pi^0(t)\|_\infty \le Ch^k,$$

possibly increasing again the constant C. □

The next lemma gives estimates for higher derivatives of $x^*(t) - x_\pi^0(t)$.

Lemma 3.3 *Suppose x^* to be sufficiently smooth and that*

$$\max_{t\in[t_i,t_{i+1}]} \|\dot{x}^*(t) - \dot{x}_\pi(t)\|_\infty \le Ch^k \tag{3.22}$$

for $x_\pi \in \mathbb{X}_\pi$ with C independent of h as $h \to 0$. Then,

$$\max_{t\in[t_i,t_{i+1}]} \|(\tfrac{d}{dt})^l (x^*(t) - x_\pi(t))\|_\infty \le Ch^{k-l+1}, \quad l = 1,\ldots,k \tag{3.23}$$

with possibly increased constant C. In particular, x_π has bounded derivatives of arbitrary order on $[t_i, t_{i+1}]$ as $h \to 0$.

Proof. This result only depends on the properties of the space $\mathbb{X}_\pi$ which is in the DAE case the same as for ordinary differential equations. Hence the result follows as in [1, Th. 5.75]. □

The second step deals with an appropriate modification of the constant β of Theorem A.1. In the present context, this contains a *stability property* of the collocation discretization.

Lemma 3.4 *The linear collocation discretization given by $DL_\pi[x^*]$ is* stable *in the sense that*

$$\|DL_\pi[x^*]^{-1}R_\pi\|_{\mathbb{X}\leftarrow\mathbb{Y}} \le \beta \tag{3.24}$$

with β independent of h.

Proof. Let $g = (f_1, f_2, v) \in \mathbb{Y}$ and consider the boundary value problem

$$DL[x^*]x = g$$

and its collocation discretization

$$DL_\pi[x^*]x_\pi = R_\pi g.$$

Although the inhomogeneity g does not have the smoothness properties required in the proof of the corresponding stability result of [15], the same proof as given there shows that the discrete problem is uniquely solvable in $\mathbb{X}_\pi$ and that

$$\max_{t\in[\underline{t},\overline{t}]} \|x_\pi(t)\|_\infty \le \beta\|g\|_{\mathbb{Y}}$$

for the solution x_π with β independent of h. To get the estimate for the derivative $\dot{x}_\pi(t)$ on $[t_i, t_{i+1}]$, we observe that

$$\begin{aligned} E_1(t_{ij})\dot{x}_\pi(t_{ij}) &= A_1(t_{ij})x_\pi(t_{ij}) + f_1(t_{ij}), \quad && j = 1,\ldots,k, \\ 0 &= A_2(s_{ij})x_\pi(s_{ij}) + f_2(s_{ij}), \quad && j = 0,\ldots,k. \end{aligned}$$

Since $x_\pi \in \mathbb{P}_{k+1,\pi}$, we can write x_π as

$$x_\pi(t) = \sum_{l=0}^{k} x_\pi(s_{il}) L_l(\tfrac{t-t_i}{h_i}),$$

cp. the proof of Lemma 3.2. Hence,

$$\begin{aligned}
A_2(t_{ij})\dot{x}_\pi(t_{ij}) &= \textstyle\sum_{l=0}^{k} A_2(t_{ij}) x_\pi(s_{il}) L_l'(\tfrac{t_{ij}-t_i}{h_i})\tfrac{1}{h_i} = \\
&= \textstyle\sum_{l=0}^{k} \Big(A_2(s_{il}) + O(h)\Big) x_\pi(s_{il}) L_l'(\rho_j)\tfrac{1}{h_i} = \\
&= \textstyle\sum_{l=0}^{k} O(h)\cdot x_\pi(s_{il}) L_j'(\rho_j)\tfrac{1}{h_i} - \sum_{l=0}^{k} f_2(s_{il}) L_l'(\rho_j)\tfrac{1}{h_i} = \\
&= \textstyle\sum_{l=0}^{k} O(h)\cdot x_\pi(s_{il}) L_l'(\rho_j)\tfrac{1}{h_i} - \sum_{l=0}^{k} \Big(f_2(s_{il}) - f_2(t_i)\Big) L_l'(\rho_j)\tfrac{1}{h_i},
\end{aligned}$$

where the latter identity follows, since for all t

$$\sum_{l=0}^{k} L_l'(\tfrac{t-t_i}{h_i}) = 0.$$

Because of $f_2 \in C^1([\underline{t},\overline{t}],\mathbb{R}^a)$, there are points $\theta_{ij} \in [t_i, t_{i+1}]$ which satisfy $f_2(s_{il}) - f_2(t_i) = h_i \sigma_l \dot{f}_2(\theta_{ij})$. Possibly increasing the constant β, we therefore have that

$$\|A_2(t_{ij})\dot{x}_\pi(t_{ij})\|_\infty \le \beta_1 \|g\|_{\mathbb{Y}} + \beta_2 \max_{t\in[t_i,t_{i+1}]} \|\dot{f}_2(t)\|_\infty \le \beta\|g\|_{\mathbb{Y}}.$$

Together with

$$\|E_1(t_{ij})\dot{x}_\pi(t_{ij})\|_\infty \le \beta\|g\|_{\mathbb{Y}},$$

it then follows as in Lemma 3.2 that

$$\|\dot{x}_\pi(t_{ij})\|_\infty \le \beta\|g\|_{\mathbb{Y}}.$$

Using again Lagrange interpolation of $\dot{x}_\pi$ at the points t_{ij}, with $\tilde{L}_j$ as in the proof of Lemma 3.2, we have

$$\dot{x}_\pi(t) = \sum_{l=1}^{k} \dot{x}_\pi(t_{il}) \tilde{L}_l(\tfrac{t-t_i}{h_i}).$$

Thus, we have

$$\max_{t\in[t_i,t_{i+1}]} \|\dot{x}_\pi(t)\|_\infty \le \beta\|g\|_{\mathbb{Y}}$$

and finally

$$\|x_\pi\|_{\mathbb{X}} \leq \beta \|g\|_{\mathbb{Y}}$$

with possibly increased constant β. Observing that the choice of β does only depend on the problem data involved in $DL[x^*]$, but not on the selected inhomogeneity $g \in \mathbb{Y}$ nor on h, the claim follows. □

In the third step, we give a suitable replacement for the Lipschitz constant γ of Theorem A.1. In the present context, this means that we must consider the dependence of the operator $DL[u]$ on u. Recall that we assume all data including the functions $\hat{F}_1$ and $\hat{F}_2$ as defined in (2.9) to be sufficiently smooth in a neighborhood of the solution x^*.

Lemma 3.5 *Let L from (3.5) be defined on a convex and compact neighborhood $\mathbb{D} \subseteq \mathbb{X}$ of x^*. Then there exists a constant γ independent of h such that*

$$\|L(x) - L(y) - DL[z](x-y)\|_{\mathbb{Y}} \leq \tfrac{1}{2}\gamma \|x-y\|_{\mathbb{X}} \Big(\|x-z\|_{\mathbb{X}} + \|y-z\|_{\mathbb{X}} \Big) \tag{3.25}$$

for all $x, y, z \in \mathbb{D}$.

Proof. See Appendix B. □

Remark 3.6 *With the same technique as in the proof of Lemma 3.5, we can show that $DL[u]$ and therefore also $DL_\pi[u] = R_\pi DL[u]$ are Lipschitz continuous with respect to u, i. e.,*

$$\|DL[x] - DL[y]\|_{\mathbb{Y}\leftarrow\mathbb{X}} \leq \gamma \|x-y\|_{\mathbb{X}} \quad \text{for all } x, y \in \mathbb{D} \tag{3.26}$$

with γ independent of h. We therefore omit a proof.

An immediate consequence of Lemmata 3.4 and 3.5 is that, although they define modifications of the constants β and γ of Theorem A.1, the crucial estimates in the proof of Theorem A.1 still hold. In particular, we have

$$\begin{aligned}
&\|x_\pi^{m+1} - x_\pi^m\|_{\mathbb{X}} = \\
&= \|DL_\pi[x^*]^{-1}[L_\pi(x_\pi^m) - L_\pi(x_\pi^{m-1}) - DL_\pi[x^*](x_\pi^m - x_\pi^{m-1})]\|_{\mathbb{X}} = \\
&= \|DL_\pi[x^*]^{-1}R_\pi[L(x_\pi^m) - L(x_\pi^{m-1}) - DL[x^*](x_\pi^m - x_\pi^{m-1})]\|_{\mathbb{X}} \leq \\
&\leq \tfrac{1}{2}\beta\gamma \|x_\pi^m - x_\pi^{m-1}\|_{\mathbb{X}} \Big(\|x_\pi^m - x^*\|_{\mathbb{X}} + \|x_\pi^{m-1} - x^*\|_{\mathbb{X}} \Big)
\end{aligned}$$

as long as $x_\pi^0, \dots, x_\pi^m, x^* \in \mathbb{D}$, and similarly (for an $x_\pi^{**} \in \mathbb{X}_\pi \cap \mathbb{D}$ with $L_\pi(x_\pi^{**}) = 0$)

$$\|x_\pi^{m+1} - x_\pi^{**}\|_\mathbb{X} \leq \tfrac{1}{2}\beta\gamma\|x_\pi^m - x_\pi^{**}\|_\mathbb{X}\Big(\|x_\pi^m - x^*\|_\mathbb{X} + \|x_\pi^{**} - x^*\|_\mathbb{X}\Big),$$

cp. also Remark A.3.

Thus, to get the claims of Theorem A.1 it only remains to discuss the assumptions of Theorem A.1 concerning the quantities $\alpha = \|x_\pi^1 - x_\pi^0\|_\mathbb{X}$ and $\hat{t} = -\|x_\pi^0 - x^*\|_\mathbb{X}$. Because of (3.21), we can choose h so small that

$$\|x_\pi^0 - x^*\|_\mathbb{X} \leq \frac{1}{2\beta\gamma}, \tag{3.27}$$

cp. Corollary A.2. Moreover, because of

$$\begin{aligned}\|x_\pi^1 - x_\pi^0\|_\mathbb{X} &= \|DL_\pi[x^*]^{-1}L_\pi(x_\pi^0)\|_\mathbb{X} = \\ &= \|DL_\pi[x^*]^{-1}R_\pi L(x^* + (x_\pi^0 - x^*))\|_\mathbb{X} \leq \\ &\leq \beta\|L(x^*) + DL[x^*](x_\pi^0 - x^*) + O(\|x_\pi^0 - x^*\|_\mathbb{X}^2)\|_\mathbb{X} \leq \\ &\leq \beta\|DL[x^*]\|_{\mathbb{Y}\leftarrow\mathbb{X}}\,\|x_\pi^0 - x^*\|_\mathbb{X} + O(\|x_\pi^0 - x^*\|_\mathbb{X}^2),\end{aligned}$$

we have

$$\|x_\pi^1 - x_\pi^0\|_\mathbb{X} \leq \tilde{C}h^k \tag{3.28}$$

with $\tilde{C}$ independent of h, and we can choose h so small that

$$\alpha = \|x_\pi^1 - x_\pi^0\|_\mathbb{X} \leq \frac{1}{9\beta\gamma} \tag{3.29}$$

and that $\overline{S}(x_\pi^0, 4\alpha) \subseteq \mathbb{D}$. It follows then inductively as in the proof of Theorem A.1 that (3.14) generates a sequence $\{x_\pi^m\}$ with

$$x_\pi^m \in \overline{S}(x_\pi^0, 4\alpha) \cap \mathbb{X}_\pi. \tag{3.30}$$

Since $\mathbb{X}_\pi \subseteq \mathbb{X}$ is closed, the sequence converges to an $x_\pi^* \in \overline{S}(x_\pi^0, 4\alpha) \cap \mathbb{X}_\pi$ with $L_\pi(x_\pi^*) = 0$. Local uniqueness follows since (3.27) and (3.29) imply that $\rho_- < \rho_+$. Finally observing that now

$$\|x^* - x_\pi^*\|_\mathbb{X} \leq \|x^* - x_\pi^0\|_\mathbb{X} + \|x_\pi^0 - x_\pi^*\|_\mathbb{X} \leq Ch^k + 4\tilde{C}h^k$$

utilizing Corollary A.2, we have arrived at the following result.

Theorem 3.7 *Let $x^* \in \mathbb{X}$ be a regular solution of $L(x) = 0$. Then, for sufficiently small h, there exists a locally unique solution $x^*_\pi \in \mathbb{X}_\pi$ of $L_\pi(x_\pi) = 0$. In particular, the estimate*

$$\|x^* - x^*_\pi\|_{\mathbb{X}} \leq Ch^k \tag{3.31}$$

holds, with C independent of h.

In the remainder of this section, we show superconvergence of the collocation method when we use special schemes in (3.7). From

$$L(x^*_\pi) = L(x^* + (x^*_\pi - x^*)) = L(x^*) + DL[x^*](x^*_\pi - x^*) + O(\|x^*_\pi - x^*\|^2_{\mathbb{X}})$$

and

$$0 = L_\pi(x^*_\pi) = R_\pi L(x^*_\pi) = DL_\pi[x^*](x^*_\pi - x^*) + R_\pi O(\|x^*_\pi - x^*\|^2_{\mathbb{X}}),$$

it follows with (3.31) and $DL_\pi[x^*]x^0_\pi = DL_\pi[x^*]x^*$ that

$$DL_\pi[x^*]x^*_\pi = DL_\pi[x^*]x^* + R_\pi O(h^{2k}) = DL_\pi[x^*]x^0_\pi + R_\pi O(h^{2k}),$$

where again the involved constants in the remainders are independent of h. Application of (3.19) and (3.24) yields

$$x^*_\pi = x^0_\pi + O(h^{2k}). \tag{3.32}$$

In particular, we have

$$\begin{aligned} x^*_\pi(t) - x^*(t) &= (x^*_\pi(t) - x^0_\pi(t)) + (x^0_\pi(t) - x^*(t)) = \\ &= x^0_\pi(t) - x^*(t) + O(h^{2k}) \end{aligned} \tag{3.33}$$

for all $t \in [\underline{t}, \overline{t}]$.

Theorem 3.8 *Let the assumptions of Theorem 3.7 hold and let $\varrho_1, \ldots, \varrho_k$ and $\sigma_0, \ldots, \sigma_k$ of (3.7) be Gauß and Lobatto nodes, respectively. Then*

$$\begin{aligned} &\text{(a)} \quad \max_{i=0,\ldots,N} \|x^*(t_i) - x^*_\pi(t_i)\|_\infty = O(h^{2k}), \\ &\text{(b)} \quad \max_{j=0,\ldots,k} \|x^*(s_{ij}) - x^*_\pi(s_{ij})\|_\infty = O(h^{k+2}) \quad \text{for } k \geq 2, \\ &\text{(c)} \quad \max_{t \in [\underline{t}, \overline{t}]} \|x^*(t) - x^*_\pi(t)\|_\infty = O(h^{k+1}), \end{aligned} \tag{3.34}$$

with constants independent of h.

Proof. Applying Theorem 3.3 and Corollary 3.1 of [15] to (3.20), we get the estimates (3.34) for $x^*(t) - x^0_\pi(t)$. Because of (3.33), they carry over to the corresponding estimates for $x^*(t) - x^*_\pi(t)$. □

Remark 3.9 *Observing stability of $DL_\pi[x^*]^{-1}R_\pi$ and Lipschitz continuity of $DL_\pi[u] = R_\pi DL[u]$ with respect to u according to Remark 3.6, the Lipschitz constant being independent of h, we can conclude existence and stability of $DL_\pi[u]^{-1}R_\pi$ for u in a sufficiently small neighborhood of x^* and thus of $DL_\pi[x^*_\pi]^{-1}R_\pi$ for sufficiently small h.*

4 Numerical realization and experiments

In the previous section, we have shown that the collocation system (3.10) has a (regular) solution x^*_π near the exact solution x^*, provided that h is sufficiently small and some (standard) regularity condition holds. But there are a number of problems to deal with (3.10) directly in order to actually compute x^*_π. First, the function Z_1 used in the definition of $\hat{F}_1$ is not known, and second, the function $\hat{F}_2$ is implicitly defined and also includes with Z_2 and K further functions that are not known. Since x^* is unknown, the iterative method (3.14) cannot be applied. Moreover, iterative processes of this kind usually have poor convergence properties.

The function Z_1 must only guarantee that the overall system has differentiation index at most one. Thus, we can use (sufficiently good) approximations $Z_{1,ij}$ to its values at the points t_{ij} without changing the solution and its regularity. Concerning $\hat{F}_2$, we must go back to the original defining equations via F_μ. This leads us to the underdetermined system

$$\begin{aligned} &\text{(a)} \quad Z_{1,ij}^T F(t_{ij}, x_\pi(t_{ij}), \dot{x}_\pi(t_{ij})) = 0, \\ &\text{(b)} \quad F_\mu(s_{ij}, x_\pi(s_{ij}), \dot{x}_{ij}, \ldots, x_{ij}^{(\mu+1)}) = 0, \\ &\text{(c)} \quad r(x_\pi(\underline{t}), x_\pi(\overline{t})) = 0 \end{aligned} \tag{4.1}$$

for the unknowns

$$(x_\pi, \dot{x}_{ij}, \ldots, x_{ij}^{(\mu+1)}) \in \mathbb{P}_{k+1,\pi} \cap C^0([\underline{t}, \overline{t}], \mathbb{R}^n) \times \mathbb{R}^{(Nk+1)(\mu+1)n} \tag{4.2}$$

with $i = 0, \ldots, N-1$, $j = 1, \ldots, k$ and $i = 0$, $j = 0$. For convenience, in the following we use the abbreviations

$$x_{ij} = x_\pi(s_{ij}), \quad y_{ij} = (\dot{x}_{ij}, \ldots, x_{ij}^{(\mu+1)}). \tag{4.3}$$

Suitable matrices $Z_{1,ij}$ can be obtained from (sufficiently good) initial guesses (x_{ij}^0, y_{ij}^0) by perturbing $F_{\mu;\dot{x},\ldots,x^{(\mu+1)}}(s_{ij}, x_{ij}^0, y_{ij}^0)$ to a matrix with rank deficiency a in order to get an approximate evaluation of $\hat{Z}_1$ along the lines of Hypothesis 2.1. Recall that $\operatorname{rank} F_{\mu;\dot{x},\ldots,x^{(\mu+1)}}(s,x,y) = (\mu+1)n - a$ if $(s,x,y) \in \mathbb{L}_\mu$. A second possibility is to project $(s_{ij}, x_{ij}^0, y_{ij}^0)$ onto $\mathbb{L}_\mu$ and to use the corresponding Jacobian there. For sufficiently small h, it is also clear that we may choose $Z_{1,ij}$ independent of j.

The equations (4.1b) guarantee that the discrete solution obeys all algebraic constraints at least at the collocation points s_{ij}. In particular, recall the discussion at the end of Section 2 on the equivalence of (4.1b) with $\hat{F}_2(s_{ij}, x_\pi(s_{ij})) = 0$, cp. also [9].

The iteration process of choice for the numerical solution of (4.1) is a Gauß-Newton-like method of the form

$$z_{m+1} = z_m - \mathcal{A}_m^+ \mathcal{F}(z_m), \quad z_m = (x_{ij}^m, y_{ij}^m), \tag{4.4}$$

when we write (4.1) as $\mathcal{F}(z) = 0$. Here $\mathcal{A}_m^+$ denotes the Moore-Penrose pseudoinverse of $\mathcal{A}_m$. In contrast to the ordinary Gauß-Newton method, we replace the Jacobian $\mathcal{F}_z(z_m)$ by a perturbed matrix $\mathcal{A}_m$ in order to get a more efficient procedure. In particular, we determine $\mathcal{A}_m$ from $\mathcal{F}_z(z_m)$ in such a way that we replace the block entries $F_{\mu;\dot{x},\ldots,x^{(\mu+1)}}(s_{ij}, x_{ij}^m, y_{ij}^m)$ by matrices of rank deficiency a (e. g., by ignoring the a smallest singular values). This decouples the determination of $\Delta y_{ij}^m = y_{ij}^{m+1} - y_{ij}^m$ for each i, j from the other corrections and leaves a linear system, representing the collocation discretization of a linear BVP, for the corrections concerning x_π^m only. Thus, we can use the techniques of [15] with solving first a number of local systems and then a global system of a structure that is known from multiple shooting methods for ordinary differential equations [1]. Having then computed the corrections for x_π^m, it remains the solution of the decoupled underdetermined linear systems for the Δy_{ij}^m. Taking the Moore-Penrose pseudoinverse to select a solution realizes the Moore-Penrose pseudoinverse of the overall system due to the decoupling. Since the applied perturbations tend to zero when z_m converges to a solution, we expect a superlinear convergence rate, see [6]. Compare also with [14] where similar techniques are used in the context of multiple shooting.

The Gauß-Newton-like procedure (4.4) has been implemented in `MATLAB` [17] as a research code. We compute an initial solution profile $z_0 = (x_{ij}^0, y_{ij}^0)$ by solving the initial value problem corresponding to a given initial value (x_{00}, y_{00}) at $t = \underline{t}$ with `GENDA` [12]. Iteration (4.4) is terminated as soon as $\|z_{m+1} - z_m\|_2 \leq \mathrm{tol}\, \|z_m\|_2$ using an appropriate tolerance $\mathrm{tol} = 10^{-8}$. All computations have been performed on a `SUN SPARC Ultra60` workstation with 360 MHz.

In the following, we present several examples which all have certain but different structures in order to demonstrate the general applicability of the new approach. Currently, except for the multiple shooting approach of [14], there are no other general methods to compare with. Compare also Remark 2.3. Clearly, exploiting the specific properties of a problem class can be expected to lead to significant efficiency improvements.

We have not included comparisons with other existing codes for solving differential-algebraic BVPs, since all of them require the problems to have a certain structure whereas we here focus on a general procedure. For a comparison with the code `COLDAE` of [3] in the linear case, see [15, 19]. The observations made there also apply here.

Example 4.1 In order to illustrate the convergence orders of Theorem 3.8, we consider the following semi-explicit problem, see [3], with known solution:

$$\begin{aligned}\dot{x}_1 &= (\varepsilon + x_2 - p_2(t))x_4 + \dot{p}_1(t),\\ \dot{x}_2 &= \dot{p}_2(t),\\ \dot{x}_3 &= x_4,\\ 0 &= (x_1 - p_1(t))(x_4 - \exp(t)).\end{aligned}$$

Choosing the boundary condition

$$x_1(0) = p_1(0) + \varepsilon, \quad x_3(0) = 1, \quad x_2(1) = p_2(1),$$

the exact solution of the BVP is given by

$$x^*(t) = \Big(\varepsilon \exp(t) + p_1(t), p_2(t), \exp(t), \exp(t)\Big).$$

In this case, the differentiation index is one and Hypothesis 2.1 is satisfied with $\mu = 0$, $d = 3$, $a = 1$. As parameters, we chose $p_1(t) = \sin(4\pi t)$, $p_2(t) = \sin(t)$, $\varepsilon = \frac{1}{2}$, and the integration for computing an initial solution profile was started with $x_{00} = (-1, 0, 0, 2)$, $\dot{x}_{00} = 0$. In Table 1, the errors

$$\mathrm{err}_i(N) = \max_{0\le i\le N} \|x(t_i) - x_i\|_2, \quad \mathrm{err}_{ij}(N) = \max_{0\le i\le N}\max_{1\le j\le k} \|x(s_{ij}) - x_{ij}\|_2$$

are given, together with corresponding orders $\log_2(\mathrm{err}_i(N/2)) - \log_2(\mathrm{err}_i(N))$ and $\log_2(\mathrm{err}_{ij}(N/2)) - \log_2(\mathrm{err}_{ij}(N))$. We clearly see that the convergence results (3.34a) and (3.34b) hold for this example. The computing times have been between 0.5 seconds (in the case $k = 4, N = 5$) and 7.4 seconds (in the case $k = 1, N = 200$).

Table 1: Errors and orders according to uniform meshes for Example 4.1

k	N	err_i	order	err_{ij}	order
1	50	0.265D-02		0.265D-02	
	100	0.662D-03	2.0	0.662D-03	2.0
	200	0.166D-03	2.0	0.166D-03	2.0
2	20	0.348D-04		0.977D-04	
	40	0.226D-05	3.9	0.613D-05	4.0
	80	0.141D-06	4.0	0.387D-06	4.0
3	10	0.196D-05		0.578D-04	
	20	0.294D-07	6.1	0.177D-05	5.0
	40	0.478D-09	5.9	0.566D-07	5.0
4	5	0.108D-05		0.148D-03	
	10	0.352D-08	8.3	0.232D-05	6.0
	20	0.132D-10	8.1	0.381D-07	5.9
5	5	0.482D-08		0.961D-05	
	10	0.391D-11	10.3	0.769D-07	7.1

Example 4.2 In [7], the model of a periodically driven electronic amplifier is given. The equations with $n = 5$ for the unknowns $(U_1, \ldots, U_5)$ read

$$\begin{aligned}
&(U_E(t) - U_1)/R_0 + C_1(\dot{U}_2 - \dot{U}_1) = 0,\\
&(U_B - U_2)/R_2 - U_2/R_1 + C_1(\dot{U}_1 - \dot{U}_2) - 0.01 f(U_2 - U_3) = 0,\\
&f(U_2 - U_3) - U_3/R_3 - C_2\dot{U}_3 = 0,\\
&(U_B - U_4)/R_4 + C_3(\dot{U}_5 - \dot{U}_4) - 0.99 f(U_2 - U_3) = 0,\\
&-U_5/R_5 + C_3(\dot{U}_4 - \dot{U}_5) = 0,
\end{aligned}$$

with

$$\begin{aligned}
&U_E(t) = 0.4\sin(200\pi t), \quad U_B = 6,\\
&f(U) = 10^{-6}(\exp(U/0.026) - 1),\\
&R_0 = 1000, \quad R_1 = \cdots = R_5 = 9000,\\
&C_1 = 10^{-6}, \quad C_2 = 2 \cdot 10^{-6}, \quad C_3 = 3 \cdot 10^{-6}.
\end{aligned}$$

This problem of differentiation index one is known to satisfy Hypothesis 2.1 with $\mu = 0$, $d = 3$, and $a = 2$. If we ask for the periodic response of the amplifier, we are led to the boundary conditions

$$U_l(0) = U_l(0.01), \quad l = 2, 3, 5,$$

thus $\underline{t} = 0$ and $\overline{t} = 0.01$. We used the initial value

$$(x_{00}, \dot{x}_{00}) = (0, V_1, V_1, U_B, 0, 0, 0, V_2, 0, 0) \in \mathbb{L}_\mu,$$

where $V_1 = U_B \frac{R_1}{R_1+R_2}$ and $V_2 = -\frac{V_1}{R_3 C_2}$.

For different k (the number of collocation points within a subinterval) and different meshes π, the presented collocation method successfully computed a periodic solution. The convergence behavior for $k = 5$ and a mesh with five uniform subintervals is given in Table 2. This computation took 10.6 seconds, about 8.0 seconds for solving the initial value problem and about 2.3 seconds for the Gauß-Newton-like iteration.

Table 2: Convergence behavior for Example 4.2

m	$\|z_{m+1} - z_m\|_2$	m	$\|z_{m+1} - z_m\|_2$
0	`0.212D+04`	5	`0.832D+01`
1	`0.280D+04`	6	`0.388D+00`
2	`0.153D+04`	7	`0.760D-03`
3	`0.325D+03`	8	`0.282D-08`
4	`0.531D+02`		

Example 4.3 A pendulum in two space dimensions is modeled by

$$\begin{aligned} &\dot p_1 = v_1, \quad \dot v_1 = 2p_1\lambda, \\ &\dot p_2 = v_2, \quad \dot v_2 = 2p_2\lambda - g, \\ &p_1^2 + p_2^2 = 1 \end{aligned}$$

with the gravity constant $g = 9.81$. The unknowns are $(p_1, p_2, v_1, v_2, \lambda)$. In [16] this problem together with the boundary conditions

$$v_2(0) = 0, \quad p_1(0.55) = 0 \tag{4.5}$$

was used to test an implementation of a multiple shooting method for DAEs with differentiation index of at most two. Since the above problem has differentiation index three, it was necessary in [16] to replace the constraint by its differentiated form

$$2p_1\dot p_1 + 2p_2\dot p_2 = 2p_1 v_1 + 2p_2 v_2 = 0$$

and to add a further boundary condition due to the introduced additional dynamics. Here we can solve this problem in its original index three formulation, where Hypothesis 2.1 is fulfilled with $\mu = 2$, $d = 2$, $a = 3$. Instead of (4.5), we also used the boundary conditions

$$v_2(0) = v_2(2.5) = 0, \tag{4.6}$$

thus seeking a periodic orbit. Observe that we must fix the phase of the solution since the problem is autonomous.

Starting in both cases with the initial value

$$\begin{aligned} x_{00} &= (1, 0.3, 0, 0, -1), & \dot{x}_{00} &= (0, 0, 0, -g, 0), \\ \ddot{x}_{00} &= (0, -g, 0, 0, 0), & x_{00}^{(3)} &= (0, 0, 0, 0, 0), \end{aligned}$$

using $k = 5$ collocation points per subinterval and a uniform mesh π with five subintervals, we obtained in about 3 seconds solutions according to Table 3.

Table 3: Results for Example 4.3

Boundary condition (4.5)		Boundary condition (4.6)	
m	$\lVert z_{m+1} - z_m \rVert_2$	m	$\lVert z_{m+1} - z_m \rVert_2$
0	`0.287D+03`	0	`0.256D+04`
1	`0.149D+03`	1	`0.136D+04`
2	`0.816D+01`	2	`0.178D+03`
3	`0.348D-01`	3	`0.407D+01`
4	`0.161D-05`	4	`0.320D-02`
		5	`0.104D-05`

Example 4.4 In [18], the model of a (two-dimensional) truck is given. It has the form of a standard multibody system

$$\begin{aligned} &\dot{p} = v, \\ &M\dot{v} = f(p, v, u, \dot{u}) - g_p(p)^T \lambda, \\ &g(p) = 0, \end{aligned}$$

where p are the (generalized) positions, v the corresponding velocities and λ the forces introduced by the constraint $g(p) = 0$. In the truck model, p and v have eleven components and λ is scalar. The differentiation index is again three and Hypothesis 2.1 is fulfilled with strangeness $\mu = 2$, $d = 20$, and $a = 3$. The (scalar) function u models the road profile and is chosen here to be

$$u(t) = \tau \sin(20\pi t).$$

Asking as in the linear case [19] for the periodic response of the system for $\tau = 0.05$, we require the boundary conditions

$$\begin{aligned} p_l(0) &= p_l(0.1), \quad l = 1, \ldots, 9, 11, \\ v_l(0) &= v_l(0.1), \quad l = 1, \ldots, 9, 11. \end{aligned}$$

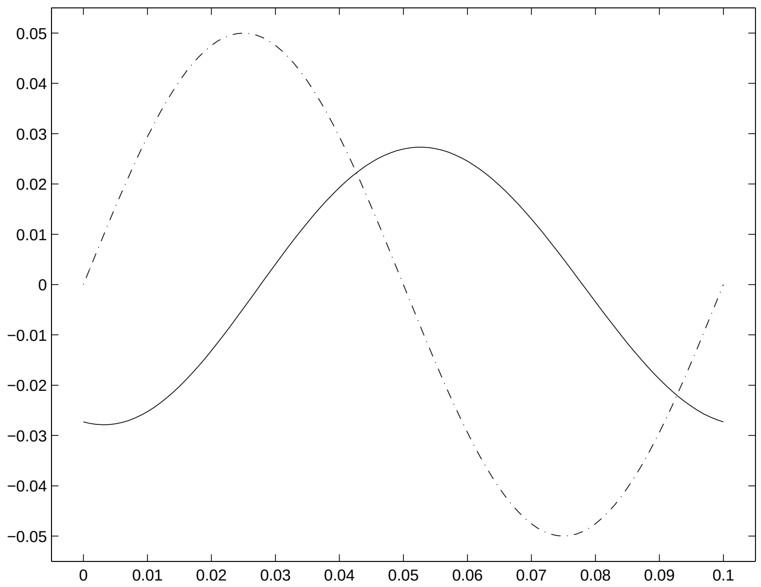

Figure 1: Road profile u -·- and response driver seat — for Example 4.4

This problem suffers from an extremely bad scaling and high nonlinearity. Therefore, we applied a (fixed) scaling to get reasonable condition numbers and used classical homotopy according to

$$\tau \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$$

to get the desired solution. The homotopy was started with the equilibrium state for $\tau = 0$. The course of the iteration procedure for $k = 5$ and a uniform mesh with five subintervals can be found in Table 4. The computation took 61 seconds, with about 10.4 seconds for every homotopy step (i. e., one Gauß-Newton-like iteration). Figure 1 shows the computed response of the driver's seat (i. e., p_{20}) in comparison to the road profile u.

Table 4: Values $\|z_{m+1} - z_m\|_2$ for the homotopy of Example 4.4

m	$\tau = 0.01$	$\tau = 0.02$	$\tau = 0.03$	$\tau = 0.04$	$\tau = 0.05$
0	0.102D+05	0.111D+05	0.114D+05	0.117D+05	0.119D+05
1	0.276D+04	0.801D+03	0.829D+03	0.920D+03	0.114D+04
2	0.303D+03	0.524D+01	0.924D+01	0.121D+02	0.122D+02
3	0.172D+01	0.190D-02	0.218D-02	0.194D-02	0.139D-02
4	0.754D-04	0.268D-08	0.751D-08	0.147D-07	0.193D-07

Example 4.5 The so-called Lotka-Volterra system is the simplest model for a predator/prey interaction and consists (in normalized form) of the two differential equations

$$\dot{x}_1 = x_1(1 - x_2), \quad \dot{x}_2 = -cx_2(1 - x_1)$$

with some constant $c > 0$. It is well-known that the quantity

$$H = c(x_1 - \log x_1) + (x_2 - \log x_2)$$

stays constant along every componentwise positive solution and that therefore every such solution is periodic, but the period T is not known. In order to compute a periodic orbit for a given value of H, we can use the above equation for H as algebraic constraint for the Lotka-Volterra system. But then the system would be overdetermined. We therefore combine the two differential equations such that the resulting relation defines a flow on the manifold defined by the algebraic constraint. Observing that we must fix the phase of the periodic orbit in order to fix a locally unique solution, we obtain the boundary value problem

$$\begin{aligned}
&(1 - x_1)\dot{x}_2 - c(1 - x_2)\dot{x}_1 + cx_2(1 - x_1)^2 + cx_1(1 - x_2)^2 = 0,\\
&c(x_1 - \log x_1) + (x_2 - \log x_2) - H = 0,\\
&x_1(0) = x_1(T), \quad x_1(0) = 1.
\end{aligned}$$

Note that now the derivatives $\dot{x}_1$ and $\dot{x}_2$ have solution dependent factors. Transforming the problem finally to unit interval and using $x_3 = H$ and $x_4 = T$ as further unknowns, the boundary value problem to solve reads

$$\begin{aligned}
&(1 - x_1)\dot{x}_2 - c(1 - x_2)\dot{x}_1 + cx_2(1 - x_1)^2 x_4 + cx_1(1 - x_2)^2 x_4 = 0,\\
&c(x_1 - \log x_1) + (x_2 - \log x_2) - x_3 = 0, \quad \dot{x}_3 = 0, \quad \dot{x}_4 = 0,\\
&x_1(0) = x_1(1), \quad x_1(0) = 1, \quad x_3(0) = H.
\end{aligned}$$

It has differentiation index one problem and satisfies Hypothesis 2.1 with $\mu = 0$, $d = 3$, and $a = 1$. Starting at

$$(x_{00}, \dot{x}_{00}) = (1.0, 0.6, 2.1, 6.0, 3.2, 0, 0, 0)$$

for the choice $c = 1$, $H = 2.2$ and using $k = 5$ together with a mesh of five uniform subintervals we obtained the solution with period $T = 6.4943$. The computation took about 3.0 seconds (from which 1.1 seconds were for the Gauß-Newton-like iteration), the convergence behavior is reported in Table 5.

Table 5: Convergence behavior for Example 4.5

m	$\|z_{m+1} - z_m\|_2$
0	0.207D+03
1	0.320D+02
2	0.498D+01
3	0.405D-01
4	0.177D-05

To summarize the numerical examples, we have demonstrated that the presented collocation methods are able to solve differential-algebraic BVP with different values of the index and different structures. Apart from [14], there are no other numerical methods that can deal with such general problems. Moreover, looking at the convergence results in Tables 1–5 we recognize the very good convergence properties of method (4.4), which cannot be distinguished from quadratic convergence.

5 Conclusions

In this paper, we have developed symmetric collocation methods for the solution of nonlinear differential-algebraic boundary value problems. No restrictions on index or structure are necessary. As in the linear case [15], Gauß-type schemes for the differential part and Lobatto-type schemes (with one more node) for the algebraic part are used. We have shown that the convergence results known for ordinary differential equations also hold in the case of differential-algebraic BVPs, including superconvergence. A Gauß-Newton-type method for the numerical solution of the underdetermined collocation systems has been implemented and used to demonstrate the applicability for several challenging examples.

References

[1] U. M. Ascher, R. Mattheij, and R. Russell. *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations.* SIAM, Philadelphia, 2nd edition, 1995.

[2] U. M. Ascher and L. R. Petzold. Projected collocation for higher-order higher-index differential-algebraic equations. *J. Comp. Appl. Math.*, 43:243–259, 1992.

[3] U. M. Ascher and R. Spiteri. Collocation software for boundary value differential-algebraic equations. *SIAM J. Sci. Comput.*, 15:938–952, 1994.

[4] S. L. Campbell. A general form for solvable linear time varying singular systems of differential equations. *SIAM J. Math. Anal.*, 18:1101–1115, 1987.

[5] S. L. Campbell and E. Griepentrog. Solvability of general differential algebraic equations. *SIAM J. Sci. Comput.*, 16:257–270, 1995.

[6] P. Deuflhard and G. Heindl. Affine invariant convergence theorems for Newton's method and extensions to related methods. *SIAM J. Numer. Anal.*, 16:1–10, 1979.

[7] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II.* Springer-Verlag, Berlin, 1991.

[8] L. W. Kantorowitsch and G. P. Akilow. *Funktionalanalysis in normierten Räumen.* Akademie-Verlag, Berlin, 2nd edition, 1978.

[9] P. Kunkel and V. Mehrmann. Regular solutions of nonlinear differential-algebraic equations and their numerical determination. *Numer. Math.*, 79:581–600, 1998.

[10] P. Kunkel and V. Mehrmann. Analysis of over- and underdetermined nonlinear differential-algebraic systems with application to nonlinear control problems. *Math. Contr. Sign. Syst.*, 14:233–256, 2001.

[11] P. Kunkel and V. Mehrmann. Index reduction for differential-algebraic equations by minimal extension. Technical Report 719-01, Institut für Mathematik, TU Berlin, D-10623 Berlin, FRG, 2001.

[12] P. Kunkel, V. Mehrmann, and I. Seufer. GENDA: A software package for the solution of GEneral Nonlinear Differential-Algebraic equations. Technical Report 730-02, Institut für Mathematik, TU Berlin, D-10623 Berlin, FRG, 2002.

[13] P. Kunkel, V. Mehrmann, and R. Stöver. Symmetric collocation for unstructured nonlinear differential-algebraic equations of arbitrary index.

Technical Report 02-12, Zentrum für Technomathematik, Universität Bremen, D-28334 Bremen, FRG, 2002.

[14] P. Kunkel, V. Mehrmann, and R. Stöver. Multiple shooting for unstructured nonlinear differential-algebraic equations of arbitrary index. Technical Report 751-02, Institut für Mathematik, TU Berlin, D-10623 Berlin, FRG, 2002.

[15] P. Kunkel and R. Stöver. Symmetric collocation methods for linear differential-algebraic boundary value problems. *Numer. Math*, 91:475–501, 2002.

[16] R. Lamour. A Shooting Method for Fully Implicit Index-2 DAEs. *SIAM J. Sci. Comput.*, 18:94-114, 1997.

[17] The MathWorks, Inc., Cochituate Place, 24 Prime Park Way, Natick, Mass., 01760. *MATLAB Version 6.0.0.88*, 2001.

[18] B. Simeon, F. Grupp, C. Führer, and P. Rentrop. A nonlinear truck model and its treatment as a multibody system. *J. Comput. Appl. Math.*, 50:523–532, 1994.

[19] R. Stöver. Collocation methods for solving linear differential-algebraic boundary value problems. *Numer. Math.*, 88:771–795, 2001.

A A generalized simplified Newton method

In this appendix, we consider the solution of a nonlinear system of equations

$$F(x) = 0 \tag{A.1}$$

with $F : \mathbb{D} \to \mathbb{R}^n$, $\mathbb{D} \subseteq \mathbb{R}^n$ open and convex, by the iteration method

$$x_{m+1} = x_m - F'(\hat{x})^{-1} F(x_m) \tag{A.2}$$

for given $\hat{x}, x_0 \in \mathbb{D}$. For such iterations the following convergence result holds, cp. [8, Ch. XVIII].

Theorem A.1 *Let $F \in C^1(\mathbb{D}, \mathbb{R}^n)$ and $\hat{x}, x_0 \in \mathbb{D}$ such that $F'(\hat{x})$ is invertible. Furthermore, let constants α, β, γ be given such that*

$$\begin{array}{ll} \text{(a)} & \|F'(\hat{x})^{-1}F(x_0)\| \leq \alpha, \\ \text{(b)} & \|F'(\hat{x})^{-1}\| \leq \beta, \\ \text{(c)} & \|F'(x) - F'(y)\| \leq \gamma\|x - y\| \text{ for all } x, y \in \mathbb{D},\ \gamma \neq 0, \\ \text{(d)} & \|x_0 - \hat{x}\| < \frac{1}{\beta\gamma}, \\ \text{(e)} & 2\alpha\beta\gamma \leq (1 + \beta\gamma\hat{t})^2 \quad \text{with } \hat{t} = -\|x_0 - \hat{x}\| \\ \text{(f)} & \overline{S}(x_0, \rho_-) \subseteq \mathbb{D}, \\ & \qquad \rho_\pm = \frac{1}{\beta\gamma}\left(1 + \beta\gamma\hat{t} \pm \sqrt{(1 + \beta\gamma\hat{t})^2 - 2\alpha\beta\gamma}\right) \end{array} \tag{A.3}$$

for some vector norm and the associated matrix norm. Then, (A.2) defines a sequence $\{x_m\}$ of points in $\overline{S}(x_0, \rho_-)$ which converges to a point x^* in $\overline{S}(x_0, \rho_-)$ satisfying $F(x^*) = 0$. There is no other solution of (A.1) in

$$\overline{S}(x_0, \rho_-) \cup (S(x_0, \rho_+) \cap \mathbb{D}).$$

In particular, for $\rho_- < \rho_+$ the solution x^* is locally unique.

Proof. A proof is given in [13]. □

Corollary A.2 *If in addition to the assumptions of Theorem A.1*

$$\|x_0 - \hat{x}\| \leq \tfrac{1}{2\beta\gamma}$$

holds, then

$$\|x^* - x_0\| \leq 4\alpha.$$

Proof. Theorem A.1 yields

$$\begin{aligned} \|x^* - x_0\| &\leq \rho_- = \\ &= \frac{2\alpha}{1 - \beta\gamma\|x_0 - \hat{x}\| + \sqrt{(1 - \beta\gamma\|x_0 - \hat{x}\|)^2 - 2\alpha\beta\gamma}} \leq \\ &= \frac{2\alpha}{1 - \beta\gamma\|x_0 - \hat{x}\|} \leq 4\alpha. \end{aligned}$$

□

Remark A.3 *Theorem A.1 holds almost verbatim in the case of an infinite dimensional Banach space problem. To avoid the argument in the proof using integration, we can simply replace (A.3c) by the assumption*

$$\|F(x) - F(y) - F'(z)(x - y)\| \leq \tfrac{1}{2}\gamma\|x - y\|\Big(\|x - z\| + \|y - z\|\Big)$$

for all $x, y, z \in \mathbb{D}$.

B Proof of Lemma 3.5

Let $x, y, z \in \mathbb{D}$, set $g = (f_1, f_2, v) = L(x) - L(y) - DL[z](x-y)$, and introduce the convex combination $u(t;s) = y(t) + s(x(t) - y(t))$ with $s \in [0,1]$. For the first component f_1 we have

$$\begin{aligned}
&\|f_1(t)\|_\infty = \\
&= \|\hat{F}_1(t, x(t), \dot{x}(t)) - \hat{F}_1(t, y(t), \dot{y}(t)) - \\
&\qquad -\hat{F}_{1;x}(t, z(t), \dot{z}(t))(x(t) - y(t)) - \hat{F}_{1;\dot{x}}(t, z(t), \dot{z}(t))(\dot{x}(t) - \dot{y}(t))\|_\infty = \\
&= \|\hat{F}_1(t, u(t;s), \dot{u}(t;s))\Big|_{s=0}^{s=1} \\
&\qquad -\hat{F}_{1;x}(t, z(t), \dot{z}(t))(x(t) - y(t)) - \hat{F}_{1;\dot{x}}(t, z(t), \dot{z}(t))(\dot{x}(t) - \dot{y}(t))\|_\infty = \\
&= \| \int_0^1 \Big[\Big(\hat{F}_{1;x}(t, u(t;s), \dot{u}(t;s)) - \hat{F}_{1;x}(t, z(t), \dot{z}(t))\Big)(x(t) - y(t)) + \\
&\qquad +\Big(\hat{F}_{1;\dot{x}}(t, u(t;s), \dot{u}(t;s)) - \hat{F}_{1;\dot{x}}(t, z(t), \dot{z}(t))\Big)(\dot{x}(t) - \dot{y}(t))\Big] ds\|_\infty \leq \\
&\leq \int_0^1 \Big[\Big(\gamma_1 \|u(t;s) - z(t)\|_\infty + \gamma_2 \|\dot{u}(t;s) - \dot{z}(t)\|_\infty\Big)\|x(t) - y(t)\|_\infty + \\
&\qquad +\Big(\gamma_3 \|u(t;s) - z(t)\|_\infty + \gamma_4 \|\dot{u}(t;s) - \dot{z}(t)\|_\infty\Big)\|\dot{x}(t) - \dot{y}(t)\|_\infty\Big] ds \leq \\
&\leq \gamma \|x - y\|_{\mathbb{X}} \int_0^1 \|u(\,\cdot\,;s) - z\|_{\mathbb{X}} ds
\end{aligned}$$

with all constants being independent of t, x, y, z and h. Analogously, we get for the second component f_2

$$\begin{aligned}
\|f_2(t)\|_\infty &= \|\hat{F}_2(t, x(t)) - \hat{F}_2(t, y(t)) - \hat{F}_{2;x}(t, z(t))(x(t) - y(t))\|_\infty = \\
&= \|\hat{F}_2(t, u(t;s))\Big|_{s=0}^{s=1} - \hat{F}_{2;x}(t, z(t))(x(t) - y(t))\|_\infty = \\
&= \| \int_0^1 \Big(\hat{F}_{2;x}(t, u(t;s)) - \hat{F}_{2;x}(t, z(t))\Big)(x(t) - y(t)) ds\|_\infty \leq \\
&\leq \gamma \|x - y\|_{\mathbb{X}} \int_0^1 \|u(\,\cdot\,;s) - z\|_{\mathbb{X}} ds,
\end{aligned}$$

possibly increasing γ. Furthermore,

$$\begin{aligned}
\|\dot{f}_2(t)\|_\infty &= \| \int_0^1 \Big[\Big(\hat{F}_{2;tx}(t, u(t;s)) + \hat{F}_{2;xx}(t, u(t;s))(\dot{u}(t;s)) - \\
&\qquad -\hat{F}_{2;tx}(t, z(t)) - \hat{F}_{2;xx}(t, z(t))(\dot{z}(t))\Big)(x(t) - y(t)) +
\end{aligned}$$

$$\begin{aligned}
&+\Big(\hat{F}_{2;x}(t,u(t;s))-\hat{F}_{2;x}(t,z(t))\Big)(\dot{x}(t)-\dot{y}(t))\Big]ds\|_\infty \leq \\
\leq\ & \int_0^1 \Big[\Big(\gamma_1\|u(t;s)-z(t)\|_\infty+\gamma_2\|u(t;s)-z(t)\|_\infty+ \\
&+\gamma_3\|\dot{u}(t;s)-\dot{z}(t)\|_\infty\Big)\|x(t)-y(t)\|_\infty+ \\
&+\gamma_4\|u(t;s)-z(t)\|_\infty\|\dot{x}(t)-\dot{y}(t)\|_\infty\Big]ds \leq \\
\leq\ & \gamma\|x-y\|_{\mathbb{X}}\int_0^1\|u(\,\cdot\,;s)-z\|_{\mathbb{X}}ds,
\end{aligned}$$

again possibly increasing γ. Finally, for v we get

$$\begin{aligned}
&\|v\|_\infty = \\
=\ & \|r(x(\underline{t}),x(\overline{t}))-r(y(\underline{t}),y(\overline{t}))- \\
&-r_{x_a}(z(\underline{t}),z(\overline{t}))(x(\underline{t})-y(\underline{t}))-r_{x_b}(z(\underline{t}),z(\overline{t}))(x(\overline{t})-y(\overline{t}))\|_\infty = \\
=\ & \|r(u(\underline{t};s),u(\overline{t};s))\Big|_{s=0}^{s=1}- \\
&-r_{x_a}(z(\underline{t}),z(\overline{t}))(x(\underline{t})-y(\underline{t}))-r_{x_b}(z(\underline{t}),z(\overline{t}))(x(\overline{t})-y(\overline{t}))\|_\infty = \\
=\ & \|\int_0^1\Big[\Big(r_{x_a}(u(\underline{t};s),u(\overline{t};s))-r_{x_a}(z(\underline{t}),z(\overline{t}))\Big)(x(\underline{t})-y(\underline{t}))+ \\
&+\Big(r_{x_b}(u(\underline{t};s),u(\overline{t};s))-r_{x_b}(z(\underline{t}),z(\overline{t}))\Big)(x(\overline{t})-y(\overline{t}))\Big]ds\|_\infty \leq \\
\leq\ & \int_0^1\Big[\Big(\gamma_1\|u(\underline{t};s)-z(\underline{t})\|_\infty+\gamma_2\|u(\overline{t};s)-z(\overline{t})\|_\infty\Big)\|x(\underline{t})-y(\underline{t})\|_\infty+ \\
&+\Big(\gamma_3\|u(\underline{t};s)-z(\underline{t})\|_\infty+\gamma_4\|u(\overline{t};s)-z(\overline{t})\|_\infty\Big)\|x(\overline{t})-y(\overline{t})\|_\infty\Big]ds \leq \\
\leq\ & \gamma\|x-y\|_{\mathbb{X}}\int_0^1\|u(\,\cdot\,;s)-z\|_{\mathbb{X}}ds,
\end{aligned}$$

and thus we have (again possibly increasing γ)

$$\begin{aligned}
\|g\|_{\mathbb{Y}} &\leq \gamma\|x-y\|_{\mathbb{X}}\int_0^1\|y+s(x-y)-z\|_{\mathbb{X}}ds = \\
&= \gamma\|x-y\|_{\mathbb{X}}\int_0^1\|s(x-z)+(1-s)(y-z)\|_{\mathbb{X}}ds \leq \\
&\leq \tfrac{1}{2}\gamma\|x-y\|_{\mathbb{X}}\Big(\|x-z\|_{\mathbb{X}}+\|y-z\|_{\mathbb{X}}\Big).
\end{aligned}$$