Boundary-Value Problems for Differential-Algebraic Equations: A Survey

René Lamour, Roswitha März, and Ewa Weinmüller

René Lamour
Humboldt-University of Berlin, Department of Mathematics, 10099 Berlin, Germany
e-mail: lamour@math.hu-berlin.de

Roswitha März
Humboldt-University of Berlin, Department of Mathematics, 10099 Berlin, Germany
e-mail: maerz@math.hu-berlin.de

Ewa Weinmüller
Vienna University of Technology, Department for Analysis and Scientific Computing, Wiedner Hauptstrasse 8-10, A-1040 Wien, Austria e-mail: e.weinmueller@tuwien.ac.at

Contents

1 Introduction

Usually, a differential-algebraic equation (DAE) has a family of solutions; to pick one of them, one has to supply additional conditions. In an initial value problem (IVP), the solution is specified by its value at a single point. A genuine boundary value problem (BVP) assigns solution and derivative values at more than one point. Most commonly, the solutions are fixed at just two points, the *boundaries*. IVPs can be seen as relatively simple special cases of BVPs.

BVPs constitute an important area of applied mathematics already for explicit ordinary differential equations (ODEs), e.g., [12]. This applies even more for DAEs. We follow [12] in mainly concentrating on two-point BVPs.

Till now, both analytical theory and numerical treatment of DAEs are mainly focused on IVPs. The more complex BVPs have not been studied with similar intensity. The related early work until 2001 is carefully summarized in [102, Section 81]. With the present paper we intend to provide an actual survey of this field.

Optimal control is one of the traditional sources of BVPs for DAEs. As it is well-known, extremal conditions for optimal control problems subject to constraints given by explicit ODEs yield BVPs for semi-explicit DAEs. If the constraints themselves are given by DAEs, the extremal conditions lead to BVPs for DAEs (e.g., [33, 54]) even more.

An important area yielding DAEs is network modeling in different application fields, for instance, electrical networks ([104, 103, 83]), and multibody systems ([45, 109]). One is interested in BVPs transforming one state or position into another, often also in periodic solutions.

DAEs in applications usually need an involved technical description and show high dimensions. Here we avoid repeating extensive case studies and prefer small, clear, possibly academic examples. We hint at some essentials by means of easy examples. We recognize features coming over from the well-known classical ODE theory, but we indicate also further difficulties emerging from the DAE context. The first example is taken from [20].

Example 1.1. Minimize the cost

$$J(x) = \int_0^{t_f} (x_3(t)^2 + (x_4(t) - R^2)^2) \mathrm{d}t$$

subject to the constraints

$$\begin{aligned} x_1'(t) + x_2(t) &= 0, & x_1(0) &= r, \\ x_2'(t) - x_1(t) - x_3(t) &= 0, & x_2(0) &= 0, \\ x_1(t)^2 + x_2(t)^2 - x_4(t) &= 0, \end{aligned}$$

with constants $r > 0$, $R > 0$. The component x_3 can be seen as a control function. For arbitrary given function x_3 the resulting components x_1, x_2, x_4 are uniquely deter-

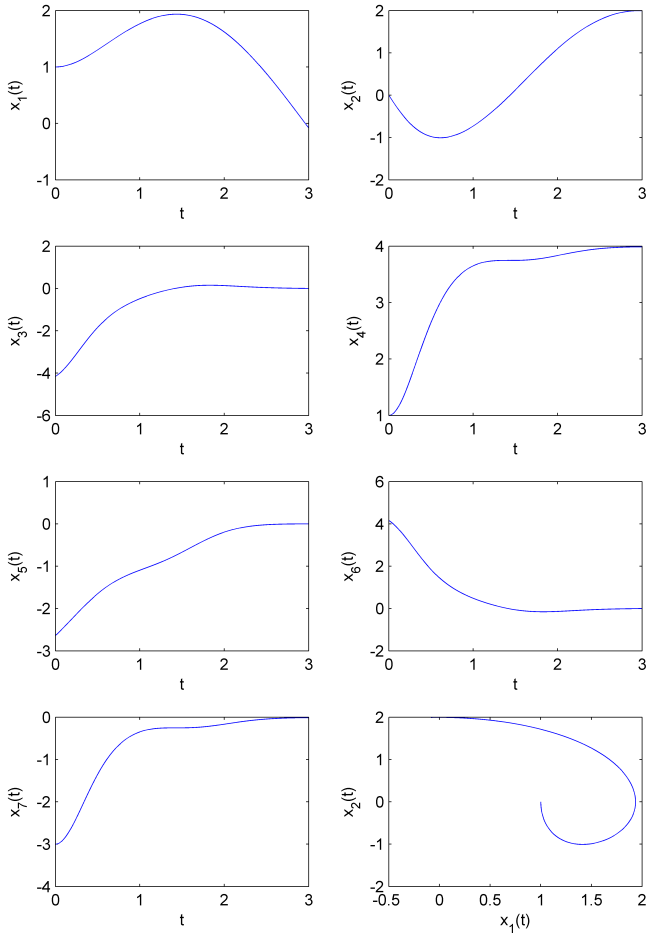

Fig. 1 Solution of the optimal BVP in Example 1.1

mined. In particular, if $x_3(t)$ vanishes identically, the remaining IVP has the unique solution $x_1(t) = r\cos t$, $x_2(t) = r\sin t$, $x_4(t) = r^2$. Then the point $(x_1(t), x_2(t))$ orbits the origin with radius r and the cost amounts to $\mathcal{J}(x) = 13.5$.
By minimizing the cost, the point $(x_1(t), x_2(t))$ becomes driven to the circle of radius

R, with low cost of $x_3(t)$. Figure 1 shows a locally optimal solution for $t_f = 3$, $r = 1$, $R = 2$, yielding the cost $\mathcal{J}(x) = 4.397$, which was generated by means of the associated extremal condition, the so-called optimality BVP,

$$\begin{aligned}
x_1'(t) + x_2(t) &= 0, & \\
x_2'(t) - x_1(t) - x_3(t) &= 0, & x_1(0) &= r, \\
x_1(t)^2 + x_2(t)^2 - x_4(t) &= 0, & x_2(0) &= 0, \\
-x_5'(t) - x_6(t) - 2x_1(t)x_7(t) &= 0, & x_5(t_f) &= 0, \\
-x_6'(t) + x_5(t) - 2x_2(t)x_7(t) &= 0, & x_6(t_f) &= 0, \\
x_6(t) + x_3(t) &= 0, & & \\
x_7(t) - x_4(t) + R^2 &= 0. & &
\end{aligned}$$

This BVP is solvable and locally well-posed, see [94, Example 6.4]. Owing to the given initial condition in the minimization problem, the optimality BVP shows separated boundary conditions. We emphasize that, for well-posedness of the optimality BVP, one necessarily needs appropriate initial conditions in the minimization problem. For instance, requiring there additionally $x_4(0) = 0$ is not a good idea.

We observe that any solution of the DAE, among them the solution of the BVP, must reside in the obvious restriction set

$$\mathcal{M}_0 = \{x \in \mathbb{R}^7 : x_1^2 + x_2^2 - x_4 = 0,\ x_6 + x_3 = 0,\ x_7 - x_4 + R^2 = 0\}.$$

Replacing the given constant R in the problem by a time-varying function $R(\cdot)$ does not change the well-posedness of the BVP. However, then one is confronted with a time-varying restriction set $\mathcal{M}_0(t)$ such that $x(t) \in \mathcal{M}_0(t)$ holds for all DAE solutions wherever they exist. □

The next example ([31], cf. [84]) shows a semi-explicit DAE describing a minimal instance of an electrical network.

Example 1.2. The DAE

$$\begin{aligned}
x_1'(t) &= -\frac{G_L}{C_1} x_1(t) + \frac{F(-(x_1(t) + x_3(t)))}{C_1}, \\
x_2'(t) &= -\frac{1}{C_2 R_Q}(x_2(t) + x_3(t) + E(t)), \\
0 &= -\frac{1}{R_Q}(x_2(t) + x_3(t) + E(t)) + F(-(x_1(t) + x_3(t))) - F(x_3(t)),
\end{aligned}$$

describes the voltage doubling network from Figure 2, where

$$E(t) = 3.95\ \sin\left(2\pi \frac{t}{T}\right) \mathrm{kV}, \quad T = 0.064, \quad F(u) = 5 \cdot 10^{-5}(e^{630u} - 1)\,\mathrm{mA}$$

and

$$C_1 = C_2 = 2.75\mathrm{nF}, \quad G_L = \frac{1}{R_L}, \quad R_Q = 0.1\mathrm{M}\Omega, \quad R_L = 10\mathrm{M}\Omega.$$

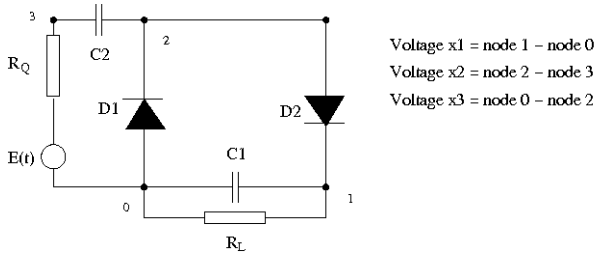

Fig. 2 Voltage doubling network in Example 1.2

We ask for a solution of this DAE which satisfies the nonseparated boundary condition

$$x_1(0) - x_1(T) = 0,$$
$$x_2(0) - x_2(T) = 0.$$

The BVP proves to be solvable and locally well-posed in its natural setting. Again, the right number of boundary conditions plays its role for well-posedness. The solution results to be T-periodic. It is displayed in Figure 3. It can be provided numerically, only.
Replacing the above boundary condition by $x(0) = x(T)$ leads to a solvable BVP, but it is no longer well-posed because of too much conditions.
Furthermore, the T-periodic solution is asymptotically stable. A fact which is checked in [84], via the eigenvalues of the monodromy matrix. Again all solutions

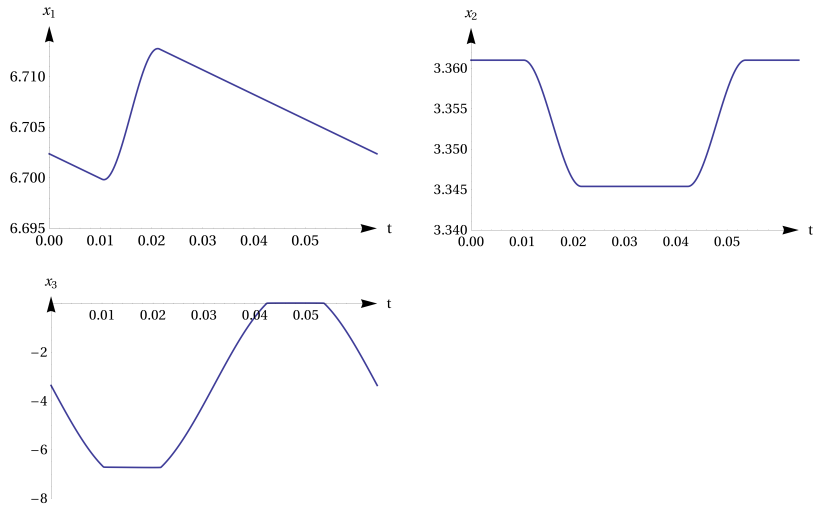

Fig. 3 T-periodic solution of the DAE in Example 1.2

of the DAE must reside in a restriction set, now in

$$\mathcal{M}_0(t) = \{x \in \mathbb{R}^3 : -\frac{1}{R_Q}(x_2 + x_3 + E(t)) + F(-(x_1 + x_3)) - F(x_3) = 0\}.$$

Though here the dimension is lower than in Example 1.1, the restriction set looks less transparent. □

Time-varying restriction sets are typical for DAEs in applications, and the solutions are not expected to feature high smoothness. From this point of view, the popular opinion that DAEs are nothing else vector fields on smooth manifolds is somehow limited. Nevertheless, corresponding case studies are helpful to gain insights.

Example 1.3. Consider the DAE

$$\begin{aligned} x_1'(t) + x_1(t) - x_2(t) - x_1(t)x_3(t) + (x_3(t) - 1)\sin t &= 0,\\ x_2'(t) + x_1(t) + x_2(t) - x_2(t)x_3(t) + (x_3(t) - 1)\cos t &= 0,\\ x_1(t)^2 + x_2(t)^2 + x_3(t) - 1 - \alpha(t) &= 0, \end{aligned}$$

with a given scalar function α, and the separated, nonlinear boundary condition

$$\begin{aligned} x_1(0) &= 0,\\ x_1(2\pi)^2 + x_2(2\pi)^2 &= 1. \end{aligned}$$

Here we have the transparent restriction set

$$\mathcal{M}_0(t) = \{x \in \mathbb{R}^3 : x_1^2 + x_2^2 + x_3 - 1 - \alpha(t) = 0\}$$

moving in $\mathbb{R}^3$. The BVP has the solution

$$x_{*1}(t) = \sin t, \quad x_{*2}(t) = \cos t, \quad x_{*3}(t) = \alpha(t).$$

This BVP will turn out to be locally well-posed, no matter how α behaves, see Example 2.6.

Replacing the boundary conditions by the new ones

$$\begin{aligned} x_1(0) - x_1(2\pi) &= 0,\\ x_2(0) - x_2(2\pi) &= 0, \end{aligned}$$

the situation changes. Assume α to be a 2π-periodic function, so that the restriction set $\mathcal{M}_0(t)$ moves periodically and each BVP solution has the property $x(0) = x(2\pi)$. We speak then shortly of a periodic BVP. Clearly, the above solution x_* of the DAE satisfies at the same time the periodic BVP.
The periodic BVP turns out to be locally well-posed for most functions α, among them $\alpha = 0$, see Example 2.6.

In contrast, the periodic BVP is no longer well-posed for $\alpha \equiv 1$. Then there is an entire family of solutions: For arbitrary parameters $c_1, c_2 \in \mathbb{R}$, $c_1^2 + c_2^2 = 1$, the

function

$$x_{**}(t) = \begin{bmatrix} c_1 \cos t + c_2 \sin t \\ c_2 \cos t - c_1 \sin t \\ 1 \end{bmatrix}$$

is a 2π-periodic solution of the DAE. Here we see a phenomenon coming over from classical BVPs in explicit ODEs. A correct number of boundary conditions is necessary but not sufficient for the well-posedness of a BVP. It is also necessary that the boundary conditions are consistent with the flow. Of course, the same remains true for DAEs.

It is quite difficult to picture the flow of a DAE. Figure 4 sketches the flow on $\mathcal{M}_0$ for the easier case $\alpha \equiv 0$. It is dominated by the asymptotically stable 2π-periodic solution

$$x_{*1}(t) = \sin t,\ x_{*2}(t) = \cos t,\ x_{*3}(t) = 0,$$

of the DAE which satisfies also our BVPs and the unstable stationary solution

$$x_{*1}(t) = 0,\ x_{*2}(t) = 0,\ x_{*3}(t) = 1.$$

□

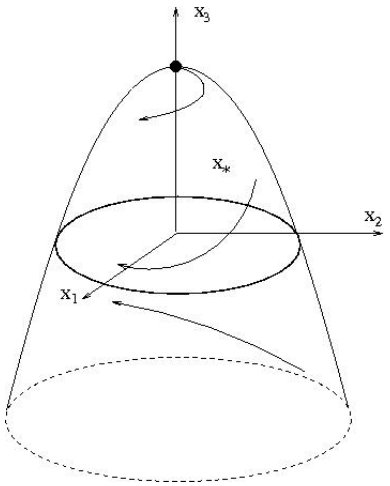

Fig. 4 Flow on the constraint set in Example 1.3 for α vanishing identically

Example 1.4. The solutions of the DAE

$$x_1'(t) + x_1(t) = 0,$$
$$x_2(t)\, x_2'(t) - x_3(t) = 0,$$
$$x_1(t)^2 + x_2(t)^2 - 1 + \frac{1}{2}\cos(\pi t) = 0,$$

reside in the set

$$\mathcal{M}_0(t) := \{x \in \mathbb{R}^3 : x_1^2 + x_2^2 - 1 + \frac{1}{2}\cos(\pi t) = 0\}.$$

A further look at this DAE makes clear that there is another set the solution values have to belong to. Namely, for any solution $x_*(.)$, differentiating the identity $x_{*1}(t)^2 + x_{*2}(t)^2 - 1 + \frac{1}{2}\cos(\pi t) = 0$ and replacing the expressions for the derivatives we obtain the new identity

$$-2x_{*1}(t)^2 + 2x_{*3}(t) - \frac{1}{2}\pi \sin(\pi t) = 0.$$

Therefore, all solution values $x_*(t)$ must also satisfy this hidden constraint, that is, they must belong to the set

$$\mathcal{H}(t) := \{x \in \mathbb{R}^3 : -2x_1^2 + 2x_3 - \frac{1}{2}\pi \sin(\pi t) = 0\}.$$

The presence of hidden constraints complicates the matter. The obvious restriction set $\mathcal{M}_0(t)$ contains points which are no longer consistent, but the consistent values must belong to the proper subset

$$\mathcal{M}_1(t) := \mathcal{M}_0(t) \cap \mathcal{H}(t) \subset \mathcal{M}_0(t).$$

Figure 5 shows $\mathcal{M}_1(t)$ for $t = 0$ and $t = \frac{1}{2}$.

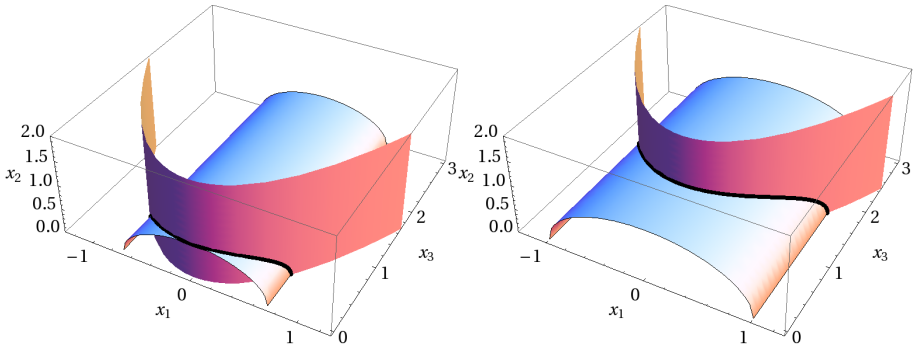

Fig. 5 Constraint set $\mathcal{M}_1$ at $t = 0$ and $t = \frac{1}{2}$ in Example 1.4

Regarding the boundary condition

$$x_1(0)-x_1(2)=\alpha,\quad |\alpha|<\frac{1}{2}(1-\mathrm{e}^{-2}),$$

the BVP has the two solutions

$$x_{*\,1}=c\mathrm{e}^{-t},\quad x_{*\,2}=(1-\frac{1}{2}\cos\pi t-c^2\mathrm{e}^{-2t})^{\frac{1}{2}},\quad x_{*\,3}=\frac{1}{4}\pi\sin\pi t+c^2\mathrm{e}^{-2t},$$

and

$$x_{**\,1}=c\mathrm{e}^{-t},\quad x_{**\,2}=-(1-\frac{1}{2}\cos\pi t-c^2\mathrm{e}^{-2t})^{\frac{1}{2}},\quad x_{**\,3}=\frac{1}{4}\pi\sin\pi t+c^2\mathrm{e}^{-2t},$$

where $c:=\alpha/(1-\mathrm{e}^{-2})$. In particular, for $\alpha=0$, thus $c=0$, the first solution component which governs the inherent dynamics becomes stationary.

The boundary condition proves to be accurately stated locally around x_*. Namely, for each arbitrary sufficiently small γ, the BVP with perturbed boundary condition

$$x_1(0)-x_1(2)=\alpha+\gamma,$$

possesses a unique solution x in the neighborhood of x_* and the inequality

$$\|x-x_*\|_\infty\leq\frac{2}{1-\mathrm{e}^{-2}}\,|\gamma|$$

is valid. This can be checked by straightforward computations. An analogous result can be derived regarding the reference solution x_{**}. Nevertheless, the BVP fails to be locally well-posed in the natural setting. Still, it will be shown to become well-posed in an special advanced setting, see Example 2.7. □

Usually, DAEs are given either in standard form

$$\mathfrak{f}(x'(t),x(t),t)=0 \tag{1}$$

or in the advanced form

$$f((Dx)'(t),x(t),t)=0, \tag{2}$$

with an extra matrix function D indicating which derivatives are actually involved. Most of the DAEs arising in applications originally show the latter form ([45, 103, 35]). For large classes of DAEs being of interest in the context of BVPs, for instance semi-explicit DAEs, the equation (1) can be also written in the form (2) as

$$\mathfrak{f}((D_{inc}x)'(t),x(t),t)=0,$$

with a constant incidence matrix D_{inc}. For instance, in Example 1.3 we can simply choose

$$D=\begin{bmatrix}1&0&0\\0&1&0\end{bmatrix},\ f(y,x,t)=\begin{bmatrix}y_1+x_1-x_2-x_1x_3+(x_3-1)\sin t\\ y_2+x_1+x_2-x_2x_3+(x_3-1)\cos t\\ x_1^2+x_2^2+x_3-1-\alpha(t)\end{bmatrix}.$$

In the present paper we deal with DAEs of the form (2), which is more comfortable from the analytic point of view ([83, 96]). Most results remain valid accordingly for the standard form (1).

The well-posed BVPs in Examples 1.1, 1.2, and 1.3 rely on regular index-1 DAEs which behave quite similarly to regular ODEs. In contrast, the solutions of any higher index DAE show an ambivalent character unlike the solutions of explicit ODEs: they are smooth with respect to the integration constant as expected coming from explicit ODEs, however, concerning perturbations of the right-hand side, the solution becomes discontinuous in the natural setting. We refer to the illustrative example [83, Example 1.5] and its functional-analytic interpretation in [96]. The discontinuity concerning the right-hand side causes well-known difficulties in numerical integration procedures and in the numerical treatment of BVPs as well.

Our exposition relies on the projector-based analysis ([83]). In particular, if not explicitly indicated otherwise, the notion *index* stands for *tractability index*. We notice to this end that, for large classes of DAEs, the tractability index coincides with the differentiation index and the perturbation index.

We see here a twofold benefit of the projector based analysis: It serves as integrative framework of the wide survey material and, at the same time, as source of new developments such as the linear BVP theory as counterpart of the classical version in [12]).

2 Analytical theory

2.1 Basic assumptions and terminology

To tie in with the general discussion in [83, 96] we deal with DAEs of the form

$$f((Dx)'(t),x(t),t) = 0, \tag{3}$$

which exhibits the involved derivative by means of an extra matrix valued function D. The function $f : \mathbb{R}^n \times \mathcal{D}_f \times \mathcal{I}_f \longrightarrow \mathbb{R}^m$, $\mathcal{D}_f \times \mathcal{I}_f \subseteq \mathbb{R}^m \times \mathbb{R}$ open, is continuous and has continuous partial derivatives f_y and f_x with respect to the first two variables $y \in \mathbb{R}^n$, $x \in \mathcal{D}_f$. The partial Jacobian $f_y(y,x,t)$ is everywhere singular. The matrix function $D : \mathcal{I}_f \to \mathcal{L}(\mathbb{R}^m,\mathbb{R}^n)$ is at least continuous, often continuously differentiable, and $D(t)$ has constant rank r on the given interval $\mathcal{I}_f$. Always, $\operatorname{im} D$ is supposed to be a $\mathcal{C}^1$-subspace varying in $\mathbb{R}^n$.

We concentrate on two-point boundary conditions

$$g(x(a),x(b)) = 0 \tag{4}$$

described by the continuously differentiable function $g : \mathcal{D}_f \times \mathcal{D}_f \to \mathbb{R}^l$ and two different points $a,b \in \mathcal{I}_f$. The number $l \leq m$ of boundary conditions will be specified below. It strongly depends on the structure of the DAE.

We are looking for classical solution of the DAE (3), that is, for functions from the function space

$$\mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m) := \{x \in \mathcal{C}(\mathcal{I},\mathbb{R}^m) : Dx \in \mathcal{C}^1(\mathcal{I},\mathbb{R}^n)\},$$

defined on an interval $\mathcal{I} \subseteq \mathcal{I}_f$, with values $x(t) \in \mathcal{D}_f, t \in \mathcal{I}$, and satisfying the DAE pointwise on $\mathcal{I}$.
Evidently, for each arbitrary given function $x \in \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m)$, with values $x(t) \in \mathcal{D}_f$, $t \in \mathcal{I} \subseteq \mathcal{I}_f$, the resulting expression

$$q(t) := f((Dx)'(t),x(t),t), \quad t \in \mathcal{I},$$

yields $q \in \mathcal{C}(\mathcal{I},\mathbb{R}^m)$. We say that this function space setting is the *natural setting* of our DAE.

The element $x_0 \in \mathcal{D}_f$ is said to be a *consistent value* of the DAE at time $t_0 \in \mathcal{I}_f$, if there is a solution $x \in \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m)$ given on an interval $\mathcal{I} \ni t_0$ such that $x(t_0) = x_0$. When dealing with BVPs (3), (4) we suppose the compact interval $\mathcal{I} = [a,b]$ and seek functions from $\mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m)$ that satisfy the DAE (3) and, additionally, the boundary condition (4).

Supposing a compact interval $\mathcal{I}$ we equip the function spaces $\mathcal{C}(\mathcal{I},\mathbb{R}^m)$ and $\mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m)$ with the norms

$$\|x\|_\infty := \max_{t\in\mathcal{I}} |x(t)|, \quad x \in \mathcal{C}(\mathcal{I},\mathbb{R}^m),$$
$$\|x\|_{\mathcal{C}^1_D} := \|x\|_\infty + \|(Dx)'\|_\infty, \quad x \in \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m),$$

respectively. This yields Banach spaces.

Definition 2.1. The DAE (3) has a *properly involved derivative*, also called *properly stated leading term*, if $\ker f_y$ is another $\mathcal{C}^1$-subspace varying in $\mathbb{R}^n$, and the transversality condition

$$\ker f_y(y,x,t) \oplus \operatorname{im} D(t) = \mathbb{R}^n, \quad (y,x,t) \in \mathbb{R}^n \times D_f \times \mathcal{I}_f, \tag{5}$$

is valid.

Below, except for Subsection 3.4 on singular problems, we always assume the DAE (3) to have a properly stated leading term. To simplify matters we further assume the nullspace $\ker f_y(y,x,t)$ to be independent of y and x. Then, the transversality condition (5) pointwise induces a projector matrix $R(t) \in \mathcal{L}(\mathbb{R}^n)$, the so-called *border projector*, such that

$$\operatorname{im} R(t) = \operatorname{im} D(t),\ \ker R(t) = \ker f_y(y,x,t),\ (y,x,t) \in \mathbb{R}^n \times \mathcal{D}_f \times \mathcal{I}_f. \tag{6}$$

Since both subspaces $\operatorname{im} D$ and $\ker f_y$ are $\mathcal{C}^1$-subspaces, the border projector function $R : \mathcal{I}_f \to \mathcal{L}(\mathbb{R}^n)$ is continuously differentiable, see [83, Lemma A.20].
Note that, if the subspace $\ker f_y(y,x,t)$ actually depends on y. then one can slightly modify the DAE by letting $\tilde{f}(y,x,t) := f(D(t)D(t)^+ y,x,t)$ such that $\ker \tilde{f}_y(y,x,t) = (\operatorname{im} D(t))^\perp$ depends on t only.

Since $D(t)$ has constant rank r, we may choose a continuous projector valued function $P_0 \in \mathcal{C}(\mathcal{I}_f, L(\mathbb{R}^m))$ such that

$$\ker P_0(t) = \ker D(t) = \ker f_y(y,x,t)D(t)$$

for all possible arguments. Denote the complementary projector function by Q_0,

$$Q_0(t) := I - P_0(t).$$

Additionally, the four conditions

$$\begin{aligned} D(t)D(t)^- D(t) &= D(t), \\ D(t)^- D(t)D(t)^- &= D(t)^-, \\ D(t)D(t)^- &= R(t), \\ D(t)^- D(t) &= P_0(t), \end{aligned}$$

determine the pointwise generalized inverse $D(t)^-$ of $D(t)$ uniquely, and the matrix function $D^-(t) := D(t)^-$ depends continuously on its argument, see [83, Proposition A.17].

A considerable part of the relevant literature (e.g. [51, 106]) restricts the interest to *semi-explicit DAEs* consisting of $m = m_1 + m_2$ equations,

$$\begin{aligned} x_1'(t) + k_1(x_1(t), x_2(t), t) &= 0, \\ k_2(x_1(t), x_2(t), t) &= 0, \end{aligned} \tag{7}$$

with $n = m_1$,

$$f(y,x,t) = \begin{bmatrix} y + k_1(x,t) \\ k_2(x,t) \end{bmatrix}, \; D(t) = \begin{bmatrix} I & 0 \end{bmatrix}, \; P_0(t) = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \; D(t)^- = \begin{bmatrix} I \\ 0 \end{bmatrix}, \; R = I.$$

Notice that special semi-explicit DAEs play their role in multibody dynamics [45]. The semi-explicit form confirms the clear significance of our solution notion. Here we seek continuous functions x having a continuously differentiable component x_1. We emphasize that there is no natural reason for requiring x_2 also to be differentiable.

Well-posedness in the sense of Hadamard in appropriate settings constitutes the classical basis of a safe numerical treatment. In view of the numerical treatment, as for most nonlinear problems, we suppose that there exist a solution to be practically approximated and we agree upon a local variant of well-posedness.

Definition 2.2. Let $x_* \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ be a solution of the BVP (3), (4), $\mathcal{I} = [a,b]$. The BVP (3), (4) is said to be *well-posed locally* around x_* in its natural setting, if the slightly perturbed BVP

$$f((Dx)'(t), x(t), t) = q(t), \quad t \in \mathcal{I}, \tag{8}$$

$$g(x(a), x(b)) = \gamma, \tag{9}$$

is locally uniquely solvable for each arbitrary sufficiently small perturbations $q \in \mathcal{C}(\mathcal{I}, \mathbb{R}^m)$ and $\gamma \in \mathbb{R}^l$, and the solution x satisfies the inequality

$$\|x - x_*\|_{\mathcal{C}_D^1} \le \kappa(|\gamma| + \|q\|_\infty), \tag{10}$$

with a constant κ. Otherwise the BVP is said to be *ill-posed* in the natural setting.

Instead of the inequality (10) one can use the somewhat simpler inequality

$$\|x - x_*\|_\infty \le \kappa(|\gamma| + \|q\|_\infty), \tag{11}$$

which is sometimes more convenient, see Remark 2.12.
The constant κ in the inequality (11) is called the *stability constant* of the BVP, e.g., in [12, 9, 51]. Here we do not share in this notation.

Representing the linear BVP

$$A(t)(Dx)'(t) + B(t)x(t) = q(t), \; t \in \mathcal{I}, \quad G_a x(a) + G_b x(b) = \gamma,$$

as operator equation $\mathcal{T}x = (q, \gamma)$ by the linear bounded operators

$$Tx := A(Dx)' + Bx, \quad \mathcal{T}x := (Tx, G_a x(a) + G_b x(b)), \quad x \in \mathcal{C}^1_D(\mathcal{I}, \mathbb{R}^m),$$
$$T \in \mathcal{L}(\mathcal{C}^1_D(\mathcal{I}, \mathbb{R}^m), \mathcal{C}(\mathcal{I}, \mathbb{R}^m)), \quad \mathcal{T} \in \mathcal{L}(\mathcal{C}^1_D(\mathcal{I}, \mathbb{R}^m), \mathcal{C}(\mathcal{I}, \mathbb{R}^m) \times \mathbb{R}^l),$$

it becomes evident that the linear BVP is well-posed, if and only if $\mathcal{T}$ is bijective, and then κ in (10) is nothing else an upper bound of $\|\mathcal{T}^{-1}\|$.

The next notion is concerned with the boundary conditions only. It is, of course, important to apply exactly the right number of conditions, neither to under specify nor to over specify. As we will see later, this task is essentially more difficult to realize for DAEs than for explicit ODEs. Also stating initial conditions accurately is a challenging task for DAEs quite unlike the case of explicit ODEs.

Definition 2.3. Let $x_* \in \mathcal{C}^1_D(\mathcal{I}, \mathbb{R}^m)$ be a solution of the BVP (3), (4), $\mathcal{I} = [a,b]$. The BVP (3), (4) has *accurately stated boundary conditions* locally around x_* if the BVP with slightly perturbed boundary conditions

$$f((Dx)'(t), x(t), t) = 0, \tag{12}$$
$$g(x(a), x(b)) = \gamma, \tag{13}$$

is uniquely solvable for each arbitrary sufficiently small $\gamma \in \mathbb{R}^l$, and the solution satisfies the inequality

$$\max_{t \in \mathcal{I}} |x(t) - x_*(t)| \leq \kappa |\gamma|, \tag{14}$$

with a constant κ.

It is evident, that x_* is locally the only solution of a BVP with accurately stated boundary conditions. In contrary, local uniqueness does not necessarily require accurately stated boundary conditions, see Example 2.1 below.

Even though, for explicit ODEs, a BVP is well-posed, exactly if its boundary conditions are accurately stated (cf.[12]), the situation is different for DAEs. Here, well-posedness implies accurately stated boundary conditions, too. However, the opposite is not true as the following example shows.

Example 2.1. Consider several BVPs (actually IVPs) for the DAE

$$\begin{aligned} x_1'(t) + x_3(t) &= 0, \\ x_2'(t) + x_3(t) &= 0, \\ x_2(t) - \sin(t-a) &= 0, \end{aligned} \tag{15}$$

and the different sets of boundary conditions

$$x_1(a) = 0,\ x_2(a) = 0,\ x_3(a) = 0, \tag{16}$$
$$x_1(a) = 0,\ x_2(a) = 0, \tag{17}$$
$$x_1(a) + \alpha x_2(a) + \beta x_3(a) = 0, \quad \alpha, \beta \in \mathbb{R}, \tag{18}$$
$$x_2(a) = 0. \tag{19}$$

The DAE possesses the general solution

$$x(t) = \begin{bmatrix} c + \sin(t-a) \\ \sin(t-a) \\ -\cos(t-a) \end{bmatrix}, \quad t \in \mathcal{I},$$

with an arbitrary constant $c \in \mathbb{R}$.

Obviously, the BVP (15), (16) fails to be solvable, and the BVP (15), (19) is satisfied by all solutions with arbitrary c.

The BVP (15), (17) and the BVP (15), (18) are both uniquely solvable, and their solutions x_* are given by $c = 0$ and $c = \beta$, respectively. However, inspecting the corresponding BVPs with perturbed boundary conditions, we learn that only the BVP (15), (18) has accurately stated boundary conditions.

To check whether the BVP (15), (18) is also well-posed we consider the fully perturbed BVP. This BVP possesses a unique solution for each $\gamma \in \mathbb{R}$ and each continuous function q having a continuously differentiable component q_3, but not for all continuous q. The solution reads

$$x(t) = \begin{bmatrix} \gamma + \sin(t-a) + q_3(t) - q_3(a) + \int_a^t (q_1(s) - q_2(s))\mathrm{d}s \\ \sin(t-a) + q_3(t) \\ q_2(t) - q_3'(t) - \cos(t-a) \end{bmatrix}, \quad t \in \mathcal{I}.$$

The difference

$$x(t) - x_*(t) = \begin{bmatrix} \gamma + q_3(t) - q_3(a) + \int_a^t (q_1(s) - q_2(s))\mathrm{d}s \\ q_3(t) \\ q_2(t) - q_3'(t) \end{bmatrix}, \quad t \in \mathcal{I},$$

can not be estimated by an inequality (10). The BVP is ill-posed in its natural setting. □

Besides the original BVP (3), (4) we consider also the DAE linearized along the reference solution x_*,

$$A_*(t)(Dx)'(t) + B_*(t)x(t) = 0, \quad t \in \mathcal{I}, \tag{20}$$

with continuous coefficients

$$\begin{aligned} A_*(t) &:= f_y((Dx_*)'(t), x_*(t), t), \\ B_*(t) &:= f_x((Dx_*)'(t), x_*(t), t), \quad t \in \mathcal{I}, \end{aligned}$$

and the linearized boundary conditions

$$G_{*a}x(a) + G_{*b}x(b) = 0, \tag{21}$$

where

$$G_{*a} := \frac{\partial g}{\partial x_a}(x_*(a), x_*(b)), \quad G_{*b} := \frac{\partial g}{\partial x_b}(x_*(a), x_*(b)).$$

The linear DAE (20) inherits the properly stated leading term from the original DAE (3). The linearized BVP (20), (21) is said to be the *variational problem* for the original BVP (3), (4) at x_* (e.g., [12, p. 90]).

Next we tie in with the notions *locally unique solution* and *isolated solution* commonly used in the context of BVPs for explicit ODEs (cf. [12]).

Definition 2.4. A solution $x_* \in \mathcal{C}^1_D(\mathcal{I}, \mathbb{R}^m)$ of the BVP (3), (4) is said to be *locally unique* if there is a "tube" around it where it is unique, i.e., there is a $\rho > 0$ such that in the class of functions

$$\left\{ x \in \mathcal{C}^1_D(\mathcal{I}, \mathbb{R}^m) : \|x - x_*\|_\infty \le \rho \right\} =: \mathcal{B}_\mathcal{C}(x_*, \rho)$$

x_* is the only solution of the BVP.

This notion is consistent with the general meaning that a solution $x_* \in \mathcal{C}^1_D(\mathcal{I}, \mathbb{R}^m)$ is locally unique if it has a neighborhood in $\mathcal{C}^1_D(\mathcal{I}, \mathbb{R}^m)$ with no further solution. Namely, if there are no further solution in $\mathcal{B}_\mathcal{C}(x_*, \rho)$, then a fortiori x_* is the only solution in the ball

$$\left\{ x \in \mathcal{C}^1_D(\mathcal{I}, \mathbb{R}^m) : \|x - x_*\|_{\mathcal{C}^1_D} \le \rho \right\} =: \mathcal{B}_{\mathcal{C}^1_D}(x_*, \rho) \subset \mathcal{B}_\mathcal{C}(x_*, \rho).$$

Conversely, assume that there is no such $\rho > 0$ as required in Definition 2.4. Then there is a sequence of solutions $x_i \in \mathcal{C}^1_D(\mathcal{I}, \mathbb{R}^m)$ of the BVP such that $\|x_i - x_*\|_\infty \xrightarrow{i\to\infty} 0$. Applying the arguments from Remark 2.12 we obtain the inequality $\|(Dx_i - Dx_*)'\|_\infty \le k_1 \|x_i - x_*\|_\infty$; and hence $\|x_i - x_*\|_{\mathcal{C}^1_D} \xrightarrow{i\to\infty} 0$. Then x_* has no neighborhood in $\mathcal{C}^1_D(\mathcal{I}, \mathbb{R}^m)$ with no further solution.

Definition 2.5. A solution $x_* \in \mathcal{C}^1_D(\mathcal{I}, \mathbb{R}^m)$ of the BVP (3), (4) is said to be *isolated* if the variational problem (20), (21) has the unique solution $x = 0$.

In the case of explicit ODEs, an isolated solution x_* of a BVP is locally unique, the BVP is well-posed if and only if the boundary conditions are accurately stated. The notion of isolatedness can be seen as practical tool to check local uniqueness and well-posedness. An explicit ODE of dimension m has m degrees of freedom, and it is beyond dispute to formulate $l = m$ boundary conditions. If then the variational problem has only the zero solution, then the boundary conditions are stated accurately, thus the BVP is locally well-posed.

A similar situation is given for regular index-1 DAEs, with $l = r = \operatorname{rank} D(t)$, e.g., [90, 55, 111, 96], cf. also Subsection 2.5 below, and for certain singular index-1 DAEs ([43]).

In general, for DAEs, it is no longer plain to secure the right number l of boundary conditions. It is further an open question to what extent the notion *isolatedly solvable* is justified in a similar sense. We refer to Remark 2.7 for further details.

2.2 The flow structure of regular linear DAEs

Each linear DAE

$$A(t)(Dx)'(t) + B(t)x(t) = q(t), \quad t \in \mathcal{I}, \tag{22}$$

which is regular with arbitrary tractability index $\mu \in \mathbb{N}$ in the sense of [83, Definition 2.25] (cf. Definition 6.2 below) and has sufficiently smooth (at least continuous) coefficients, can be decoupled into its two structurally characteristic parts, namely the *inherent explicit regular ODE* (IERODE) and the *algebraic part housing all differentiations*, by means of certain smartly constructed continuous projector valued functions beginning with P_0. If $P_0, \ldots, P_{\mu-1} \in \mathcal{C}(\mathcal{I}, \mathcal{L}(\mathbb{R}^m))$ are those *fine decoupling projector functions* for the DAE (3), then the products

$$\Pi_{can} := (I - \mathcal{H}_0)\Pi_{\mu-1}, \quad \Pi_{\mu-1} := P_0 \cdots P_{\mu-1}, \quad D\Pi_{can}D^- = D\Pi_{\mu-1}D^-, \tag{23}$$

with a coefficient $\mathcal{H}_0$ described in terms of the coefficients A, D, B in Appendix 6.1.2, are also projector valued functions. In particular, Π_{can} has a special meaning independent of the choice of the corresponding factors (e.g., [83, Section 2.4]). Namely, for every $t \in \mathcal{I}$ it holds that

$$\begin{aligned} \operatorname{im} \Pi_{can}(t) &= \{x(t) \in \mathbb{R}^m : x \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m), \quad A(Dx)' + Bx = 0\}, \\ \ker \Pi_{can}(t) &= \ker \Pi_{\mu-1}(t) = \ker P_0(t) + \cdots + \ker P_{\mu-1}(t). \end{aligned}$$

Both subspaces $\operatorname{im} \Pi_{can}(t)$ and $\ker \Pi_{can}(t)$ are independent of the choice of the admissible projector functions $P_0, \ldots, P_{\mu-1}$ (e.g., [83, Chapter 2]). The subspace $\operatorname{im} \Pi_{can}(t)$ represents the *linear space of all consistent values* at time t of the homogeneous DAE. On the other hand, $\ker \Pi_{can}(t)$ is such that

$$x \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m), \quad A(Dx)' + Bx = 0, \quad \text{and} \quad x(t) \in \ker \Pi_{can}(t)$$

imply x to vanish identically.

The projector function Π_{can} is said to be the *canonical projector function* associated with the DAE (22). Π_{can} has constant rank; denote

$$l := \operatorname{rank} \Pi_{can}(t) = \operatorname{rank} \Pi_{\mu-1}(t), \quad t \in \mathcal{I}. \tag{24}$$

The rank l can be computed by means of the matrix function sequence supporting the regularity notion, see [83, Section 7.4], also Definitions 6.1, 6.2 below.

In the simpler case of constant coefficients A, D, B, the projector matrix Π_{can} takes the role of the *spectral projector* of the matrix pair $\{AD, B\}$ ([83, Section 1.4]).

The canonical projector function depends strongly on the index. In particular, the canonical projector function of a regular index-1 DAE (22) is given by the subspaces

$$\begin{aligned}\operatorname{im}\Pi_{can}(t) = S_0(t) &:= \{z \in \mathbb{R}^m : B(t)z \in \operatorname{im}A(t) = \operatorname{im}A(t)D(t)\},\\ \ker\Pi_{can}(t) = N_0(t) &:= \ker D(t) = \ker A(t)D(t),\end{aligned}$$

but for all regular higher-index DAEs the intersection $N_0(t) \cap S_0(t)$ is no longer a trivial one.

The following example describes the canonical projector function of semi-explicit index-1 DAEs in more detail.

Example 2.2. We have

$$A(t) = \begin{bmatrix} I \\ 0 \end{bmatrix},\ D(t) = \begin{bmatrix} I\ 0 \end{bmatrix},\ B(t) = \begin{bmatrix} B_{11}(t)\ B_{12}(t) \\ B_{21}(t)\ B_{22}(t) \end{bmatrix},$$

with $B_{22}(t)$ remaining nonsingular,

$$\begin{aligned}\operatorname{im}\Pi_{can}(t) = S_0(t) &:= \{z \in \mathbb{R}^m : B_{21}(t)z_1 + B_{22}(t)z_2 = 0\},\\ \ker\Pi_{can}(t) = N_0(t) &:= \{z \in \mathbb{R}^m : z_1 = 0\},\end{aligned}$$

and hence

$$\Pi_{can}(t) = \begin{bmatrix} I & 0 \\ -B_{22}(t)^{-1}B_{21}(t)\ 0 \end{bmatrix}.$$

We observe that $\Pi_{can}(t)$ often is far from being symmetric, and the subspaces are far from being orthogonal, and $|\Pi_{can}(t)|_2$ can become large. In the particular instance $m_1 = m_2$, $B_{21}(t) = I$, $B_{22}(t) = \alpha I$, $\alpha > 0$ small, if $\alpha > 0$ tends to zero, the angle between the subspaces $N_0(t)$ and $S_0(t)$ becomes more and more acute, and $|\Pi_{can}(t)|_2$ becomes larger and larger. □

In the following, we suppose the DAE (22) to be regular with index $\mu \in \mathbb{N}$. We omit the less interesting case $\mu = 0$.
For arbitrary fixed $\bar{t} \in \mathcal{I}$, there is a unique matrix function $X(\cdot, \bar{t})$ satisfying the IVP ([83, Section 2.6])

$$A(t)(DX)'(t) + B(t)X(t) = 0, \quad t \in \mathcal{I}, \quad X(\bar{t}) = \Pi_{can}(\bar{t}). \tag{25}$$

The columns of $X(\cdot, \bar{t})$ are functions from $\mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$. $X(t, \bar{t})$ is named *maximal-size fundamental solution matrix normalized at* $\bar{t}$. It can be determined also by the IVP

$$A(t)(DX)'(t) + B(t)X(t) = 0, \quad t \in \mathcal{I}, \quad \Pi_{\mu-1}(\bar{t})(X(\bar{t}) - I) = 0, \tag{26}$$

with initial conditions built by arbitrary admissible projector functions. This is considerably easier to realize in practice than providing the canonical projector $\Pi_{can}(\bar{t})$ and fine decoupling projectors (cf. [83]).

For DAEs, different kind of fundamental solution matrices make sense, in particular so-called *maximal size* and *minimal size* ones (cf. [28, 29, 83]). The minimal size fundamental solution is rectangular with full column-rank l, the maximal size (shortly: maximal) fundamental solution has m columns. The great advantage of the

latter consists in useful group properties to describe the flow ([83, Section 2.6], also Remark 2.4).

In contrast to regular ODEs with always nonsingular fundamental solution matrices, any fundamental solution matrix of a regular DAE fails to be nonsingular.

We have (e.g., [83, Section 2.6])

$$\operatorname{im} X(t,\bar{t}) = \operatorname{im} \Pi_{can}(t), \quad \ker X(t,\bar{t}) = \ker \Pi_{can}(\bar{t}), \quad \operatorname{rank} X(t,\bar{t}) = l. \tag{27}$$

In the particular case of a regular constant coefficient DAE in Weierstraß-Kronecker form

$$\begin{bmatrix} I_l & 0 \\ 0 & \mathcal{N} \end{bmatrix} x' + \begin{bmatrix} W & 0 \\ 0 & I_{m-l} \end{bmatrix} x = q, \tag{28}$$

with a nilpotent matrix N, it simply results that

$$\Pi_{can} = \begin{bmatrix} I_l & 0 \\ 0 & 0 \end{bmatrix}, \quad X(t,\bar{t}) = \begin{bmatrix} \mathrm{e}^{-(t-\bar{t})W} & 0 \\ 0 & 0 \end{bmatrix}.$$

In the general case, the (maximal) fundamental solution matrix $X(t,\bar{t})$ can be described by

$$X(t,\bar{t}) = \Pi_{can}(t) D(t)^{-} U(t,\bar{t}) D(\bar{t}) \Pi_{can}(\bar{t}), \tag{29}$$

whereby $U(t,\bar{t})$ denotes the classical nonsingular fundamental solution matrix of the IERODE

$$u' - (D\Pi_{\mu-1}D^{-})'u + D\Pi_{\mu-1}G_{\mu}^{-1}B_{\mu}D^{-}u = D\Pi_{\mu-1}G_{\mu}^{-1}q \tag{30}$$

normalized by the condition $U(\bar{t},\bar{t}) = I$. Recall that the matrix functions G_μ and B_μ are built from the DAE coefficients A, D, B and G_μ is nonsingular (cf. Definitions 6.1, 6.2 below).

The generalized inverse $X(t,\bar{t})^{-}$ of $X(t,\bar{t})$ being determined by the four relations

$$XX^{-}X = X, \quad X^{-}XX^{-} = X^{-}, \quad XX^{-} = \Pi_{can}(t), \quad X^{-}X = \Pi_{can}(\bar{t}),$$

shows the structure

$$X(t,\bar{t})^{-} = \Pi_{can}(\bar{t}) D(\bar{t})^{-} U(t,\bar{t})^{-1} D(t) \Pi_{can}(t).$$

For all $t_1, t_2, t_3 \in \mathcal{I}$ we have that

$$X(t_1,t_2)^{-} = X(t_2,t_1), \quad X(t_1,t_2)X(t_2,t_3) = X(t_1,t_3). \tag{31}$$

The general solution of the DAE (22), with admissible right-hand side q, can now be expressed as

$$x(t) = X(t,\bar{t})c + x_q(t), \quad t \in \mathcal{I}, \tag{32}$$

whereby $x_q \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ is the unique solution of the IVP ([83, Theorem 2.52])

$$A(Dx)' + Bx = q, \quad \Pi_{can}(\bar{t})x(\bar{t}) = 0, \tag{33}$$

and $c \in \mathbb{R}^m$ is a free constant. It follows that

$$x(t) = X(t,\bar{t})c + x_q(t) = X(t,\bar{t})\Pi_{can}(\bar{t})c + x_q(t), \quad t \in \mathcal{I},$$
$$x(\bar{t}) = X(\bar{t},\bar{t})c + x_q(\bar{t}) = \Pi_{can}(\bar{t})c + x_q(\bar{t}).$$

Obviously, only the component $\Pi_{can}(\bar{t})c$ serves as effective integration constant. The complementary component $(I - \Pi_{can}(\bar{t}))c$ has no impact on the solution. The dynamical degree of freedom results as $l = \operatorname{rank}\Pi_{can}(\bar{t})$.

Take a closer look at the solution x_q of the IVP (33), which has a quite involved structure. Let q be admissible and the function u_q be the classical solution of the explicit ODE (30) that satisfies the initial condition $u(\bar{t}) = 0$. By means of fine decoupling projector functions we obtain the coefficients applied below ([83, Section 2.4], also Appendix 6.1.2) from the given coefficients A, D, B and then we determine consecutively

$$\begin{aligned}
v_{\mu-1} &= \mathcal{L}_{\mu-1}q,\\
v_{\mu-2} &= \mathcal{L}_{\mu-2}q - \mathcal{N}_{\mu-2,\mu-1}(Dv_{\mu-1})',\\
&\cdots\\
v_1 &= \mathcal{L}_1 q - \sum_{l=2}^{\mu-1}\mathcal{N}_{1,l}(Dv_l)' - \sum_{l=3}^{\mu-1}\mathcal{M}_{1,l}v_l,\\
v_0 &= \mathcal{L}_0 q - \sum_{l=1}^{\mu-1}\mathcal{N}_{0,l}(Dv_l)' - \sum_{l=2}^{\mu-1}\mathcal{M}_{0,l}v_l - \mathcal{H}_0 D^- u_q.
\end{aligned}$$

Introduce further

$$\tilde{v}_0 = v_0 + \mathcal{H}_0 D^- u_q.$$

We have $v_0 = \tilde{v}_0$ in case of completely decoupling projector functions. We emphasize that for obtaining $v_{\mu-2}$ one has to differentiate the term $Dv_{\mu-1} = D\mathcal{L}_{\mu-1}q$ and so on. That means, an admissible right-hand side q is basically continuous, possibly with certain additional smoothness properties. We refer to [83, Section 2.4] for a detailed description.

Inspecting the decoupling procedure (Appendix 6.1.2) we find that $\Pi_{can}v_i = 0$ for $i = 0, \ldots, \mu-1$. We introduce the additional function

$$v_q := \tilde{v}_0 + v_1 + \cdots + v_{\mu-1}. \tag{34}$$

Regarding the identity $D\Pi_{can}D^- u_q = u_q$ we then obtain the relations

$$\begin{aligned}
x_q &= D^- u_q + v_q - \mathcal{H}_0 D^- u_q = (I - \mathcal{H}_0)D^- u_q + v_q = \Pi_{can}D^- u_q + v_q,\\
(I - \Pi_{can})x_q &= v_q,\\
D\Pi_{can}x_q &= D\Pi_{can}D^- u_q = u_q, \quad \Pi_{can}x_q = \Pi_{can}D^- u_q = D^- u_q.
\end{aligned}$$

The solution component $(I-\Pi_{can})x_q$ is fully fixed by the part $(I-\Pi_{can})G_\mu^{-1}q$ of the right-hand side q. Furthermore, we derive the useful representations

$$\begin{aligned}\Pi_{can}(t)x_q(t) &= \Pi_{can}(t)D(t)^- \int_{\bar{t}}^t U(t,s)D(s)\Pi_{can}(s)G_\mu^{-1}(s)q(s)\mathrm{d}s \\ &= \int_{\bar{t}}^t X(t,s)G_\mu^{-1}(s)q(s)\mathrm{d}s\end{aligned}$$

and

$$x_q(t) = \int_{\bar{t}}^t X(t,s)G_\mu^{-1}(s)q(s)\mathrm{d}s + v_q(t), \quad t \in \mathcal{I}.$$

In summary, the general solution of the DAE (22) reads

$$x(t) = X(t,\bar{t})c + \int_{\bar{t}}^t X(t,s)G_\mu^{-1}(s)q(s)\mathrm{d}s + v_q(t), \quad t \in \mathcal{I}, \tag{35}$$

and the consistent values at $\bar{t}$ have the form

$$x(\bar{t}) = \Pi_{can}(\bar{t})c + v_q(\bar{t}). \tag{36}$$

Comparing with the general solution of an explicit ODE, the first and second terms of the general DAE solution (35) have counterparts, however, in the DAE solution there emerges the additional new term v_q.

For each fixed right-hand side q, and thus fixed v_q, the flow of the regular DAE (22) is restricted to the time-varying affine subspace

$$\begin{aligned}\mathcal{M}_{\mu-1}(t) &= \{x + v_q(t) : x \in \operatorname{im}\Pi_{can}(t)\} \\ &= \{\Pi_{can}(t)c + x_q(t) : c \in \mathbb{R}^m\},\end{aligned}$$

which precisely consists of all consistent values at time t.

We recall that, in all higher index cases, for obtaining v_q one has to carry out certain differentiations of parts of q. Therefore, an admissible right-hand side q has to be smooth enough. Solely for index-1 DAEs, the space of admissible functions coincides with the continuous function space $\mathcal{C}(\mathcal{I},\mathbb{R}^m)$. For all higher-index DAEs, the spaces of admissible functions $\mathcal{C}^{\mathrm{ind}\,\mu}(\mathcal{I},\mathbb{R}^m)$ represent proper nonclosed subsets of the continuous function space, [83, 96], also Appendix 6.1.4. This fact constitutes the *ambivalent character of the solutions of higher-index DAEs*: they are as smooth as expected coming from explicit ODEs with respect to the integration constant $\Pi_{can}(\bar{t})c$, but, in strict contrast to the ODE case, they behave discontinuously concerning the right-hand side.

We refer to [83, Example 1.5] and its functional-analytic interpretation in [96] for a deeper insight. The discontinuity concerning the right-hand side causes well-known difficulties in numerical integration procedures.

We take a closer look to the special cases of index-1 and index-2 DAEs (22) (cf. [83, pp. 104-107] for the specification for semi-explicit systems).

Index-1 DAE: Let (22) be regular with tractability index 1.
Form $G_0 := AD$, $r_0 = \operatorname{rank} G_0 = \operatorname{rank} D < m$, $\Pi_0 = P_0$ and $G_1 := G_0 + BQ_0$. G_1 remains nonsingular. The DAE decoupling reads

$$\begin{aligned}(Dx)' - R'Dx + DG_1^{-1}BD^-Dx &= DG_1^{-1}q,\\ Q_0x + Q_0G_1^{-1}BD^-Dx &= Q_0G_1^{-1}q,\\ v_q = \tilde{v}_0 &= Q_0G_1^{-1}q,\\ x &= (I-\mathcal{H}_0)D^-Dx + Q_0G_1^{-1}q.\end{aligned}$$

We have here $u = Dx$, further $\mathcal{H}_0 = Q_0G_1^{-1}BP_0$. The dynamical degree of freedom is $l = r_0$. The canonical projector $\Pi_{can}(t) = (I-\mathcal{H}_0(t))\Pi_0(t)$ is actually the projector onto

$$S_0(t) := \{z \in \mathbb{R}^m : B(t)z \in \operatorname{im} G_0(t)\} \quad \text{along} \quad \ker G_0(t).$$

The DAE is solvable for each arbitrary $q \in \mathcal{C}(\mathcal{I},\mathbb{R}^m)$.

Index-2 DAE: Let (22) be regular with tractability index 2.
Form $G_0 := AD$, $r_0 = \operatorname{rank} G_0 = \operatorname{rank} D < m$, $\Pi_0 = P_0$, $G_1 := G_0 + BQ_0$, $r_1 = \operatorname{rank} G_1 < m$. Owing to the index-2 property the decomposition $\mathbb{R}^m = S_1(t) \oplus \ker G_1(t)$ is valid, with

$$S_1(t) := \{z \in \mathbb{R}^m : B_1(t)z \in \operatorname{im} G_1(t)\}.$$

We choose $P_1(t)$ to be the projector onto $S_1(t)$ along $\ker G_1(t)$. Then we form $\Pi_1 = P_0P_1$, $B_1 := BP_0 - G_1D^-(D\Pi_1D^-)'D\Pi_0$, and $G_2 := G_1 + B_1Q_1$. G_2 remains nonsingular. The DAE decoupling results in

$$\begin{aligned}(D\Pi_1x)' - (D\Pi_1D^-)'D\Pi_1x + DG_2^{-1}B_1D^-D\Pi_1x &= D\Pi_1G_2^{-1}q,\\ v_1 &= \Pi_0Q_1G_2^{-1}q,\\ \tilde{v}_0 &= Q_0P_1G_2^{-1}q + Q_0Q_1D^-(D\Pi_0Q_1G_2^{-1}q)',\\ v_q &= \tilde{v}_0 + v_1,\\ x &= (I-\mathcal{H}_0)D^-D\Pi_1x + v_q.\end{aligned}$$

We have here $u = D\Pi_1x$. The dynamical degree of freedom is $l = r_0 + r_1 - m$. The coupling coefficient $\mathcal{H}_0$ is now more elaborated,

$$\mathcal{H}_0 = Q_0P_1G_2^{-1}B\Pi_1 + Q_0P_1D^-(D\Pi_1D^-)'D\Pi_1.$$

The DAE is solvable for precisely each arbitrary

$$q \in \{w \in \mathcal{C}(\mathcal{I},\mathbb{R}^m) : D\Pi_0Q_1G_2^{-1}w \in \mathcal{C}^1(\mathcal{I}\mathbb{R}^n)\} =: \mathcal{C}^{\text{ind }2}(\mathcal{I},\mathbb{R}^m),$$

which is a proper nonclosed subset in $\mathcal{C}(\mathcal{I},\mathbb{R}^m)$. We take a closer look at the size-2 Hessenberg DAE.

Example 2.3. For the Hessenberg size 2 system of $m_1 + m_2 = m$ equations, $m_2 \le m_1$,

$$x_1' + B_{11}x_1 + B_{12}x_2 = q_1,$$
$$B_{21}x_1 = q_2,$$

with nonsingular product $B_{21}B_{12}$, we obtain $r_0 = m_1, r_1 = m_1$, $l = m_1 - m_2$, and

$$\Pi_{can} = \begin{bmatrix} I - \Omega & 0 \\ B_{12}^-(B_{11} - \Omega')(I - \Omega) & 0 \end{bmatrix}, \quad \Omega = B_{12}B_{12}^-, \quad B_{12}^- := (B_{21}B_{12})^{-1}B_{21}.$$

Further the projectors

$$P_0 = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, P_1 = \begin{bmatrix} I - \Omega & 0 \\ B_{12}^- & I \end{bmatrix},$$

provide a fine decoupling, $D\Pi_{\mu-1}D^- = I - \Omega$, and

$$D\Pi_0 Q_1 G_2^{-1} = [0 \quad B_{12}(B_{21}B_{12})^{-1}].$$

The set of admissible right-hand sides is

$$\mathcal{C}^{ind\,2}(\mathcal{I}, \mathbb{R}^m) = \{q \in \mathcal{C}(\mathcal{I}, \mathbb{R}^m) : B_{12}(B_{21}B_{12})^{-1}q_2 \in \mathcal{C}^1(\mathcal{I}, \mathbb{R}^{m_2})\}.$$

2.3 Accurately stated two-point boundary conditions

This section provides solvability statements for the BVPs

$$A(Dx)' + Bx = q, \quad G_a x(a) + G_b x(b) = \gamma. \tag{37}$$

The DAE is supposed to be regular with $l := \operatorname{rank}\Pi_{can}(a) = \operatorname{rank}\Pi_{\mu-1}(a)$ on the compact interval $\mathcal{I} = [a, b]$. The right-hand side q is supposed to be admissible such that the DAE has a solution in $\mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ (cf. [83, Subsubsection 2.6.4]). The boundary condition is given by the matrices $G_a, G_b \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^l)$, which is in full accordance with the number of free integration constants as described in the previous section.

We follow the well-known classical lines to treat BVPs for ODEs (e.g., [12]). We apply the general solution expression (35) with $\bar{t} = a$,

$$x(t) = X(t,a)c + \int_a^t X(t,s)G_\mu^{-1}(s)q(s)\mathrm{d}s + v_q(t), \quad t \in \mathcal{I}. \tag{38}$$

and insert it into the boundary condition. This yields an equation system for c, namely

$$(G_a X(a,a) + G_b X(b,a))c = \hat{\gamma}, \tag{39}$$
$$\hat{\gamma} := \gamma - \gamma_q - G_b \int_a^b X(b,s) G_\mu(s)^{-1} q(s) \mathrm{d}s,$$
$$\gamma_q := G_a v_q(a) + G_b v_q(b).$$

Now it is evident that the so-called *solvability matrix*

$$S := G_a X(a,a) + G_b X(b,a) \tag{40}$$

actually plays the key role for solvability of the BVP. By construction, it holds that $\ker \Pi_{can}(a) \subseteq \ker S$. This fits to the fact that the components $(I - \Pi_{can}(a))c$ do not at all matter for the DAE solutions. The boundary condition must precisely fix the component $\Pi_{can}(a)c$. Consequently, we have to request that $\ker S = \ker \Pi_{can}(a)$. If this is given, then S has full row-rank l. Then we introduce the generalized inverse S^- of S by

$$SS^-S = S, \quad S^-SS^- = S^-, \quad SS^- = I, \quad S^-S = \Pi_{can}(a), \tag{41}$$

and further the so-called *Green's matrix function* of the BVP

$$\mathcal{G}(t,s) := \begin{cases} X(t,a)S^- G_a X(a,a) X(s,a)^-, & \text{if } t \geq s \\ -X(t,a)S^- G_b X(b,a) X(s,a)^-, & \text{if } t < s. \end{cases} \tag{42}$$

After the idea of conditioning constants for classical BVPs (e.g., [12]) we denote

$$\kappa_1 := \max_{t \in \mathcal{I}} |X(t,a)S^-|, \quad \kappa_2 := \sup_{s,t \in \mathcal{I}} |\mathcal{G}(t,s)|, \quad \kappa_3 := \max_{t \in \mathcal{I}} |\Pi_{can}(t) G_\mu(t)^{-1}|.$$

As in the classical ODE case, the expressions $X(t,a)S^-$ and $\mathcal{G}(t,s)$ do not change if one uses an arbitrary $\bar{t} \in \mathcal{I}$ instead of $\bar{t} = a$. The first two quantities κ_1 and κ_2 are counterparts of the classical conditioning constants for ordinary BVPs. The extra quantity κ_3 is independent of the boundary condition; for an explicit ODE we would have $\Pi_{can}(t) G_\mu(t)^{-1} \equiv I$, thus $\kappa_3 = 1$.
In general, the expression $\Pi_{can}(t) G_\mu(t)^{-1}$ represents a generalized inverse of $G_\mu(t)\Pi_{can}(t) = G_0(t)\Pi_{can}(t)$.
Inspecting the regularity notion one observes that scaling a given regular DAE by $G_\mu(t)^{-1}$ and using the same admissible projector functions for the scaled DAE again, leads to $G_\mu(t) \equiv I$ and $\kappa_3 := \max_{t \in \mathcal{I}} |\Pi_{can}(t)|$ for the scaled version. As pointed out in Subsection 2.2, the canonical projector function Π_{can} is an essential inherent feature of the DAE.

Theorem 2.1. *Let the DAE in (37) be regular with index $\mu \in \mathbb{N}$ on the interval $\mathcal{I} = [a,b]$ and $l = \operatorname{rank} \Pi_{can}(a)$. Π_{can} is the canonical projector function of the DAE. Given are the matrices $G_a, G_b \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^l)$. Then the following assertions are true:*

(1) *The BVP (37) is uniquely solvable for each arbitrary $\gamma \in \mathbb{R}^l$ and each arbitrary admissible right-hand side q, if and only if the conditions*

$$\operatorname{im}[G_a\, G_b] = \mathbb{R}^l \quad \textit{and} \quad \ker S = \ker \Pi_{can}(a) \tag{43}$$

are valid.

(2) *If (43) is satisfied, then the BVP solution can be represented as*

$$x(t) = X(t,a)S^{-}(\gamma - \gamma_q) + \int_a^b \mathcal{G}(t,s)G_\mu(s)^{-1}q(s)\mathrm{d}s + v_q(t),$$

by means of the fundamental solution matrix normalized at $\bar{t} = a$ (25), the solvability matrix (40), Green's matrix function (42), the function v_q defined by (34), γ_q given in (39), and the matrix function G_μ constructed via Definition 6.1.

(3) *If (43) is satisfied, then the BVP solution can be estimated by*

$$\max_{t\in\mathcal{I}} |x(t)| \le \kappa_1 |\gamma - \gamma_q| + \kappa_2\kappa_3 \max_{t\in\mathcal{I}} |q(t)| + \max_{t\in\mathcal{I}} |v_q(t)|.$$

(4) *If (43) is satisfied, then the BVP (37) has accurately stated boundary condition in the sense of Definition 2.3.*

(5) *Let (43) be satisfied. Then the BVP (37) is well-posed in its natural setting, if and only if $\mu = 1$, and ill-posed otherwise.*

We mention that the first condition in (43) is a consequence of the second one, since $\ker S = \ker \Pi_{can}(a)$ implies $\operatorname{rank} S = l$ thus $\mathbb{R}^l = \operatorname{im} S \subseteq \operatorname{im}[G_a\, G_b] \subseteq \mathbb{R}^l$. Here we explicitly indicate that condition because of its practical meaning.

Proof. Let $\gamma \in \mathbb{R}^l$ be given, q be admissible, and $\hat{\gamma} := \gamma - G_a v_q(a) - G_b v_q(b) - G_b \int_a^b X(b,s)G_\mu(s)^{-1}q(s)\mathrm{d}s$. Owing to condition (43), the equation $Sc = \hat{\gamma}$ yields $\Pi_{can}(a)c = S^{-}\hat{\gamma}$, hence a solution of the BVP. The BVP solution is unique, since the homogenous BVP has the zero solution only.

Conversely, if all BVPs are uniquely solvable, then S must have full rank, and $\ker S = \ker \Pi_{can}(a)$ must be valid for reasons of dimensions. The first assertion is verified.

The assertions (2), (3), and (4) can be proved by straightforward standard calculations.

By Definition 2.2, well-posedness necessarily requires the inequality $\|v_q\|_\infty \le k\|q\|_\infty$, but that is valid exactly for the case $\mu = 1$, with $v_q = \mathcal{L}_0 q$. This proves assertion (5). □

In particular, condition (43) serves as a criterion indicating whether the initial conditions are stated accurately.

Corollary 2.1. *Let the DAE be regular, $l = \operatorname{rank} \Pi_{can}(a)$ and $C \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^l)$. Then the IVP*

$$A(Dx)' + BX = q, \quad Cx(a) = \gamma \tag{44}$$

is uniquely solvable for each arbitrary $\gamma \in \mathbb{R}^l$ and each arbitrary admissible right-hand side q, if and only if $\ker C \cap \operatorname{im} \Pi_{can}(a) = \{0\}$ is valid.

Proof. This is a special BVP with solvability matrix $S = CX(a,a) = C\Pi_{can}(a)$. □

The most natural way to state initial conditions is letting $\ker C = \ker \Pi_{can}(a)$ which directly implies $\ker C \cap \operatorname{im} \Pi_{can}(a) = \{0\}$. By this, the initial condition is immediately directed to the IERODE.

In contrast, for practical reasons one can be interested in prescribing other components. Then one has to take into account that the condition $\ker C \cap \operatorname{im} \Pi_{can}(a) = \{0\}$ possibly requires additional regularity conditions concerning the DAE as in the following example.

Example 2.4. Consider the semi-explicit system with $m_1 + m_2 = m$ equations

$$\begin{aligned} x_1' + B_{11}x_1 + B_{12}x_2 &= q_1, \\ B_{21}x_1 + B_{22}x_2 &= q_2. \end{aligned}$$

Let B_{22} be nonsingular such that the DAE is regular with index 1 and $l = m_1$,

$$\Pi_{can}(a) = \begin{bmatrix} I & 0 \\ -B_{22}(a)^{-1}B_{21}(a) & 0 \end{bmatrix}.$$

For $C = [C_1\, C_2]$ we compute $S = C\Pi_{can}(a) = [C_1 - C_2B_{22}(a)^{-1}B_{21}(a), 0]$. This makes clear that letting $C = [I\, 0]$ is the natural choice of initial conditions.
Put, in contrast, $m_1 = m_2$ and $C = [0\, I]$ yielding $S = C\Pi_{can}(a) = [-B_{22}(a)^{-1}B_{21}(a), 0]$. Now for accurate initial conditions it is necessary that also B_{21} is nonsingular. □

Our next small example demonstrates how the condition $\ker C \cap \operatorname{im} \Pi_{can}(a) = \{0\}$ restricts the possible choice of the initial condition in a reasonable way.

Example 2.5. Consider the semi-explicit index-2 system

$$\begin{aligned} x_1' + x_1 &= 0, \\ x_2' + x_1 + x_3 &= 0, \\ x_2 + x_4 &= 1, \\ x_4 &= 1 + \sin t, \end{aligned}$$

yielding the canonical projector

$$\Pi_{can}(a) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

We have $l = \operatorname{rank} \Pi_{can}(a) = 1$, thus we state the initial condition using the matrix

$$C = \begin{bmatrix} c_1 & c_2 & c_3 & c_4 \end{bmatrix}.$$

The condition $\ker C \cap \operatorname{im} \Pi_{can}(a) = \{0\}$ is satisfied exactly if $c_1 \neq c_3$. Therefore, the initial condition $Cx(a) = \gamma$ is accurately stated, if and only if $c_1 \neq c_3$. A look at the DAE shows that this condition is reasonable. If $c_1 = c_3$, then the condition $Cx(a) = \gamma$

represents a certain consistency requirement, but the free integration constant is no longer fixed. □

The structure of the fundamental solution matrix $X(t,a)$ given by (29) tempts to consider the associated BVP induced for the IERODE (30).

We rewrite the solvabilty matrix S as

$$\begin{aligned} S &= G_a X(a,a) + G_b X(b,a) \\ &= G_a \Pi_{can}(a) D(a)^- U(a,a) D(a) \Pi_{can}(a) + G_b \Pi_{can}(b) D(b)^- U(b,a) D(a) \Pi_{can}(a) \\ &= \underbrace{\left(G_a \Pi_{can}(a) D(a)^- U(a,a) + G_b \Pi_{can}(b) D(b)^- U(b,a) \right)}_{=: S_{IERODE}} D(a) \Pi_{can}(a) \\ &=: S_{IERODE} D(a) \Pi_{can}(a). \end{aligned} \tag{45}$$

By construction, owing to the property

$$\Pi_{can}(t) D(t)^- U(t,a) = \Pi_{can}(t) D(t)^- U(t,a) D(a) \Pi_{can}(a) D(a)^-$$

it results that

$$\begin{aligned} & S_{IERODE} D(a) \Pi_{can}(a) D(a)^- = S_{IERODE}, \\ & \ker D(a) \Pi_{can}(a) D(a)^- \subseteq \ker S_{IERODE}, \\ & \operatorname{rank} S_{IERODE} \leq l. \end{aligned}$$

The solvability matrix S has rank l exactly if $S_{IERODE} \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^l)$ has rank l, and equivalently, if $\ker S_{IERODE} = \ker D(a) \Pi_{can}(a) D(a)^-$.
Let the additional matrix $C_a \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^{n-l})$ be such that

$$\ker C_a = \operatorname{im} D(a) \Pi_{can}(a) D(a)^-.$$

Then, C_a has rank $n-l$, and the classical inherent BVP

$$u' - (D \Pi_{\mu-1} D^-)' u + D \Pi_{\mu-1} G_\mu^{-1} B_\mu D^- u = D \Pi_{\mu-1} G_\mu^{-1} q, \tag{46}$$

$$C_a u(a) = 0, \tag{47}$$

$$G_a \Pi_{can}(a) D(a)^- u(a) + G_b \Pi_{can}(b) D(b)^- u(b) = \hat{\gamma} \tag{48}$$

is uniquely solvable and well-posed. This yields the further representation of the solution of the BVP (37), namely

$$x = D^- u + v_q,$$

with the solution u of the BVP (46)-(48). We summarize what we have in the next proposition.

Proposition 2.1. *Let the DAE in (37) be regular with index μ and $l = \operatorname{rank} \Pi_{can}(a)$. Given are the matrices $G_a, G_b \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^l)$. Then the BVP (37) is uniquely solvable*

for each arbitrary $\gamma \in \mathbb{R}^l$ and each arbitrary admissible right-hand side q, if and only if the homogeneous version of the classical inherent BVP (46)-(48) has the zero solution only.

If one is able to provide v_q by analytically performing the differentiations, and if the IERODE is available, then it remains only to solve the classical well-posed BVP (46)–(48).

It is noteworthy that the IERODE (46) which lives in $\mathbb{R}^n$ can be condensed to a so-called *essential underlying* ODE living in $\mathbb{R}^l$,

$$\eta' + W\eta = \rho_q,$$

by letting $\eta = \Gamma_l u$, with a suitable transformation $\Gamma_l \in \mathcal{C}^1(\mathcal{I}, \mathcal{L}(\mathbb{R}^n, \mathbb{R}^l))$ ([83, Theorem 4.5]). Then, condition (47) becomes redundant and the boundary condition (48) transforms via $u = \Gamma_l^- \eta$.

In [102, Section 13], for linear DAEs, a gradual index reduction procedure is established, which comprises analytical transformations and differentiations. Thereby, the given linear BVP for the DAE is reduced to a BVP for an explicit ODE, which is in essence a condensed version of our BVP (46)–(48).

A comparable approach consists in forming analytically the derivative array system, extracting a relevant index-0 or index-1 DAE from the derivative array system, and then turning to the regularized form for further investigations as in [111].

2.4 Conditioning constants and dichotomy

Already in the classical theory of ordinary BVPs it is established (cf. [12]) that the key quantity for well-conditioning of a BVP is κ_2. There are problems where κ_1 is moderate but κ_2 can be made arbitrary large. Moreover, for though a scaling of the boundary condition does not change the solution, the quantity κ_1 changes. Namely, if we multiply the boundary condition by the nonsingular matrix $L \in \mathcal{L}(\mathbb{R}^l)$, we arrive at $X(t,a)S^- L^{-1}$ instead of $X(t,a)S^-$.

For appropriately scaled boundary conditions the quantity κ_1 can be bounded by κ_2. Furthermore, there is a close relation between dichotomy, appropriately boundary conditions, and moderate size of κ_2. We are going to adapt these well-known classical results to the case of DAEs. The following lemma allows an useful scaling of the boundary conditions.

Lemma 2.1. *Given is the matrix $[B_a\, B_b] \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^l)$ with full row-rank l, $k_a := \operatorname{rank} B_a \le l$, $k_b := \operatorname{rank} B_b \le l$.*
Then $k_a + k_b \ge l$ and there are orthogonal matrices $Q_a, Q_b \in \mathcal{L}(\mathbb{R}^m)$, and $V \in \mathcal{L}(\mathbb{R}^l)$, and a nonsingular $R \in \mathcal{L}(\mathbb{R}^l)$ such that

$$B_a \Pi_{can}(a) = V \begin{bmatrix} I_{l-k_b} & & \\ & \Delta_a & 0 \cdots 0 \\ & & 0 \end{bmatrix} Q_a, \quad B_b \Pi_{can}(b) = V \begin{bmatrix} 0 & & \\ & \Delta_b & 0 \cdots 0 \\ & & I_{l-k_a} \end{bmatrix} Q_b,$$

whereby the blocks $\Delta_a, \Delta_b \in \mathcal{L}(\mathbb{R}^{k_a+k_b-l})$ have diagonal form with diagonal elements belonging to the interval $[0,1]$.

Proof. First, applying a Householder factorization we obtain

$$[\tilde{B}_a\, \tilde{B}_b] := R^{-1}[B_a\, B_b] = [I_l \underbrace{0}_{m-l} \underbrace{0}_{m}]H,$$

with orthogonal H, such that $[\tilde{B}_a\, \tilde{B}_b]$ has orthogonal rows and $\tilde{B}_a\tilde{B}_a^* + \tilde{B}_b\tilde{B}_b^* = I$. Then we apply the singular value decomposition $\tilde{G}_a^* = Q_a^*\Sigma_a^*V^*$, with $\Sigma_a = [D_a\, 0]$ and

$$D_a = \begin{bmatrix} \sigma_1 & & & & & & \\ & \sigma_2 & & & & & \\ & & \ddots & & & & \\ & & & \sigma_{k_a} & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{bmatrix} \in \mathcal{L}(\mathbb{R}^l), \quad \sigma_1 \geq \cdots \geq \sigma_{k_a} > 0.$$

It follows that $\tilde{B}_a\tilde{B}_a^* = V^*D_a^2V$ and $\tilde{B}_b\tilde{B}_b^* = V^*(I-D_a^2)V$. It holds that $1-\sigma_i^2 \geq 0$, since $\tilde{B}_b\tilde{B}_b^*$ is positive semi-definite. Because of $\operatorname{rank}\tilde{B}_b\tilde{B}_b^* = k_b$ it must hold that $l-k_b \leq k_a$ and $\sigma_1 = \cdots = \sigma_{l-k_b} = 1$. Finally we obtain the factorization $\tilde{B}_b^* = Q_b^*\Sigma_b^*V$ with $\Sigma_b = [D_b\ 0]$, $D_b = \operatorname{diag}(0, \Delta_b, I_{l-k_b})$. Δ_b is absent if $k_a+k_b = l$. For $k_a+k_b \geq l+1$, the values $(1-\sigma_{l-k_b+i}^2)^{\frac{1}{2}}$, $i = 1,\ldots,l-k_b-k_a$, form the diagonal of Δ_b. □

Theorem 2.2. *Let the DAE in (37) be regular with index μ and $l = \operatorname{rank}\Pi_{can}(a)$. Π_{can} denotes the canonical projector function of the DAE. Given are the matrices $G_a, G_b \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^l)$. Let condition (43) be valid. Then the following assertions hold:*

(1) *The matrix function ϕ defined by $\phi(t) := X(t,a)S^-$, $t \in \mathcal{I}$, is the minimal fundamental solution matrix of the DAE associated with the BVP such that*

$$S_{BVP} := G_a\phi(a) + G_b\phi(b) = I.$$

Thereby $X(t,a)$ denotes the fundamental solution matrix normalized at a (see (25)), and S is the solvability matrix (40).

(2) *The Green's function can be represented as*

$$\mathcal{G}(t,s) := \begin{cases} \phi(t)G_a\phi(a)\phi(s)^-, & \text{if } t \geq s \\ -\phi(t)G_b\phi(b)\phi(s)^-, & \text{if } t < s, \end{cases}$$

with the generalized inverse $\phi(t)^- = SX(t,a)^-$ satisfying the four conditions

$$\phi\phi^-\phi = \phi, \quad \phi^-\phi\phi^- = \phi^-, \quad \phi\phi^- = \Pi_{can}, \quad \phi^-\phi = I.$$

(3) *The boundary conditions can be scaled so that*

$$G_a\Pi_{can}(a) = \begin{bmatrix} I_{l-k_b} & & \\ & \Delta_a & 0\cdots 0 \\ & 0 & \end{bmatrix} Q_a, \quad G_b\Pi_{can}(b) = \begin{bmatrix} 0 & & \\ & \Delta_b & 0\cdots 0 \\ & & I_{l-k_a} \end{bmatrix} Q_b,$$

with orthogonal matrices $Q_a, Q_b \in \mathcal{L}(\mathbb{R}^m)$, $k_a := \operatorname{rank} G_a\Pi_{can}(a)$, *and* $k_b := \operatorname{rank} G_b\Pi_{can}(b)$. *The blocks* $\Delta_a, \Delta_b \in \mathcal{L}(\mathbb{R}^{k_a+k_b-l})$ *have diagonal form with diagonal elements belonging to the interval* $(0,1)$, *and* $\Delta_a^2 + \Delta_b^2 = I$.

(4) *If the boundary conditions are scaled as described in (3), then it holds that*

$$|\phi(t)|_2 \le |\mathcal{G}(t,a)|_2 + |\mathcal{G}(t,b)|_2, \quad t \in \mathcal{I},$$

which leads to $\kappa_1 \le 2\kappa_2$ *when applying the Euclidean and spectral norms.*

Proof. (1) $\phi(t)$ has full column-rank l since $\phi(t)z = 0$, i.e., $X(t,a)S^-z = 0$, implies $S^-z = (I - \Pi_{can}(a))S^-z$, thus $z = SS^-z = S(I - \Pi_{can}(a))S^-z = 0$.
(2) can be shown by straightforward calculation.
(3) Writing

$$S = G_aX(a,a) + G_bX(b,a) = [G_a\Pi_{can}(a)\ G_b\Pi_{can}(b)] \begin{bmatrix} X(a,a) \\ X(b,a) \end{bmatrix} \tag{49}$$

makes clear that the factor $[B_a\, B_b] := [G_a\Pi_{can}(a)\ G_b\Pi_{can}(b)]$ has also full row-rank l. We apply Lemma 2.1 and scale by V^{-1}.
(4) We recall that

$$\begin{aligned}
|\phi(t)|_2 &= \left|\phi(t)\begin{bmatrix} I & 0 & 0 \\ 0 & \Delta_a^2 & 0 \\ 0 & 0 & 0 \end{bmatrix}\right|_2 + \left|\phi(t)\begin{bmatrix} 0 & 0 & 0 \\ 0 & \Delta_b^2 & 0 \\ 0 & 0 & I \end{bmatrix}\right|_2 \le \left|\phi(t)\begin{bmatrix} I & 0 & 0 \\ 0 & \Delta_a & 0 \\ 0 & 0 & 0 \end{bmatrix}\right|_2 + \left|\phi(t)\begin{bmatrix} 0 & 0 & 0 \\ 0 & \Delta_b & 0 \\ 0 & 0 & I \end{bmatrix}\right|_2 \\
&= \left|\phi(t)\begin{bmatrix} I & 0 & 0\,0\ldots 0 \\ 0 & \Delta_a & 0\,0\ldots 0 \\ 0 & 0 & 0\,0\ldots 0 \end{bmatrix}\right|_2 + \left|\phi(t)\begin{bmatrix} 0 & 0 & 0\,0\ldots 0 \\ 0 & \Delta_b & 0\,0\ldots 0 \\ 0 & 0 & I\,0\ldots 0 \end{bmatrix}\right|_2 \\
&= \left|\phi(t)\begin{bmatrix} I & 0 & 0\,0\ldots 0 \\ 0 & \Delta_a & 0\,0\ldots 0 \\ 0 & 0 & 0\,0\ldots 0 \end{bmatrix} Q_a\right|_2 + \left|\phi(t)\begin{bmatrix} 0 & 0 & 0\,0\ldots 0 \\ 0 & \Delta_b & 0\,0\ldots 0 \\ 0 & 0 & I\,0\ldots 0 \end{bmatrix} Q_b\right|_2 \\
&= |\phi(t)G_a\Pi_{can}(a)|_2 + ||\phi(t)G_b\Pi_{can}(b)|_2 = ||\mathcal{G}(t,a)|_2 + ||\mathcal{G}(t,b)|_2\,.
\end{aligned}$$

□

For dichotomic explicit ODEs, one obtains a moderate conditioning quantity κ_2, if the asymptotically noninreasing mode is fixed by boundary conditions at the left border of the interval and the asymptotically nondecreasing mode is fixed at the right boundary. In other words, the conditioning constants, if they have moderate size, indicate that the boundary conditions fit well into the dynamics of the ODE. For dichotomic DAEs the situation is quite similar. To be more precise we quote the dichotomy notion ([83, Definition 2.56]).

Definition 2.6. The regular DAE (22) with index μ is said to be *dichotomic*, if there are constants $K, \alpha, \beta \geq 0$ and a nontrivial projector (not equal to the zero or identity matrix) $P_{dich} \in \mathcal{L}(\mathbb{R}^m)$ such that $P_{dich} = \Pi_{can}(a)P_{dich} = P_{dich}\Pi_{can}(a)$, and the following inequalities apply for all $t, s \in \mathcal{I}$:

$$|X(t,a)P_{dich}X(s,a)^-| \leq K\mathrm{e}^{-\alpha(t-s)}$$
$$|X(t,a)(I-P_{dich})X(s,a)^-| \leq K\mathrm{e}^{-\beta(s-t)}.$$

If $\alpha, \beta > 0$ one speaks of an *exponential dichotomy*.

The notion is independent of the choice of the reference point a; one can use any other point $\bar{t} \in \mathcal{I}$. An equivalent definition works with the minimal fundamental solution ϕ and the projector $P_{\phi,dich} = SP_{dich}S^- \in \mathcal{L}(\mathbb{R}^l)$:

$$|\phi(t)P_{\phi,dich}\phi(s)^-| \leq K\mathrm{e}^{-\alpha(t-s)},$$
$$|\phi(t)(I-P_{\phi,dich})\phi(s)^-| \leq K\mathrm{e}^{-\beta(s-t)}.$$

Theorem 2.3. *Let the DAE in (37) be regular with index μ and $l = \operatorname{rank}\Pi_{can}(a)$. Π_{can} denotes the canonical projector function of the DAE. Given are the matrices $G_a, G_b \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^l)$. Let condition (43) be valid. Let the DAE be dichotomic and let the boundary condition be such that*

$$G_a\Pi_{can}(a)(I-P_{dich}) = 0, \quad G_b\Pi_{can}(b)P_{dich} = 0. \tag{50}$$

Then the Green's function satisfies the inequalities

$$|\mathcal{G}(t,s)| \leq K\mathrm{e}^{-\alpha(t-s)} \quad for \quad s \leq t,$$
$$|\mathcal{G}(t,s)| \leq K\mathrm{e}^{-\beta(s-t)} \quad for \quad s > t.$$

Proof. The conditions (50) can be rewritten as

$$G_a\phi(a)(I-P_{\phi,dich}) = 0, \quad G_b\phi(b)P_{\phi,dich} = 0.$$

We derive

$$\begin{aligned} G_b\phi(b) &= G_b\phi(b)(I-P_{\phi,dich}) = (I-G_a\phi(a))(I-P_{\phi,dich}) \\ &= I - P_{\phi,dich} - G_a\phi(a) + G_a\phi(a)P_{\phi,dich}, \end{aligned}$$

thus $P_{\phi,dich} = G_a\phi(a)P_{\phi,dich}$. Then we compute for $s < t$

$$\begin{aligned} \mathcal{G}(t,s) &= \phi(t)G_a\phi(a)\phi(s)^- = \phi(t)G_a\phi(a)P_{\phi,dich}\phi(s)^- \\ &= \phi(t)P_{\phi,dich}\phi(s)^-, \end{aligned}$$

which yields

$$|\mathcal{G}(t,s)| = |\phi(t)P_{\phi,dich}\phi(s)^-| \leq K\mathrm{e}^{-\alpha(t-s)}.$$

The part $s > t$ is proven analogously. □

We emphasize that the concerns mentioned in [12] related to the fact that dichotomy of ODEs is thought for infinite intervals to feature the asymptotic flow behavior apply likewise for DAEs, too.

2.5 Nonlinear BVPs

The solutions of linear regular DAEs always exist on the entire given interval $\mathcal{I} = [a,b]$. We are able to precisely describe all these solutions. In particular, if $x_* \in C_D^1(\mathcal{I},\mathbb{R}^m)$ satisfies the regular index-μ DAE

$$A(t)(Dx)'(t) + B(t)x(t) - q(t) = 0, \quad t \in \mathcal{I}, \tag{51}$$

and the matrix $C \in \mathcal{L}(\mathbb{R}^m,\mathbb{R}^l)$ describing the initial condition

$$Cx(a) = Cz, \quad z \in \mathbb{R}^m, \tag{52}$$

satisfies the condition $\ker C = \ker \Pi_{can}(a)$, then the solutions of all IVPs (51), (52) are given on the entire interval e.g., by

$$x(t,z) = x_*(t) + X(t,a)(z - x_*(a)), \quad t \in \mathcal{I}. \tag{53}$$

The nonlinear regular DAE (3), that is,

$$f((Dx)'(t),x(t),t) = 0 \tag{54}$$

is much more difficult to deal with. If $x_* \in C_D^1(\mathcal{I},\mathbb{R}^m)$ satisfies this DAE on the entire interval $\mathcal{I} = [a,b]$, we form the linearized DAE

$$A_*(t)(Dx)'(t) + B_*(t)x(t) = 0, \quad t \in \mathcal{I}. \tag{55}$$

If the graph of the reference function x_* resides within an index-μ regularity region of the DAE, then the linear DAE (55) is also regular with index μ and shares with (54) further characteristics, see Appendix 6.1.3. We then denote by $\Pi_{*\,can}$ and $X_*(t,a)$ the canonical projector function associated with (55), and the fundamental solution matrix of (55) normalized by $X_*(a,a) = \Pi_{*\,can}(a)$.
After the idea of (53) we form the function

$$\tilde{x}(t,z) = x_*(t) + X_*(t,a)(z - x_*(a)), \quad t \in \mathcal{I}, \tag{56}$$

with values $\tilde{x}(t,z) \in \mathcal{D}_f$ for all z sufficiently close to $x_*(a)$. This function fulfills the condition

$$C_*\tilde{x}(a,z) = C_*z, \quad z \in \mathbb{R}^m, \tag{57}$$

with any matrix $C_* \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^l)$ such that $\ker C_* = \ker \Pi_{*can}(a)$. In the nonlinear case, the function $\tilde{x}$ satisfies the DAE only approximately. We have

$$\max_{t \in \mathcal{I}} |f((D\tilde{x})'(t,z), \tilde{x}(t,z), t)| = o(|z - x_*(a)|)$$

for all z close enough to $x_*(a)$.
Regarding the boundary condition (4), i.e.,

$$g(x(a), x(b)) = 0, \tag{58}$$

and introducing the solvability matrix of the linearized BVP,

$$S_* := G_{*a} X_*(a,a) + G_{*b} X_*(b,a), \tag{59}$$

we may derive that

$$g(\tilde{x}(a,z), \tilde{x}(b,z)) = S_*(z - x_*(a)) + o(|z - x_*(a)|).$$

If the linearized BVP has accurately stated boundary conditions, then the property (57) implies $S_*(z - x_*(a)) = 0$, and hence $|g(\tilde{x}(a,z), \tilde{x}(b,z))| = o(|z - x_*(a)|)$. In summary, the function $\tilde{x}$ satisfies the BVP approximately for all sufficiently small $z - x_*(a)$.

The last consideration raises the expectation that solutions of nonlinear DAEs can be provided under reasonable conditions, at least that there exist solutions neighboring to a given reference solution on the entire interval.

For index-1 and index-2 DAEs useful perturbation results are available which ensure the existence of DAE solutions satisfying perturbed initial conditions on the entire interval and allow the shooting approach and an sensitivity analysis. In case of higher-index DAEs the hitherto known respective results are much too restrictive. We describe more details in the next two parts.

As yet, there is a lack of precise general conditions ensuring the existence of solution. In the literature the existence of solutions is usually assumed, either frankly by a comprehensive solvability notion (e.g. in [38]) or somewhat covertly in special hypotheses (e.g. in [75, 74]), cf. Remark 2.7 for details. In [5] solvability of multipoint BVPs for special weakly nonlinear index-1 DAEs is proved by means of Schauder's fixed point theorem.

In contrast to the flow of a regular ODE that propagates in the entire $\mathbb{R}^m$, the flow of a DAE (54) is restricted to certain lower-dimensional subsets determined by the so-called obvious constraint and possibly additional hidden constraints which are quite difficult to recognize. In any case, the solution values at time t must reside within the *obvious constraint set* (cf. [83, pp. 317–318])

$$\begin{aligned}
\mathcal{M}_0(t) :&= \{x \in \mathcal{D}_f : \exists y \in \mathbb{R}^n : f(y,x,t) = 0\} \\
&= \{x \in \mathcal{D}_f : \exists y \in \mathbb{R}^n : y \in \operatorname{im} D(t), f(y,x,t) = 0\} \\
&= \{x \in \mathcal{D}_f : \exists! y \in \mathbb{R}^n : y \in \operatorname{im} D(t), f(y,x,t) = 0\}.
\end{aligned}$$

We note that also the obvious constraint set is not necessarily clearly manifested in fact, as e.g., in Example 1.2.

2.5.1 BVPs well-posed in the natural setting

Theorem 2.4. *Let $x_* \in \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m)$ satisfy the BVP (54), (58), $r := \operatorname{rank} D(a)$. Then the following assertions are equivalent:*

(1) *The original nonlinear BVP is locally well-posed in the natural setting.*
(2) *The linearized along x_* BVP is well-posed in the natural setting.*
(3) *The linearized DAE is regular with index 1, and the linearized BVP has accurately stated boundary conditions with $l = r$.*
(4) *The graph of x_* resides in a index-1 regularity region of the DAE (54), and the linearized BVP has accurately stated boundary conditions with $l = r$.*
(5) *x_* is an isolated solution of the BVP, $l = r$, and the linearized DAE is regular with index 1.*

Proof. We first formulate the DAE (54) and the BVP (54), (58) as the operator equations $F(x) = 0$ and $\mathcal{F}(x) = 0$ in Banach spaces, with $F : \operatorname{dom} F \subseteq \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m) \to \mathcal{C}(\mathcal{I},\mathbb{R}^m)$, $\mathcal{F} : \operatorname{dom} F \subseteq \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m) \to \mathcal{C}(\mathcal{I},\mathbb{R}^m) \times \mathbb{R}^l$,

$$(Fx)(t) := f((Dx)'(t),x(t),t),\ t \in \mathcal{I}, \quad x \in \operatorname{dom} F,$$
$$\mathcal{F}x := (Fx, g(x(a),x(b))), \quad x \in \operatorname{dom} F.$$

The definition domain $\operatorname{dom} F$ is a neighborhood of x_* in $\mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m)$ (e.g. [89, 83, 96]). F and thus $\mathcal{F}$ are Fréchet differentiable,

$$F'(x_*)x = A_*(Dx)' + B_*x, \quad x \in \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m).$$

The linear equation $\mathcal{F}'(x_*)x = 0$ represents the homogenous version of the linearized along x_* BVP.
(1)→(2): In the context of nonlinear functional analysis, local well-posedness of the equation $\mathcal{F}x = 0$ means that $\mathcal{F}$ is a local diffeomorphism at x_*. Then the derivative $\mathcal{F}'(x_*)$ is necessarily a homeomorphism. In turn, the boundedness of $(\mathcal{F}'(x_*))^{-1}$ means that the linearized BVP is well-posed.
(2)→(3): This is a consequence of Theorem 2.1(5).
(3)→(4): Consider the matrix function

$$G(y,x,t) := f_y(y,x,t)D(t) + f_x(y,x,t)Q_0(t), \quad y \in \mathbb{R}^n, x \in \mathcal{D}_f, t \in \mathcal{I}_f.$$

Owing to the index-1 property of the linearized DAE,

$$G((Dx_*)'(t),x_*(t),t) := A_*(t)D(t) + B_*(t)Q_0(t), \quad t \in \mathcal{I},$$

remains nonsingular. Since the interval $\mathcal{I}$ and thus the graph are compact, there is an open neighborhood $\mathcal{N}_* \subseteq \mathbb{R}^n \times \mathcal{D}_f \times \mathcal{I}_f$ of the graph so that $G(y,x,t)$ is nonsingular

also on $\mathcal{N}_*$. This means, that $\mathcal{N}_*$ is actually an index-1 regularity region.
(4)→(1): Here the linearized DAE is regular with index 1 and its boundary conditions are stated accurately. This means that $\mathcal{F}'(x_*)$ is a homeomorphism and $\mathcal{F}$ is a local diffeomorphism.
(5)⇄(3):This is a direct consequence of Definitions 2.4 and 2.5. □

Example 2.6. We continue considering Example 1.3. The homogenous DAE linearized along the solution x_* reads

$$\begin{bmatrix}1&0\\0&1\\0&0\end{bmatrix}\left(\begin{bmatrix}1&0&0\\0&1&0\end{bmatrix}x\right)'(t)+\begin{bmatrix}1-\alpha(t) & -1 & 0\\ 1 & 1-\alpha(t) & 0\\ 2\sin t & 2\cos t & 1\end{bmatrix}x(t)=0,\quad t\in\mathcal{I}=[a,b],$$

where $a=0$ and $b=2\pi$. Compute

$$Q_0=\begin{bmatrix}1&0&0\\0&1&0\\0&0&0\end{bmatrix},\quad G_{*\,1}(t)=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}.$$

The linearized DAE has index 1 owing to the nonsingularity of $G_{*\,1}(t)$. We obtain the canonical projector function

$$\Pi_{*\,can}(t)=\begin{bmatrix}1&0&0\\0&1&0\\-2\sin t&-2\cos t&0\end{bmatrix},$$

and the homogenous IERODE

$$u'(t)+\begin{bmatrix}1-\alpha(t) & -1\\ 1 & 1-\alpha(t)\end{bmatrix}u(t)=0,$$

with the fundamental solution matrix

$$U_*(t,0)=\mathrm{e}^{-\int_0^t(1-\alpha(s))\mathrm{d}s}\begin{bmatrix}\cos t & \sin t\\ -\sin t & \cos t\end{bmatrix}.$$

The fundamental solution matrix of the linearized DAE results as

$$\begin{aligned}X_*(t,0)&=\Pi_{*\,can}(t)\begin{bmatrix}1&0\\0&1\\0&0\end{bmatrix}U_*(t,0)\begin{bmatrix}1&0&0\\0&1&0\end{bmatrix}\Pi_{*\,can}(0)\\&=\begin{bmatrix}1&0\\0&1\\-2\sin t&-2\cos t\end{bmatrix}U_*(t,0)\begin{bmatrix}1&0&0\\0&1&0\end{bmatrix}.\end{aligned}$$

The linearization of the nonlinear boundary condition leads to

$$G_{*\,a} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad G_{*\,b} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 2 & 0 \end{bmatrix},$$

thus

$$S_* = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix} U_*(2\pi,0) \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2\mathrm{e}^{-\int_0^{2\pi}(1-\alpha(s))\mathrm{d}s} & 0 \end{bmatrix}.$$

This proves that the linearized boundary condition are accurately stated, and hence the linearized BVP and also the nonlinear BVP are well-posed.

The BVP with periodic boundary condition in Example 1.3 leads to

$$G_{*\,a} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad G_{*\,b} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix},$$

thus

$$S_* = \begin{bmatrix} 1-\mathrm{e}^{-\int_0^{2\pi}(1-\alpha(s))\mathrm{d}s} & 0 & 0 \\ 0 & 1-\mathrm{e}^{-\int_0^{2\pi}(1-\alpha(s))\mathrm{d}s} & 0 \end{bmatrix},$$

so that the periodic BVP is well-posed if $\int_0^{2\pi}(1-\alpha(s))\mathrm{d}s \neq 0$. In particular, this is the case for identically vanishing α as drafted in Figure 4 in Example 1.3 in the Introduction.

If $\int_0^{2\pi}(1-\alpha(s))\mathrm{d}s = 0$, the BVP is no longer well-posed. If $\alpha(t) \equiv 1$, then, for arbitrary parameters $c_1, c_2 \in \mathbb{R}$, $c_1^2 + c_2^2 = 1$, the functions given as

$$x_{**}(t) = \begin{bmatrix} c_1 \cos t + c_2 \sin t \\ c_2 \cos t - c_1 \sin t \\ 1 \end{bmatrix}$$

are 2π-periodic and satisfy the DAE. □

Theorem 2.4 clearly points out that *only BVPs for index-1 DAEs can be well-posed in the natural setting*. This fact is in full concert with the general computational experience. At this place we allude to a peculiar definition of well-posed BVPs in [75, 74], which seemingly says that also BVPs for higher index DAEs could be well-posed. We refer to Remarks 2.6 and 2.7 for a further discussion.

We concentrate now briefly on index-1 problems which are well understood for a long time. So the next perturbation results are nothing else than useful updates of [89, Theorem 4]. We refer to [83, Part II] for a recent elaborate exposition.

Theorem 2.5. *Let $x_* \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ satisfy the DAE (54), and the linearized along x_* DAE (55) be regular with index 1. Let $\Pi_{*\,can}$ denote the canonical projector function of the linear DAE (55). Let the matrix $C \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^l)$, $l = r$, be such that*

$$\ker C \cap \operatorname{im} \Pi_{*can}(a) = \{0\} \tag{60}$$

Then the IVP

$$f((Dx)'(t), x(t), t) = 0, \quad t \in \mathcal{I}, \quad Cx(a) = Cz. \tag{61}$$

has a locally unique solution $x(\cdot; a, z) \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ *for each arbitrary* $z \in \mathbb{R}^m$, $|\Pi_{*can}(a)(z - x_*(a))|$ *sufficiently small.*
Moreover, there exists the sensitivity matrix

$$X(t,z) := \frac{\partial}{\partial z} x(t; a, z)$$

with columns in $\mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$*, and it satisfies the variational equation*

$$f_y((Dx)'(t;a,z), x(t;a,z), t)(DX)'(t,z) + f_x((Dx)'(t;a,z), x(t;a,z), t)X(t,z) = 0,$$
$$C(X(a,z) - I) = 0.$$

Proof. The assertion follows from the implicit function theorem applied to the equation $\mathcal{H}(x,z) = 0$,

$$\mathcal{H}(x,z) := (Fx, C(x(a) - z)), \quad x \in \operatorname{dom} F,\ z \in \mathbb{R}^m,$$

with the differential-algebraic operator F from the proof of Theorem 2.4. □

An *index-1 regularity region* $\mathcal{G}$ of the DAE (54) is an open connected subset of the definition domain of f characterized by the nonsingularity of the matrix function

$$G(y,x,t) := f_y(y,x,t)D(t) + f_x(y,x,t)Q_0(t), \quad y \in \mathbb{R}^n, x \in \mathcal{D}_f, t \in \mathcal{I}_f.$$

on $\mathcal{G}$, or, equivalently, by the decomposition

$$\mathbb{R}^m = S(y,x,t) \oplus \ker D(t), \tag{62}$$
$$S(y,x,t) := \{z \in \mathbb{R}^m : f_x(y,x,t)z \in \operatorname{im} f_y(y,x,t)\}.$$

or, equivalently, by a regular matrix pencil $\lambda f_y(y,x,t)D(t) + f_x(y,x,t)$ with Kronecker index 1 (e.g., [83, Part II]).

The decomposition (62) defines the *canonical projector function* Π_{can} of the index-1 DAE by

$$\operatorname{im} \Pi_{can}(y,x,t) = S(y,x,t), \quad \ker \Pi_{can}(y,x,t) = \ker D(t).$$

It holds that

$$\Pi_{can}(y,x,t) = I - Q_0(t)G^{-1}(y,x,t)f_x(y,x,t).$$

Formulating the initial condition in (61) by a matrix C that has the same nullspace as $D(a)$ or choosing $C = D(a)$ makes condition (60) to be trivially fulfilled. This

means that the initial condition is directed promptly to the dynamical component and yields the following practically most useful special case of Theorem 2.5.

Corollary 2.2. *The assertions of Theorem 2.5 remain valid, if the condition (60) is replaced by the easier condition*

$$\ker C = \ker D(a).$$

For the further analysis the decoupled form (e.g. [55, 83])

$$u'(t) - R'(t)u(t) = D(t)\omega(u(t),t), \tag{63}$$

$$x(t) = D(t)^{-}u(t) + Q_0(t)\omega(u(t),t), \tag{64}$$

of the index-1 DAE (54) is approved to be useful. The decoupling function $w = \omega(u,t)$ is uniquely defined from the equation

$$f(D(t)w, D(t)^{-}u + Q_0(t)w, t) = 0$$

locally around a reference solution solution $x_*(\cdot)$ or points $\bar{x} \in \mathcal{M}_0(\bar{t})$ by the implicit function theorem. The function ω is continuous and has the continuous partial derivative ([83, Theorem 4.5])

$$\omega_u(u,t) = -(G^{-1}f_x)(D(t)\omega(u,t), D(t)^{-}u + Q_0(t)\omega(u,t), t).$$

We additionally quote a solvability result from [83, Theorem 4.11]:

Theorem 2.6. *Given is the DAE (54) with the index-1 regularity region $\mathcal{G}$, and $(y_0,x_0,t_0) \in \mathcal{G}$. Then, if additionally $x_0 \in \mathcal{M}_0(t_0)$, the DAE possesses a solution $x_* \in \mathcal{C}^1_D(\mathcal{I}_*,\mathbb{R}^m)$ defined at least on a neighborhood $\mathcal{I}_* \subseteq \mathcal{I}_f$ of t_0 and passing through $x_*(t_0) = x_0$. The solution $x_* \in \mathcal{C}^1_D(\mathcal{I}_*,\mathbb{R}^m)$ is locally unique.*

The solution x_* from Theorem 2.6 can be continued at least as long as its graph resides in the regularity region. It also may happen that a solution crosses the border of a maximal regularity region ([83, Section 3.3]).

2.5.2 BVPs well-posed in an advanced setting

By Theorem 2.4, BVPs for higher-index DAEs are essentially different since they are never well-posed in the natural setting – even if the boundary conditions are accurately stated. In some situations, a weaker well-posedness by means of an adapted image space Y with stronger topology instead of the continuous function space might be helpful, but, as described in detail in [96], one should be highly cautious concerning the actual practical meaning. The following can be seen as quite straightforward generalization of index-2 results from [96, Subsection 4.3.3] and [83, Section 3.9] for the case of arbitrary higher index.

Definition 2.7. Let $x_* \in \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m)$ be a solution of the BVP (3), (4), $\mathcal{I} = [a,b]$. Let $Y \subseteq \mathcal{C}(\mathcal{I},\mathbb{R}^m)$ be a complete normed linear space, and $\|q\|_Y \geq \|q\|_\infty$, $q \in Y$. The

BVP (3), (4) is said to be *well-posed in the advanced setting with image space Y* locally around x_*, if the slightly perturbed BVP

$$f((Dx)'(t),x(t),t) = q(t), \quad t \in \mathcal{I}, \tag{65}$$

$$g(x(a),x(b)) = \gamma, \tag{66}$$

is locally uniquely solvable for each arbitrary sufficiently small perturbations $q \in Y$, $\gamma \in \mathbb{R}^l$, and the solution x satisfies the inequality

$$\|x - x_*\|_{\mathcal{C}^1_D} \le \kappa(|\gamma| + \|q\|_Y), \tag{67}$$

with a constant κ. Otherwise the BVP is said to be *ill-posed* in the advanced Y-setting.

Instead of the inequality (67) one can use the somewhat simpler inequality

$$\|x - x_*\|_\infty \le \kappa(|\gamma| + \|q\|_Y), \tag{68}$$

which is sometimes more convenient, see Remark 2.12.

Representing the linear BVP

$$A(t)(Dx)'(t) + B(t)x(t) = q(t),\ t \in \mathcal{I}, \quad G_a x(a) + G_b x(b) = \gamma, \tag{69}$$

with a regular index-μ DAE, as operator equation $\mathcal{T}x = (q,\gamma)$ by the linear bounded operators

$$\begin{aligned} &Tx := A(Dx)' + Bx, \quad \mathcal{T}x := (Tx, G_a x(a) + G_b x(b)), \quad x \in \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m),\\ &T \in \mathcal{L}(\mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m),Y), \quad \mathcal{T} \in \mathcal{L}(\mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m),Y\times\mathbb{R}^l), \quad Y = \mathcal{C}^{ind\,\mu}(\mathcal{I},\mathbb{R}^m), \end{aligned}$$

we know (cf. Appendix 6.1.4) that the linear BVP is well-posed in the advanced setting with Y, if and only if $\mathcal{T}$ is bijective, and then κ in (10) is nothing else an upper bound of $\|\mathcal{T}^{-1}\|_{[Y\times\mathbb{R}^l \to \mathcal{C}^1_D]}$.

Theorem 2.7. *Let $x_* \in \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m)$ satisfy the DAE (54). Let the linearized along x_* DAE (55) be regular with index μ, Π_{*can} denotes the canonical projector function of the linear DAE (55), and $l = \operatorname{rank}\Pi_{*can}(a)$. Let Y_* denote the associated Banach space of admissible right-hand sides with the norm $\|\cdot\|_{Y_*}$.*
Assume that there exists a radius $\rho > 0$ such that

$$x \in \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m),\ \|x - x_*\|_{\mathcal{C}^1_D} \le \rho \implies f(Dx)'(\cdot),x(\cdot),\cdot) \in Y_*. \tag{70}$$

Then the following assertions are valid:

(1) *Let x_* also satisfy the boundary condition (58). Then the BVP (55), (58) is locally well-posed in the advanced setting with Y_* if and only if x_* is an isolated solution.*

(2) *Let the matrix $C \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^l)$ be such that*

$$\ker C \cap \operatorname{im} \Pi_{*can}(a) = \{0\} \tag{71}$$

Then the IVP

$$f((Dx)'(t), x(t), t) = 0, \quad t \in \mathcal{I}, \quad Cx(a) = Cz. \tag{72}$$

has a locally unique solution $x(\cdot; a, z) \in \mathcal{C}^1_D(\mathcal{I}, \mathbb{R}^m)$ for each arbitrary $z \in \mathbb{R}^m$ with sufficiently small difference $|\Pi_{*can}(a)(z - x_*(a))|$. Moreover, there exists the sensitivity matrix

$$X(t, z) := \frac{\partial}{\partial z} x(t; a, z)$$

with columns in $\mathcal{C}^1_D(\mathcal{I}, \mathbb{R}^m)$, and it satisfies the variational equation

$$f_y((Dx)'(t; a, z), x(t; a, z), t)(DX)'(t, z) + f_x((Dx)'(t; a, z), x(t; a, z), t)X(t, z) = 0,$$
$$C(X(a, z) - I) = 0.$$

Proof. We again formulate the DAE (54) and the BVP (54), (58) as the operator equations $F(x) = 0$ and $\mathcal{F}(x) = 0$ in Banach spaces, this time, owing to condition (70), with definition domain $\operatorname{dom} F = \{x \in \mathcal{C}^1_D(\mathcal{I}, \mathbb{R}^m) : \|x - x_*\|_{\mathcal{C}^1_D} < \rho\}$ and advanced image spaces,

$$F : \operatorname{dom} F \subseteq \mathcal{C}^1_D(\mathcal{I}, \mathbb{R}^m) \to Y_*, \quad \mathcal{F} : \operatorname{dom} F \subseteq \mathcal{C}^1_D(\mathcal{I}, \mathbb{R}^m) \to Y_* \times \mathbb{R}^l,$$
$$(Fx)(t) := f((Dx)'(t), x(t), t),\ t \in \mathcal{I}, \quad \mathcal{F}x := (Fx, g(x(a), x(b)), \quad x \in \operatorname{dom} F.$$

Regarding condition (70) and Appendix 6.1.4, the operators F and $\mathcal{F}$ can be shown to be Fréchet-differentiable also in this setting, and

$$\mathcal{F}'(x_*)x = (A_*(Dx)' + B_*x,\ G_{*a}x(a) + G_{*b}x(b)), \quad x \in \mathcal{C}^1_D(\mathcal{I}, \mathbb{R}^m).$$

(1) The composed map $\mathcal{F}$ is a local diffeomorphism if and only if $\mathcal{F}'(x_*) \in \mathcal{L}(\mathcal{C}^1_D(\mathcal{I}, \mathbb{R}^m), Y_* \times \mathbb{R}^l)$ is bijective. Since $\mathcal{F}'(x_*)$ is surjective by construction of Y_*, bijectivity becomes equivalent with injectivity. In turn, $\mathcal{F}'(x_*)$ is injective exactly if the solution x_* is isolated.
(2) The assertion follows from the implicit function theorem applied to the equation $\mathcal{H}(x, z) = 0$, with $\mathcal{H}(x, z) := (Fx, C(x(a) - z)), \quad x \in \operatorname{dom} F,\ z \in \mathbb{R}^m$. □

Example 2.7. We turn once again to Example 1.4 and take x_* as reference solution. Similar arguments will then apply to the case of the second solution x_{**}. Inspecting the matrix function sequence we know that the DAE has two maximal regularity regions, both with characteristics $r_0 = r_1 = 2$, $r_2 = 3$, $\mu = 2$, and $l = 1$. The border of the regularity regions is given by the plane $x_2 = 0$. It holds that $x_*(t) \geq \frac{1}{4}$ for all $t \in [0, 2]$, so that the graph of x_* resides within an index-2 regularity region. Then, Theorem 2.4 excludes well-posedness in the natural setting.
The homogenous BVP linearized along x_*

$$\begin{bmatrix} 1 & 0 \\ 0 & x_{*2} \\ 0 & 0 \end{bmatrix} \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} x \right)' + \begin{bmatrix} 1 & 0 & 0 \\ 0 & x'_{*2} & 0 \\ 2x_{*1} & 2x_{*2} & 0 \end{bmatrix} x = 0,$$
$$x_1(0) - x_1(2) = 0.$$

has only the trivial solution, and hence x_* is an isolated solution. The linearized DAE inherits from the nonlinear original the characteristics $r_0 = r_1 = 2$, $r_2 = 3$, $\mu = 2$, and $l = 1$. Inspecting the admissible right-hand sides of the linearized DAE we find that

$$\mathcal{C}_*^{index2}(\mathcal{I}, \mathbb{R}^3) = \{q \in \mathcal{C}(\mathcal{I}, \mathbb{R}^3) : q_3 \in \mathcal{C}^1(\mathcal{I}, \mathbb{R})\}$$

does not depend on x_*. We set $Y = \mathcal{C}_*^{index2}(\mathcal{I}, \mathbb{R}^3)$. Equipped with the norm

$$\|q\|_Y := \|q\|_\infty + \|q'_3\|_\infty,$$

Y is a Banach space. Furthermore, for each arbitrary $x \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^3)$ and

$$q(t) := f((Dx)'(t), x(t), t) = \begin{bmatrix} x'_1(t) + x_1(t) \\ x_2(t)x'_2(t) - x_3(t) \\ x_1(t)^2 + x_2(t)^2 - 1 + \frac{1}{2}\cos \pi t \end{bmatrix}, \quad t \in \mathcal{I},$$

it results that $q \in Y$. Finally, owing to Theorem 2.7 the nonlinear BVP proves to be well-posed in the advanced setting with image space Y. In particular, the perturbed BVPs with sufficiently small $|\gamma|$ and $\|q\|_\infty + \|q'_3\|_\infty$ are locally uniquely solvable and the solutions satisfy the inequality

$$\|x - x_*\|_\infty \leq \kappa(|\gamma| + \|q\|_\infty + \|q'_3\|_\infty).$$

□

Although Theorem 2.7 sounds promising there are serious objections to it concerning the relevance for practical computations:

(1) The advanced image space Y_* and its norm are rarely available in practice.
(2) The higher the index μ the more unsuitable is the norm $\|\cdot\|_{Y_*}$ for practical needs, see [96, Section 2].
(3) Condition (70) seems to be quite acceptable. However, in the light of possible variations of $\operatorname{im} F'(x)$ with x (see [96, Example 4.3]), there are more restrictions on the classes of nonlinear DAEs the higher the index is.

The situation turns out to remain more or less acceptable only in the easier index-2 case, as already demonstrated by Example 2.7. The general solution of a linear regular index-2 DAE is established in Subsection 2.2, in particular, the canonical projector function is given there. In Example 2.3 the particular case of index-2 DAEs in Hessenberg form is specified.

The subspace

$$\mathcal{C}^{ind\,2}(\mathcal{I},\mathbb{R}^m) := \{w \in \mathcal{C}(\mathcal{I},\mathbb{R}^m) : D\Pi_0 Q_1 G_2^{-1} w \in \mathcal{C}^1(\mathcal{I},\mathbb{R}^n)\}.$$

serves as set of admissible right-hand sides of the linear index-2 DAE (51). The dynamical degree of freedom amounts to $l = r_0 + r_1 - m$. It becomes clear that the linear BVP (69) for an index-2 DAE, with accurately stated boundary condition is well-posed in the advanced setting with $Y = \mathcal{C}^{index2}(\mathcal{I},\mathbb{R}^m)$.
In [8], for linear Hessenberg index-2 DAEs, an inequality like (68) is obtained and the constant κ is called *stability constant*. Further, if κ is of moderate size, the BVP is said to be well-conditioned. In essence, in our context this means well-posedness in the advanced setting, and moderate conditioning constants.

Accordingly, if the linearized DAE (55) is regular with index 2, then the associated set of admissible right-hand sides is given by

$$Y_* = \mathcal{C}_*^{ind\,2}(\mathcal{I},\mathbb{R}^m) := \{w \in \mathcal{C}(\mathcal{I},\mathbb{R}^m) : D\Pi_0 Q_{*1} G_{*2}^{-1} w \in \mathcal{C}^1(\mathcal{I},\mathbb{R}^n)\}.$$

The asterisk-index indicates the possible dependence of the reference solution x_*.

For index-2 DAEs (54), we are aware of more transparent sufficient criteria for condition (70) to be valid. Namely, if the structural restriction (cf. [92], [83, Subsection 3.9.2])

$$W_{*0}(t)\{f(y,x,t) - f(0,P_0(t)x,t)\} \in \operatorname{im} W_{*0}(t)B_*(t)Q_0(t), \tag{73}$$

with $W_{*0}(t) = I - A_*(t)A_*(t)^-$, or, equivalently,

$$f(y,x,t) - f(0,P_0(t)x,t) \in \operatorname{im} G_{*1}(t), \tag{74}$$

is satisfied, then condition (70) is guarantied. Fortunately, often the subspaces $\operatorname{im} A_*(t)$ and $\operatorname{im} G_{*1}(t)$ are actually independent of the reference solution x_*. Index-2 DAEs in Hessenberg form serve as particular instance of DAEs satisfying condition (73).

Formulating the initial condition in (72) by a matrix C that has the nullspace as $\ker C = \ker \Pi_{*can}(a) = \ker \Pi_{*\mu-1}(a)$ makes condition (71) to be trivially fulfilled. This means that the initial condition is directed promptly to the dynamical component and yields the following useful assertion.

Corollary 2.3. *The second assertion of Theorem 2.7 remains valid, if the condition (71) is replaced by the simpler condition*

$$\ker C = \ker \Pi_{*\mu-1}(a) = \ker D(a) \oplus \ker G_{*1}(a). \tag{75}$$

Example 2.8. For the index-2 DAE in Hessenberg form in Example 2.3 we obtain $\ker \Pi_{*can}(a) = \{z \in \mathbb{R}^{m_1+m_2} : z_1 = \Omega_* z_1\}$, with $\Omega_* = B_{*12}B_{*12}^-$. □

In contrast to the index-1 case in Corollary 2.2, now also the matrix C in formula (75) depends on x_*, in fact. This foreshadows one of the challenging difficulties concerning higher index DAEs, the determination of consistent initial values.

2.6 Other boundary conditions

As established for explicit ODEs in [12], various conditions are applied to fix solutions in different applications, for instance, multipoint conditions, integral conditions, and separated conditions, and BVPs of different forms can be converted to each other. The same happens for DAEs. Here we address some of the related topics.

We call attention to the fact that the dynamical degree of freedom $l \leq m$ of a regular DAE strongly depends on the structure of this special DAE. In the context of the projector based analysis (cf. Appendix 6.1) l is determined as

$$l = m - \sum_{i=0}^{\mu-1} (m - r_i). \tag{76}$$

Another way providing l using derivative arrays is described in [38]. Evidently the number of initial or boundary conditions must be chosen accordingly.

Except for the index-1 case, where $l = r_0 = r = \operatorname{rank} D(a)$, the number l is rarely a priori available. Usually, l has to be computed (e.g. [83, Chapter 7], [38]).

In the present paper we decide on mainly stating the boundary condition in $\mathbb{R}^l$ (following [38, 8, 41, 22], [12, p. 474]) and most notably accenting that the right number of conditions should be given.

In contrast, it is also just and equitable to state the boundary condition in $\mathbb{R}^m$ (e.g. [89, 90, 55, 4, 5]) and so to emphasize that l has to be determined. Then, a consistency condition has to be respected. We address this topic in Subsubsection 2.6.2 below.

For practical computations it is recommended to regard the relation

$$\ker \Pi_{can} = \ker \Pi_{\mu-1} = N_0 + \cdots + N_{\mu-1},$$

which is an inherent property of all admissible matrix function sequences for a regular DAE. It is much easier to calculate some admissible sequence than to provide the canonical projector function by a completely decoupling sequence (cf. [83]). In general, the canonical projector function is of great avail in theory, however, though there are constructive approaches, as yet, there are no efficient means to provide it practically.

2.6.1 General boundary conditions in $\mathbb{R}^l$

The most general linear condition for fixing solutions of DAEs is given by a linear bounded map as

$$\Gamma x = \gamma, \quad \Gamma : \mathcal{C}(\mathcal{I}, \mathbb{R}^m) \to \mathbb{R}^l.$$

In particular, this comprises IVPs, two-point BVPs, multipoint BVPs, and problems with integral condition by

$$\Gamma x := Cx(a),$$
$$\Gamma x := G_a x(a) + G_b x(b),$$
$$\Gamma x := \sum_{i=0}^{s} G_i x(\eta_i), \quad a = \eta_0 < \cdots < \eta_s = b,$$
$$\Gamma x := \int_a^b G(t)x(t)\mathrm{d}t,$$

respectively. The notion of well-posedness and accurately stated boundary condition can be immediately resumed.

Supposing a regular index-μ DAE

$$A(Dx)' + Bx = q \tag{77}$$

and applying the solution representation (32) with $\bar{t} = a$ we see that $\Gamma x = \gamma$ actually means $\Gamma X(\cdot, a)c = \gamma - \Gamma x_q$. The *solvability matrix*

$$S := \Gamma X(\cdot, a) \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^l)$$

inherits the property $\ker \Pi_{can}(a) = \ker X(t,a) \subseteq \ker S$. The general BVP for (77) has *accurately stated boundary condition* exactly if (cf. (43))

$$\operatorname{im} \Gamma = \mathbb{R}^l, \quad \ker S = \ker \Pi_{can}(a). \tag{78}$$

The general linear BVP is *well-posed in the natural setting* exactly if the boundary condition is accurately stated and, furthermore, the DAE has index 1. Then, one has simply $l = \operatorname{rank} D(a)$, cf. Subsubsection 2.5.1.

Nonlinear versions of those well-posed BVPs are treated e.g. in [41, 22].

It might be often convenient to utilize for problems originally given with different boundary conditions well approved software written for two-point BVPs – as is common practice for regular ODEs (cf. [12]).

For a BVP with integral condition one introduces the additional continuously differentiable function y by

$$y(t) = \int_a^t G(s)x(s)\mathrm{d}s.$$

The augmented two-point BVP

$$\begin{bmatrix} A & 0 \\ 0 & I \end{bmatrix} \Big(\begin{bmatrix} D & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \Big)' + \begin{bmatrix} B & 0 \\ -G & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} q \\ r \end{bmatrix}, \tag{79}$$

$$\begin{bmatrix} 0 & I \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x(a) \\ y(a) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} x(b) \\ y(b) \end{bmatrix} = \begin{bmatrix} \psi \\ \gamma \end{bmatrix}, \tag{80}$$

is uniquely solvable for each arbitrary $q \in \mathcal{C}^{ind\,\mu}(\mathcal{I},\mathbb{R}^m)$, $r \in \mathcal{C}(\mathcal{I},\mathbb{R}^l)$, $\gamma,\psi \in \mathbb{R}^l$, if and only if condition (78) is valid. If so, then the augmented BVP with $r=0$ and $\psi=0$ reproduces as x-component the solution of the original BVP.

A multipoint BVP with given points $a=\eta_0<\cdots<\eta_s=b$ can be converted by linear changes of the variable t mapping each subinterval $[\eta_{i-1},\eta_i]$ to $[0,1]$. Introduce the functions x_i,A_i,D_i,B_i,q_i, all given on the interval $[0,1]$, by

$$x_i(\tau)=x(t)=x(\eta_{i-1}+\tau(\eta_i-\eta_{i-1})),\quad t=\eta_{i-1}+\tau(\eta_i-\eta_{i-1}),\quad \tau\in[0,1],$$

and so on. Then we turn to the sm-dimensional two-point BVP on $[0,1]$,

$$A_i\,\tfrac{1}{\eta_i-\eta_{i-1}}\,\tfrac{\mathrm{d}}{\mathrm{d}\tau}(D_ix_i)+B_ix_i=q_i,\quad i=1,\ldots,s, \tag{81}$$

$$C_i(x_i(1)-x_{i+1}(0))=0,\ i=1,\ldots,s-1,\quad \sum_{i=0}^{s-1}G_ix_{i+1}(0)+G_sx_s(1)=\gamma. \tag{82}$$

It is evident that the augmented DAE (81) is regular with index μ and dynamical degree of freedom is sl, if the original DAE (77) is regular with index μ and dynamical degree of freedom l. If we choose matrices $C_i\in\mathcal{L}(\mathcal{I},\mathbb{R}^l)$ such that $\ker C_i=\ker\Pi_{can}(\eta_i)$, we put the right number of boundary conditions. It is straightforward to prove that the boundary conditions (82) are accurately stated if the original BVP has accurately stated boundary condition, i.e., if

$$\ker S=\ker\Pi_{can}(a),\quad S:=\sum_{i=0}^{s}G_iX(\eta_i,a)\in\mathcal{L}(\mathbb{R}^m,\mathbb{R}^l).$$

Replacing in (82) the matrices C_i by the identity $I\in\mathcal{L}(\mathbb{R}^m)$ and so requiring the $l+(s-1)m$ boundary conditions

$$x_i(1)=x_{i+1}(0),\quad i=1,\ldots,s-1,\quad \sum_{i=0}^{s-1}G_ix_{i+1}(0)+G_sx_s(1)=\gamma. \tag{83}$$

leads to a consistent overdetermined problem.

BVPs for explicit ODEs with so-called *switching points* are discussed in [12]. In the case of DAEs, this corresponds in some sense to the BVP (81), (82) with unknown points $\eta_1,\ldots,\eta_{s-1}$. Till now it remains open whether the usual trick to introduce constant functions η_i by adding the trivial differential equations $\eta_i'=0$, $i=1,\ldots,s-1$, can be here also adapted to work.

2.6.2 General boundary conditions in $\mathbb{R}^m$

Often one formulates IVPs with the initial condition

$$x(a) = x_a \in \mathbb{R}^m, \tag{84}$$

This makes good sense for regular ODEs. For DAEs, this initial condition fails to be accurately stated. Such an IVP is solvable if and only if x_a is a consistent value, otherwise the IVP is overdetermined. Recall that the number of initial conditions should be chosen in accordance with the dynamical degree of freedom $l < m$ of the DAE.

Consider the general BVP for the DAE (77) with boundary conditions stated in $\mathbb{R}^m$,

$$\Gamma x = \gamma, \quad \Gamma : \mathcal{C}(\mathcal{I}, \mathbb{R}^m) \to \mathbb{R}^m, \tag{85}$$

whereby Γ is a linear bounded map describing initial, two-point boundary, multipoint boundary, and integral conditions as in Subsection 2.6.2. Let the DAE be regular with index μ. The so-called *solvability matrix* (also: *shooting matrix*)

$$S := \Gamma X(\cdot, a) \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^m)$$

has the properties

$$\ker \Pi_{can}(a) \subseteq \ker S, \quad \operatorname{rank} S \leq l.$$

We represent the BVP as operator equation $\mathcal{T}x = (q, \gamma)$ by means of Γ and the additional bounded linear operators (cf. Appendix 6.1.4)

$$\begin{aligned} &\mathcal{T} : \mathcal{C}^1_D(\mathcal{I}, \mathbb{R}^m) \to \mathcal{C}(\mathcal{I}, \mathbb{R}^m)) \times \mathbb{R}^m, && \mathcal{T}x := (Tx, \Gamma x), \\ &T : \mathcal{C}^1_D(\mathcal{I}, \mathbb{R}^m) \to \mathcal{C}(\mathcal{I}, \mathbb{R}^m)), && Tx := A(Dx)' + Bx. \end{aligned}$$

The subspace $\operatorname{im}\Gamma \subseteq \mathbb{R}^m$ has necessarily finite dimension. Also the nullspace of $\mathcal{T}$ is finite-dimensional, more precisely,

$$\ker \mathcal{T} = \{X(\cdot, a)c : c \in \ker S \cap \operatorname{im} \Pi_{can}(a)\}, \quad \dim \ker \mathcal{T} = l - \operatorname{rank} S.$$

The boundary condition (85) is said to be *accurately stated* if and only if

$$\operatorname{im} S = \operatorname{im} \Gamma, \quad \ker S = \ker \Pi_{can}(a). \tag{86}$$

Condition (86) requires $\operatorname{rank} S = l$ and $\dim \operatorname{im} \Gamma = l$. If the condition (86) is valid, then the operator $\mathcal{T}$ is a bijection between $\mathcal{C}^1_D(\mathcal{I}, \mathbb{R}^m)$ and $\mathcal{C}^{ind\,\mu}(\mathcal{I}, \mathbb{R}^m) \times \operatorname{im}\Gamma$. In comparison with the basic Definition 2.3 now the role of $\mathbb{R}^l$ is resumed by the l-dimensional subspace $\operatorname{im}\Gamma \subseteq \mathbb{R}^m$. The condition $\gamma \in \operatorname{im}\Gamma$ can be seen as trivial consistency condition.

If, additionally, the DAE has index 1, then $\mathcal{C}(\mathcal{I}, \mathbb{R}^m) = \mathcal{C}^{ind\,\mu}(\mathcal{I}, \mathbb{R}^m)$, and $\mathcal{T}$ has a bounded inverse. Then the inequality

$$\|x\|_\infty \leq \|x\|_{\mathcal{C}^1_D} \leq \|\mathcal{T}^{-1}\|(\|q\|_\infty + |\gamma|)$$

is satisfied by each arbitrary pair $(q,\gamma) \in \mathcal{C}(\mathcal{I},\mathbb{R}^m) \times \operatorname{im}\Gamma$ and the solution $x = \mathcal{T}^{-1}(q,\gamma)$. Then the BVP is said to be *well-posed* – in accordance with the basic Definition 2.2, with $\operatorname{im}\Gamma$ substituting $\mathbb{R}^l$.

Nonlinear versions of well-posed two-point BVPs for standard form index-1 DAEs and boundary conditions stated in $\mathbb{R}^m$ are treated, e.g., in [89, 55, 87, 91]. Linear and nonlinear multipoint BVPs for index-1 DAEs are studied [6, 4, 5]. Recall that in several early papers after [55] one speaks of *transferable* DAEs instead of (regular) index-1 DAEs. In [6, 4, 5], the BVP for an transferable DAE is said to be *regular* if the condition (86) is satisfied, and *irregular* otherwise. We do not resume this notation.

In [6] it is shown that well-posedness of multipoint BVPs persists under some special small perturbations.

If the DAE is regular with index 1, but the condition (86) does no longer hold, then the operator $\mathcal{T}$ has the closed image $\operatorname{im}\mathcal{T} = \mathcal{C}(\mathcal{I},\mathbb{R}^m) \times \operatorname{im} S$ with finite codimension $m - \operatorname{rank} S$ (cf.[4]). This means that $\mathcal{T}$ is actually a Fredholm operator (Noether operator in [4]) with $\operatorname{ind}_{fredholm} = l - m = -(m - r_0) = -\dim\ker D(a)$. A representation of the general solution of such an BVP including the resulting consistency condition is developed in [4] by means of projectors onto $\ker\mathcal{T}$ and $\operatorname{im}\mathcal{T}$.

We emphasize that for higher-index DAEs this approach does no longer work since then $\operatorname{im}\mathcal{T}$ fails to be closed in the given natural setting (cf. [96]).

By the initial condition of the form

$$Cx(a) = Cz, \quad \text{with} \quad z \in \mathbb{R}^m,$$

mostly written as $C(x(a) - z) = 0$, one trivially ensures the consistency condition $Cz =: \gamma \in \operatorname{im} C$. The component of z belonging to the nullspace of C does not impact the solution of the IVP.

The condition (86) simplifies here to $\ker C \cap \operatorname{im}\Pi_{can}(a) = \{0\}$. Recall that $\ker\Pi_{\mu-1}(a) = \ker\Pi_{can}(a)$ is valid for arbitrary admissible projector functions.

An important special case is given, if C is any matrix $C \in \mathcal{L}(\mathbb{R}^m,\mathbb{R}^m)$ such that $\ker C = \ker\Pi_{\mu-1}(a)$. Then this condition is evidently satisfied. In particular, one can put $C = \Pi_{\mu-1}(a)$ and $C = \Pi_{can}(a)$.

2.6.3 Separated boundary conditions

The boundary condition

$$G_a x(a) + G_b x(b) = \gamma \tag{87}$$

is said to be separated, if

$$G_a = \begin{bmatrix} G_{a,1} \\ 0 \end{bmatrix}, \quad G_b = \begin{bmatrix} 0 \\ G_{b,2} \end{bmatrix}.$$

Separated boundary conditions turn out to be pleasant. Exploiting this structure the computational costs of shooting algorithms can be reduced ([38]) and, furthermore, if the boundary conditions are placed in accordance with a dichotomy (see Theorem 2.3), then the conditioning constant κ_2 is moderate, thus the BVP is well-conditioned. Moreover, transfer methods relying on the description of solution subspaces (cf. Subsection 5.2) can be applied.

If the boundary condition (87) fails to be separated, then the BVP can be converted to an augmented BVP with separated boundary condition by the same trick used for ODEs, see [12, Section 1.1]. For this one can utilize if either G_a or G_b is rank deficient

Consider the BVP with boundary condition (87) of the form

$$G_a = \begin{bmatrix} G_{a,1} \\ G_{a,2} \end{bmatrix}, \quad G_b = \begin{bmatrix} 0 \\ G_{b,2} \end{bmatrix} \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^l), \quad G_{b,2} \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^{l-s}), \quad \gamma = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix},$$

with $0 \le s \le l$, for the regular index-μ DAE

$$A(Dx)' + Bx = q. \tag{88}$$

Introduce the additional function $z \in \mathcal{C}^1(\mathcal{I}, \mathbb{R}^{l-s})$ and add the equation $z' = 0$ to the DAE. The resulting DAE

$$\begin{bmatrix} A & 0 \\ 0 & I \end{bmatrix} \Big(\begin{bmatrix} D & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} \Big)' + \begin{bmatrix} B & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} = \hat{q} \tag{89}$$

is regular with index μ, too. The dynamical degree of freedom is $\hat{l} = l + l - s$. State for (89) separated boundary condition

$$\begin{bmatrix} G_a & K \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x(a) \\ z(a) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ G_{b,2} & I \end{bmatrix} \begin{bmatrix} x(b) \\ z(b) \end{bmatrix} = \hat{\gamma}, \quad K = \begin{bmatrix} 0 \\ -I \end{bmatrix} \in \mathcal{L}(\mathbb{R}^{l-s}, \mathbb{R}^l). \tag{90}$$

Letting

$$\hat{q} = \begin{bmatrix} q \\ 0 \end{bmatrix}, \quad \hat{\gamma} = \begin{bmatrix} \gamma_1 \\ 0 \\ \gamma_2 \end{bmatrix},$$

the function z becomes constant, thus $z(a) = z(b)$. Then the boundary condition (90) yields

$$\begin{aligned} G_{a,1} x(a) &= \gamma_1, \\ G_{a,2} x(a) - z(a) &= 0, \\ G_{b,2} x(b) + z(b) &= \gamma_2, \end{aligned}$$

and hence $G_a x(a) + G_b x(b) = \gamma$. Therefore, then the x-component of the solution of the BVP (89), (90) reproduces the solution of the original BVP (87), (88). If the boundary condition of the original BVP are stated accurately, then so are the boundary condition of the augmented one.

Another possibility to convert a BVP to a new one with separated boundary conditions is Moszyński's trick ([99]). We adapt this tool for converting the BVP (87), (88) to the augmented BVP on the half interval $[a, \frac{a+b}{2}]$,

$$\begin{bmatrix} A(t) & 0 \\ 0 & A(a+b-t) \end{bmatrix} \Big(\begin{bmatrix} D(t) & 0 \\ 0 & D(a+b-t) \end{bmatrix} \hat{x}(t) \Big)' + \begin{bmatrix} B & 0 \\ 0 & 0 \end{bmatrix} = \hat{q} \tag{91}$$

with

$$\hat{x}(t) = \begin{bmatrix} x(t) \\ x(a+b-t) \end{bmatrix}, \quad \hat{q}(t) = \begin{bmatrix} q(t) \\ q(a+b-t) \end{bmatrix}, \quad t \in [a, \frac{a+b}{2}],$$

and the separated boundary condition

$$\begin{bmatrix} G_a & G_b \\ 0 & 0 \end{bmatrix} \hat{x}(a) + \begin{bmatrix} 0 & 0 \\ C_{\frac{a+b}{2}} & -C_{\frac{a+b}{2}} \end{bmatrix} \hat{x}(b) = \begin{bmatrix} \gamma \\ 0 \end{bmatrix}, \tag{92}$$

with a matrix $C_{\frac{a+b}{2}} \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^l)$ such that $\ker C_{\frac{a+b}{2}} = \ker \Pi_{can}(\frac{a+b}{2})$. This manipulation changes neither the index of the DAE nor the accurateness of the boundary condition. The new solvability matrix is

$$\hat{S} = \begin{bmatrix} G_a X(a,a) & G_b X(b,a) \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ C_{\frac{a+b}{2}} X(\frac{a+b}{2}, a) & -C_{\frac{a+b}{2}} X(\frac{a+b}{2}, a) \end{bmatrix} \in \mathcal{L}(\mathbb{R}^{2m}, \mathbb{R}^{2l}). \tag{93}$$

The inclusion $\ker \Pi_{\mu-1}(a) \times \ker \Pi_{\mu-1}(a) \subseteq \ker \hat{S}$ is a consequence of the respective property of the fundamental solution matrix. On the other side, $\hat{S}z = 0$ yields $G_a X(a,a) z_1 + G_b X(b,a) z_2 = 0$ and $\Pi_{can}(a)(z_1 - z_2) = 0$, thus $(G_a X(a,a) + G_b X(b,a)) z_1 + G_b X(b,a)(z_2 - z_1) = S z_1 = 0$. Therefore, if $\ker S = \ker \Pi_{can}(a)$ then it follows that $z_1 \in \ker \Pi_{can}(a)$, further $z_1 \in \ker \Pi_{can}(a)$, and finally

$$\ker \hat{S} = \ker \Pi_{\mu-1}(a) \times \ker \Pi_{\mu-1}(a).$$

2.7 Further references, comments, and open questions

Remark 2.1 ($\mathcal{C}^1$-solutions). Often in the literature one insists on $\mathcal{C}^1$-solutions. This is less appropriate from a functional-analytic viewpoint as shown in detail in [96]. In any case, the basic structural characteristics of the DAE such as index, characteristic values, and regularity regions, are independent of the smoothness of the wanted solutions. *Occasional* additional smoothness requirements concerning the data imply each existing $\mathcal{C}^1_D$-solution also to belong to class $\mathcal{C}^1$.

The *axiomatic* use of $\mathcal{C}^1$-solutions, e.g., in [75, 74], necessitates additional smoothness requirements in principle. For instance, in the linear index-1 system, to ensure surjectivity in the respective setting $\mathcal{C}^1(\mathcal{I}, \mathbb{R}^m) \to \mathcal{C}(\mathcal{I}, \mathbb{R}^{m_1}) \times \mathcal{C}^1(\mathcal{I}, \mathbb{R}^{m_2})$,

$$\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} x'(t) + \begin{bmatrix} B_{11}(t) & B_{12}(t) \\ B_{21}(t) & B_{22}(t) \end{bmatrix} x(t) = \begin{bmatrix} q_1(t) \\ q_2(t) \end{bmatrix},$$

with $B_{22}(t)$ remaining nonsingular, they (have to) suppose that B_{21}, B_{22} as well as q_2 are continuously differentiable. Therefore, in this approach, q_2 can not serve as a control function being only continuous.

Remark 2.2 (The class of DAEs). Relations between DAEs in standard form (1) and DAEs showing a properly involved derivative (2) have been discussed at great length in [83, 96]. The setting with properly involved derivative indicates solutions from $\mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$. We emphasize that this are classical solutions; they satisfy the DAE at all points $t \in \mathcal{I}$. The present paper is concerned with a classical analytical theory and the respective numerical treatment.

We do not consider generalized solutions. To this end we mention that, for special DAEs, measurable solutions satisfying the DAE a.e. on $\mathcal{I}$ and distributional solutions are treated, e.g., in [56, 83, 54, 96, 112].

Remark 2.3 (Regularization). The structure of solutions of BVPs for certain linear index-1 and index-2 DAEs is investigated in [64] via regularization by singular perturbations. In particular, it is discussed how consistent boundary conditions can be stated. Already these case studies show the immense complexity of that approach. Further related profound studies concerning classes of linear and nonlinear BVPs are reported in [57, 58, 46, 59].

Remark 2.4 (Fundamental solution matrices). Given is a regular linear DAE (22) with index μ, $l = \operatorname{rank} \Pi_{can}(a)$, and $l \le k \le m$. Each matrix function $X : \mathcal{I} \to \mathcal{L}(\mathbb{R}^k, \mathbb{R}^m)$ with columns from $\mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$, satisfying

$$A(DX)' + BX = 0,$$

is said to be a *fundamental solution matrix* of the DAE if the relation

$$\operatorname{im} X(t) = \operatorname{im} \Pi_{can}(t), \quad t \in \mathcal{I},$$

is valid. One speaks of *maximal(-size)* and *minimal(-size) fundamental solution matrices* if $k = m$ and $k = l$, respectively. These notions have been introduced in [28] for index-1 DAEs in standard form and in [29] for properly stated DAEs with index 1 and index 2, and in [83] for regular DAEs with arbitrary index.

Given a maximal-size fundamental solution matrix X, a time $\bar{t} \in \mathcal{I}$, and a matrix $C \in \mathcal{L}(\mathbb{R}^l, \mathbb{R}^m)$ with full column-rank l such that

$$\ker X(\bar{t}) \cap \operatorname{im} C = \{0\}, \tag{94}$$

then the product XC is a minimal-size fundamental solution matrix and $X(\bar{t})C$ represents a basis of $\operatorname{im} \Pi_{can}(\bar{t})$. Namely, $X(\bar{t})Cz = 0$ implies $Cz = 0$, thus $z = 0$.

If, the maximal-size fundamental solution matrix X is normalized at $\bar{t}$ by $X(\bar{t}) = \Pi_{can}(\bar{t})$, then the above condition (94) simplifies to

$$\ker \Pi_{can}(\bar{t}) \cap \operatorname{im} C = \{0\}.$$

Conversely, if the given fundamental solution matrix X has minimal size and the matrix $C \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^l)$ has full row-rank l, the product XC is a maximal-size fundamental solution matrix. For getting an at $\bar{t}$ normalized one, we choose the special $C = X(\bar{t})^-$ being the generalized inverse of $X(\bar{t})$ defined by

$$\begin{aligned} X(\bar{t})X(\bar{t})^-X(\bar{t}) = X(\bar{t})^-, \quad X(\bar{t})^-X(\bar{t})X(\bar{t})^- = X(\bar{t})^-, \\ X(\bar{t})X(\bar{t})^- = \Pi_{can}(\bar{t}), \quad X(\bar{t})^-X(\bar{t}) = I. \end{aligned}$$

We have then $X(\bar{t})C = X(\bar{t})X(\bar{t})^- = \Pi_{can}(\bar{t})$ in fact.

A considerable part of the relevant former literature relies on minimal-size fundamental solution matrices, e.g. [38], whereas normalized maximal-size fundamental solution matrices are used in other parts, e.g. [55]. We mention that maximal-size fundamental solution matrices are applied for obtaining general solution representations for linear time-invariant DAEs by means of Drazin inverses and Wong sequences (e.g., [55, 112]).

The relations between the different fundamental solution matrices of a given DAE and those of the adjoint DAE are studied in [28, 29, 26]. A generalization for arbitrary index DAEs is open so far, it seems to be possible in the light of the projector based analysis.

Remark 2.5 (Shooting approach). The solvability matrix is often named *shooting matrix*. The shooting approach by maximal fundamental solution matrices for obtaining solvability results is already applied for nonlinear index-1 DAEs in [89, 55] and for linear standard form DAEs with arbitrary index in [91]. Here we present a comprehensive generalization for linear DAEs with arbitrary index by means of the projector-based analysis given in [83], which is straightforward within this framework. We also adress nonlinear DAEs.

Supposing in essence the solution structure (32) by a special involved solvability notion for linear DAEs, the shooting approach is justified for linear arbitrary index (standard form) DAEs in [38]. It is also pointed out that one has to provide the correct number of boundary conditions l, whereby l is determined by the investigation of the derivative array system. In contrast to our approach, an arbitrary minimal fundamental solution matrix $\psi(t,a) \in \mathcal{L}(\mathbb{R}^l, \mathbb{R}^m)$ which has full column-rank is applied in [38] instead of the maximal solution matrix $X(t,a) \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^m)$. Because of the relations $X(t,a)W = \psi(t,a)$, with a full column-rank constant matrix $W \in \mathcal{L}(\mathbb{R}^l, \mathbb{R}^m)$, this yields the quadratic solvabilty matrix

$$\tilde{S} := G_a\psi(a,a) + G_b\psi(b,a) = (G_aX(a,a) + G_bX(b,a))W = SW,$$

which depends on W, that is on the chosen basis $\psi(a)$ of $\operatorname{im}\Pi_{can}(a)$. Nevertheless, $\tilde{S}$ is nonsingular, if and only if $\ker S = \ker \Pi_{can}(a)$.

The approach in [111] repeats and extends that of [38] on the slightly different background of the strangeness-index regularization concept. In particular, a basis $\psi(a)$ of the subspace $\operatorname{im}\Pi_{can}(a)$ with orthogonal columns is constructed.

Remark 2.6 (Well-posedness). Well-posedness and ill-posedness are traditionally named *correctness* and *noncorrectness* in the Russian literature.

The Definition 2.2 constitutes a local specification of Hadamard's well-posedness notion. It has been used, e.g., in [55, 90]. Actually, it says that the operator representing the BVP in its natural setting as operator equation is a local diffeomorphism at x_*. (cf. [90, 96]). General nonlinear BVPs for index-1 DAEs with accurately stated boundary conditions are shown to be well-posed in [90] and the ill-posedness for DAEs with higher index is indicated.

In [111] one can find a further proof of well-posedness in the natural setting for the so-called *regularized BVP*, which consists of a special form index-1 DAE and appropriate boundary conditions.
In [75] well-posedness of BVPs for index-1 DAEs in reduced form (100), (101) is obtained in the setting (cf. also Remarks 2.1 and 2.7)

$$\mathcal{C}^1(\mathcal{I},\mathbb{R}^m) \to \mathcal{C}(\mathcal{I},\mathbb{R}^l) \times \mathcal{C}^1(\mathcal{I},\mathbb{R}^a) \times \mathbb{R}^l,$$

which looks about $\mathcal{C}^1$- solutions.

A different well-posed notion purpose-built for Hessenberg form DAEs describing multibody systems is agreed upon in [51]. There certain components of the perturbation are set to be zero.

Remark 2.7 (Isolated solvability). We conjecture that, if the solution x_* of the BVP is located within a regularity region of the DAE, then x_* is locally unique exactly if it is isolated in the sense of Definition 2.5.

Till now, explicit proofs are known for the general index-1 case and also for higher-index cases under certain structural restrictions. Such a result is obtained in [51] for periodic solutions of multibody DAEs. The hitherto applied structural restrictions become more and more annoying with increasing index, see Subsection 2.5, [83, Remark 4.5]. It is open to what extent one can do without those restrictions.

Of course, if the original DAE can be reduced locally around the wanted solution x_* to an index-1 DAE possessing the same solutions as the original DAE, and if x_* is an isolated solution of the reduced BVP, then x_* is at the same time an locally unique solution of the original BVP. Unfortunately, this fine idea is not quite easy to be predicated on precise criteria. With the notion of a *regular solution of the BVP* the authors of [75] attempt to provide such a criterion. We take a closer look.

In [75, 74], nonlinear BVPs

$$\mathfrak{f}(x'(t),x(t),t) = 0, \quad t \in \mathcal{I} = [a,b], \tag{95}$$

$$g(x(a),x(b)) = 0, \tag{96}$$

with sufficiently smooth data, are addressed by means of the strangeness-index reduction framework. A solution is defined to be a sufficiently smooth function $x_* \in \mathcal{C}(\mathcal{I},\mathbb{R}^m)$ satisfying the system

$$\mathfrak{f}(x_*'(t), x_*(t), t) = 0, \quad t \in \mathcal{I}, \tag{97}$$

$$\mathfrak{f}_\mu(\mathcal{P}_*(t), x_*(t), t) = 0, \quad t \in \mathcal{I}, \tag{98}$$

$$g(x_*(a), x_*(b)) = 0, \tag{99}$$

where f_μ denotes the derivative array function and $\mathcal{P}_* : \mathcal{I} \to \mathbb{R}^{m(\mu+1)}$ is some smooth function that coincides with $x_*'(t)$ in its first m components. Under quite involved hypotheses, there are functions Z_{*1}, Z_{*2}, and K_*, all depending on x_*, such that the reduced system of $l + a = m$ equations

$$\hat{\mathfrak{f}}_1(x'(t), x(t), t) = 0, \tag{100}$$

$$\hat{\mathfrak{f}}_2(x(t), t) = 0, \quad t \in \mathcal{I}, \tag{101}$$

results, with

$$\hat{\mathfrak{f}}_1(x^1, x, t) := Z_{*1}(t)^T \mathfrak{f}(x^1, x, t),$$
$$\hat{\mathfrak{f}}_2(x, t) := Z_{*2}(t)^T \mathfrak{f}_\mu(K_*(x, t), x, t).$$

The reduced system has index 0 if $a = 0$, and otherwise index 1.

Then the solution x_* is said to be a *regular solution of the original BVP* ([75]), if the linearized at x_* *reduced* BVP has the trivial solution only. In our context this means that the reduced BVP is locally well-posed in the related setting (cf. Remark 2.6). However, this does not say that the original BVP is well-posed! On this background the claim ([75]) that *the original BVP* (95), (96) *takes the form* of the operator equation $\mathcal{F}(x) = 0$, with $\mathcal{F}$ acting in Banach spaces X and Y (perhaps $X = \mathcal{C}^1(\mathcal{I}, \mathbb{R}^m)$, $Y = \mathcal{C}(\mathcal{I}, \mathbb{R}^l) \times \mathcal{C}^1(\mathcal{I}, \mathbb{R}^a) \times \mathbb{R}^l$),

$$\mathcal{F}(x)(t) := \begin{bmatrix} \hat{\mathfrak{f}}_1(x'(t), x(t), t) \\ \hat{\mathfrak{f}}_2(x(t), t) \\ g(x(a), x(b)) \end{bmatrix}, \quad t \in \mathcal{I},$$

with a bijective Fréchet derivative $\mathcal{F}'(x_*)$, becomes rather misleading.

As the specific feature of derivative array approaches all involved derivatives are prepared analytically. This is, in the linear case, comparable to preparing analytically the functions v_q in Subsections 2.2, 2.3.

Remark 2.8 (Segregation of solution subspaces by means of the adjoint equation). Similarly as it is well-known for explicit ODEs, any affine linear subspace of solutions within the whole solution set of a regular index-μ DAE, $\mu = 1$ or $\mu = 2$, can be segregated by means of solutions of the homogeneous adjoint DAE ([27, 101, 100, 30]). Thereby, the interval $\mathcal{I}$ is not necessarily compact. The generalization for arbitrary high index seems to be possible. We quote the main result from [30].

Let the DAE (51) be regular with index 1 or index 2, and the right-hand side q be admissible, $l = \operatorname{rank} \Pi_{can}(a)$, $1 \le k \le l$, $s = l - k$.

Then a set $\mathfrak{L} \subset \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ is a k-dimensional affine-linear subspace of solutions of the DAE if and only if it is described by

$$L(t) = \{x \in \mathbb{R}^m : Y(t)^* A(t) D(t) x + v(t) = 0,\ x \in \mathcal{M}_{\mu-1}(t)\}, \quad t \in \mathcal{I},$$

with matrix functions $Y : \mathcal{I} \to \mathcal{L}(\mathbb{R}^s, \mathbb{R}^m)$, $v : \mathcal{I} \to \mathbb{R}^s$ such that $\operatorname{rank} Y(t) = s$ and

$$\begin{aligned} -D^*(A^*Y)' + B^*Y &= 0, \\ v' + Y^*q &= 0. \end{aligned}$$

Linear BVPs for explicit ODEs with separated boundary conditions can be successfully solved by so-called transfer methods. Relying on the above representation, correspondent methods can be created for DAEs, see Subsection 5.2.

Remark 2.9 (Conditioning constants). In [12], dealing with BVPs for explicit ODEs, the constant κ in the inequality (11) and the quantities κ_1, κ_2 introduced in Subsection 2.3 are called *conditioning constants*. If they have moderate size, then one speaks of *well-conditioned* BVPs.

Our presentation in Subsections 2.3 and 2.4 are generalizations of the results in [87, 88] by means of the projector-based analysis from [83].

It should be emphasized once again, that there is an essentially different meaning of the conditioning constants for index-1 DAEs and higher-index DAEs. For index-1 DAEs, the quantities $\kappa_1, \kappa_2, \kappa_3$ can be seen as specification of κ, which is in turn actually a bound of the inverse of the operator representing the BVP in its natural setting.

In case of higher index DAEs the BVP is necessarily ill-posed in the natural setting such that a constant κ no longer does exist, but $\kappa_1, \kappa_2, \kappa_3$ do exist and can show moderate size. Therefore, for higher-index DAEs, it may happen that a BVP can be ill-posed but well-conditioned! One should avoid confusions! Here, a well-conditioned BVP is given, if the boundary condition fits well to the dynamic part of the DAE. Thereby, the index does not matter.

Special studies concerning the conditioning constants of IVPs for linear Hessenberg index-1 and index-2 DAEs, their sensitivity with respect to several small parameters are described in [114]. The conditioning of BVPs for index-2 Hessenberg systems is adressed in [8] by means of the reduction to the *essential underlying ODE*. In essence, in our context this means well-posedness in the advanced index-2 setting along with conditioning constants $\kappa_1, \kappa_2, \kappa_3$ of moderate size.

Remark 2.10 (Structural restriction in Theorem 2.7). Consider the nonlinear DAE (54). If the reference solution x_* resides within an index-μ regularity region, then the linearized along x_* DAE (55) is regular with index μ, too, see Appendix 6.1.3.

The converse is true only for the index-1 case. If the linearized along x_* DAE (55) is regular with index 1, then there is a neighborhood of the graph of x_* being an index-1 regularity region.
In contrast, it may well happen that (55) is regular with index $\mu \geq 2$, but there is no regularity region housing the graph of x_*, e.g., [95, 83].

If $\mu = 2$, Then the additional condition (70) ensures that x_* resides in a regularity region ([92]). We think that the same is true also for arbitrary $\mu \geq 2$, but a correct proof is not yet available.

There arises a challenging question: To what extend the objectionable condition (70) could be replaced by the requirement that x_* resides within a regularity region? Till now, no idea is in sight.

Remark 2.11 (Scaling of the DAE). It is convenient to analyze the regular implicit ODE $A(t)x'(t)+B(t)x(t) = 0$ in the explicit form $x'(t)+A(t)^{-1}B(t)x(t) = 0$. However, for practical computations one usually prefers the implicit form.

As mentioned in Subsection 2.3, the scaling of a given regular index-μ DAE by G_μ^{-1} leads, for the scaled DAE, to $G_\mu = I$. As for regular implicit ODEs, it is unlikely that this fact could be qualified to practical consequences.

Nevertheless, some useful basic scalings would be welcome for both implicit regular ODEs and regular DAEs. As yet, there is no idea in sight.

Remark 2.12 (Inequalities (11) and (10)). In the context of BVPs for explicit ODEs (e.g., [12]), with good cause, one commonly uses the practically more convenient norm $\|\cdot\|_\infty$ instead of $\|\cdot\|_{\mathcal{C}^1}$. Analogously, one is allowed to replace the inequality (10) by the simpler inequality (11) by the following arguments: The inequality (10) immediately implies (11), that is,

$$\|x-x_*\|_\infty \leq \|x-x_*\|_{\mathcal{C}^1_D} \leq \kappa(|\gamma|+\|q\|_\infty).$$

Conversely, (10) follows from (11). Namely, for $x \in \mathcal{B}_{\mathcal{C}^1_D}(x_*,\rho)$, the identities

$$f((Dx_*)'(t),x_*(t),t) = 0, \quad f((Dx)'(t),x(t),t) = q(t), \quad t \in \mathcal{I},$$

imply

$$A_{[x,x_*]}(t)(Dx-Dx_*)'(t)+B_{[x,x_*]}(t)(x(t)-x_*(t)) = q(t), \quad t \in \mathcal{I}, \tag{102}$$

with uniformly bounded coefficients

$$A_{[x,x_*]}(t) := \int_0^1 f_y((Dx_*)'(t)+s((Dx)'(t)-(Dx_*)'(t),x_*(t)+s(x(t)-x_*(t)),t)\mathrm{d}s,$$
$$B_{[x,x_*]}(t) := \int_0^1 f_x((Dx_*)'(t)+s((Dx)'(t)-(Dx_*)'(t),x_*(t)+s(x(t)-x_*(t)),t)\mathrm{d}s.$$

We have $\operatorname{rank}A_{[x,x_*]}(t) \leq \operatorname{rank}A_*(t) = \operatorname{rank}D(t) = r$ because of $\ker A_{[x,x_*]}(t) = \ker R(t)$ and, if ρ is sufficiently small, $\operatorname{rank}A_{[x,x_*]}(t) \geq \operatorname{rank}A_*(t) = \operatorname{rank}D(t) = r$, since

$$A_{[x,x_*]}(t) = A_*(t)+R_{[x,x_*]}(t), \quad |R_{[x,x_*]}(t)| \leq k_0\|x-x_*\|_{\mathcal{C}^1_D} \leq k_0\rho,$$

and hence

$$\ker A_{[x,x_*]}(t) = \ker A_*(t) = \ker R(t) = \operatorname{im}D(t), \quad \operatorname{rank}A_{[x,x_*]}(t) = r, \quad t \in \mathcal{I}.$$

Choosing a continuous generalized inverse $A_{[x,x_*]}(t)^-$ such that $A_{[x,x_*]}(t)^- A_{[x,x_*]}(t) = R(t)$ and multiplying equation (102) by $A_{[x,x_*]}(t)^-$ leads to

$$R(t)(Dx - Dx_*)'(t) + A_{[x,x_*]}(t)^- B_{[x,x_*]}(t)(x(t) - x_*(t)) = A_{[x,x_*]}(t)^- q(t), \quad t \in \mathcal{I},$$

further

$$\begin{aligned}(Dx - Dx_*)'(t) - R'(t)(Dx - Dx_*)(t) + A_{[x,x_*]}(t)^- B_{[x,x_*]}(t)(x(t) - x_*(t)) \\ = A_{[x,x_*]}(t)^- q(t), \quad t \in \mathcal{I},\end{aligned}$$

and then

$$\|(Dx - Dx_*)'\|_\infty \leq k_1 \|(x - x_*\|_\infty + k_2 \|q\|_\infty.$$

Regarding (11) we finally obtain

$$\|x - x_*\|_{\mathcal{C}_D^1} \leq (k_1 + 1)\kappa(|\gamma| + \|q\|_\infty) + k_2 \|q\|_\infty \leq K(|\gamma| + \|q\|_\infty).$$

The same arguments apply also to the respective inequalities associated with well-posedness in advanced settings.

3 Collocation methods for well-posed BVPs

Piecewise polynomial collocation is an accepted method to approximately solve classical well-posed BVPs for regular ODEs. Several general purpose codes are implemented, which have been successfully applied to a great variety of practical problems. For instance, the package COLSYS ([10]) and its later modification COLNEW ([21, 12]) can be used to solve multipoint boundary value problems for mixed-order systems of explicit ODEs. This leads to the idea to treat additional constraints, i.e., derivative-free equations, as zero-order ODEs as it is done in [62] for semi-explicit DAEs

$$x_1'(t) + k_1(x_1(t), x_2(t), t) = 0, \tag{103}$$
$$k_2(x_1(t), x_2(t), t) = 0, \tag{104}$$

with index 1. The package COLDAE ([13]) also plays on this approach, but now for a wider class of DAEs. The MATLAB code BVPSUITE ([15]) is designed to solve systems of implicit ordinary differential equations of arbitrary order including order zero, which includes an implicit version of (103).

We restrict our interest to two-point BVPs and refer to Subsection 2.6 for other boundary conditions.

As pointed out in Subsection 2.5, BVPs for DAEs may be locally well-posed in different senses: in the natural setting, in the advanced setting and in the setting associated to the special reduced form, see Remark 2.6,

$$\mathfrak{f}_1(x'(t), x(t), t) = 0, \tag{105}$$
$$\mathfrak{f}_2(x(t), t) = 0, \tag{106}$$

which inter alia arises by reduction from derivative array systems (e.g., [75]). Additionally to the regular DAEs we consider also singular index-1 DAEs featuring a singular inherent explicit ODE. In the latter case, it is more difficult to state the boundary conditions and to achieve well-posedness.

The semi-explicit DAE (103), (104) indicates the different smoothness of the first and second components, which can be reasonably resumed for their approximations (e.g., [62, 73, 22, 42, 13]). A useful generalization of this class of DAE is given by DAEs with properly involved derivatives

$$f((Dx)'(t), x(t), t) = 0, \tag{107}$$

which satisfy the basic assumption from Subsection 2.1, and, additionally,

$$\operatorname{im} D(t) = \mathbb{R}^n, \quad t \in [a,b] = \mathcal{I}, \tag{108}$$

which leads to the border projector $R(t) \equiv I$. Then the enlarged DAE

$$f(u'(t),x(t),t) = 0, \tag{109}$$
$$u(t) - D(t)x(t) = 0. \tag{110}$$

features partitioned variables. For each solution x_* of (107), the pair (Dx_*, x_*) solves the enlarged DAE. Conversely, if (u_*, x_*) is a solution of the enlarged DAE, then the component x_* is a solution of (107).

Furthermore, the enlarged DAE is regular with index 1, exactly if the original DAE is regular with index 1. It can be checked by straightforward computations, that, in the index-1 case, both DAEs have the same IERODE

$$u'(t) = D(t)\omega(u(t),t) \tag{111}$$

and the dynamical degree of freedom $l = n = \operatorname{rank} D(a)$. Thereby, ω is the decoupling function introduced in Subsubsection 2.5.1 for index-1 problems.

With the boundary condition

$$g(x(a),x(b)) = 0, \tag{112}$$

a wellposed BVP (107), (112), yields a well-posed BVP (109), (110), (112), and vice versa.

As pointed out in [61], [83, Chapter 5], in the context of integration methods, it is reasonable to turn to models with constant border projector, so-called *numerically qualified DAEs* and to arrange numerical approximations via the enlarged DAE. Owing to the time-invariance of the border projector, the methods are transferred to the IERODE with no mutation. Otherwise the methods might change substantially, for instance, the implicit Euler method might be converted into its explicit counterpart.

For the collocation methods, we define meshes

$$\pi := \{a = t_0 < t_1 < \cdots < t_i < t_{i+1} < \ldots < t_N = b\},$$

with stepsizes $h_i := t_{i+1} - t_i,\ i = 0,\ldots,N-1$. We allow equidistant meshes $h_i = h,\ i = 0,\ldots,N-1$, and non-uniform meshes which have a limited variation in the step sizes, i.e.,

$$h := \max_{i=0,\ldots,N-1} h_i \le \kappa \min_{i=0,\ldots,N-1} h_i,$$

with a general constant κ.

In each subinterval $[t_i, t_{i+1}]$ we insert s collocation points $\tau_{ik} := t_i + h_i\rho_k$, $k = 1,\ldots,s$, using s distinct canonical points

$$0 \le \rho_1 < \cdots < \rho_s \le 1.$$

A grid with equidistant interior collocation points is illustrated in Figure 6.

We denote by $\mathcal{B}^j_{\pi,s}$ the linear space of vector-valued functions with j components given on $[a,b]$ so that, according to the mesh π, each component is a piecewise polynomial function of degree $\le s$. To be precise, we agree upon right continuity at the mesh points $t_0,\ldots,t_{N-1}$.

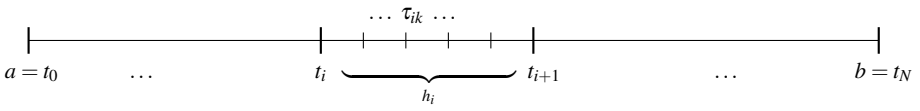

Fig. 6 The computational grid

An attractive feature of collocation schemes is their possible high accuracy at the mesh points $t_0, \ldots, t_N$, called *superconvergence* ([39]). For classical BVPs in regular first order ODEs one usually approximates the solution by continuous piecewise polynomial functions. This leads to a uniform error order s. Depending on the canonical collocation points, the order at the mesh points can be higher. More precisely, if there is an integer $s < s_+ \leq 2s$, and the canonical collocation points $\rho_1 < \cdots < \rho_s$ satisfy the orthogonality relations

$$\int_0^1 t^j \prod_{i=1}^{s} (t - \rho_i) \mathrm{d}t = 0, \quad j = 0, \ldots, s_+ - s, \tag{113}$$

then s_+ is the superconvergence order in the context of nonstiff regular explicit ODEs. For instance, one has $s_+ = 2s$ for Gauß schemes, $s_+ = 2s - 1$ for Radau schemes, and $s_+ = 2s - 2$ for Lobatto schemes ([39]).

There is a variety of possible collocation approaches for DAEs. As emphasized in [13], collocating the differential components by continuous piecewise polynomial functions and allowing generally discontinuous piecewise polynomial functions for the algebraic components is most natural, see Subsubsection 3.1.1 and Subsection 3.3. Alternative approaches suppose continuous (Approach A in Subsubsection 3.1.2, Approach C in Subsubsection 3.1.4, and Subsection 3.2) or discontinuous (Approach B in Subsubsection 3.1.3) piecewise polynomial functions uniformly for all components.
In contrast to regular ODEs, any solution x_* of a DAE proceeds within the so-called obvious constraint set of the DAE, $x_*(t) \in \mathcal{M}_0(t)$ for all t. This leads to the extra question in the context of DAEs whether the approximation values $x_\pi(t_i)$ are consistent, that means $x_\pi(t_i) \in \mathcal{M}_0(t_i)$.

3.1 BVPs being well-posed in the natural setting

Let the BVP (107), (112), satisfy the basic assumptions described in Subsection 2.1, let the DAE have a properly involved derivative, and let (108) be valid. Let x_* denote the wanted solution, and $u_* = Dx_*$.
Theorem 2.4 provides exact criteria for the local well-posedness in the natural setting. Therefore, we assume that the DAE is regular with index 1, the boundary conditions are stated accurately and $l = \operatorname{rank} D(a)$.

3.1.1 Partitioned component approximation

We continue considering the well-posed BVP (107), (112) by means of the enlarged version (109), (110), (112). Let $u_\pi \in \mathcal{B}^n_{\pi,s} \cap \mathcal{C}(\mathcal{I},\mathbb{R}^n)$ and $x_\pi \in \mathcal{B}^m_{\pi,s-1}$ serve as approximations of u_* and x_*, respectively. The required continuity of u_π means

$$u_\pi(t_i^-) = u_\pi(t_i), \quad i = 1,\dots,N-1, \tag{114}$$

and therefore, we have to determine $n(s+1)N + msN - n(N-1) = (n+m)sN + n$ remaining unknowns. The boundary condition (112) yields

$$g(x_\pi(a), x_\pi(b)) = \gamma, \tag{115}$$

which contains n equations. To create a balanced system, we apply the $(n+m)sN$ collocation equations

$$f(u'_\pi(\tau_{ik}), x_\pi(\tau_{ik}), \tau_{ik}) = 0, \tag{116}$$

$$u_\pi(\tau_{ik}) - D(\tau_{ik}) x_\pi(\tau_{ik}) = 0, \quad k = 1,\dots,s,\ i = 0,\dots,N-1. \tag{117}$$

If $\rho_1 = 0$, then $u'_\pi(\tau_{i1})$ is the right derivative at $\tau_{i1} = t_i$, if $\rho_s = 1$, then $u'_\pi(\tau_{is})$ is defined as left derivative at $\tau_{is} = t_{i+1}$.

By means of the decoupling function the scheme (116), (117) transforms to

$$x_\pi(\tau_{ik}) = D(\tau_{ik})^- u_\pi(\tau_{ik}) + Q_0(\tau_{ik}) \omega(u_\pi(\tau_{ik}), \tau_{ik}), \tag{118}$$

$$u'_\pi(\tau_{ik}) = D(\tau_{ik}) \omega(u_\pi(\tau_{ik}), \tau_{ik}), \quad k = 1,\dots,s,\ i = 0,\dots,N-1. \tag{119}$$

On the other hand, we are given the solution representation (cf. (63), (64))

$$x_*(t) = D(t)^- u_*(t) + Q_0(t) \omega(u_*(t), t), \tag{120}$$

$$u'_*(t) = D(t) \omega(u_*(t), t), \quad t \in [a,b]. \tag{121}$$

In particular, u_* satisfies the IERODE (111). Obviously, the collocation scheme (116), (117), (114) results in the classical collocation scheme for the IERODE subject to the boundary conditions. Therefore, u_π is uniquely determined, and, in turn, x_π is also unique by (118).

The next theorem represents a straightforward extension of [22, Theorem 3.2] which concerns semi-explicit index-1 DAEs. It can be proved analogously.

Theorem 3.1. *Let the BVP (107), (112) be well-posed locally around its solution x_* in the natural setting. Let condition (108) hold and the data of the DAE be sufficiently smooth for respective order conditions.*
Then, for the collocation scheme (116), (117), (115), (114), the following statements hold:

(1) *There is a $h_* > 0$, such that, for meshes with $h \le h_*$, there exists a unique collocation solution u_π, x_π in the sufficiently close neighborhood of u_*, x_*.*

(2) *With a sufficiently good initial guess, the collocation solution can be generated by the Newton method, which converges quadratically.*
(3) *It holds that*

$$\|x_* - x_\pi\|_\infty = O(h^s), \quad \|u_* - u_\pi\|_\infty = O(h^s).$$

(4) *If there is an integer $s < s_+ \le 2s$, and the canonical collocation points satisfy the orthogonality relations (113), then the superconvergence property*

$$\max_{i=0,\dots,N} |u_*(t_i) - u_\pi(t_i)| = O(h^{s_+})$$

holds for the smooth component.
(5) *If $\rho_1 = 0$, $\rho_s = 1$, then the approximations become smoother. More precisely, $u_\pi \in \mathcal{B}^n_{\pi,s} \cap \mathcal{C}^1(\mathcal{I},\mathbb{R}^n)$ and $x_\pi \in \mathcal{B}^m_{\pi,s-1} \cap \mathcal{C}(\mathcal{I},\mathbb{R}^n)$.*
(6) *For Lobatto points the superconvergence applies to all components,*

$$\max_{i=0,\dots,N} |x_*(t_i) - x_\pi(t_i)| = O(h^{2s-2}).$$

Except for methods with canonical points $\rho_1 = 0$, $\rho_s = 1$, such as Lobatto methods, the generated values at mesh points $x_\pi(t_i)$ do not necessarily belong to the obvious constraint $\mathcal{M}_0(t_i)$, that means, they may fail to be consistent. This might be seen to be a drawback. For methods with canonical points $\rho_1 > 0$, $\rho_s = 1$, such as the Radau IIA method, one obtains $x_\pi(t_i) \in \mathcal{M}_0(t_i)$ for $i > 0$. This is widely appreciated in the context of numerical integration.

3.1.2 Uniform approach A

Again we consider the well-posed BVP (107), (112) by means of the enlarged version (109), (110), (112). Now we approximate all components by *continuous* piecewise polynomials of the *same degree*. Let $u_\pi \in \mathcal{B}^n_{\pi,s} \cap \mathcal{C}(\mathcal{I},\mathbb{R}^n)$ and $x_\pi \in \mathcal{B}^m_{\pi,s} \cap \mathcal{C}(\mathcal{I},\mathbb{R}^m)$ serve as approximations of u_* and x_*, respectively. The required continuity means

$$u_\pi(t_i^-) = u_\pi(t_i), \quad x_\pi(t_i^-) = x_\pi(t_i), \quad i = 1,\dots,N-1, \tag{122}$$

and we have to determine $(n+m)(s+1)N - (n+m)(N-1) = (n+m)(sN+1)$ further coefficients. The boundary condition (112) contains n equations. We now apply the $(n+m)sN$ collocation equations and the boundary conditions

$$f(u'_\pi(\tau_{ik}), x_\pi(\tau_{ik}), \tau_{ik}) = 0, \tag{123}$$
$$u_\pi(\tau_{ik}) - D(\tau_{ik}) x_\pi(\tau_{ik}) = 0, \quad k = 1,\dots,s,\ i = 0,\dots,N-1, \tag{124}$$
$$g(x_\pi(a), x_\pi(b)) = \gamma, \tag{125}$$

with $\rho_1 > 0$. If $\rho_s = 1$, then $u'_\pi(\tau_{is})$ is defined as the left derivative at $\tau_{is} = t_{i+1}$.

By inspection, we see that m further conditions are necessary to close the system for the numerical treatment and these additional conditions have to be consistent with the original DAEs. For this purpose we introduce a matrix function $\tilde{W}(y,x,t) \in \mathcal{L}(R^m, \mathbb{R}^{m-n})$ such that $\ker \tilde{W}(y,x,t) = \operatorname{im} f_y(y,x,t)$ and complete the above scheme by the following $n+(m-n)=m$ equations:

$$D(a)x_\pi(a) - u_\pi(a) = 0, \quad \tilde{W}(u'_\pi(a), x_\pi(a), a) f(u'_\pi(a), x_\pi(a), a) = 0. \tag{126}$$

Observe that $\rho_1 = 0$ would lead to $\tau_{01} = t_0 = a$ and cause the second part of the consistency condition (126) and the collocation (123) for $i = 0$, $k = 1$ to become redundant.

If the DAE is given with separated derivative free equations

$$\begin{aligned} f_1((D(t)x(t))', x(t), t) &= 0, \\ f_2(x(t), t) &= 0, \end{aligned}$$

where f_1 and f_2 have n and $m-n$ components, respectively, then we can augment the scheme by

$$D(a)x_\pi(a) - u_\pi(a) = 0, \quad f_2(x_\pi(a), a) = 0.$$

We observe that, $\rho_1 = 0$ yields $\tau_{01} = t_0 = a$. Again, the equations (123), (124) can be decoupled,

$$x_\pi(\tau_{ik}) = D(\tau_{ik})^- u_\pi(\tau_{ik}) + Q_0(\tau_{ik}) \omega(u_\pi(\tau_{ik}), \tau_{ik}), \tag{127}$$

$$u'_\pi(\tau_{ik}) = D(\tau_{ik}) \omega(u_\pi(\tau_{ik}), \tau_{ik}), \quad k = 1, \ldots, s, \; i = 0, \ldots, N-1. \tag{128}$$

Therefore, the related equations from (123), (124), (122), (125) result in the classical collocation scheme for the IERODE, and hence, u_π is uniquely determined. In turn, for given u_π, the approximation x_π is uniquely determined by the conditions (127), (126) together with the continuity conditions (122).

The following theorem is a byproduct of the investigations in [71, 43] which were originally devoted to problems featuring a singularity at $t = a$. An analogous result is valid for $\rho_s < 1$ instead of $\rho_1 > 0$, if one states condition (126) accordingly at the right interval end b.

Theorem 3.2. *Under the assumptions of Theorem 3.1 , the following statements hold for the collocation scheme (123), (124), (125), (122), (126) with $\rho_1 > 0$:*

(1) *There is a $h_* > 0$, such that, for meshes with $h \le h_*$, there exists a unique collocation solution u_π, x_π in the sufficiently close neighborhood of u_*, x_*.*
(2) *For a sufficiently good initial guess, the collocation solution can be generated by the Newton method, which converges quadratically.*
(3) *It holds*

$$\|x_* - x_\pi\|_\infty = O(h^s), \quad \|u_* - u_\pi\|_\infty = O(h^s).$$

At the time being, there is only an experimental observation of the superconvergence properties described below. The analysis of this aspect of the collocation is still

missing. For collocation points satisfying (113) the following observation have been made:

$$\|x_* - x_\pi\|_\infty = O(h^s), \quad \|u_* - u_\pi\|_\infty = O(h^{s+1})$$

and

$$|u_*(t_i) - u_\pi(t_i)| = 0(h^{s+}), \quad i = 0, \ldots, N.$$

In case of a sufficiently smooth solution x_*, its global error for s equidistant collocation points is $O(h^s)$ uniformly in t, while for Gauß and Radau points, the global error is $O(h^{s+1})$ uniformly in t. For the global error concerning the part u, the superconvergence order seems to hold, at least for Radau points. Clearly, when the solution of the problem is not sufficiently smooth, order reductions are observed, in line with classical collocation theory.

Example 3.1. The BVP

$$\begin{bmatrix} 1 & -t & t^2 \\ 0 & 1 & -t \\ 0 & 0 & 0 \end{bmatrix} x'(t) + \begin{bmatrix} 1 & -(t+1) & t^2+2t \\ 0 & -1 & t-1 \\ 0 & 0 & 1 \end{bmatrix} x(t) = \begin{bmatrix} 0 \\ 0 \\ \sin t \end{bmatrix}, \quad t \in \mathcal{I} = [0,1],$$
$$x_1(0) = 1,$$
$$x_2(1) - x_3(1) = \mathrm{e},$$

serves as test problem in [38]. The unique solution is

$$x_1(t) = \mathrm{e}^{-t} + t\mathrm{e}^t, \quad x_2(t) = \mathrm{e}^t + t\sin t, \quad x_3(t) = \sin t.$$

We used the equivalent formulation of the DAE with properly stated leading term

$$\begin{bmatrix} 1 & -t \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -t \end{bmatrix} x \right)' (t) + \begin{bmatrix} 1 & -(t+1) & t^2+t \\ 0 & -1 & t \\ 0 & 0 & 1 \end{bmatrix} x(t) = \begin{bmatrix} 0 \\ 0 \\ \sin t \end{bmatrix}.$$

The DAE is regular with index 1, the boundary conditions are accurately stated, and the BVP is well-posed. In [38] the implicit midpoint rule is applied, and it is

s=1,uniform		gex_π		gex_{unif}	
N	h	error	order	error	order
40	0.025	2.92676e-04	1.999	4.53226e-04	1.976
80	0.0125	7.31839e-05	2.000	1.14392e-04	1.986
160	0.00625	1.82969e-05	2.000	2.87355e-05	1.993
320	0.00313	4.57429e-06	2.000	7.09675e-06	2.018

s=1,uniform		geu_π		geu_{unif}	
N	h	error	order	error	order
40	0.025	2.02790e-04	2.000	4.53226e-04	1.976
80	0.0125	5.06967e-05	2.000	1.14392e-04	1.986
160	0.00625	1.26741e-05	2.000	2.87355e-05	1.993
320	0.00313	3.16853e-06	2.000	7.09675e-06	2.018

Table 1

reported that the error behaves consistently as $O(h^2)$. Tables 1–6 show the results generated by the collocation scheme (122)–(125), for $s = 1, 2, 3$, each with uniform and Gauß collocation. gex_π and geu_π denote the maximal global errors in the mesh

s=1,gaussian		gex_π		gex_{unif}	
N	h	error	order	error	order
40	0.025	2.92676e-04	1.999	4.53226e-04	1.976
80	0.0125	7.31839e-05	2.000	1.14392e-04	1.986
160	0.00625	1.82969e-05	2.000	2.87355e-05	1.993
320	0.00313	4.57429e-06	2.000	7.09675e-06	2.018

s=1,gaussian		geu_π		geu_{unif}	
N	h	error	order	error	order
40	0.025	2.02790e-04	2.000	4.53226e-04	1.976
80	0.0125	5.06967e-05	2.000	1.14392e-04	1.986
160	0.00625	1.26741e-05	2.000	2.87355e-05	1.993
320	0.00313	3.16853e-06	2.000	7.09675e-06	2.018

Table 2

s=2,uniform		gex_π		gex_{unif}	
N	h	error	order	error	order
40	0.025	5.48222e-05	2.000	5.48222e-05	2.000
80	0.0125	1.37054e-05	2.000	1.37054e-05	2.000
160	0.00625	3.42635e-06	2.000	3.42635e-06	2.000
320	0.00313	8.56588e-07	2.000	8.56588e-07	2.000

s=2,uniform		geu_π		geu_{unif}	
N	h	error	order	error	order
40	0.025	5.48222e-05	2.000	5.48222e-05	2.000
80	0.0125	1.37054e-05	2.000	1.37054e-05	2.000
160	0.00625	3.42635e-06	2.000	3.42635e-06	2.000
320	0.00313	8.56588e-07	2.000	8.56588e-07	2.000

Table 3

s=2,gaussian		gex_π		gex_{unif}	
N	h	error	order	error	order
40	0.025	1.59617e-05	2.000	1.59617e-05	2.000
80	0.0125	3.99043e-06	2.000	3.99043e-06	2.000
160	0.00625	9.97608e-07	2.000	9.97608e-07	2.000
320	0.00313	2.49402e-07	2.000	2.49402e-07	2.000

s=2,gaussian		geu_π		geu_{unif}	
N	h	error	order	error	order
40	0.025	2.98818e-09	4.000	1.29681e-06	2.984
80	0.0125	1.86771e-10	4.000	1.62452e-07	2.997
160	0.00625	1.16747e-11	4.000	2.04121e-08	2.993
320	0.00313	7.24754e-13	4.010	2.52902e-09	3.013

Table 4

s=3,uniform		gex_π		gex_{unif}	
N	h	error	order	error	order
40	0.025	4.22512e-09	3.999	4.93475e-09	3.976
80	0.0125	2.64144e-10	4.000	3.10976e-10	3.988
160	0.00625	1.65843e-11	3.993	1.93570e-11	4.006
320	0.00313	1.04050e-12	3.994	1.21325e-12	3.996

s=3,uniform		geu_π		geu_{unif}	
N	h	error	order	error	order
40	0.025	2.96391e-09	4.000	4.93475e-09	3.976
80	0.0125	1.85253e-10	4.000	3.10976e-10	3.988
160	0.00625	1.15774e-11	4.000	1.93570e-11	4.006
320	0.00313	7.23421e-13	4.000	1.21281e-12	3.996

Table 5

points, and gex_{unif} and geu_{unif} are discrete maxima taken over 1000 equidistributed points. □

s=3,gaussian		gex_π		gex_{unif}	
N	h	error	order	error	order
40	0.025	2.50651e-09	3.999	2.77352e-09	3.993
80	0.0125	1.56677e-10	4.000	1.74531e-10	3.990
160	0.00625	9.78662e-12	4.001	1.09468e-11	3.995
320	0.00313	6.16396e-13	3.989	6.83453e-13	4.002

s=3,gaussian		geu_π		geu_{unif}	
N	h	error	order	error	order
40	0.025	2.13163e-14	5.965	2.77352e-09	3.993
80	0.0125	8.88178e-16	4.585	1.74530e-10	3.990
160	0.00625	1.77636e-15	-1.000	1.09472e-11	3.995
320	0.00313	5.32907e-15	-1.585	6.83009e-13	4.003

Table 6

3.1.3 Uniform approach B

Any regular index-1 DAE in standard form

$$E(t)x'(t)+F(t)x(t)=q(t) \tag{129}$$

can be reformulated as regular index-1 DAE with properly stated leading term

$$A(t)(Dx)'(t)+B(t)x(t)=q(t) \tag{130}$$

by means of a proper factorization $E=AD$, and $B=E-AD'$. The BVP for (129) and the boundary condition

$$G_a x(a)+G_b x(b)=\gamma \tag{131}$$

is well-posed in the natural setting exactly if this is the case for the BVP (130), (131).

This time we approximate the solution x_* of the linear well-posed BVP by a possibly discontinues $x_\pi \in \mathcal{B}^m_{\pi,s}$ and consider the system

$$A(\tau_{ik})(Dx_\pi)'(\tau_{ik})+B(\tau_{ik})x_\pi(\tau_{ik})=q(\tau_{ik}), \quad k=1,\dots,s,\ i=0,\dots,N-1, \tag{132}$$

$$D(t_i)(x_\pi(t_i^-)-x_\pi(t_i))=0, \quad i=1,\dots,N-1, \tag{133}$$

$$x_\pi(t_i)\in\mathcal{M}_0(t_i), \quad i=0,\dots,N-1, \tag{134}$$

$$G_a x_\pi(a)+G_b x_\pi(b)=\gamma, \tag{135}$$

which consists of the usual smN collocation conditions (132), $(N-1)n$ continuity conditions applying only to the component Dx_π which approximates the smooth solution component Dx_*, further, $N(m-n)$ consistency conditions (134), and the n boundary conditions. Altogether one has $(s+1)Nm$ conditions to determine all $(s+1)Nm$ coefficients of x_π.

If $\rho_1=0$, then (132) already contains the condition $x_\pi(\tau_{01})=x_\pi(t_0)\in\mathcal{M}_0(t_0)$ and the equation (134) with $i=0$ is redundant.

For $\rho_1>0$, the approximation x_π is uniquely determined. It should be emphasized that x_π is not necessarily continuous, but the product Dx_π is so. The values $x_\pi(t_1),\dots,x_\pi(t_N)$ are consistent by construction. In case of $\rho_s=1$, in particular for Radau IIa, x_π is continuous in $t_1,\dots,t_N$.

This approach partly reflects ideas of both Subsubsections 3.1.2 and 3.1.1. It was introduced and studied in [22, 23, 24] for BVPs in standard form DAEs with the aim to preserve superconvergence properties of Gauß and Radau collocations. The system originally proposed in [22, p. 39] reads:

$$E(\tau_{ik})x'_\pi(\tau_{ik}) + F(\tau_{ik})x_\pi(\tau_{ik}) = q(\tau_{ik}), \quad k = 1,\dots,s,\ i = 0,\dots,N-1, \tag{136}$$

$$E_1(t_i)(x_\pi(t_i^-) - x_\pi(t_i)) = 0, \quad i = 1,\dots,N-1, \tag{137}$$

$$F_2(t_i)x_\pi(t_i) - q_2(t_i) = 0, \quad i = 0,\dots,N-1, \tag{138}$$

$$G_a x_\pi(a) + G_b x_\pi(b) = \gamma, \tag{139}$$

whereby the transformation

$$S(t)E(t) = \begin{bmatrix} E_1(t) \\ 0 \end{bmatrix}, \quad S(t)F(t) = \begin{bmatrix} F_1(t) \\ F_2(t) \end{bmatrix}, \quad S(t)q(t) = \begin{bmatrix} q_1(t) \\ q_2(t) \end{bmatrix},$$

is applied. Since $\operatorname{rank} E_1(t) = n$, this corresponds to the factorization

$$E(t) = S(t)^{-1} \begin{bmatrix} E_1(t) \\ 0 \end{bmatrix} = (S(t)^{-1} \begin{bmatrix} I \\ 0 \end{bmatrix}) E_1(t) =: A(t)D(t)$$

Consequently, equations (133)–(135) coincide with (137)–(139), respectively. The relation

$$E(\tau_{ik})x'_\pi = A(\tau_{ik})D(\tau_{ik})x'_\pi = A(\tau_{ik})(Dx_\pi)'(\tau_{ik}) - A(\tau_{ik})D'(\tau_{ik})x_\pi(\tau_{ik})$$

is valid for the right derivatives. This shows that also (132) and (136) coincide. The next theorem is a consequence of [22, Theorem 5.11].

Theorem 3.3. *Let the linear BVP (129), (131) be well-posed in the natural setting. Let the data of the DAE be sufficiently smooth for respective order conditions. Then, for the collocation scheme (116)–(139), with $\rho_1 > 0$, the following statements hold:*

(1) *There is a $h_* > 0$, such that, for meshes with $h \le h_*$, there exists a unique collocation solution x_π.*

(2) *It holds*

$$\|x_* - x_\pi\|_\infty = O(h^{\min(s+1, s_+)}).$$

(3) *For Radau and Gauß points the superconvergence order holds,*

$$\max_{i=0,\dots,N} |x_*(t_i) - x_\pi(t_i)| = O(h^{s_+}).$$

The method is applied in [22] to well-posed nonlinear BVPs

$$\begin{aligned} \mathfrak{f}(x'(t), x(t), t) &= 0, \quad t \in [a,b], \\ g(x(a), x(b)) &= 0. \end{aligned}$$

To this aim, it is supposed that there is a transformation S depending at most on x and t such that

$$S(x,t)\mathfrak{f}_{x'}(x', x, t) = \begin{bmatrix} E_1(x', x, t) \\ 0 \end{bmatrix}, \quad \operatorname{rank} E_1(x', x, t) = n.$$

Then it follows that the second part of $S\mathfrak{f}$ is independent of x' ([22, Lemma 7.1]), and thus

$$S(x,t)\mathfrak{f}(x',x,t) =: \begin{bmatrix} F_1(x',x,t) \\ F_2(x,t) \end{bmatrix}.$$

Finally the corresponding collocation scheme reads:

$$\mathfrak{f}(x'_\pi(\tau_{ik}),x_\pi(\tau_{ik}),\tau_{ik}) = 0, \quad k = 1,\dots,s, \qquad i = 0,\dots,N-1, \tag{140}$$

$$E_1(x'_\pi(\tau_{ik}),x_\pi(\tau_{ik}),\tau_{ik})(x_\pi(t_i^-) - x_\pi(t_i)) = 0, \qquad i = 1,\dots,N-1, \tag{141}$$

$$F_2(x_\pi(\tau_{ik}),\tau_{ik}) = 0, \qquad i = 0,\dots,N-1, \tag{142}$$

$$g(x_\pi(a),x_\pi(b)) = 0. \tag{143}$$

For $s > 2$, Theorem 3.3 applies accordingly also to this nonlinear case, in particular the desired superconvergence properties for Radau and Gauß points are reached, see [22, Theorems 7.5 and 7.6]. If the function $\mathfrak{f}$ is linear in x', this is also valid for $s = 2$.

3.1.4 Uniform approach C

As proposed in [111], one can approximate the solution x_* of the linear well-posed BVP (129), (131) by an continuous piecewise polynomial function $x_\pi \in \mathcal{B}^m_{\pi,s} \cap \mathcal{C}(\mathcal{I},\mathbb{R}^m)$ using the system

$$E(\tau_{ik})x'_\pi(\tau_{ik}) + F(\tau_{ik})x_\pi(\tau_{ik}) = q(\tau_{ik}), \quad k = 1,\dots,s,\ i = 0,\dots,N-1, \tag{144}$$

$$x_\pi(t_i^-) - x_\pi(t_i) = 0, \qquad i = 1,\dots,N-1, \tag{145}$$

$$F_2(a)x_\pi(a) - q_2(a) = 0, \tag{146}$$

$$G_a x_\pi(a) + G_b x_\pi(b) = \gamma, \tag{147}$$

or, equivalently (cf. Subsubsection 3.1.3), by

$$A(\tau_{ik})(Dx_\pi)'(\tau_{ik}) + B(\tau_{ik})x_\pi(\tau_{ik}) = q(\tau_{ik}), \quad k = 1,\dots,s,\ i = 0,\dots,N-1, \tag{148}$$

$$x_\pi(t_i^-) - x_\pi(t_i) = 0, \qquad i = 1,\dots,N-1, \tag{149}$$

$$x_\pi(a) \in \mathcal{M}_0(a), \tag{150}$$

$$G_a x_\pi(a) + G_b x_\pi(b) = \gamma. \tag{151}$$

Again, one has to determine $(s+1)Nm$ coefficients of x_π by means of the $(s+1)mN$ collocation conditions, $m(N-1)$ continuity conditions, the consistency condition with $m-n$ equations, and the n boundary conditions. We see, that the number of unknown coefficients and the number of conditions are the same. In [111], the discussion is restricted to the case

$$\rho_1 > 0,\ \rho_s = 1,$$

and Radau methods are in the focus of interest. We quote results given in [111, Sätze 5.1, 5.2, and 5.3].

Theorem 3.4. *Let the linear BVP (129), (131) be well-posed in the natural setting. Let E and F be twice continuously differentiable.*
Then, for the collocation scheme (144)–(147), with $\rho_1 > 0$ and $\rho_s = 1$, the following statements hold:

(1) *There is a $h_* > 0$, such that, for meshes with $h \le h_*$, there exists a unique collocation solution $x_\pi \in \mathcal{B}^m_{\pi,s} \cap \mathcal{C}(\mathcal{I}, \mathbb{R}^m)$.*
(2) *If the data of the DAE is sufficiently smooth, then*

$$\|x_* - x_\pi\|_\infty = O(h^s).$$

(3) *For Radau points the superconvergence order holds,*

$$\max_{i=0,\dots,N} |x_*(t_i) - x_\pi(t_i)| = O(h^{2s-1}).$$

3.2 Partitioned equations

For the DAE (105), (106) featuring explicitly the derivative-free equation one has the option to apply different collocation points in the first and second equations. This is proposed in [76, 75, 74] by combining the Gauß scheme with s points for the first equation and the Lobatto scheme with $s+1$ points for the second one.
The BVP for the DAE (105), (106), with n and $m-n$ equations, and the boundary condition (112) is now assumed to be well-posed in the modified setting with pre-image space $\mathcal{C}^1(\mathcal{I}, \mathbb{R}^m)$ and image space $\mathcal{C}(\mathcal{I}, \mathbb{R}^n) \times \mathcal{C}^1(\mathcal{I}, \mathbb{R}^{m-n}) \times \mathbb{R}^n$, see Remarks 2.1 and 2.6. We discuss here the case when $m-n > 0$. This means that the DAE has index 1.

The linear BVP for the partitioned index-1 DAE with n and $m-n$ equations

$$E_1(t)x'(t) + F_1(t)x(t) = q_1(t), \tag{152}$$
$$F_2(t)x(t) = q_2(t), \tag{153}$$

is treated in [76] by means of the symmetric scheme

$$E_1(\tau_{ik})x'_\pi(\tau_{ik}) + F_1(\tau_{ik})x_\pi(\tau_{ik}) = q_1(\tau_{ik}), \quad k = 1,\dots,s,\ i = 0,\dots,N-1, \tag{154}$$
$$F_2(\tau^L_{ik})x_\pi(\tau^L_{ik}) = q_2(\tau^L_{ik}), \quad k = 0,\dots,s,\ i = 0,\dots,N-1, \tag{155}$$
$$T_2(t_i)^*(x_\pi(t_i^-) - x_\pi(t_i)) = 0, \qquad i = 1,\dots,N-1, \tag{156}$$
$$G_a x_\pi(a) + G_b x_\pi(b) = \gamma. \tag{157}$$

with Gauß points $0 < \rho_1 < \cdots < \rho_s < 1$ and Lobatto points $0 = \rho^L_0 < \cdots < \rho^L_s = 1$. The matrix $T_2(t) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ has, by construction, full column-rank n and satisfies the condition $F_2(t)T_2(t) = 0$ for all $t \in [a,b]$.

To compute the $m(s+1)N$ unknowns of $x_\pi \in \mathcal{B}^m_{\pi,s}$ one has $sNn+(s+1)N(m-n)+n(N-1)+n=(s+1)Nm$ conditions so that the system is balanced. The following theorem combines parts of [76, Theorems 3.1, 3.2, and 3.3].

Theorem 3.5. *Let the linear BVP (152), (153), (112) be well-posed in the modified index-1 setting. Let E and F be twice continuously differentiable.*
Then, if h is sufficiently small, the following statements hold:

(1) *There is a unique continuous collocation solution $x_\pi \in \mathcal{B}^m_{\pi,s} \cap \mathcal{C}(\mathcal{I},\mathbb{R}^m)$ that satisfies the collocation conditions (154) and (155), the boundary condition (112) as well as the consistency conditions (156).*

(2) *It holds*
$$\|x_* - x_\pi\|_\infty = O(h^s).$$

(3) *If the data of the DAE is sufficiently smooth, then superconvergence order holds,*
$$\max_{i=0,\dots,N} |x_*(t_i) - x_\pi(t_i)| = O(h^{2s}).$$

Condition (156) is no longer mentioned in Theorem 3.5. It only ensures the continuity of the differential component, similar to condition (137). In fact, (156) could be replaced by the easier conditions (137). Namely, for each fixed $1 \le i \le N-1$, one has from (155) the equations

$$\begin{aligned} 0 &= F_2(t_i)x_\pi(\tau^L_{i-1\,s}) + q_2(t_i) = F_2(t_i)x_\pi(t_i^-) + q_2(t_i),\\ 0 &= F_2(t_i)x_\pi(\tau^L_{i\,0}) + q_2(t_i) \;\; = F_2(t_i)x_\pi(t_i) + q_2(t_i), \end{aligned}$$

thus $F_2(t_i)(x_\pi(t_i^-) - x_\pi(t_i)) = 0$. Regarding, additionally, one of the two conditions

$$T_2(t_i)^*(x_\pi(t_i^-) - x_\pi(t_i)) = 0, \quad E_1(t_i)(x_\pi(t_i^-) - x_\pi(t_i)) = 0,$$

implies $x_\pi(t_i^-) - x_\pi(t_i) = 0$, since both matrices,

$$\begin{bmatrix} T_2(t_i)^* \\ F_2(t_i) \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} E_1(t_i) \\ F_2(t_i) \end{bmatrix},$$

are nonsingular.

The approach is extended in [75, 74] to nonlinear BVPs with partitioned DAEs (105), (106) by means of the scheme

$$\mathfrak{f}_1(x'_\pi(\tau_{ik}), x_\pi(\tau_{ik}), \tau_{ik}) = 0, \quad k = 1,\dots,s, \quad i = 0,\dots,N-1, \tag{158}$$

$$\mathfrak{f}_2(x_\pi(\tau^L_{ik}), \tau^L_{ik})) = 0, \quad k = 0,\dots,s, \quad i = 0,\dots,N-1, \tag{159}$$

$$g(x_\pi(a), x_\pi(b)) = 0. \tag{160}$$

For the above scheme, a result analogous to Theorem 3.5 is given. The continuity conditions are now hidden in the claim concerning the continuity of $x_\pi \in \mathcal{B}^m_{\pi,s}$. The

convergence and error investigations in [76, 75, 74] are solely directed to the partitioned index-1 DAE, which is seen there as reduced system of a general arbitrary index DAE satisfying a series of hypotheses, see Remark 2.7. The collocation procedure described in [75, 74] is strongly interlinked with the reduction procedure via the derivative array system. Possible errors in the reduction procedure are neglected.

3.3 BVPs for index-2 DAEs

BVPs for higher-index DAEs are ill-posed in the natural setting even though the boundary conditions are accurately stated—this is the clear message of Theorem 2.4. Fortunately, for a large class of index-2 DAEs, the BVPs with accurately stated boundary conditions become well-posed in the advanced setting, see Subsubsection 2.5.2. In this case, the associated inequality (68) reads:

$$\|x - x_*\|_\infty \leq \kappa(\,|\gamma| + \|q\|_\infty + \|(DQ_{*1}G_{*2}^{-1}q)'\|_\infty). \tag{161}$$

The linear Hessenberg system of m_1 and $m_2 \leq m_1$ equations,

$$\begin{aligned} x_1'(t) + B_{11}(t)x_1(t) + B_{12}(t)x_2(t) &= q_1(t), \\ B_{21}(t)x_1(t) &= q_1(t), \end{aligned}$$

with sufficiently smooth coefficients and $B_{21}(t)B_{12}(t)$ remaining nonsingular, belongs to this class, cf., Example 2.3. We have to provide $l = m_1 - m_2$ boundary conditions

$$G_a x(a) + G_b x(b) = \gamma.$$

For boundary conditions which are accurately stated, the homogeneous linear BVP has the trivial solution $x_* = 0$, only. For the solutions of the inhomogeneous linear BVPs the inequality (161) simplifies to

$$\begin{aligned} \|x - x_*\|_\infty &\leq \kappa(\,|\gamma| + \|q\|_\infty + \|(B_{12}(B_{21}B_{12})^{-1}q_2)'\|_\infty) \\ &\leq \tilde{\kappa}(\,|\gamma| + \|q\|_\infty + \|q_2'\|_\infty). \end{aligned} \tag{162}$$

A direct investigation of the linear index-2 DAE by the linear decoupling makes evident that the first solution component x_1 is actually independent of the term q_2'. A related inequality is derived in [7], and the BVP is said to be well-conditioned, if $\tilde{\kappa}$ has moderate size.
Here, it should be again emphasized that the notions *well-posed*, *stable*, and *well-conditioned* are used in different places with different meanings, cf., Remarks 2.9 and 2.6.

In particular, the inequality (162) applies to the nonlinear index-2 DAE

$$x_1'(t) + b_1(x_1(t), x_2(t), t) = 0, \tag{163}$$
$$b_2(x_1(t), t) = 0, \tag{164}$$

with $B_{ij}(t)$ replaced by the partial derivatives $B_{*ij}(t) := \frac{\partial b_i}{\partial x_j}(x_*(t), t)$, and nonlinear boundary conditions.

Regarding the discretization of index-2 problems, errors in the derivative-free equation (164) can be significantly amplified, at least by a factor h^{-1}. Therefore, it is a good idea to keep the defects in this equations reasonable small. For this purpose, so-called *projected Runge–Kutta methods* and *projected collocation* are introduced in [8, 7].

It is proposed to complete the standard collocation methods locally at fixed time points by an additional backward projection onto the constraint given by equation (164). More precisely, let t_l be fixed and $x_{l,1}, x_{l,2}$ denote already computed approximations of $x_1(t_l)$, $x_2(t_l)$. The defect $b_2(x_{l,1}, t_l)$ represents the deviation of the given approximation away from the obvious constraint. If $b_2(x_{l,1}, t_l) \neq 0$, a new approximation $x_{l,1}^{new}, x_{l,2}^{new}$ is constructed such that

$$b_2(x_{l,1}^{new}, t_l) = 0. \tag{165}$$

This is accomplished by the ansatz

$$x_{l,1}^{new} := x_{l,1} + B_{12}(x_{l,1}^{new}, x_{l,2}^{new}, t_l)\lambda_l, \tag{166}$$
$$x_{l,2}^{new} := x_{l,2}, \tag{167}$$

where $B_{ij} := \frac{\partial b_i}{\partial x_j}$. If the given approximation are sufficiently accurate, then the values x_l^{new} and λ_l are locally uniquely determined by (165)-(167). A Newton step starting from the initial guess $x_l^{new,(0)} = x_l$, $\lambda_l^{(0)} = 0$ yields

$$x_{l,1}^{new,(1)} = x_{l,1} - F_l b_2(x_{l,1}, x_{l,2}, t_l), \tag{168}$$

where F_l denotes $B_{12}(B_{21}B_{12})^{-1}$ taken at $(x_{l,1}, x_{l,2}, t_l)$. The $m_1 \times m_1$ matrix $\Omega_l := F_l B_{21}(x_{l,1}, x_{l,2}, t_l)$ represents a projector, and hence, formula (168) means in more detail

$$\Omega_l x_{l,1}^{new,(1)} = \Omega_l x_{l,1} - F_l b_2(x_{l,1}, x_{l,2}, t_l),$$
$$(I - \Omega_l) x_{l,1}^{new,(1)} = (I - \Omega_l) x_{l,1},$$

which shows that the particular Ω-component is affected, only. The $(I - \Omega_l)$-component corresponds to the IERODE, cf., Example 2.3, thus the true differential component is not changed.

In contrast to the index-1 case, the accurate number of boundary conditions is now $m_1 - m_2$. For completing the collocation schemes one has always to find addi-

tional m_2 conditions. The usual choice is $b_2(x(a),a) = 0$ and $\rho_1 > 0$. This seems to exclude uniform approaches for the different components.

Completing a collocation scheme at the mesh points $t_i > a$, by equations corresponding to (165)-(167) has proved its value in various cases. If the BVP is locally well-posed (in the advanced setting) with a moderate $\tilde{\kappa}$ and the problem data is sufficiently smooth, then, owing to [7, Theorem 3.3], there are locally unique approximations $x_{\pi,1} \in \mathcal{B}^{m_1}_{\pi,s}$ and $x_{\pi,2} \in \mathcal{B}^{m_2}_{\pi,s-1}$ satisfying the projected collocation scheme and the error estimates

$$\begin{aligned} \|x_{*,1} - x_{\pi,1}\|_\infty &= O(h^{\min(s+1,s_+)}), \quad \|x_{*,2} - x_{\pi,2}\|_\infty = O(h^s), \\ |x_{*,1}(t_i) - x_{\pi,1}(t_i)| &= O(h^{s_+}), \quad i = 0,\ldots,N, \end{aligned}$$

hold. In contrast to the results for index-1 problems, now $x_{\pi,1}$ is generally discontinues due to the backward projection.

The projected collocation is extended to some more general semi-explicit index-2 DAEs ([13, 11]), whereby the components to be changed by projections are locally identified by means of a singular value decompositions. This procedure is called *selective projected collocation*.

The package COLDAE ([13]) includes the options to treat BVPs for index-2 DAEs in Hessenberg form by *projected collocation* and for more general semi-explicit index-2 DAEs by *selective projected collocation*.

3.4 BVPs for singular index-1 DAEs

In recent years, motivated by numerous applications a lot of efforts has been put into the analysis and numerical treatment of BVPs in ODEs which can exhibit singularities (e.g., [12, 16, 17, 68, 71, 43] and references therein). Such problems are typically given as

$$t^\alpha u'(t) = M(t)u(t) + h(u(t),t), \quad t \in (0,1], \quad g(u(0),u(1)) = 0, \tag{169}$$

with $\alpha \geq 1$. For $\alpha = 1$, one speaks of a *singularity of the first kind*. For instance, a singularity of the first kind may come from a reduction of a PDE to an ODE owing to cylindrical or spherical symmetry. Naturally, DAEs may feature those singularities more than ever, as it is the case in the following two examples.

Example 3.2. The DAE taken from [71],

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix} \left(\begin{bmatrix} 1 & -1 \end{bmatrix} x \right)'(t) + \begin{bmatrix} 2 & 0 \\ 0 & t+2 \end{bmatrix} x(t) = 0, \tag{170}$$

has index 1 on the interval $(0,1]$ and yields there the inherent ODE

$$t\,u'(t) = -(2t+4)u(t), \quad u(t) = x_1(t) - x_2(t), \tag{171}$$

showing a singularity of the first kind. The inherent ODE (171) possesses the general solution

$$u(t) = c_0\, \mathrm{e}^{-2t} t^{-4},$$

with a constant c_0. Except for the trivial solution, that is, for $c_0 = 0$, all solutions of the inherent ODE grow unboundedly for $t \to 0$. The canonical projector of the DAE (170)

$$\Pi_{can}(t) = I - Q_{can}(t) = \begin{bmatrix} 1 + \frac{2}{t} & -\frac{2+t}{t} \\ \frac{2}{t} & 1 - \frac{2+t}{t} \end{bmatrix}$$

is unbounded for $t \to 0$. The general DAE solution is given by

$$x(t) = \Pi_{can}(t) \begin{bmatrix} 1 \\ 0 \end{bmatrix} u(t) = \Pi_{can}(t) \begin{bmatrix} 1 \\ 0 \end{bmatrix} c_0\, \mathrm{e}^{-2t} t^{-4} = c_0\, \mathrm{e}^{-2t} t^{-4} \begin{bmatrix} 1 + \frac{2}{t} \\ \frac{2}{t} \end{bmatrix}.$$

Except for the case $c_0 = 0$, the DAE solutions are unbounded. By means of the condition $D(0)x(0) = 0$ one picks up the only bounded solution. □

Example 3.3. The DAE (cf. [98]),

$$\begin{bmatrix} t & 0 \\ 0 & t \\ 0 & 0 \end{bmatrix} \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} x \right)'(t) + \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} x(t) = q(t), \tag{172}$$

has index 1 on the interval $(0, 1]$ and yields the inherent ODE

$$t\, u'(t) = \begin{bmatrix} -1 & 1 \\ -1 & 0 \end{bmatrix} u(t) + \begin{bmatrix} q_1(t) - 2q_3(t) \\ q_2(t) \end{bmatrix}, \quad u(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}. \tag{173}$$

The canonical projector is now constant,

$$\Pi_{can}(t) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix},$$

and it trivially has a continuous extension for $t \to 0$. All solutions of the DAE (172) can be expressed as

$$x(t) = \Pi_{can}(t) \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} u(t) + \begin{bmatrix} 0 \\ 0 \\ q_3(t) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & -1 \end{bmatrix} u(t) + \begin{bmatrix} 0 \\ 0 \\ q_3(t) \end{bmatrix},$$

which shows that the bounded solutions of the inherent ODE (173) correspond to the bounded solutions of the DAE (172). □

In the context of classical singular BVPs (169), seeking a solution being continuous on the closed interval, one has to state the boundary conditions in a special

smart way depending on the spectrum of the matrix $M(0)$ (see [40, 68]). In case of DAEs, this procedure becomes a much more tough job. Below, we bring out the mathematical background of the case when the DAE represents an index-1 DAE with a singularity at $t = 0$ and the inherent ODE is singular with a singularity of the first kind. We deal with the BVP

$$f((D(t)x(t))', x(t), t) = 0, \quad t \in (0,1], \tag{174}$$

$$G_a x(0) + G_b x(1) = \gamma, \tag{175}$$

where, as before, $f(y,x,t) \in \mathbb{R}^m$, $D(t) \in \mathbb{R}^{n\times m}$, $y \in \mathbb{R}^n$, $x \in \mathcal{D}$, with $\mathcal{D} \subseteq \mathbb{R}^m$ open, $t \in [0,1]$, $n \le m$, and the data f, f_y, f_x, D are assumed to be at least continuous on their definition domains. Moreover, now we require that

$$\ker f_y(y,x,t) = \{0\}, \quad (y,x,t) \in \mathbb{R}^n \times \mathcal{D} \times (0,1], \tag{176}$$

$$\operatorname{im}(D(t)) = \mathbb{R}^n, \quad t \in [0,1]. \tag{177}$$

Conditions (176) and (177) mean that the matrix $D(t)$ has again full row rank n on the closed interval, but $f_y(y,x,t)$ has full column rank n on $\mathbb{R}^n \times \mathcal{D} \times (0,1]$ only. At $t = 0$ the matrix $f_y(y,x,t)$ may undergo a rank drop as it is the case for the DAE (172). The structural conditions (176) and (177) guarantee that the system (174) has a properly stated leading term at least on $\mathbb{R}^n \times \mathcal{D} \times (0,1]$, with the border-projector function $R(t) = I$.

Let the boundary condition (175) be such that

$$G_a = B_0 D(0), \quad G_b = B_1 D(1), \quad B_0, B_1 \in \mathcal{L}(\mathbb{R}^n),$$

which will result in a BVP for the inherent ODE with respect to the component Dx.

Put $\mathcal{I} = [0,1]$. We are looking for a solution of the BVP (174),(175) which belongs at least to the function space $\mathcal{C}(\mathcal{I}, \mathbb{R}^m) \cap \mathcal{C}_D^1((0,1], \mathbb{R}^m)$.

The further structure of the boundary conditions (175) which is necessary and sufficient for the BVP (174)–(175) to become well-posed in a special sense will be specified in the course of the discussion. Here we do without function space settings, but adopt the understanding of well-posed BVPs common in the framework of singular ODEs (e.g., [68]). Though first existence and uniqueness results are given for a special class of singular DAEs in [98], more general solvability assertions justifying well-posedness of BVPs in appropriate function spaces are missing till now. As we will see, well-posedness in this special sense incorporates aspects of well-conditioning.

We put

$$N_0(t) := \ker D(t), \quad t \in [0,1],$$

and note that

$$\ker f_y(y,x,t)D(t) = N_0(t), \quad (y,x,t) \in \mathbb{R}^n \times \mathcal{D} \times (0,1],$$
$$\ker f_y(y,x,t)D(t) \supset N_0(t), \quad (y,x,t) \in \mathbb{R}^n \times \mathcal{D} \times \{0\}.$$

Below, the pointwise generalized inverse D^- of D is defined as in the regular case in Subsection 2.1.

In [71, 43], well-posed BVPs in linear and nonlinear index-1 DAEs featuring inherent ODEs with a singularity of the first kind are specified and approximated by polynomial collocation. It is shown that for a well-posed BVP having a sufficiently smooth solution the global error of the collocation scheme converges with the order $O(h^s)$, where s is the number of collocation points. Superconvergence cannot be expected in general due to the singularity, not even for the differential components of the solution. We outline the main results; for proofs and technical details, we refer to [71, 43].

3.4.1 Linear case

Following the lines of [71] we first decouple the DAE in order to formulate sufficient conditions ensuring a singularity of the first kind for the inherent ODE and then well-posed boundary conditions. Consider the linear DAE

$$A(t)(Dx)'(t) + B(t)x(t) = q(t), \quad t \in (0,1], \tag{178}$$

and assume that the DAE is regular with index 1 on $(0,1]$. Here, $A(t)$ may undergo a rank drop at $t = 0$. We have from (176),(177) that

$$\ker A(t) = \{0\}, \qquad t \in (0,1], \tag{179}$$
$$\operatorname{im} D(t) = \mathbb{R}^n, \qquad t \in [0,1]. \tag{180}$$

We decouple the DAE on the interval $(0,1]$ as described in Subsection 2.2. With Q_0 being a continuous projector function onto $\ker D$, and $P_0 := I - Q_0$ we form

$$G_0(t) := A(t)D(t), \quad t \in [0,1], \tag{181}$$
$$G_1(t) := G_0(t) + B(t)Q_0(t), \quad t \in [0,1]. \tag{182}$$

Owing to the index-1 property, the matrix $G_1(t)$ is nonsingular for $t \in (0,1]$. Now we assume $G_1(0)$ to be singular.

If $A(t)$, and therefore $G_0(t)$, undergoes a rank drop at $t = 0$, as in Example 3.3, then $G_1(0)$ is necessarily singular. Applying the classification of critical points arising in DAEs from [105, 97, 103, 83], in this case, $t = 0$ represents a critical point of type 0. As in Example 3.2, it may happen that $G_0(t)$ has constant rank on the closed interval $\mathcal{I}$, but $G_1(0)$ is singular. Then $t = 0$ is said to be a critical point of type 1–A.

We incorporate the case where the inherent ODE associated with (178) exhibits a singularity of the first kind. To this end, we decouple the solution of DAE (178) on $(0,1]$ into the differential component Dx and the algebraic component Q_0x. While $u = Dx$ satisfies the inherent explicit ODE,

$$u'(t) + D(t)G_1^{-1}(t)B(t)D(t)^- u(t) = D(t)G_1^{-1}(t)q(t), \quad t \in (0,1], \tag{183}$$

the algebraic component is given by

$$Q_0(t)x(t) = -Q_0(t)G_1^{-1}(t)B(t)D(t)^- u(t) + Q_0(t)G_1^{-1}(t)q(t), \quad t \in (0,1]. \quad (184)$$

If $u(t)$ represents the general solution of the inherent ODE (183). then the general solution of the DAE (178) can be expressed as

$$x(t) = D(t)^- u(t) + Q_0(t)x(t) = \Pi_{can}(t)D(t)^- u(t) + Q_0(t)G_1^{-1}(t)q(t), \quad t \in (0,1],$$

whereby

$$\Pi_{can}(t) = I - Q_0(t)G_1^{-1}(t)B(t), \quad t \in (0,1],$$

is the canonical projector function. We are interested in solutions being at least *continuous on the whole interval* $[0,1]$, The asymptotic behavior of the ODE (183) related to a singularity of the first kind arises when $G_1(0)$ is singular but $tG_1^{-1}(t)$ has a continuous extension on $[0,1]$. Then, we can rewrite the matrix $D(t)G_1^{-1}(t)B(t)D(t)^-$ and obtain

$$D(t)G_1^{-1}(t)B(t)D(t)^- =: -\frac{1}{t}M(t), \quad (185)$$

where $M \in \mathcal{C}([0,1], \mathcal{L}(\mathbb{R}^n))$. For the subsequent existence and uniqueness analysis we require $M \in \mathcal{C}^1([0,1], \mathcal{L}(\mathbb{R}^n))$ which means that the problem data needs to be appropriately smooth. Denoting the right-hand side of (183) by $p(t)$ we arrive at the inherent explicit ODE of the form

$$u'(t) = \frac{1}{t}M(t)u(t) + p(t), \quad t \in (0,1]. \quad (186)$$

As mentioned before, we are interested in bounded solutions x and therefore u needs to be at least in $\mathcal{C}([0,1], \mathbb{R}^n)$. It turns out that the smoothness of u depends on the smoothness of p and, additionally, the eigen-structure of $M(0)$. The theoretical background for this problem class, where $p \in \mathcal{C}([0,1], \mathbb{R}^n)$, is discussed in detail in [40]. In order to use this standard theory, we assume that $G_1^{-1}(t)q(t)$ and thus $p(t)$ are continuous in the whole interval $[0,1]$. Then, by [40], the bounded solutions of the ODE (186) can be represented in the form

$$u(t) = Ec + tf(t), \quad t \in [0,1], \quad (187)$$

where the columns of the matrix E form a basis of $\ker M(0)$ and $f \in \mathcal{C}([0,1]), \mathbb{R}^n)$. Next we provide conditions to guarantee that, given a bounded solution $u(t)$, the solution $x(t)$ of the DAE resulting via (184) is also bounded.

Proposition 3.1. *Let the DAE (178) be regular with index 1 on* $(0,1]$ *and satisfy conditions (179), (180), and let the coefficients be sufficiently smooth.*
Let $G_1(0)$ *be singular, but the matrix functions*

$$tG_1^{-1}(t), \quad G_1^{-1}(t)q(t), \quad Q_0(t)G_1^{-1}(t)B(t)D(t)^- E, \quad t \in (0,1], \quad (188)$$

have continuous extensions on the closed interval $[0,1]$*,*

$$[tG_1^{-1}(t)]^{ext}, \quad [G_1^{-1}(t)q(t)]^{ext}, \quad [Q_0(t)G_1^{-1}(t)B(t)D(t)^-E]^{ext}.$$

Then the inherent explicit ODE of the DAE exhibits a singularity of the first kind and each bounded solution of the DAE has the form

$$\begin{aligned} x(t) =& [\Pi_{can}(t)D(t)^-E]^{ext}c + [t\Pi_{can}(t)]^{ext}D(t)^-f(t) + Q_0(t)[G_1(t)^{-1}q(t)]^{ext}, \\ & t \in [0,1], \end{aligned}$$

with a constant $c \in \mathbb{R}^{n_0}$*,* $n_0 := n - \operatorname{rank} M(0)$.

If the matrix $M(0)$ is nonsingular, then E disappears. In this case, the last term in (188) vanishes identically and has trivially the continuous extension.
If the canonical projector $Q_{can}(t) = I - \Pi_{can}(t)$ has a continuous extension on $[0,1]$, which is possible if $t = 0$ is critical point of type 0, see Example 3.3, then also the term $Q_0(t)G_1^{-1}(t)B(t)D(t)^-E = Q_{can}(t)D(t)^-E$ has the continuous extension.

The inherent ODE (186) is augmented by the boundary conditions

$$B_a u(0) + B_b u(1) = \gamma. \tag{189}$$

These boundary conditions have to be chosen such that a well-posed singular boundary value problem results for u. In [71], the attention is focused on boundary value problems for singular ODE systems (186) which can equivalently be expressed as a well-posed initial value problem with initial conditions at $t^* = 0$ or terminal conditions at $t^* = 1$. This means a restriction on the spectrum of the matrix $M(0)$ from (185), see [69, 72], for a detailed explanation of this fact. The reason for the above assumption is that a shooting argument is applied in the course of the analysis of polynomial collocation approximation.

A singular initial value problem posed at $t^* = 0$ for the differential equation (186) is well-posed if and only if the spectrum of $M(0)$ contains no eigenvalues with positive real parts and the initial value satisfies $u(0) \in \ker M(0)$. A singular terminal value problem posed at $t^* = 1$ is well-posed if and only if the spectrum of $M(0)$ contains no eigenvalues with negative real parts and the invariant subspace associated with the eigenvalue zero coincides with the nullspace of $M(0)$ ([40, 69]).

Under the assumptions of Proposition 3.1, polynomial collocation methods are analyzed in[70, 71]. The meshes π are specified as before in this section. Motivated by the singularity, the collocation points are chosen in the interior of the subintervals, with $\rho_1 > 0$ and $\rho_s < 1$. We approximate x and u by continuous piecewise polynomial functions $x_\pi \in \mathcal{B}^m_{\pi,s} \cap \mathcal{C}(\mathcal{I}, \mathbb{R}^m))$ and $u_\pi \in \mathcal{B}^n_{\pi,s} \cap \mathcal{C}(\mathcal{I}, \mathbb{R}^n))$ as in Subsubsection 3.1.1. The numerical scheme defining x_π and u_π has the form

$$A(\tau_{ik})u'_\pi(\tau_{ik}) + B(\tau_{ik})x_\pi(\tau_{ik}) = q(\tau_{ik}), \tag{190}$$

$$D(\tau_{ik})x_\pi(\tau_{ik}) - u_\pi(\tau_{ik}) = 0, \qquad k = 1,\dots,s,\ i = 0,\dots,N-1, \tag{191}$$

$$B_0u_\pi(0) + B_1u_\pi(1) = \gamma. \tag{192}$$

As in Subsubsection 3.1.1, further conditions are necessary to close the system for the numerical computations. We choose these additional conditions as,

$$B(0)x_\pi(0) - q(0) \in \lim_{t\to 0^+} \operatorname{im}(A(t)), \quad u_\pi(0) = D(0)x_\pi(0), \tag{193}$$

or

$$B(1)x_\pi(1) - q(1) \in \operatorname{im}(A(1)), \quad u_\pi(1) = D(1)x_\pi(1). \tag{194}$$

The convergence results in case of a singular inherent ODE are quite similar to the regular index-1 DAE case. Owing to the assumptions of Proposition 3.1, for arbitrary collocation points, stage order s uniformly in t is ensured in case that the solutions of the DAE and the inherent ODE, respectively, are sufficiently smooth,

$$\|u_* - u_\pi\|_\infty = O(h^s), \quad \|x_* - x_\pi\|_\infty = O(h^s).$$

Note that for Gauß collocation points the superconvergence behavior $O(h^{2s})$ in π does not hold in general, a well known fact in the context of singular ODEs. Rather, the orders

$$\|u_* - u_\pi\|_\infty = O(h^{s+1})$$

hold. If the BVP for the inherent ODE is a terminal value or boundary value problem, the analysis in [71] additionally requires

$$Q_{can} \in \mathcal{C}([0,1], \mathcal{L}(\mathbb{R}^m)) \tag{195}$$

to ensure this optimal convergence behavior. If the assumptions (188) and (195) are violated, order reductions in the algebraic components might occur. Especially, order reductions can be due to the behavior of the canonical projector $Q_{can}(t)$ for $t \to 0^+$, in the case when Q_{can} becomes unbounded in this limit. We illustrate this important aspect by the next example picked from [70, 71]. Therein, we highlight additional order reductions in the sense that the stage order is no longer observed.

Example 3.4. We consider the following four-dimensional semi-explicit DAE

$$A(Dx)' + \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22}(t) \end{bmatrix} x(t) = q(t), \tag{196}$$

with

$$A = \begin{bmatrix} I \\ 0 \end{bmatrix}, \quad D = [I\ \ 0], \quad D^- = \begin{bmatrix} I \\ 0 \end{bmatrix},$$

$$B_{11} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad B_{12} = \begin{bmatrix} 3 & -1 \\ -2 & 1 \end{bmatrix}, \quad B_{21} = \begin{bmatrix} 1 & 1 \\ 2 & 3 \end{bmatrix}, \quad B_{22}(t) = \begin{bmatrix} t & 0 \\ 0 & \frac{t}{5} \end{bmatrix}.$$

This yields

$$G_1(t) = \begin{bmatrix} I & B_{12} \\ 0 & B_{22}(t) \end{bmatrix}, G_1(t)^{-1} = \begin{bmatrix} I & -B_{12}B_{22}(t)^{-1} \\ 0 & B_{22}(t)^{-1} \end{bmatrix}, \quad B_{22}(t)^{-1} = \frac{1}{t}\begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix}$$

which shows that $tG_1(t)^{-1}$ has a continuous extension onto $[0,1]$. In contrast, the canonical projector

$$Q_{can}(t) = Q_0 G_1^{-1}(t)B(t) = \begin{bmatrix} 0 & 0 \\ B_{22}(t)^{-1}B_{21} & I \end{bmatrix} \tag{197}$$

is unbounded on $(0,1]$. Moreover, it holds that

$$DG_1^{-1}(t)B(t)D^- = B_{11} - B_{12}B_{22}(t)^{-1}B_{21} = -\frac{1}{t}\begin{bmatrix} -7 & -12 \\ 8 & 13 \end{bmatrix} =: -\frac{1}{t}M.$$

Since M is nonsingular, we have $E = 0$, and the matrix function $Q_0G_1(t)^{-1}B(t)D^-E = 0$ has trivially a continuous extension on $[0,1]$. We consider the continuously differentiable solution

$$x(t) = \begin{bmatrix} t^\gamma \sin(t) \\ t^\delta \mathrm{e}^t \\ \cos(t) \\ t^\ell \mathrm{e}^{-t} \end{bmatrix},$$

with parameters specified below. The respective right-hand side $q(t)$ is such that $G_1(t)^{-1}q(t)$ is actually continuous on $[0,1]$. In summary, all three matrix function in (188) possess the requested continuous extensions on $[0,1]$.

The matrix M has the eigenvalues 1 and 5. Since they are both positive, we may state a well-posed terminal problem prescribing the values of the differential components $x_1(t)$ and $x_2(t)$ at $t = 1$.

Therefore, we consider system (196) subject to the boundary conditions

$$x_1(1) = \sin(1), \quad x_2(1) = \mathrm{e}^1.$$

The additional conditions

$$\begin{aligned} x_1(1) + x_2(1) + x_3(1) &= q_3(1), \\ 2x_1(1) + 3x_2(1) + \frac{1}{5}x_4(1) &= q_4(1) \end{aligned}$$

are consistent boundary conditions for the algebraic components to complete the collocation scheme used in BVPSUITE. These condition simply reflect the obvious constraint at time $t = 1$.

Note, that we solve a terminal value problem which is more likely to show order reductions when $Q_{can}(t)$ becomes unbounded when $t \to 0$.

Problem 1: Set $\ell = 3$, $\gamma = 1$, $\delta = 1$. All solution components are smooth.

Problem 2: Set $\ell = \frac{5}{2}$, $\gamma = \frac{6}{5}$, $\delta = \frac{5}{2}$. The differential components x_1 and x_2 become unsmooth.

		$s=3$				$s=4$			
		gex		*geu*		*gex*		*geu*	
Problem	**Collocation**	π	π_{coll}	π	π_{coll}	π	π_{coll}	π	π_{coll}
Problem 1	equidistant	3	3	4	4	4	4	4	4
	Gaussian	3	3	4	4	4	4	5	5
Problem 2	equidistant	*0.3*	*0.3*	1.2	1.2	*0.3*	*0.3*	1.2	1.2
	Gaussian	*0.3*	*0.3*	1.3	1.3	*0.3*	*0.3*	1.2	1.2

Table 7 Problems 1 and 2: Experimentally observed convergence rates for different collocation schemes with $s = 3,4$, cf. [71] for details. Here, the global error in x is denoted by *gex* and the global error in u by *geu*. π means that the maximum of the global error was calculated using its values at the mesh points in π. We denote by π_{coll} the union of the mesh points and the collocation points. Then, π_{coll} indicates that the maximum of the global error is computed using its values at points in π_{coll}. Order reductions are highlighted in italic.

The numerical results obtained by means of BVPSUITE for this example are given in Table 7. For more details see [70, Tables 192 to 200, 228 to 236]. For the case when the differential solution components, $u(t)$, are smooth no order reduction is observed, although the projection matrix (197) is unbounded for $t \to 0$.
In Problem 2 we observe order reductions due to the fact that the canonical projector (197) is unbounded for $t \to 0$. One would expect to see the convergence order $O(h^{2.5})$ owing to the properties of x, especially the differential components. However, one looses approximately one additional power of h which can be attributed to the $O(1/t)$ behavior of $Q_{can}(t)$. □

3.4.2 Nonlinear Problem

Now we turn to the nonlinear BVP (174), (175). We assume the DAE to be regular with index 1 all overall for $t > 0$, but allow a critical point at the left boundary which causes a singularity in the inherent nonlinear ODE. In [43], the case when the inherent ODE system is singular with a singularity of the first kind is studied and polynomial collocation applied to the original DAE system is analyzed. It is shown that for a certain class of well-posed boundary value problems in DAEs having a sufficiently smooth solution, the global error of the collocation scheme converges uniformly with the stage order. Due to the singularity, superconvergence at the mesh points does not hold in general. We outline some aspects from [43].

Regarding the experience with conditions (188) for linear BVPs, it is assumed that

$$tG_1(y,x,t)^{-1} \tag{198}$$

has a continuous extension for $t \to 0$, where

$$\begin{aligned} &G_0(y,x,t) := f_y(y,x,t)D(t), \\ &G_1(y,x,t) := G_0(y,x,t) + f_x(y,x,t)Q_0(t), \quad (y,x,t) \in \mathbb{R}^n \times \mathcal{D} \times [0,1]. \end{aligned}$$

Additionally, to prevent the additional difficulties caused by unbounded canonical projectors known in the linear case, in [43] the canonical projector function Π_{can} along $\ker D$ given by

$$\Pi_{can}(y,x,t) := I - Q_0(t)G_1(y,x,t)^{-1} f_x(y,x,t)$$

is assumed to remain bounded for $t \to 0$. The following practical criterion of the latter property is given in [43]. Let $W(y,x,t) \in \mathbb{R}^m$ denote the orthoprojector matrix onto $\operatorname{im} f_y(y,x,t)^{\perp}$, pointwise for all arguments. Since $f_y(y,x,t)$ has constant rank n for $t > 0$, $W(y,x,t)$ depends continuously on its arguments for $t > 0$. We assume that W has a continuous extension W^{ext} for $t \to 0$, such that, for $t > 0$,

$$W^{ext}(y,x,t) = W(y,x,t).$$

We emphasize that, due to a possible rank drop of $f_y(y,x,t)$ at $t = 0$, in general $W^{ext}(y,x,0) \neq W(y,x,0)$, but $W^{ext}(y,x,0) f_y(y,x,0) = 0$. Then the canonical projector function Π_{can} has a continuous extension exactly if

$$\operatorname{rank} \begin{bmatrix} W^{ext}(y,x,0) f_x(y,x,0) \\ D(0) \end{bmatrix} = m. \tag{199}$$

An inspection of Examples 3.2 and 3.3 confirms this criterion.

To apply standard linearization arguments, the BVP (174)–(175) is supposed to possesses a solution $x_\star \in \mathcal{C}_D^1([0,1],\mathbb{R}^m)$ and the linearization of the DAE (174) along $x_\star$,

$$A_\star(t)(D(t)z(t))' + B_\star(t)z(t) = 0, \quad t \in (0,1], \tag{200}$$

is considered. Since the matrix

$$G_{\star 1}(t) := A_\star(t)D(t) + B_\star(t)Q_0(t) = G_1((D(t)x_\star(t))', x_\star(t), t)$$

is nonsingular for $t \in (0,1]$, the linear DAE (200) is regular with tractability index 1 on the interval $(0,1]$. Thus the linearized BVP can be treated as in Subsubsection 3.4.1.

In analogy to Definition 2.5 , one says that the solution $x_\star$ of the BVP (174)–(175) is *isolated* if and only if its linearization

$$\begin{aligned} &A_\star(t)(D(t)z(t))' + B_\star(t)z(t) = 0, \quad t \in (0,1], \\ &B_0 D(0)z(0) + B_1 D(1)z(1) = 0, \end{aligned}$$

has only the trivial solution. In this case, as common in the theory of singular explicit ODEs (e.g., [69, 68]), also the nonlinear BVP (174), (175) is said to be *well-posed* in [43].

The decoupling function $\omega : \mathcal{D}_\omega \times (0,1] \to \mathbb{R}^m$ and the decoupled form (cf., (63), (64)) of the nonlinear DAE (174),

$$u'(t) = D(t)\omega(u(t),t), \quad t \in (0,1]. \tag{201}$$

$$x(t) = D(t)^{-}u(t) + Q_0(t)\omega(u(t),t), \quad t \in (0,1], \tag{202}$$

can be used for $t > 0$ in order to specify the inherent explicit ODE associated with the nonlinear DAE.

To apply the standard analysis for singular boundary value problems, cf. [40, 68], it is assumed that the decoupling function ω satisfies

$$D(t)\omega(u,t) = \frac{1}{t}M(t)u + q(u,t), \quad u \in \mathcal{D}_\omega, \quad t \in (0,1], \tag{203}$$

where the $n \times n$ matrix function M and the function q are appropriately smooth for $t \to 0$. Note that in [43] a special class of quasi-linear DAEs is shown to meet the conditions 198), (203), as well as to feature a bounded canonical projector function.

This yields the BVP

$$u'(t) = \frac{1}{t}M(t)u(t) + q(u(t),t), \quad t \in (0,1], \tag{204}$$

$$B_0u(0) + B_bu(1) = \gamma. \tag{205}$$

In turn, the linearization of the last BVP reads,

$$\zeta'(t) = D(t)\omega_u(u_\star(t),t)\zeta(t) = \frac{1}{t}M_\star(t)\zeta(t), \quad t \in (0,1], \tag{206}$$

$$B_0\zeta(0) + B_1\zeta(1) = 0, \tag{207}$$

with

$$M_*(t) := -tD(t)G_{*,1}(t)^{-1}B_*(t)D(t)^{-}, \quad t \in (0,1].$$

We can now specify the necessary and sufficient conditions for the linear ODE problem (206)–(207) to have only the trivial solution. It was shown in [40] that the form of the boundary conditions (207) which guarantee that (206)–(207) has only the trivial solution depends on the spectral properties of the coefficient matrix $M_\star(0)$. Note that (203) implies

$$M_\star(t) = M(t) + tg_u(u_\star(t),t), \quad t \in (0,1]$$

and therefore $M_\star(0) = M(0)$. To avoid fundamental modes of (206) which have the form $\cos(\sigma\ln(t)) + i\sin(\sigma\ln(t))$, we assume that zero is the only eigenvalue of $M(0)$ on the imaginary axis.

Now, let R_+ denote the projection onto the invariant subspace which is associated with eigenvalues of $M(0)$ which have strictly positive real parts. Let Q_M be a projection onto the kernel of $M(0)$. Finally, define

$$U := R_+ + Q_M, \quad V := I - U, \tag{208}$$

The BVP (206)–(207) is well-posed if and only if the boundary conditions (207) can equivalently be written as ([40])

$$V\zeta(0)=0, \quad R_+\zeta(1)=0, \quad Q_M\zeta(0)=0, \text{ or } Q_M\zeta(1)=0. \tag{209}$$

The first set of homogeneous initial conditions specified in (209) are necessary and sufficient for ζ to be continuous on the closed interval $[0,1]$.

The polynomial collocation methods (uniform approach A) described in Subsubsection 3.1.2 are used in [43] to approximate the solution of well-posed singular nonlinear BVPs (174), (175). The basic collocation scheme

$$\begin{aligned}
u_\pi(t_i^-)-u_\pi(t_i)&=0, \quad i=1,\ldots,N-1,\\
x_\pi(t_i^-)-x_\pi(t_i)&=0, \quad i=1,\ldots,N-1,\\
f(u'_\pi(\tau_{ik}),x_\pi(\tau_{ik}),\tau_{ik})&=0, \quad k=1,\ldots,s,\ i=0,\ldots,N-1\\
u_\pi(\tau_{ik})-D(\tau_{ik})x_\pi(\tau_{ik})&=0, \quad k=1,\ldots,s,\ i=0,\ldots,N-1,\\
B_0u_\pi(a)+B_1u_\pi(b))&=\gamma,
\end{aligned}$$

is completed by the consistency conditions

$$D(a)x_\pi(a)-u_\pi(a)=0, \quad W^{ext}(u'_\pi(a),x_\pi(a),a)f(u'_\pi(a),x_\pi(a),a)=0.$$

By means of the analytical decoupling and the commutativity of discretization and decoupling, one obtains a classical collocation scheme for the component u_π. According to [68, Theorem 3.1], there exists a unique collocation solution $u_\pi\in\mathcal{B}^n_{\pi,s}\cap\mathcal{C}([0,1],\mathbb{R}^n)$, under the assumptions that the underlying analytical problem is well-posed with sufficiently smooth data, and that the mesh is sufficiently fine. Finally, $x_\pi\in\mathcal{B}^m_{\pi,s}\mathcal{C}([0,1],\mathbb{R}^m)$ is uniquely specified by its values at all collocation points, see (127), and the consistency conditions. It results that

$$\|x_*-x_\pi\|_\infty=O(h^s), \quad \|u_*-u_\pi\|_\infty=O(h^s).$$

3.5 Defect-based a posteriori error estimation for index-1 DAEs

When designing error estimation procedures, one usual has different choices. One of the most popular is a very robust and easy to implement $h-h/2$ strategy, where the basic method is carried out first on a given, not necessarily uniform, grid and then repeated on a grid with doubled number of subintervals. This procedure is used often in software for boundary value problems in ODEs and DAEs, for instance, in COLNEW, COLDAE, see [13]. Since this procedure in context of collocation methods is quite expensive, it seems reasonable to look for cheaper alternatives.
Here, we describe a computationally efficient a posteriori error estimator for collocation solutions to linear index-1 DAEs in properly stated formulation proposed in

[18]. The procedure is based on a modified defect correction principle, extending an established technique from the ODE context to the DAE case. The resulting error estimate is proved to be asymptotically correct and tested in numerical experiments with IVPs. For all technical details, we refer the reader to [18].

Let us consider a regular index-1 DAE with properly stated leading term

$$A(t)(Dx)'(t) + B(t)x(t) = q(t), \quad t \in [a,b], \tag{210}$$

satisfying the general assumptions in Subsection 2.1, and, additionally, condition (108) yielding the border projector $R = I$. Moreover, here we assume the coefficient D to be even constant. Otherwise one can turn to the enlarged version according to (109), (110) of the DAE under consideration.

We consider a well-posed BVP (cf. Subsection 2.3) for the DAE (210) and the collocation equations

$$A(\tau_{ik})u'_\pi(\tau_{ik}) + B(\tau_{ik})x_\pi(\tau_{ik}) = q(\tau_{ik}), \tag{211}$$

$$Dx_\pi(\tau_{ik}) - u_\pi(\tau_{ik}) = 0, \quad k = 1,\ldots,s,\ i = 0,\ldots,N-1, \tag{212}$$

with

$$s \text{ even}, \quad \rho_s = 1.$$

Note, in particular, that $\rho_s = 1$ is essential for the analysis. This ensures in a natural way stability of the integration schemes, cf. [60, 83] for a more detailed discussion. We also assume that s is even, which will be necessary to guarantee the asymptotic correctness of our error estimator to be defined in Section 3.5.2.

The focus is now on the effective design and analysis of an asymptotically correct a posteriori error estimator for collocation solutions to (210), with a uniform, 'black box' treatment of the differential and algebraic components, and an appropriate handling of the case where $D(t)$ is not constant. The generalization of the method and its analysis for DAEs with a singular inherent ODE can be found in [19].

3.5.1 The main idea of the defect-based error estimation

A posteriori error estimation in ODEs based on the defect correction principle is an old idea originally due to Zadunaisky [115] and further developed by Stetter [110]. In the context of regular and singular ODEs, this approach was refined and analyzed in [16, 17] and implemented in [14]. In particular, for a special realization of the defect, an efficient, asymptotically correct error estimator, the QDeC estimator, was designed in [16] for collocation solutions on arbitrary grids. These ideas have been extended to the DAE context in [18], which appears not to be straightforward because of the coupling between differential and algebraic components. In abstract notation, the basic structure of a defect-based estimator can be described as follows: Consider a numerical solution ξ_π which approximates the vector of exact solution values x^*_π, $\xi_\pi \approx x^*_\pi$, for a problem

$$F(x(t)) = 0, \quad t \in [a,b], \tag{213}$$

on a grid π. Define the *defect* $d = d(t)$ by interpolating ξ_π by a continuous piecewise polynomial function $p(t)$ of degree $\leq s$ and substituting $p(t)$ into (213),

$$d(t) := F(p(t)), \quad t \in [a,b]. \tag{214}$$

Obviously, $p(t)$ is the exact solution to a *neighboring problem*

$$F(x(t)) = d(t) \tag{215}$$

related to the original problem (213). Now we use a procedure of low effort (typically a low order scheme), the so-called *auxiliary scheme* $\tilde{F}$, to obtain approximate discrete solutions $\tilde{x}_\pi$ and $\tilde{x}_\pi^{def}$ for both the original and neighboring problems on the grid π, i.e., $\tilde{F}(\tilde{x}_\pi) = 0$ and $\tilde{F}(\tilde{x}_\pi^{def}) = d_\pi$, where d_π is an appropriate restriction of $d(t)$ to the grid π.

Since (213) and (215) differ only by the (presumably) small defect d, we expect that

$$\varepsilon_\pi := \tilde{x}_\pi^{def} - \tilde{x}_\pi \tag{216}$$

is a good estimate for the global error

$$e_\pi := \xi_\pi - x_\pi^*. \tag{217}$$

In other terms,

$$\begin{aligned} e_\pi := \xi_\pi - x_\pi^* &\approx F^{-1}(d) - F^{-1}(0) \\ &\approx \tilde{F}^{-1}(d_\pi) - \tilde{F}^{-1}(0) = \tilde{x}_\pi^{def} - \tilde{x}_\pi = \varepsilon_\pi. \end{aligned} \tag{218}$$

This is exactly the procedure originally proposed in [110]. However, in concrete applications, the auxiliary scheme $\tilde{F}$ and a suitable representation for the defect d_π have to be carefully chosen. In particular, in [16] collocation for the ODE case was considered. For $\tilde{F}$ chosen as the backward Euler scheme, it was shown that a modified version of the pointwise defect (214) has to be used in order to obtain an asymptotically correct estimator for the error of a given collocation approximation $x_\pi(t)$ yielding ξ_π. In the following section this approach (the 'QDeC estimator') is described in more detail and will be extended to the DAE case.

3.5.2 The QDeC estimator for DAEs

Now we apply the procedure described in Section 3.5.1 to the linear DAE (210). In addition to the collocation method, we use a scheme of backward Euler type over the collocation nodes as an auxiliary method. Let $h_{ik} := \tau_{ik} - \tau_{i,k-1}$ and consider the grid function ε_{ik} satisfying the auxiliary scheme

$$A(\tau_{ik})\,\frac{D\varepsilon_{ik}-D\varepsilon_{i,k-1}}{h_{ik}}+B(\tau_{ik})\varepsilon_{ik}=\bar{d}_{ik}, \tag{219}$$

with homogeneous initial condition $\varepsilon_{0,0}=0$ and the backward Euler scheme playing the role of $\tilde{F}$. According to (214), the straightforward, classical way to define the defect $\bar{d}_{ik}$ would be to substitute $x_\pi(t)$ into (210) in the pointwise sense,

$$d(t):=A(t)(Dx_\pi)'(t)+B(t)x_\pi(t)-q(t),\quad t\in[a,b], \tag{220}$$

and using the pointwise defect $\bar{d}_{ik}:=d(\tau_{ik})$ in (219). However, as has been pointed out in [16] in the ODE context, this procedure does not lead to successful results. For collocation this is obvious: Since, by definition of the collocation solution (211), the defect $d(\tau_{ik})$ which enters the backward Euler scheme, vanishes at each point τ_{ik} ($i=0\ldots N-1$, $k=1\ldots s$), the error estimate $\varepsilon(\tau_{ik})$ would always be zero.

In slight variation of the procedure introduced in [16], we now define a modified defect via the integral means

$$\bar{d}_{ik}:=\sum_{l=0}^{s}\alpha_{kl}d(\tau_{il})=\frac{1}{h_{ik}}\int_{\tau_{i,k-1}}^{\tau_{ik}}d(t)\mathrm{d}t+\mathcal{O}(h^{s+1}), \tag{221}$$

for $i=0,\ldots,N-1$, $k=1,\ldots,s$, where the α_{kl} are quadrature coefficients for the integral means in (221), i.e.,

$$\alpha_{kl}=\frac{1}{\rho_k-\rho_{k-1}}\int_{\rho_{k-1}}^{\rho_k}L_l(t)dt,\quad k=1\ldots s,\ l=0\ldots s, \tag{222}$$

with the Lagrange polynomials L_l of degree s, such that $L_l(\rho_k)=\delta_{kl}$. Note that, in contrast to collocation at s nodes in each subinterval excluding the left endpoint t_i, we now include the additional node $\tau_{i0}:=t_i+h_i\rho_0$ with $\rho_0=0$, for the polynomial quadrature defining (221).

The following result is proved in [18].

Theorem 3.6. *While the global error of the collocation method (211) is of order h^s, i.e.,*

$$e(t)=x_\pi(t)-x_*(t)=\mathcal{O}(h^s), \tag{223}$$

the error estimate of the global error (223) based on the modified defect (221) and the auxiliary scheme (219) is asymptotically correct, i.e.,

$$\varepsilon_{ij}-e(\tau_{ij})=\mathcal{O}(h^{s+1}). \tag{224}$$

Example 3.5. We consider the initial value problem

$$\begin{bmatrix}\mathrm{e}^t\\ \mathrm{e}^t\end{bmatrix}\left(\begin{bmatrix}1&0\end{bmatrix}x\right)'(t)+\begin{bmatrix}\mathrm{e}^t(1+\cos^2 t) & \cos^2 t\\ \mathrm{e}^t(-1+\cos^2 t) & -\cos^2 t\end{bmatrix}x(t)=\begin{bmatrix}\sin^2 t(1-\cos t)-\sin t\\ \sin^2 t(-1-\cos t)-\sin t\end{bmatrix}, \tag{225}$$

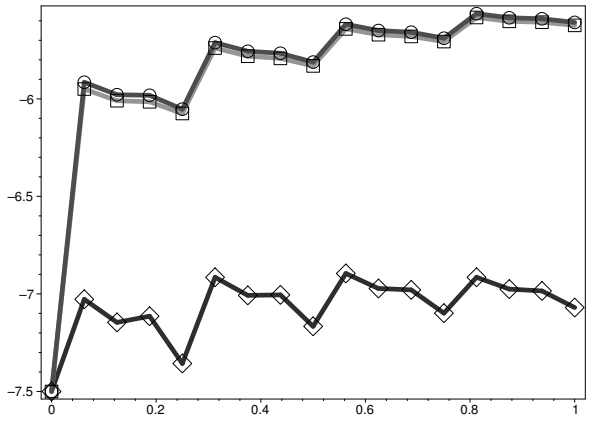

Fig. 7 $\log_{10}$ - plot for first solution component, $N = 4$ of Example 3.5

○ ... error $|e_1(t)| = |x_{\pi,1}(t) - x_{*,1}(t)|$
□ ... error estimate $|\varepsilon_1(t)|$
◇ ... error of error estimate $|\varepsilon_1(t) - e_1(t)|$

on $[a,b] = [0,1]$, with initial condition $x_1(0) = 1$. We use a realization of our method in MATLAB, based on collocation at equidistant points with $s = 4$, on $N = 2,4,8,16,32$ subintervals of length $1/N$. In the following tables, the asymptotical order $\varepsilon - e = \mathcal{O}(h^{s+1})$ is clearly visible; see also Figure 7.

- First solution component, at $t = 1$:

N	e	ord$_e$	$\varepsilon - e$	ord$_{\varepsilon-e}$
4	−2.466e–06	3.8	8.513e–08	4.6
8	−1.634e–07	3.9	2.989e–09	4.8
16	−1.051e–08	4.0	9.886e–11	4.9
32	−6.664e–10	4.0	3.180e–12	5.0

- First solution component, maximum absolute values over all collocation points $\in [0,1]$:

N	e	ord$_e$	$\varepsilon - e$	ord$_{\varepsilon-e}$
4	2.732e–06	4.0	1.272e–07	5.3
8	1.711e–07	4.0	3.578e–09	5.2
16	1.074e–08	4.0	1.074e–10	5.1
32	6.734e–10	4.0	3.311e–12	5.0

- Second solution component, at $t = 1$:

N	e	ord_e	$\varepsilon - e$	$\mathrm{ord}_{\varepsilon-e}$
4	2.906e–05	3.8	−7.927e–07	4.6
8	1.522e–06	3.9	−2.783e–08	4.8
16	9.788e–08	4.0	−9.206e–10	4.9
32	6.205e–09	4.0	−2.961e–12	5.0

□

3.6 Further references, comments, and open questions

Remark 3.1. In essence, for $s = 3$ and Lobatto points $\rho_1 = 0, \rho_2 = \frac{1}{2}, \rho_3 = 1$, Theorem 3.1 reflects results obtained in [73, 41, 42] in a quite different way using a rigorous functional-analytic discretization theory. This work applies to DAEs $f(Px)'(t), x(t), t) = 0$ showing a constant projector matrix instead of the matrix function D in (107), which allows to restrict the consideration directly to $u_* = Px_*$, $v_* = (I-P)x_*$ and their approximations. [41, Theorem 4.13] provides superconvergence order 4. Moreover, a stability inequality is verified and global error estimations by defect correction are provided.

Remark 3.2. The early work [113] deals with BVPs providing periodical solutions. A special collocation method using trigonometrical polynomials is developed.

Remark 3.3. Here, we did not regard the possible implementations of the various collocation approaches for BVPs in DAEs. Of course, the special ansatz of the piecewise polynomial functions x_π, the arrangement of the finite-dimensional nonlinear equations to be solved, the linear and nonlinear equation solvers play an important role and the error estimates and mesh control as well.
As noted, e.g., in [89, 55], if integration methods approved for regular ODEs are applied to index-1 DAEs, then additional stability conditions might appear. In particular, the implicit midpoint rule applied to the simple equation $x(t) = 0$, $t \in [0,1]$, leads, in the worst case, to a linear growth of the involved perturbations. It is unclear whether and to what extend those effects can be resolved.
Concerning the different collocation approaches to DAEs, till now it remains generally open which versions will prove to be more favorable. This question is closely related to the aspects of possible implementations.

Remark 3.4. Singularities of the flow of an DAE might be caused by a singular inherent ODE as in Subsection 3.4, but also by the other components of a DAE, see [102, 103, 83]. In the context of the projector based DAE analysis, *regular points*

are supported by several constant-rank conditions. By definition, for *critical points* at least one of these rank conditions is violated. In general, among critical points might be so-called harmless ones ([83, 44]), however, this does not happen for singular index-1 DAEs.
Attempts to detect DAE singularities in practice are reported in [49, 50]. First solvability results justifying the notion *well-posed BVPs for singular index-1 DAEs* are proved in [98].

Remark 3.5. Linear BVPs in DAEs are treated in [56] by means of *least squares collocation*, which represents a special method created for ill-posed problems. It is an open question whether such approaches could be advanced to become practicable for a considerable class of BVPs.

Remark 3.6. The projected collocation is adapted in [51] to work for BVPs associated with periodic motions in multibody system dynamics. The collocation scheme is applied to an index-2 formulation of the related DAE. Besides the projections at the meshpoints, an extra boundary projection is introduced.

Remark 3.7. The idea of backward projection has been used for numerical integration of regular ODEs and index-1 DAEs for maintaining given invariants numerically, e.g., [52, 106, 108]. A generalization of backward projection and selective backward projection as *projected defect correction* is developed in [93] for a quite large class of nonlinear index-2 DAEs. We conjecture, that it would work also for general regular index-2 DAEs (107) satisfying (108). Further, this way, projected collocation for the corresponding BVPs possibly could work well.

4 Shooting methods

The *shooting method* or *initial-value adjusting method* - a description used in very former publications - is a classical method to solve two-point boundary value problems (TPBVP) but also multi-point boundary value problems for ODEs and DAEs. The first appeared papers dealing with DAEs and shooting methods are [89], [55], [38], [79]. The idea is to imbed the BVP into a family of IVPs, with unknown initial values, and then to seek among them the true one.
We consider the TPBVP

$$f((Dx)'(t),x(t),t) = 0, \quad t \in [a,b] \tag{226}$$

$$g(x(a),x(b)) = 0. \tag{227}$$

We assume that the DAE (226) is regular with index μ and that the TPBVP has a locally unique solution x_*. Set $z_* := x_*(a)$.

As it is well-known, for explicit ODEs there is a neighborhood $\mathcal{N}_* \subseteq \mathbb{R}^m$ around z_* so that all IVPs with the initial condition $x(a) = z \in \mathcal{N}_*$ are uniquely solvable, their solutions exist on the entire interval $[a,b]$ and depend smoothly on z.

In contrast, for DAEs, the extra condition $z \in \mathcal{M}_{\mu-1}(a)$ is necessary for solvability, whereby the associated set of consistent initial values $\mathcal{M}_{\mu-1}(a)$ is a lower dimensional subset of $\mathbb{R}^m$. For linear DAEs, an explicit theoretical description is given in Subsection 2.2. Generally, no direct description is available, except for the index-1 case, where $\mathcal{M}_0(a)$ is the obvious restriction set.

We try to overcome this difficulty by formulating the corresponding IVPs with the initial condition

$$C(x(a) - z) = 0, \quad z \in \mathcal{N}_* \tag{228}$$

with an appropriate singular matrix $C \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^l)$. As shown in Section 2, linear IVPs have unique solutions existing on $[a,b]$, if C is such that

$$\ker C = \ker \Pi_{can} = \ker \Pi_{\mu-1}, \tag{229}$$

and IVPs in nonlinear index-1 DAEs are uniquely solvable with solutions existing on $[a,b]$, if

$$\ker C = \ker \Pi_{can} = \ker \Pi_0 = \ker D(a).$$

For nonlinear higher index DAEs the situation is much more difficult since then C itself might become solution dependent.

If the IVPs (226), (228) are uniquely solvable on $[a,b]$, then one looks for a z such that the boundary condition (227) is satisfied. This is the basic idea of the shooting method.

4.1 Solution of linear DAEs

Suppose a properly stated linear regular DAE with index μ. We consider the linear TPBVP (37) and a related IVP

$$A(Dx)' + Bx = q, \tag{230}$$
$$C(x(a) - z) = 0 \tag{231}$$

and C is chosen fulfilling (229) with given value z. The solution of the IVP is represented in (35) as

$$x(t) = X(t,a)z + \int_a^t X(t,s)G_\mu^{-1}(s)q(s)\mathrm{d}s + v_q(t)$$

and we discover using the structure of X (cf. (29)) from

$$X(t,a)z = X(t,a)D(a)^- D(a)\Pi_{\mu-1}(a)z$$

that the solution x depends for a given right-hand side q from the initial value $\xi := D(a)\Pi_{\mu-1}(a)z$ only and not from the whole vector z. The component $(I - \Pi_{can}(a))z$ does not matter at all.

We denote the solution of an IVP (230), (231) by $x(t;a,\xi)$. This means that we implicitly assume for the moment, that we know the solution also at $t = a$ (This is the difficult problem of computing consistent initial values, which is discussed later on). Thus $x(a;a,\xi) = x(a) = X(a,a)\xi + v_q(a)$. At $t = b$ we have with the general solution expression (38) that

$$x(b;a,\xi) = x(b) = X(b,a)D^-(a)\xi + \int_a^b X(b,s)G_\mu^{-1}(s)q(s)\mathrm{d}s + v_q(b).$$

$x(\cdot;a,\xi)$ solves the DAE (230) and to solve the TPBVP (37) also the boundary condition has to be fulfilled. The relation to determine ξ is given by

$$\begin{aligned} G_a x(a) + G_b x(b) = {} & G_a(X(a,a)D^-(a)\xi + v_q(a)) + G_b(X(b,a)D^-(a)\xi \\ & + \int_a^b X(b,s)G_\mu^{-1}(s)q(s)\mathrm{d}s + v_q(b)) = \gamma. \end{aligned} \tag{232}$$

We obtain the linear system

$$\underbrace{(G_a X(a,a) + G_b X(b,a))}_{=S} D^-(a)\xi = \hat{\gamma}$$

with $\hat{\gamma} = \gamma - G_a v_q(a) - G_b(\int_a^b X(b,s)G_\mu^{-1}(s)q(s)\mathrm{d}s + v_q(b))$ (cf. (39)). Theorem 2.1 provides a unique initial value $D^-(a)\xi$. The solution of the IVP (230) and the initial condition

$$D(a)\Pi_{\mu-1}(a)(x(a)-D^-(a)\xi)=0,$$

i.e., $C=D(a)\Pi_{\mu-1}(a)$, has the solution of the TPBVP represented by the solution of an IVP with the initial value ξ. Because of $C(x(a)-D^-(a)\xi)=D(a)\Pi_{\mu-1}(a)(x(a)-D^-(a)\xi)=0$ follows $D(a)\Pi_{\mu-1}(a)x(a)=D(a)\Pi_{\mu-1}(a)D^-(a)\xi=\xi$.

For the practical application of the shooting method two of our assumptions are difficult to realize. First, the used choice of the matrix C in the initial condition differs usually from $D(a)\Pi_{can}(a)$ (see Remark 4.1) and second, in general, the integration codes do not provide consistent initial values, i.e., the full vector $x(a)$. But in contrast to IVPs we have to know the whole vector $x(a)$, to evaluate the boundary condition (227). Additionally, consistent initial values are very helpful to start an integration itself.

4.1.1 Computation of consistent initial values

The computation of consistent initial values in the index-μ case is a nontrivial task. In the literature we find several papers, which focus on that topic using various ways to compute consistent initial values. [79] and [47] propose for index-1 DAEs the use of the tractability index concept, [2], [36] and [67] assume a semiexplicit structure of the DAE, which makes the computation much easier. [53] considers special structured index-2 DAEs, which are reduced to index 1 by differentiation.

We investigate proper formulated linear index-μ DAEs. We have to compute at an interesting time point $\bar{t}$ vectors $y:=(D^-(Dx)')(\bar{t})$ and $v:=(I-\Pi_{\mu-1}(\bar{t}))x(\bar{t})$. These values have with known value $\xi:=D(\bar{t})\Pi_{\mu-1}(\bar{t})x(\bar{t})$ at least to fulfill

$$ADy+B(D^-\xi+v)=q(\bar{t}). \tag{233}$$

Because of $\operatorname{rank} D=r_0$ and $\operatorname{rank}(I-\Pi_{\mu-1})=m-l$ we have to determine $d:=r_0+m-l$ unknowns but we have m natural conditions only. Using the dynamical degree l (cf. (76)) we see that for $\mu=1$ we have $d=m$ and if $\mu>1$ we obtain $d>m$, i.e. we need additional conditions to compute consistent initial values. These additional conditions are the so-called hidden constraints, which are computed by differentiating suitable relations.

We define an operator $\mathcal{I}_\mu$, which computes for linear index-μ DAEs y and v depending from a known ξ as

$$\begin{pmatrix} y \\ v \end{pmatrix}=\mathcal{I}_\mu(\xi,\bar{t}). \tag{234}$$

We demonstrate the operator $\mathcal{I}_\mu$ for index-1 and index-2 DAEs.
The index-1 case:

We have to compute $d=r_0+m-l=m$ values. We define y as before and $v:=(I-P_0(\bar{t}))x(\bar{t})=Q_0(\bar{t})x(\bar{t})$. Eq. (233) looks

$$ADy + B(D^-\xi + v) = q(\bar{t}), \tag{235}$$
$$Q_0 y + P_0 v = 0. \tag{236}$$

Eq. (236) ensures that y and v lie in the right subspaces. The initial condition reads with $C = D(\bar{t})$ as

$$D(\bar{t})(x(\bar{t}) - z) = 0$$

with given z. The Jacobian matrix of (235), (236) with respect to y, v is the regular matrix $J_1 := \begin{pmatrix} G_0 & B \\ Q_0 & P_0 \end{pmatrix}$. Using the inverse $J_1^{-1} = \begin{pmatrix} P_0 G_1^{-1} & Q_0 - P_0 G_1^{-1} B P_0 \\ Q_0 G_1^{-1} & (I - Q_0 G_1^{-1} P_0) P_0 \end{pmatrix}$ we obtain

$$\begin{pmatrix} y \\ v \end{pmatrix} = J_1^{-1} \begin{pmatrix} q - BD^-\xi \\ 0 \end{pmatrix} =: \mathcal{I}_\mu(\xi).$$

With

$$v = Q_0 G_1^{-1}(q - BD^-\xi) \tag{237}$$

we obtain $\frac{\partial v}{\partial \xi} = -Q_0 G_1^{-1} B D^- = -\mathcal{H}_0$ for index 1 (cf. (291)).

The index-2 case:

The number of unknowns is $d = r_0 + m - l = 2m - r_1$. We are looking as in the index-1 case for $y = D^-(Dx)'(\bar{t})$ and now $v := (I - \Pi_1(\bar{t}))x(\bar{t})$. In contrast to the index-1 case we have to add a relation to describe the hidden constraint (cf. [83, Ch. 2.10.3 and 10.2.2.1]). For that reason we differentiate the equation

$$W_1 B x = W_1 q$$

resulting from the multiplication of (230) by the projector W_1 projecting along $\operatorname{im} G_1$ and we obtain with $W_1 B Q_0 = 0$

$$W_1 B \underbrace{D^-(Dx)'}_{=y} + (W_1 B D^-)' D x = (W_1 q)'.$$

This leads to the system

$$ADy + B(D^-\xi + v) = q(\bar{t}), \tag{238}$$
$$W_1 B y + (W_1 B D^-)'(\xi + Dv) = (W_1 q)'(\bar{t}), \tag{239}$$
$$Q_0 y + \Pi_1 v = 0. \tag{240}$$

We solve Eqn. (238)–(240) explicitly for a given value $\xi = D\Pi_1 D^-\xi$. Multiplying (238) by $Q_1 G_2^{-1}$ provides using the relations $v = (I - \Pi_1)v$ from (240) and the admissible projector $Q_1 = Q_1 G_2^{-1} B_1$, which realizes a fine decoupling,

$$Q_1G_2^{-1}Bv = Q_1G_2^{-1}q(\bar{t}) - \underbrace{Q_1G_2^{-1}BD^-D}_{Q_1}P_1D^-\xi \quad \text{and we obtain}$$

$$Q_1v = Q_1G_2^{-1}q(\bar{t}),$$

i.e., $P_0v = \underbrace{\Pi_1v}_{=0} + P_0Q_1v = P_0Q_1G_2^{-1}q(\bar{t})$. The multiplication of (239) by $Q_1G_2^{-1}$ results because of $Q_1G_2^{-1}W_1 = Q_1G_2^{-1}$ and $Dv = DP_0v$ as

$$\underbrace{Q_1G_2^{-1}BP_0}_{=Q_1}y = Q_1G_2^{-1}((W_1q)'(\bar{t}) - (W_1BD^-)'(\xi + Dv)).$$

From Eq. (238) we obtain by scalation with G_2^{-1}

$$\begin{aligned}
G_2^{-1}G_0y + G_2^{-1}BQ_0v &= G_2^{-1}(q(\bar{t}) - BD^-(\xi + Dv)),\\
(\Pi_1 - Q_0Q_1)y + Q_0v &= G_2^{-1}(q(\bar{t}) - BD^-(\xi + Dv)),\\
\Pi_1y + Q_0v &= G_2^{-1}(q(\bar{t}) - BD^-(\xi + Dv)) + Q_0Q_1y. \qquad (241)
\end{aligned}$$

Multiplying Eqn. (241) by Π_1 respectively Q_0, we obtain

$$\begin{aligned}
\Pi_1y &= \Pi_1G_2^{-1}(q(\bar{t}) - BD^-(\xi + Dv)), \text{ respectively}\\
Q_0v &= Q_0G_2^{-1}(q(\bar{t}) - BD^-(\xi + Dv)) + Q_0Q_1y.
\end{aligned}$$

Summarizing the components of y and v we obtain

$$\begin{aligned}
y &= \Pi_1G_2^{-1}(q(\bar{t}) - BD^-(\xi + Dv)) + P_0Q_1G_2^{-1}((W_1q)'(\bar{t}) - (W_1BD^-)'(\xi + Dv)),\\
v &= (I - \Pi_1)G_2^{-1}q(\bar{t}) - Q_0G_2^{-1}BD^-(\xi + Dv) + Q_0Q_1y. \qquad (242)
\end{aligned}$$

Later on we will need the relation between v and ξ. With (242) we obtain

$$\frac{\partial v}{\partial \xi} = -Q_0G_2^{-1}BD^- - Q_0Q_1G_2^{-1}(W_1BD^-)' = -\mathcal{H}_0 \qquad (243)$$

for the index-2 case (cf. Appendix (291)).

Lemma 4.1. *The linear DAE* (233) *has index 2 and let v be the solution of Eqn.* (238)–(240). *We choose the fine decoupling projector* $Q_1 = Q_1G_2^{-1}B_1$ *and assume that* $Q_0Q_1G_2^{-1}, Q_0Q_1D^- \in \mathcal{C}^1$ *then* $(D^- - \frac{\partial v}{\partial \xi})D\Pi_1 = \Pi_{can,2}$.

Proof. Using Eq. (243) we consider

$$(D^- - \frac{\partial v}{\partial \xi})D\Pi_1 = (D^- - Q_0G_2^{-1}BD^- - Q_0Q_1G_2^{-1}(W_1BD^-)')D\Pi_1.$$

It holds that $Q_0G_2^{-1}B\Pi_1 = Q_0(P_1 + \underbrace{Q_1}_{=0})G_2^{-1}B\Pi_1 = Q_0P_1G_2^{-1}B\Pi_1$ and

$$
\begin{aligned}
Q_0Q_1G_2^{-1}(W_1BD^-)'D\Pi_1 &= Q_0((Q_0Q_1D^-)' - (Q_0Q_1G_2^{-1})'W_1BD^-)D\Pi_1D^-D\Pi_1,\\
&= -\underbrace{Q_0Q_1D^-}_{=-Q_0P_1D^-}(D\Pi_1D^-)'D\Pi_1\\
&\qquad - Q_0(Q_0Q_1G_2^{-1})'W_1G_2\underbrace{Q_1G_2^{-1}BD^-D\Pi_1}_{=0}
\end{aligned}
$$

because of $W_1 = W_1G_2Q_1G_2^{-1}$. Now we have with (243) the representation

$$
\begin{aligned}
(D^- - \frac{\partial v}{\partial \xi})D\Pi_1 &= (D^- - (Q_0P_1G_2^{-1}BD^- + Q_0P_1D^-(D\Pi_1D^-)'))D\Pi_1,\\
&= (D^- - \mathcal{H}_0D^-)D\Pi_1 = (I-\mathcal{H}_0)D^-D\Pi_1 = \Pi_{can,2}.
\end{aligned}
$$

□

The relation described in Lemma 4.1 between the v-component of $\mathcal{I}_\mu$ and the canonical projector also holds for arbitrary index μ.

Lemma 4.2. *We consider the regular index-μ DAE* (230)*. We choose fine decoupling projectors $Q_0, Q_1, \ldots, Q_{\mu-1}$ (cf. Subsection 6.1.2) then*

$$
\frac{\partial \mathcal{I}_{\mu,v}}{\partial \xi}(\xi,\bar{t})) = \frac{\partial \mathcal{I}_{\mu,v}}{\partial \xi}(\xi,\bar{t}))D(\bar{t})D^-(\bar{t}) \quad \textit{and} \tag{244}
$$

$$
(D^-(\bar{t}) - \frac{\partial \mathcal{I}_{\mu,v}}{\partial \xi}(\xi,\bar{t}))D(\bar{t})\Pi_{\mu-1}(\bar{t}) = \Pi_{can}(\bar{t}). \tag{245}
$$

hold.

Proof. We are interested in the v-component of $\mathcal{I}_\mu$ only. It holds that $v = (I - \Pi_{\mu-1}x)$ and $v = v_0 + \cdots + v_{\mu-1}$. We refer to the decomposition (291) which explicitly represent the components $v_i,\ i = 0,\ldots,\mu-1$. For a fine decoupling (291) specializes to $\mathcal{H}_1,\ldots,\mathcal{H}_{\mu-1} = 0$. We observe that v_0 depends on ξ only and therefore $\frac{\partial \mathcal{I}_{\mu,v}}{\partial \xi} = \mathcal{H}_0D^-$. This relation shows (244). With $(D^-(\bar{t}) - \mathcal{H}_0D^-(\bar{t}))D(\bar{t})\Pi_{\mu-1}(\bar{t}) = (I-\mathcal{H}_0)D^-(\bar{t})D(\bar{t})\Pi_{\mu-1}(\bar{t}) = \Pi_{can}(\bar{t})$ (cf. Subsection 6.1.2) the proof is done. □

The realization of algorithms to compute consistent initial values using (291) is very expensive. For higher index systems it would be helpful to take advantage from a given structure like Hessenberg form etc.

4.1.2 Single shooting

Here we deal with linear regular index-μ DAEs. In contrast to the ODE-case a shooting method consists not only in the integration of the DAE but also in providing consistent initial values. In [38] we find that the "Knowledge of the solution manifold

... is required ... at the initial time point $t_0 = a$ ...". The shooting method proposed in [79] combined the computation of consistent initial values with the shooting procedure for index-1 DAEs. We generalize this idea to index-μ DAEs.
We consider the TPBVP (37) and a related IVP

$$A(Dx)' + Bx = q, \tag{246}$$
$$C(x(a) - z) = 0. \tag{247}$$

The solution of the IVP (246), (247) at $t = b$ is applying (35) given by

$$x(b;a,u) = X(b,a)D^-(a)\xi + \int_a^b X(b,s)G_\mu^{-1}(s)q(s)\mathrm{d}s + v_q(b)$$

and at $t = a$ we obtain from (35) $x(a) = X(a,a)D^-(a)\xi + v_q(a)$. The boundary condition fixes the unknown ξ we are looking for

$$G_a(X(a,a)D^-(a)\xi + v_q(a)) + G_b x(b;a,\xi) = 0, \tag{248}$$
$$(I - D(a)\Pi_{\mu-1}(a)D^-(a))\xi = 0 \tag{249}$$

and (249) fixes that $\xi \in \mathrm{im}\, D(a)\Pi_{\mu-1}(a)$. But for a realization of (248) we have to know $v_q(a)$ too. Therefore we combine (248) with the equations describing consistent initial values at $t = a$ (cf. (234))

$$\begin{pmatrix} y \\ v \end{pmatrix} - \mathcal{I}_\mu(\xi, a) = 0. \tag{250}$$

Lemma 4.3. *Let the BVP* (37) *be uniquely solvable and the admissible projectors* Q_i, $0 \le i \le \mu - 1$ *realize a fine decoupling. The Jacobian matrix of* (248)–(250) *with respect to* ξ, y, v *has full column rank.*

Proof. The Jacobian matrix is given by

$$J_\mu = \begin{pmatrix} (G_a + G_b X(b,a))D^-(a) & 0 & G_a \\ I - D(a)\Pi_{\mu-1}(a)D^-(a) & 0 & 0 \\ \frac{\partial \mathcal{I}_{\mu,y}}{\partial \xi} & I & 0 \\ \frac{\partial \mathcal{I}_{\mu,v}}{\partial \xi} & 0 & I \end{pmatrix}.$$

We consider the equation $J_\mu \begin{pmatrix} z_\xi \\ z_y \\ z_v \end{pmatrix} = 0$ and we show that $z = 0$. If (37) is uniquely solvable then $\ker S = \ker \Pi_{\mu-1}(a)$ (cf. Theorem 2.1). We obtain $z_v = -\frac{\partial \mathcal{I}_{\mu,v}}{\partial u} z_\xi = -\frac{\partial \mathcal{I}_{\mu,v}}{\partial u} D(a)D^-(a)z_\xi$ using (244). From the second equation of $J_\mu z = 0$ we have the relation $z_\xi = D(a)\Pi_{\mu-1}(a)D^-(a)z_\xi$ and therefore

$$(G_a(D^-(a) - \frac{\partial \mathcal{I}_{\mu,v}}{\partial \xi} D(a)D^-(a)) + G_b X(b,a)D^-(a))z_\xi = 0,$$
$$(G_a(D^-(a) - \frac{\partial \mathcal{I}_{\mu,v}}{\partial \xi})D(a)\Pi_{\mu-1}(a)D^-(a)) + G_b X(b,a)D^-(a))z_\xi = 0,$$

Applying Lemma 4.2 $(D^-(a) - \frac{\partial \mathcal{I}_{\mu,v}}{\partial \xi})D(a)\Pi_{\mu-1}(a) = \Pi_{can}(a) = X(a,a)$, we consider $SD^-(a)z_\xi = 0$ which leads to $\Pi_{\mu-1}(a)D^-(a)z_\xi = 0$ and finally to $z_\xi = 0$. Applying the last two equations results $z_y = 0,\ z_v = 0$. □

The implementation of a single shooting method for index-μ DAEs requires an algorithm to compute consistent initial values and an integration method to solve an IVP and to compute the fundamental matrix $X(b,a)$.

The algorithmic procedure solving a BVP by single shooting method starts with an initial guess z_0. Consistent initial values are computed obtaining the related values ξ_0, v_0, y_0. We solve the IVP (246)-(247) and obtain the solution $x(b;a,u_0)$. The correction $\Delta\xi, \Delta v$ are the solutions of the linear system

$$\begin{pmatrix} (G_a + G_b X(b,a))D^-(a) & G_a \\ I - D(a)\Pi_{\mu-1}(a)D^-(a) & 0 \\ \mathcal{H}_0 & I \end{pmatrix} \begin{pmatrix} \Delta\xi \\ \Delta v \end{pmatrix} = \begin{pmatrix} G_a(D^-(a)\xi_0 + v_0) + G_b x(b;a,\xi_0) - \gamma \\ 0 \\ 0 \end{pmatrix}. \tag{251}$$

The solution of the TPBVP (37) at $t = a$ is $x(a) = D^-(a)(\xi_0 - \Delta\xi) + v_0 - \Delta v$.
It is straightforward that the relation for $\Delta\xi$ finally looks like

$$\begin{aligned}(G_a X(a,a) + G_b X(b,a))D^-(a)\Delta\xi &= SD^-(a)\Delta\xi \\ &= G_a(D^-(a)\xi_0 + v_0) + G_b x(b;a,\xi_0) - \gamma.\end{aligned}$$

The rectangular coefficient matrix can be arranged in such a way that may handle with quadratic matrices. We have to combine the first two rows of equation system (251) (cf. for the index-2 case [77]), because the first row contains the l boundary conditions and the second row the $m-l$-dimensional subspace condition for $\Delta\xi$.

4.1.3 Multiple shooting

The single shooting has also for DAEs the disadvantages known from the ODE case. The chosen (unknown) initial value may not have a calculable solution of the IVP over the whole interval $[a,b]$. We overcome that by the multiple shooting method. The idea of multiple shooting is the subdivision of $[a,b]$ into smaller subintervals

$$a = t_0 < t_1 < \ldots < t_{N-1} < t_N = b.$$

The aim is the reduction of the sensitivity of the initial value problems by shorter integration intervals and a smaller condition number of the resulting coefficient matrix of the linear systems compared with the single shooting coefficient matrix (cf.

(251)).
We discuss here the case of multiple forward (parallel) shooting only. Methods shooting in different direction are analogously applicable like in the ODE case (cf. [78]).
On every subinterval $[t_{j-1},t_j],\ j\in[1,N]$ we solve an IVP. At a matching points t_j we require continuity of the dynamic component u of the solution (cf. Section 2.2). We obtain

$$D\Pi_{\mu-1}(t_j)(D^-(t_j)\xi_j - x(t_j;t_{j-1},\xi_{j-1})) = 0, \quad 1\le j\le N-1 \quad \text{or shorter} \tag{252}$$

$$u_j - D\Pi_{\mu-1}(t_j)x(t_j;t_{j-1},\xi_{j-1}) = 0, \tag{253}$$

with $\xi_j := u(t_j)$ and from the boundary condition

$$G_a(D^-\xi_0 + v_0) + G_b x(t_N;t_{N-1},\xi_{N-1}) = 0. \tag{254}$$

The unknowns are $(\xi_0,\xi_1,\ldots,\xi_{N-1},v_0,y_0)$, i.e., that we have to extend the system by the computation of consistent initial values at t_0 to determine v_0,

$$\begin{pmatrix} y_0 \\ v_0 \end{pmatrix} - \mathcal{I}_\mu(\xi_0,t_0) = 0 \tag{255}$$

and the restriction of ξ_i to the subspace $\operatorname{im} D\Pi_{\mu-1}(t_i),\ i=0,\ldots,N-1$ by

$$(I - D(t_i)\Pi_{\mu-1}(t_i)D^-(t_i))\xi_i = 0 \tag{256}$$

as in the single shooting case. For an implementation of the multiple shooting methods for DAEs these additional equations computing consistent initial values are necessary at every shooting point. (This was mentioned the first time in [38]).
We obtain the following Jacobian matrix of the system (253)–(256) with respect to $\xi_0,\ldots,\xi_{N-1},y_0,v_0$ using the abbreviations $Y(t_j,t_i) := D(t_j)\Pi_{\mu-1}(t_j)X(t_j,t_i)D^-(t_i)$ and $\pi_{\mu-1}(t_j) := D(t_j)\Pi_{\mu-1}(t_j)D^-(t_j)$

$$J_\mu = \begin{bmatrix} G_aD^-(t_0) & & & & & & G_bX(t_N,t_{N-1})D^-(t_{N-1}) & 0 & G_a \\ -Y(t_1,t_0) & \pi_{\mu-1}(t_1) & & & & & & & \\ & -Y(t_2,t_1) & \pi_{\mu-1}(t_2) & & & & & & \\ & & \ddots & \ddots & & & & & \\ & & & -Y(t_{N-1},t_{N-2}) & \pi_{\mu-1}(t_{N-1}) & & & & \\ I-\pi_{\mu-1}(t_0) & & & & & & & & \\ & I-\pi_{\mu-1}(t_1) & & & & & & & \\ & & I-\pi_{\mu-1}(t_2) & & & & & & \\ & & & \ddots & & & & & \\ & & & & I-\pi_{\mu-1}(t_{N-1}) & & & & \\ \frac{\partial \mathcal{I}_{\mu,y_0}}{\partial \xi_0} & & & & & & & I & 0 \\ \frac{\partial \mathcal{I}_{\mu,v_0}}{\partial \xi_0} & & & & & & & 0 & I \end{bmatrix} \tag{257}$$

There is, for practical reasons, the possibility to compress J_μ mixing (252) with (256). We obtain

$$\bar{J}_\mu = \begin{bmatrix} G_aD^-(t_0) & & & & & & G_bX(t_N,t_{N-1})D^-(t_{N-1}) & 0 & G_a \\ -Y(t_1,t_0) & I & & & & & & & \\ & -Y(t_2,t_1) & I & & & & & & \\ & & \ddots & \ddots & & & & & \\ & & & -Y(t_{N-1},t_{N-2}) & I & & & & \\ I-\pi_{\mu-1}(t_0) & & & & & & & & \\ \frac{\partial \mathcal{I}_{\mu,y_0}}{\partial u_0} & & & & & & & I & 0 \\ \frac{\partial \mathcal{I}_{\mu,v_0}}{\partial u_0} & & & & & & & 0 & I \end{bmatrix} \tag{258}$$

The use of (253) for computing of the Jacobian matrix leads immediately to (258). As for the single shooting method, we show a regularity condition for the matrix (257).

Lemma 4.4. *Let the BVP* (37) *be uniquely solvable and the admissible projectors* Q_i, $0 \le i \le \mu-1$ *realize a fine decoupling.*
The interval $[a,b]$ *is subdivided into N subintervals*

$$a = t_0 < t_1 < \ldots < t_{N-1} < t_N = b,$$

then the Jacobian matrix (257) *has full column rank.*

Proof. To show the column regularity of J_μ we consider $J_\mu z = 0$ with $z = (z_0^T, z_1^T, \ldots, z_{N-1}^T, z_y^T, z_v^T)^T$. Because of $\pi_{\mu-1}(t_i)z_i = D(t_i)X(t_i,t_{i-1})D^-(t_{i-1})z_{i-1}$ for $i = 1,\ldots,N-1$, the second up to the N th equation leads to $\pi_{\mu-1}(t_{N-1})z_{N-1} = D(t_{N-1})X(t_{N-1},t_0)D^-(t_0)z_0$. Using this result, the first equation looks

$$(G_a + G_b X(t_N, t_0))D^-(t_0)z_0 + G_a z_v = 0$$

and the last but one equation gives $z_v = -\frac{\partial \mathcal{I}_{\mu,v}}{\partial \xi_0} z_0$. From the last equation we obtain $z_0 = \pi_{\mu-1}(t_0)z_0$ which leads for the first equation to

$$(G_a X(t_0, t_0) + G_b X(t_N, t_0))D^-(t_0)z_0 = SD^-(t_0)z_0 = 0.$$

From (41) we have that then $\Pi_{\mu-1}(t_0)D^-(t_0)z_0 = 0$, therefore $z_0 = 0$ and successively using (256) $z_i = 0,\ i = 1, \ldots, N-1$ and at last follows that $z_v = 0$, $z_y = 0$. □

We are interested in the relation of the multiple shooting method of an DAE with the inherent ODE. For that we use the v-component of (255) in (254) and we consider the system (252) and (254). Its Jacobian matrix looks

$$S_{mult} = \begin{bmatrix} G_a X(t_0,t_0)D^-(t_0) & & & & G_b X(t_N,t_{N-1})D^-(t_{N-1}) \\ 0_{n-l,n} & & & & \\ -Y(t_1,t_0) & \pi_{\mu-1}(t_1) & & & \\ & -Y(t_2,t_1) & \pi_{\mu-1}(t_2) & & \\ & & \ddots & \ddots & \\ & & -Y(t_{N-1},t_{N-2}) & \pi_{\mu-1}(t_{N-1}) & \end{bmatrix}. \tag{259}$$

For (259) we have the representation $S_{mult} = \Pi_l S_{mult,ODE} \Pi_r$ with

$$\Pi_l = \begin{bmatrix} I_n & & & \\ & \pi_{\mu-1}(t_1) & & \\ & & \ddots & \\ & & & \pi_{\mu-1}(t_{N-1}) \end{bmatrix},$$

$$S_{mult,ODE} = \begin{bmatrix} G_a \Pi_{can}(t_0) D^-(t_0) & & & & G_b \Pi_{can}(t_N) D^-(t_N) U(t_N, t_{N-1}) \\ C_a & & & & \\ -U(t_1,t_0) & I & & & \\ & \ddots & \ddots & & \\ & & -U(t_{N-1},t_{N-2}) & & I \end{bmatrix},$$

$$\Pi_r = \begin{bmatrix} \pi_{can,\mu-1}(t_0) & & & \\ & \pi_{can,\mu-1}(t_1) & & \\ & & \ddots & \\ & & & \pi_{can,\mu-1}(t_{N-1}) \end{bmatrix}$$

with $\pi_{can,\mu-1} = D\Pi_{can,\mu-1}D^-$. The matrix $S_{mult,ODE}$ has the known structure of the Jacobian matrix of the parallel shooting method for ODEs, here the inherent ODE, and is related to the TPBVP (46)–(48) with $C_a = K^{-1}(I - \pi_{can,\mu-1}(t_0))$. K is chosen such that $C_a \in \mathbb{R}^{n-l}$ (see [77]). Its inverse is given by

$$S_{mult,ODE}^{-1} = \begin{bmatrix} U(t_0,t_0)S_{ODE}^{-1} & \bar{\mathcal{G}}(t_0,t_1) & \cdots & \bar{\mathcal{G}}(t_0,t_{N-1}) \\ \vdots & \vdots & & \vdots \\ U(t_{N-1},t_0)S_{ODE}^{-1} & \bar{\mathcal{G}}(t_{N-1},t_1) & \cdots & \bar{\mathcal{G}}(t_{N-1},t_{N-1}) \end{bmatrix}$$

with the nonsingular matrix $S_{ODE} = \begin{bmatrix} S_{IERODE} \\ C_a \end{bmatrix}$ and the Green's function

$$\bar{\mathcal{G}}(t,s) = \begin{cases} U(t,t_0)S_{ODE}^{-1} \begin{bmatrix} G_a \Pi_{can} D^-(t_0) \\ C_a \end{bmatrix} U(s,t_0)^{-1}, & t \geq s \\ -U(t,t_0)S_{ODE}^{-1} \begin{bmatrix} G_b \Pi_{can} D^-(t_N) \\ 0_{n-l,n} \end{bmatrix} U(t_N,t_0)U(s,t_0)^{-1}, & t < s \end{cases}$$

(cf. for the Green's function (42) and for S_{IERODE} (45)).
The factors Π_l and Π_r are projectors and have projectors on its diagonal. The eigenvalues of projectors are 0 or 1, which allows an appropriate estimation $||\Pi_l|| \leq K$ and $||\Pi_r|| \leq K$ with moderate K. This makes clear that we obtain an estimation of the condition number of S_{mult} as cond $S_{mult} \approx ||S_{mult,ODE}||\,||S_{mult,ODE}^{-1}||$ of the same structure as in the ODE case (cf. [12]).

Theorem 4.1. *The reflexive inverse S_{mult}^- of the multiple shooting matrix S_{mult} is given by*

$$S_{mult}^- = \Pi_r S_{mult,ODE}^{-1} \Pi_l = diagD \begin{bmatrix} X(t_0,t_0)S^- & \mathcal{G}(t_0,t_1) & \cdots & \mathcal{G}(t_0,t_{N-1}) \\ \vdots & \vdots & & \vdots \\ X(t_{N-1},t_0)S^- & \mathcal{G}(t_{N-1},t_1) & \cdots & \mathcal{G}(t_{N-1},t_{N-1}) \end{bmatrix} diagD^-$$

with $diagD := \mathrm{diag}\,(D(t_0),\ldots,D(t_{N-1}))$ and $diagD^- := \mathrm{diag}\,(D(t_0)^-,\ldots,D(t_{N-1})^-)$.

Proof. We have to show the reflexivity properties $S_{mult} = S_{mult}S_{mult}^-S_{mult}$ and $S_{mult}^- = S_{mult}^-S_{mult}S_{mult}^-$. We consider $S_{mult}^-S_{mult} = \Pi_r S_{mult,ODE}^{-1}\Pi_l\Pi_l S_{mult,ODE}\Pi_r$. It holds that

$$(I-\Pi_l)S_{mult,ODE}\Pi_r = 0.$$

This follows from $U(t,s)D(s)\Pi_{can,\mu-1}(s) = D(t)\Pi_{can,\mu-1}(t)D(t)^-U(t,s)D(s)\Pi_{can,\mu-1}(s)$ (cf. [83, (2.82)]) and $D\Pi_{can,\mu-1}D^- = D\Pi_{\mu-1}D^-$ (cf. (23)). We obtain $S_{mult}^-S_{mult} = \Pi_r$ which proves the assertion. □

The relation of Theorem 4.1 was shown for index-1 DAEs in [87].

4.2 Nonlinear index-1 DAEs

The most realizations of shooting methods are done for index-1 DAEs or for DAEs reduced to index-1. A reduction of the index is mostly done applying the differentiation index concept ([38]) or the strangeness index concept ([111], [74]). In the latter,

the realization of the shooting procedure is strongly interlocked with the reduction from the derivative array system. [53] investigated a special structured index-2 DAE, which is reduced to index 1 by differentiation. See also Remark 2.5. In [79], [80], [37], [47]and [25] the shooting method is investigated for index-1 DAEs. We find many papers considering very special applications. [80] and [25] focus on periodic BVPs. The necessary conditions of optimal control problems are investigated and shooting methods applied in [37], [32], [53], and [63]. A lot of papers are dealing with single problems in sciences and technique which are then solved by shooting methods.

We consider the TPBVP (3), (4). We subdivide the interval $[a,b]$ into N subintervals $a = t_0 < t_1 < \cdots < t_N = b$. At every subinterval we have to integrate and to compute consistent initial values. The IVP at a point $\bar{t}$ is represented by

$$D(\bar{t})(x(\bar{t}) - \bar{\alpha}) = 0$$

for given $\bar{\alpha}$. The computation of consistent initial values at $\bar{t}$ using $y := D^-(\bar{t})(Dx)'(\bar{t})$ can be done by the solution of the equations

$$f(D(\bar{t})y, P_0(\bar{t})\bar{\alpha} + Q_0 v, \bar{t}) = 0, \tag{260}$$

$$Q_0 y + P_0 v = 0, \tag{261}$$

which have in the index-1 case the nonsingular Jacobian matrix (cf. for a related proof [83, Lemma 4.12])

$$\begin{bmatrix} f_y D & f_x Q_0 \\ Q_0 & P_0 \end{bmatrix}.$$

The matching conditions are given by

$$D(t_i)(D^-(t_i)\xi_i - x(t_i; t_{i-1}, \xi_{i-1})) = 0 \quad \text{for } i = 1, \ldots, N-1. \tag{262}$$

The system to solve consists of (4) as

$$g(D^-(t_0)\xi_0 + v_0, x(b; t_{N-1}, \xi_{N-1})) = 0, \tag{263}$$

the matching conditions (262) and the determination of v_0 using (260), (261) at $\bar{t} = t_0$.

The Jacobian matrix with respect to $\xi_0, \xi_1, \ldots, \xi_{N-1}, y_0, v_0$ is related to (257) with the linearization (cf. Subsection 2.5) of g and $Y_*(t_j, t_i) := D(t_j)X_*(t_j, t_i)D^-(t_i)$

$$J_1 = \begin{bmatrix}
G_{*a}D^-(t_0) & & & & & & G_{*b}X_*(t_N,t_{N-1})D^-(t_{N-1}) & 0 & G_{*a} \\
-Y_*(t_1,t_0) & R(t_1) & & & & & & & \\
 & -Y_*(t_2,t_1) & R(t_2) & & & & & & \\
 & & \ddots & \ddots & & & & & \\
 & & & -Y_*(t_{N-1},t_{N-2}) & & R(t_{N-1}) & & & \\
I-R(t_0) & & & & & & & & \\
 & I-R(t_1) & & & & & & & \\
 & & & \ddots & & & & & \\
 & & & & & I-R(t_{N-1}) & & & \\
-P_0G_1^{-1}f_xD^-(t_0) & & & & & & & I & 0 \\
-Q_0G_1^{-1}f_xD^-(t_0) & & & & & & & 0 & I
\end{bmatrix}$$

with $R(t_i) := D(t_i)D^-(t_i)$. The column regularity of J_1 follows from Lemma 4.4. All techniques solving nonlinear overdetermined systems are applicable. As mentioned above also a formulation as square system is possible, which results in a nonsingular Jacobian matrix of the system.

If the DAE is represented with a full rank matrix D the system dimension decreases because of $R(t) \equiv I$. This holds because of the nonsingularity of DD^T ($D = DD^-D \Rightarrow DD^- = I$), i.e., all equations related to (256) vanish.
Very often a semi-explicit structure of f is assumed (see (7)). Semi-explicit structure means that $D = [I\ 0]$ and $D^- = \begin{bmatrix} I \\ 0 \end{bmatrix}$. Therefore $R = DD^- = I_{m_1}$ and $I - R = 0$. This reduces the dimension of J_1 drastically, because the blocks, e.g. $Y_*(t_j,t_i)$, have now dimension $m_1 \times m_1$ and not the full dimension of the DAE $m \times m$ and D has also full rank.
A semiexplicit structure of the DAE is assumed, e.g., in [67], [106], [63].

4.3 Further references, comments, and open questions

Remark 4.1 (Take advantage of (partially) separated boundary conditions). If the boundary condition $G_a x(a) + G_b x(b) = \gamma$ are structured such that a part is separated at $t = a$ we should use advantage of such explicitly required initial values. This can be done by using for shooting an adapted inital value condition $C(x(a) - z) = 0$ which includes the separated boundary conditions. For DAEs up to index 2 a proposal can be found in [48], [81].
The advantage of partially separated boundary conditions is also considered in [38]. Here a possible reduction of "the number of IVPs to be solved" is discussed.

Remark 4.2 (Avoiding inconsistent values for semiexplicit index-1 DAEs). In [34], (cf. also [45]) for semiexplicit index-1 DAEs, a special way to avoid the computation of consistent initial values at every shooting point is proposed. Considered is the DAE

$$y' = f(t,y,u,p)$$
$$0 = g(t,y,u,p)$$

The algebraic condition $g(t,y,u,p) = 0$ is replaced at every shooting interval by $g(t,y,u,p) - g(r_j, s_j^y, s_j^u, p) = 0$, where $y(r_j) = s_j^y, u(r_j) = s_j^u$ describes the current values of the Newton iteration values of the jth interval. Additionally it is secured that $g(r_j, s_j^y, s_j^u, p) \to 0$ over the Newton iteration.

Remark 4.3 (Realizations for higher index DAEs). Very few papers investigate higher index DAEs directly, i.e., without an index reduction.
In [77] a shooting method for index-2 DAEs in standard formulation is proposed; the necessary differentiation for calculating consistent initial values are realized by finite differences.
Consistent initial values for Hessenberg index-2 and index-3 DAEs using boundary value methods are considered in [3] and for general index 3 DAEs in [86].
The computation of consistent initial values of index-2 DAEs in standard formulation using the tractability index concept is considered in [48] and for properly stated index-2 DAEs in [81].

5 Miscellaneous

5.1 Periodic solutions

Periodic solutions of DAEs are studied in the context of applications in multibody system dynamics and circuit simulation, e.g., [113, 25, 51, 107]. As for explicit ODEs, one can provide periodic solutions via BVPs with periodic boundary conditions.

As pointed out already in [80], when formulating periodic boundary conditions, one should try for a well-posed BVP and regard the accurate number of boundary conditions. In contrast to the classical ODE case, the full condition $x(0)-x(T)=0$ is overdetermined for DAEs, cf. our Examples 1.2, 1.3.

In full analogy to explicit ODEs, the right number of boundary conditions is necessary but not sufficient for well-posedness, cf. Example 1.3. The boundary conditions must be consistent with the flow.

For autonomous DAEs one applies the usual trick to introduce the auxiliary equation $T'=0$ for the unknown period T and an additional boundary condition for fixing the phase (e.g., [80, 51]).

Lyapunov stability criteria for periodic solutions of index-1 and index-2 DAEs are provided in [84, 85] by means of an appropriate generalization of the Floquet theory. Thereby the maximal normalized fundamental solution matrix plays its role yielding the monodromy matrix and Floquet exponents. Note that certain structural conditions restrict the class of index-2 DAEs in [85]. In essence, from an actual point of view, these conditions ensure that the reference solution belongs to an index-2 regularity region. We conjecture that the respective results remain valid if the structural conditions are replaced by assuming the reference solution to proceed in a stability region.

5.2 Abramov transfer method

The Abramov transfer method is extended to BVPs for index-1 DAEs in [27, 100] and for index-2 DAEs in [101, 30]. We do not go into detail, but explain the main idea for the case of explicit ODEs only.

It is well-known that the solution space $\mathcal{M}(t)\subset\mathbb{R}^m$ of the classical IVP

$$x'(t)+B(t)x(t)=0,\quad t\in[a,b], \tag{264}$$

$$C_a x(a)=0, \tag{265}$$

with $C_a\in\mathcal{L}(\mathbb{R}^m,\mathbb{R}^k)$, $\operatorname{rank}C_a=k\le m$, can be described by the relation

$$y_a(t)^*x(t)=y_a(a)^*x(a)=0,$$

if the matrix-valued function y_a solves the IVP

$$y'(t) - B(t)^* y(t) = 0, \quad t \in [a,b], \tag{266}$$
$$y(a) = C_a^*. \tag{267}$$

The subspace $\mathcal{M}(t) = \ker y(t)^* = (\operatorname{im} y(t))^\perp$ has dimension $m-k$. BVPs for (264) and separated boundary conditions

$$C_a x(a) = 0, \quad C_b x(b) = 0 \tag{268}$$

can be traced back to the linear system

$$y_a(t)^* x(t) = 0,$$
$$y_b(t)^* x(t) = 0,$$

by solving an IVP and a terminal value problem for the adjoint equation. We emphasize that there is no need for well-posedness of the BVP. As a byproduct one gathers a constructive criterion of uniquely solvability.

Generally the adjoint ODE is not easier to integrate than the original ODE. The idea behind the Abramov transfer method ([1]) consists in a continuous orthogonalization by demanding $y^* y' = 0$ and turning to the nonlinear equation

$$y'(t) - (I - y(t)(y(t)^* y(t))^{-1} y(t)^*) B(t)^* y(t) = 0, \quad t \in [a,b], \tag{269}$$

instead of (266). The equation (269) has nice theoretical and practical solvability properties. Slightly modified versions of this approach apply to inhomogeneous BVPs. To provide an opinion of the capability of the Abramov transfer method we mention the test problem [12, p. 121],

$$x'(t) - \begin{bmatrix} -\lambda\cos(2\omega t) & \omega + \lambda\sin(2\omega t) \\ -\omega + \lambda\sin(2\omega t) & \lambda\cos(2\omega t) \end{bmatrix} x(t) = 0, \quad t \in [0,\pi],$$

with the fundamental solution matrix

$$X(t) = \begin{bmatrix} \cos(\omega t) & \sin(\omega t) \\ -\sin(\omega t) & \cos(\omega t) \end{bmatrix} \begin{bmatrix} \mathrm{e}^{-\lambda t} & 0 \\ 0 & \mathrm{e}^{\lambda t} \end{bmatrix}.$$

As noted in [12], the Riccati method does not work well for $\lambda = 1$ and greater ω, whereas it performs well for $\omega = 1$ and greater λ. In [100, 101] it is recorded that the Abramov transfer method provides good results for ω from 1 to 1000 and λ from 1 to 200.

5.3 Finite-difference methods

For classical BVPs in explicit ODEs, finite-difference methods generally turn out to be less efficient than collocation methods. The same is true for BVPs in DAEs. We will take only a quick look at the topic.

Diverse one-step and multi-step finite-difference schemes for approximating the solution of the BVP

$$\begin{aligned} f((Dx)'(t),x(t),t) &= 0, \quad t \in [a,b], \\ g(x(a),x(b)) &= 0, \end{aligned}$$

on a grid $\pi : a = t_0 < \cdots < t_N = b$ have been studied already in [89]. For well-posed BVPs, thus for regular index-1 DAEs, stability inequalities and convergence results are provided by means of the well-known discretization theory developed in [65, 66]. From the difference approach concerning the DAE on each subinterval, one generally obtains mN equations for determining the unknowns $x_0,\ldots,x_N$. In contrast to the case of explicit ODEs, the boundary condition yields $n = \operatorname{rank} D(a) < m$ conditions, and hence, one needs additional $m-n$ consistency equations to obtain a balanced scheme. In comparison to the case of explicit ODEs, also certain extra stability conditions are needed.

Finite-difference methods for index-1 DAEs in standard form have been treated in [55] accordingly.

Respective convergence results have been offered in [38] for smoothly solvable linear BVPs with no restriction concerning the DAE index. Instead, the availability of a globally $O(h^s)$-convergent method for solving the corresponding IVPs is postulated and the additionally needed consistency conditions are supposed to be given by means of a derivative array system.

A further detailed convergence proof is described in [111] for linear index-1 DAEs with separated derivative-free equations.

In general, it seems that multi-step methods may be affected by varying inherent subspaces and one-step methods perform better (e.g., [91, p.169]).

5.4 Newton-Kantorovich iterations

Newton–Kantorovich iteration methods applied to BVPs for index-1 and -index-2 DAEs are studied in [92, 101], see also [96].

The BVP

$$f((Dx)'(t),x(t),t) = 0, \quad t \in [a,b] = \mathcal{I}, \tag{270}$$

$$g(x(a),x(b)) = 0, \tag{271}$$

can be formulated as operator equation (cf. the proofs of Theorems 2.4 and 2.7). Let $\mathcal{D}_F \subseteq \mathcal{D}_f$ be open. We associate with the DAE (270) the nonlinear operator

$$\begin{aligned} F &: \operatorname{dom} F \subseteq \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m) \to \mathcal{C}(\mathcal{I},\mathbb{R}^m), \\ \operatorname{dom} F &:= \{x \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m) : x(t) \in \mathcal{D}_F \text{ for all } t \in \mathcal{I}\}, \\ (Fx)(t) &:= f((Dx)'(t),x(t),t), \quad t \in \mathcal{I}, \quad x \in \operatorname{dom} F, \end{aligned} \tag{272}$$

such that the DAE (270) is represented as the operator equation

$$Fx = 0. \tag{273}$$

F is said to be a *nonlinear differential-algebraic* operator. The operator equation (273) reflects the classical view on a DAE: the solutions belong to $\mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$ and satisfy the DAE pointwise for all $t \in \mathcal{I}$. The arguments in [96] enable us to speak of the *natural* Banach space setting.

The operator F *is Fréchet differentiable* and the map $F'(x_*)$ defined by

$$F'(x_*)x = A_*(Dx)' + B_*x, \quad x \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m),$$

is the Fréchet derivative of F at x_*. The linear operator equation

$$F'(x_*)x = q$$

stands now for the *linearization* of the original DAE at x_*, that is, for the linear DAE

$$A_*(Dx)' + B_*x = q. \tag{274}$$

The composed operator

$$\begin{aligned} \mathcal{F} &: \operatorname{dom} F \subseteq \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m) \to \mathcal{C}(\mathcal{I},\mathbb{R}^m) \times \mathbb{R}^{m-l}, \\ \mathcal{F}x &:= (Fx,\ g(x(a),x(b))), \quad x \in \operatorname{dom} F, \end{aligned} \tag{275}$$

is Fréchet differentiable since F is so. The equation $\mathcal{F}x = 0$ represents the BVP (270), (271), whereas the equation $\mathcal{F}x = (q,\gamma)$ is the operator form of the perturbed BVP

$$f((D(t)x(t))',x(t),t) = q(t),\ t \in \mathcal{I}, \quad g(x(a),x(b)) = \gamma. \tag{276}$$

Suppose that the composed operator $\mathcal{F}$ associated with the BVP is a local diffeomorphism at $x_* \in \operatorname{dom}\mathcal{F}$ and $\mathcal{F}(x_*) = 0$, then the well-known Newton–Kantorovich iteration

$$x_{k+1} = x_k - \mathcal{F}'(x_k)^{-1}\mathcal{F}(x_k), \quad k \geq 0, \tag{277}$$

can be applied to approximate x_*. If the initial guess x_0 is sufficiently close to x_*, then these iterations are well-defined and x_k tends to x_*. Practically, one solves the linear equations

$$\mathcal{F}'(x_k)z = -\mathcal{F}(x_k), \quad k \geq 0, \tag{278}$$

and, having the solution z_{k+1} of the linear problem (278), one puts

$$x_{k+1} = x_k + z_{k+1}. \tag{279}$$

The linear problem (278) represents the linear BVP

$$f_y((Dx_k)'(t),x_k(t),t))(Dz)'(t)+f_x((Dx_k)'(t),x_k(t),t))z(t)=-f((Dx_k)'(t),x_k(t),t)),\quad t\in\mathcal{I},$$
$$Ga(x_k(a),x_k(b))z(a)+G_b(x_k(a),x_k(b))z(b)=-g(x_k(a),x_k(b)),$$

with partial derivatives G_a, G_b of the function g with respect to its first and second arguments.

Mostly, a damping parameter is incorporated, and instead of (279) one applies

$$x_{k+1}=x_k+\alpha_{k+1}z_{k+1},\quad \text{with}\quad \alpha_{k+1}\in(0,1]. \tag{280}$$

Usually the damping parameter is chosen so that the residuum $\mathcal{F}(x_{k+1})$ becomes smaller in some sense, that is

$$\|\mathcal{F}(x_{k+1})\|_{res}<\|\mathcal{F}(x_k)\|_{res},$$

with a suitable measure of the residuum, for instance,

$$\|\mathcal{F}(x)\|_{res}:=\|\mathcal{F}(x)\|=\|F(x)\|_\infty+|g(x(a),x(b))|$$
$$\text{and}\quad \|\mathcal{F}(x)\|_{res}^2:=\|F(x)\|_{L^2}^2+|g(x(a),x(b))|^2.$$

Sufficient conditions for the composed operator $\mathcal{F}$ to be a local diffeomorphism in the natural setting are described in [96, Subsubsection 4.3.2]. Then the BVP is well-posed in the natural setting and the DAE has index 1, see Subsubsection 2.5.1.

In [96, Subsubsection 4.3.3] and Subsubsection 2.5.2 one finds conditions for BVPs for a class of index-2 problems being well-posed in an advanced setting.

Next we take a look at the differentiable functional

$$J(x):=\frac{1}{2}\|F(x)\|_{L^2}^2+\frac{1}{2}|g(x(a),x(b))|^2,\quad x\in\operatorname{dom}\mathcal{F}. \tag{281}$$

Of course, the problem to solve the equation $\mathcal{F}(x)=0$ can be regarded as the problem to minimize this functional.
For $x\in\operatorname{dom}\mathcal{F}$ and $z\in\mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$, the directional derivative reads

$$J'(x)z=(F'(x)z,F(x))_{L^2}$$
$$+\langle\, b_a(x(a),x(b))z(a)+b_e(x(a),x(b))z(b),\, b(x(a),x(b))\,\rangle.$$

If $x^0\in\operatorname{dom}\mathcal{F}$ is fixed, $\mathcal{F}(x^0)\neq 0$, and if there exists a solution z_N of the linear equation,

$$\mathcal{F}'(x^0)z=-\mathcal{F}(x^0),\quad k\geq 0, \tag{282}$$

then it results that

$$J'(x^0)z_N=-\|F(x^0)\|_{L^2}^2-|g(x^0(a),x^0(b))|^2<0$$

thus $J(x^0 + \alpha z_N) < J(x^0)$ for all sufficiently small $\alpha > 0$. Therefore, the so-called Newton direction z_N serves as descent direction. Constructing a descent method by applying Newton directions is essentially the same as the damped Newton–Kantorovich iteration. This works under the conditions described above, that is, for index-1 and a restricted class of index-2 problems (cf.,[92, 101].

In [101] the Newton–Kantorovich iteration has been applied in combination with the Abramov transfer method for solving the linear BVPs, with different succes. Though the linear BVPs could be solved succesfully, the intermediate processing to prepare the next iteration could not be managed in an efficient way. Though a collocation solver for the linear BVPs seems to be less accurate than the transfer method, because of a possibly much better intermediate processing from one iteration level to the next one, the Newton–Kantorovich iteration combined with collocation can be expected to work well for the mentioned classes of DAEs. No related practical experience is reported till now.

Following [96], for equations $\mathcal{F}(x) = 0$ involving higher index differential-algebraic operators F, there are two principal difficulties concerning Newton descent and Newton–Kantorovich iteration:

1. The linear equation (278) resp. (282) is essentially ill-posed and might not be solvable. Changing to least-squares solutions does not make great sense, since the linearizations $\mathcal{F}'(x)$ are not normally solvable.

2. For an essentially ill-posed problem a small residuum $\mathcal{F}(x_k)$ does not mean that x_k is close to a solution, see [83, Section 1.1].

Among the methods for ill-posed problems one finds generalizations of Newton-like methods using outer inverses. Instead of the unbounded inverse $\mathcal{F}(x_k)^{-1}$ in (277) one uses a bounded outer inverse. Such an outer inverse is provided by [96, Theorem 4.2]. It seems, no practical experience is available in this context till now.

6 Appendix

6.1 Basics concerning regular DAEs

We collect basic facts on the DAE

$$f((Dx)'(t),x(t),t)=0, \tag{283}$$

which exhibits the involved derivative by means of an extra matrix valued function D. The function $f:\mathbb{R}^n\times\mathcal{D}_f\times\mathcal{I}_f\longrightarrow\mathbb{R}^m$, $\mathcal{D}_f\times\mathcal{I}_f\subseteq\mathbb{R}^m\times\mathbb{R}$ open, is continuous and has continuous partial derivatives f_y and f_x with respect to the first two variables $y\in\mathbb{R}^n$, $x\in\mathcal{D}_f$. The partial Jacobian $f_y(y,x,t)$ is everywhere singular. The matrix function $D:\mathcal{I}_f\to\mathcal{L}(\mathbb{R}^m,\mathbb{R}^n)$ is continuously differentiable and $D(t)$ has constant rank r on the given interval $\mathcal{I}_f$. Then, $\operatorname{im}D$ is a $\mathcal{C}^1$-subspace in $\mathbb{R}^n$. We refer to [83] for proofs, motivation, and more details.

6.1.1 Regular DAEs, regularity regions

The DAE (283) is assumed to have a properly stated leading term. To simplify matters we further assume the nullspace $\ker f_y(y,x,t)$ to be independent of y. Then, the transversality condition (5) pointwise induces the continuously differentiable (see [83, Lemma A.20]) *border projector* $R:\mathcal{D}_f\times\mathcal{I}_f\to\mathcal{L}(\mathbb{R}^n)$ given by

$$\operatorname{im}R(x,t)=\operatorname{im}D(t),\ \ker R(x,t)=\ker f_y(y,x,t),\ (y,x,t)\in\mathbb{R}^n\times\mathcal{D}_f\times\mathcal{I}_f. \tag{284}$$

Next we depict the notion of regularity regions of a DAE (283). For this aims we introduce *admissible matrix function sequences* and associated projector functions (cf. [83]). Denote

$$\begin{aligned}
A(x^1,x,t) &:= f_y(D(t)x^1+D'(t)x,x,t)\in\mathcal{L}(\mathbb{R}^n,\mathbb{R}^m),\\
B(x^1,x,t) &:= f_x(D(t)x^1+D'(t)x,x,t)\in\mathcal{L}(\mathbb{R}^m),\\
G_0(x^1,x,t) &:= A(x^1,x,t)D(t)\in\mathcal{L}(\mathbb{R}^m),\\
B_0(x^1,x,t) &:= B(x^1,x,t)\in\mathcal{L}(\mathbb{R}^m)\quad\text{for } x^1\in\mathbb{R}^m, x\in\mathcal{D}_f, t\in\mathcal{I}_f.
\end{aligned}$$

The transversality condition (5) implies $\ker G_0(x^1,x,t)=\ker D(t)$. We introduce projector valued functions $Q_0,P_0,\Pi_0\in\mathcal{C}(\mathcal{I}_f,\mathcal{L}(\mathbb{R}^m))$ such that for all $t\in\mathcal{I}_f$

$$\operatorname{im}Q_0(t)=N_0(t):=\ker D(t),\quad \Pi_0(t):=P_0(t):=I-Q_0(t). \tag{285}$$

Since D has constant rank, the orthoprojector function onto N_0 is as smooth as D. Therefore, as Q_0 we can choose the orthoprojector function onto N_0 which is even continuously differentiable. Next we determine the generalized inverse $D(x,t)^-$ of $D(t)$ pointwise for all arguments by

$$D(x,t)^- D(t) D(x,t)^- = D(x,t)^-,$$
$$D(t) D(x,t)^- D(t) = D(t),$$
$$D(x,t)^- D(t) = P_0(t),$$
$$D(t) D(x,t)^- = R(x,t).$$

The resulting function D^- is continuous, if P_0 is continuously differentiable then so is also D^-.

Definition 6.1. Let the DAE (283) have a properly involved derivative. $\mathcal{G} \subseteq \mathcal{D}_f \times \mathcal{I}_f$ be open connected.
For the given level $\kappa \in \mathbb{N}$, we call the sequence $G_0, \ldots, G_\kappa$ an *admissible matrix function sequence* associated with the DAE (283) on the set $\mathcal{G}$, if it is built pointwise for all $(x,t) \in \mathcal{G}$ and all arising $x^j \in \mathbb{R}^m$ by the rule:
set $G_0 := AD$, $B_0 := B$, $N_0 := \ker G_0$,
for $i \geq 1$:

$$G_i := G_{i-1} + B_{i-1} Q_{i-1}, \tag{286}$$
$$N_i := \ker G_i, \quad \widehat{N}_i := (N_0 + \cdots + N_{i-1}) \cap N_i,$$
$$\text{find a complement } X_i \text{ such that } N_0 + \cdots + N_{i-1} = \widehat{N}_i \oplus X_i,$$
$$\text{choose a projector } Q_i \text{ such that } \operatorname{im} Q_i = N_i \text{ and } X_i \subseteq \ker Q_i,$$
$$\text{set } P_i := I - Q_i,\ \Pi_i := \Pi_{i-1} P_i,$$
$$B_i := B_{i-1} P_{i-1} - G_i D^- (D \Pi_i D^-)' D \Pi_{i-1}, \tag{287}$$

and, additionally,

(a) the matrix function G_i has constant rank r_i on $\mathbb{R}^{mi} \times \mathcal{G}$, $i = 0, \ldots, \kappa$,

(b) the intersection $\widehat{N}_i$ has constant dimension $u_i := \dim \widehat{N}_i$ there,

(c) the product function Π_i is continuous and $D \Pi_i D^-$ is continuously differentiable on $\mathbb{R}^{mi} \times \mathcal{G}$, $i = 0, \ldots, \kappa$.

The projector functions $Q_0, \ldots, Q_\kappa$ linked with an admissible matrix function sequence are said to be *admissible* themselves.

An admissible matrix function sequence G_0, …, G_κ is said to be *regular admissible*, if

$$\widehat{N}_i = \{0\} \quad \text{for all} \quad i = 1, \ldots, \kappa.$$

Then, also the projector functions $Q_0, \ldots, Q_\kappa$ are called *regular admissible*.
The numbers $r_0 = \operatorname{rank} G_0, \ldots, r_\kappa = \operatorname{rank} G_\kappa$ and $u_1, \ldots, u_\kappa$ are named *characteristic values* of the DAE on $\mathcal{G}$.

To shorten the wording we often speak simply of *admissible projector functions* having in mind the admissible matrix function sequence built with these admissible projector functions. Admissible projector functions are always cross-linked with their matrix function sequence. Changing a projector function yields a new matrix function sequence.

We refer to [83] for many useful properties of the admissible matrix function sequences. It always holds that

$$r_0 \le \cdots \le r_{\kappa-1} \le r_\kappa.$$

The notion of *characteristic values* makes sense, since these values are independent of the special choice of admissible projector functions and invariant under regular transformations.

In case of a linear constant coefficient DAE, the construct simplifies to a sequence of matrices. In particular, the second term in the definition of B_i disappears. It is long-known that a pair $\{E,F\}$ of $m \times m$ matrices E,F is regular with Kronecker index μ exactly if an admissible sequence of matrices starting with $G_0 = AD = E$, $B_0 := F$ yields

$$r_0 \le \cdots \le r_{\mu-1} < r_\mu = m. \tag{288}$$

Thereby, neither the factorization nor the special choice of admissible projectors do matter. The characteristic values describe the structure of the Weierstraß–Kronecker form : we have $l = \sum_{j=0}^{\mu-1}(m - r_j)$ and the nilpotent part N contains altogether $s = m - r_0$ Jordan blocks, among them $r_i - r_{i-1}$ Jordan blocks of order i, $i = 1,\ldots,\mu$, see [83, Corollary 1.32].

For linear DAEs with time-varying coefficients, the term $(\cdot)'$ in (287) means the derivative in time, and all matrix functions are functions in time. In general, the term $(\cdot)'$ in (287) stands for the total derivative in jet variables and then the matrix function G_i depends on the basic variables $(x,t) \in \mathcal{G}$ and, additionally, on the jet variables $x^1,\ldots,x^{i+1} \in \mathbb{R}^m$. Owing to the total derivative $(D\Pi_i D^-)'$ the new variable $x^{i+2} \in \mathbb{R}^m$ comes in at this level, see [83, Section 3.2].

Owing to the constant-rank conditions, the terms $D\Pi_i D^-$ are basically continuous. It may happen, for making these terms continuously differentiable, that the data function f must satisfy additional smoothness requirements. A precise description of those smoothness is much too involved and an overall sufficient condition, say $f \in \mathcal{C}^m$, is much too superficial. To indicate that there might be additional smoothness demands we restrict us to the wording *f is sufficiently smooth*.

The next definition ties regularity up to the inequalities (288) and so generalizes regularity of matrix pencils for time-varying linear DAEs as well as for nonlinear DAEs. We emphasize that regularity is supported by several constant-rank conditions.

Definition 6.2. Let the DAE (283) have a properly involved derivative. Let $\mathcal{G} \subseteq \mathcal{D}_f \times \mathcal{I}_f$ be an open, connected subset. The DAE (283) is said to be

(1) *regular on $\mathcal{G}$ with tractability index 0*, if $r_0 = m$,
(2) *regular on $\mathcal{G}$ with tractability index μ*, if an admissible matrix function sequence exists such that (288) is valid on $\mathcal{G}$.
(3) *regular on $\mathcal{G}$*, if it is, on $\mathcal{G}$, regular with any index (i.e., case (1) or (2) apply).

The open connected subset $\mathcal{G}$ is called a *regularity region* or *regularity domain*.
A point $(\bar{x},\bar{t}) \in \mathcal{D}_f \times \mathcal{I}_f$ is a *regular point*, if there is a regularity region $\mathcal{G} \ni (\bar{x},\bar{t})$.

If $\mathcal{D} \subseteq \mathcal{D}_f$ is an open subset and $\mathcal{I} \subseteq \mathcal{I}_f$ is a compact subinterval, then the DAE (283) is said to be regular on $\mathcal{D} \times \mathcal{I}$, if there is a regularity region $\mathcal{G}$ such that $\mathcal{D} \times \mathcal{I} \subset \mathcal{G}$.

Example 6.1 (Regularity regions). We write the DAE

$$\begin{aligned} x_1'(t) + x_1(t) &= 0, \\ x_2(t)x_2'(t) - x_3(t) &= 0, \\ x_1(t)^2 + x_2(t)^2 - 1 - \gamma(t) &= 0, \end{aligned}$$

in the form (283), with $n = 2,\ m = k = 3$,

$$f(y,x,t) = \begin{bmatrix} y_1 + x_1 \\ x_2 y_2 - x_3 \\ x_1^2 + x_2^2 - \gamma(t) - 1 \end{bmatrix}, \quad f_y(y,x,t) = \begin{bmatrix} 1 & 0 \\ 0 & x_2 \\ 0 & 0 \end{bmatrix},$$
$$D(t) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix},$$

for $y \in \mathbb{R}^2$, $x \in \mathcal{D}_f = \mathbb{R}^3$, $t \in \mathcal{I}_f = \mathbb{R}$.
The derivative is properly involved on the open subsets $\mathbb{R}^2 \times \mathcal{G}_+$ and $\mathbb{R}^2 \times \mathcal{G}_-$, $\mathcal{G}_+ := \{x \in \mathbb{R}^3 : x_2 > 0\} \times \mathcal{I}_f$, $\mathcal{G}_- := \{x \in \mathbb{R}^3 : x_2 < 0\} \times \mathcal{I}_f$. We have there

$$G_0 = AD = \begin{bmatrix} 1 & 0 & 0 \\ 0 & x_2 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad B_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & x_2^1 & -1 \\ 2x_1 & 2x_2 & 0. \end{bmatrix}.$$

Letting

$$Q_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \text{yields} \quad G_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2x_2 & -1 \\ 0 & 0 & 0 \end{bmatrix}.$$

G_1 is singular but has constant rank. Since $N_0 \cap N_1 = \{0\}$ we find a projector function Q_1 such that $N_0 \subseteq \ker Q_1$. We choose

$$Q_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{1}{x_2} & 0 \end{bmatrix},\ P_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & -\frac{1}{x_2} & 1 \end{bmatrix},\ \Pi_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},\ D\Pi_1 D^- = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

and obtain $B_1 = B_0 P_0 Q_1$, and then

$$G_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2x_2 + x_2^1 & -1 \\ 0 & 2x_2 & 0 \end{bmatrix}.$$

The matrix $G_2 = G_2(x^1, x, t)$ is nonsingular for all arguments (x^1, x, t) with $x_2 \neq 0$. The admissible matrix function sequence terminates at this level. The open connected subsets $\mathcal{G}_+$ and $\mathcal{G}_-$ are regularity regions, here both with characteristics $r_0 = 2$, $r_1 = 2$, $r_2 = 3$, and tractability index $\mu = 2$. □

For regular DAEs, all intersections $\widehat{N}_i$ are trivial ones, thus $u_i = 0$, $i \geq 1$. Namely, because of the inclusions

$$\widehat{N}_i \subseteq N_i \cap N_{i+1} \subseteq N_{i+1} \cap N_{i+2} \subseteq \cdots \subseteq N_{\mu-1} \cap N_\mu,$$

for reaching a nonsingular G_μ, which means $N_\mu = \{0\}$, it is necessary to have $\widehat{N}_i = \{0\}$, $i \geq 1$. This is a useful condition for checking regularity in practice.

Observe that each open connected subset of a regularity region is again a regularity region. A regularity region consist of regular points having uniform characteristics. The union of regularity regions is, if it is connected, a regularity region, too. Further, the nonempty intersection of two regularity regions is also a regularity region. Only regularity regions with uniform characteristics may yield nonempty intersections. *Maximal regularity regions* are then bordered by so-called critical points. Solutions may cross the borders of maximal regularity regions and undergo there bifurcations et cetera, see examples in [83, 95, 82]. No doubt, much further research is needed to elucidate these phenomena.

6.1.2 The structure of linear DAEs

The general DAE (283) captures linear DAEs

$$A(t)(Dx)'(t) + B(t)x(t) - q(t) = 0 \tag{289}$$

as $f(y,x,t) := A(t)y + B(t)x - q(t)$, $t \in \mathcal{I}_f$. Now, admissible matrix function sequences depend only on time t; and hence, we speak on *regularity intervals* instead of regions. A regularity interval is open by definition. We say that the linear DAE with properly leading term is *regular on the compact interval* $[t_a, t_e]$, if there is an accommodating regularity interval, or equivalently, if all points of $[t_a, t_e]$ are regular. If the linear DAE is regular on the interval $\mathcal{I}$, then it is also regular on each subinterval of $\mathcal{I}$ with the same characteristics. This sounds as a triviality; however, there is a continuing profound debate about some related questions, cf. [96, Subsection 4.4].

If the linear DAE (289) is regular on the interval $\mathcal{I}$, then (see [83, Section 2.4]) it can be decoupled by admissible projector functions into an *inherent explicit regular ODE* (IERODE)

$$u' - (D\Pi_{\mu-1}D^-)'u + D\Pi_{\mu-1}G_\mu^{-1}B_\mu D^- u = D\Pi_{\mu-1}G_\mu^{-1}q \tag{290}$$

and a triangular subsystem of several equations including differentiations

$$\begin{bmatrix} 0 & \mathcal{N}_{01} & \cdots & \mathcal{N}_{0,\mu-1} \\ & 0 & \ddots & \vdots \\ & & \ddots & \mathcal{N}_{\mu-2,\mu-1} \\ & & & 0 \end{bmatrix} \begin{bmatrix} 0 \\ (Dv_1)' \\ \vdots \\ (Dv_{\mu-1})' \end{bmatrix} \tag{291}$$
$$+ \begin{bmatrix} I & \mathcal{M}_{01} & \cdots & \mathcal{M}_{0,\mu-1} \\ & I & \ddots & \vdots \\ & & \ddots & \mathcal{M}_{\mu-2,\mu-1} \\ & & & I \end{bmatrix} \begin{bmatrix} v_0 \\ v_1 \\ \vdots \\ v_{\mu-1} \end{bmatrix} + \begin{bmatrix} \mathcal{H}_0 \\ \mathcal{H}_1 \\ \vdots \\ \mathcal{H}_{\mu-1} \end{bmatrix} D^- u = \begin{bmatrix} \mathcal{L}_0 \\ \mathcal{L}_1 \\ \vdots \\ \mathcal{L}_{\mu-1} \end{bmatrix} q.$$

The subspace $\operatorname{im} D\Pi_{\mu-1}$ is an invariant subspace for the IERODE (290).
This structural decoupling is associated with the decomposition

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

The coefficients are continuous and explicitly given in terms of an admissible matrix function sequence as

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

with

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

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

The IERODE is always uncoupled of the second subsystem, but the latter is tied to the IERODE (290) if among the coefficients $\mathcal{H}_0,\dots,\mathcal{H}_{\mu-1}$ is at least one who does not vanish. One speaks about a *fine decoupling*, if $\mathcal{H}_1=\cdots=\mathcal{H}_{\mu-1}=0$, and about a *complete decoupling*, if $\mathcal{H}_0=0$, additionally. A complete decoupling is given, exactly if the coefficient $\mathcal{K}$ vanishes identically.

If the DAE (289) is regular and the original data are sufficiently smooth, then the DAE (289) is called *fine*. Fine DAEs possess always fine and complete decouplings, see [83, Subsection 2.4.3] for the constructive proof. The coefficients of the IERODE as well as the so-called *canonical projector function* $\Pi_{can}=(I-\mathcal{H}_0)\Pi_{\mu-1}$ are independent of the special choice of the fine decoupling projector functions.

It is noteworthy that, if $Q_0,\dots,Q_{\mu-1}$ generate a complete decoupling for a constant coefficient DAE $Ex'(t)+Fx(t)=0$, then $\Pi_{\mu-1}$ is the spectral projector of the matrix pencil $\{E,F\}$. This way, the projector function $\Pi_{\mu-1}$ associated with a complete decoupling of a fine time-varying DAE represents the generalization of the spectral projector.

6.1.3 Linearizations

Given is now a reference function $x_*\in\mathcal{C}_D^1(\mathcal{I}_*,\mathbb{R}^m)$ on an individual interval $\mathcal{I}_*\subseteq\mathcal{I}_f$, whose values belong to $\mathcal{D}_f$. For each such reference function (here not necessarily a solution!) we may consider the linearization of the (283) along x_*, that is, the linearized DAE

$$A_*(t)(Dx)'(t)+B_*(t)x(t)=q(t),\quad t\in\mathcal{I}_*, \tag{292}$$

with coefficients

$$A_*(t):=f_y((Dx_*)'(t),x_*(t),t),\quad B_*(t):=f_x((Dx_*)'(t),x_*(t),t),\quad t\in\mathcal{I}_*.$$

The linear DAE (292) inherits from the nonlinear DAE (283) the properly stated leading term.

We denote by $\mathcal{C}_{ref}^m(\mathcal{G})$ the set of all $\mathcal{C}^m$ functions x_*, defined on individual intervals $\mathcal{I}_{x_*}$, and with graph in $\mathcal{G}$, that is, $(x_*(t),t)\in\mathcal{G}$ for $t\in\mathcal{I}_{x_*}$. Clearly, then we have also $x_*\in\mathcal{C}_D^1(\mathcal{I}_{x_*},\mathbb{R}^m)$. By the smoothness of the reference functions x_* and the function f we ensure that also the coefficients A_* and B_* are sufficiently smooth for regularity.

Next we adapt the necessary and sufficient regularity condition from [83, Theorem 3.33] to our somewhat simpler situation.

Theorem 6.1. *Let the DAE (283) have a properly involved derivative and let f be sufficiently smooth. Let $\mathcal{G}\subseteq\mathcal{D}_f\times\mathcal{I}_f$ be an open connected set. Then the following statements are valid:*

(1) *The DAE (283) is regular on $\mathcal{G}$ if the linearized DAE (292) along each arbitrary reference function $x_*\in\mathcal{C}_{ref}^m(\mathcal{G})$ is regular, and vice versa.*

(2) *If the DAE (283) is regular on $\mathcal{G}$ with tractability index μ and characteristic values $r_0 \le \cdots \le r_{\mu-1} < r_\mu = m$, then all linearized DAEs (292) along reference functions $x_* \in \mathcal{C}^m_{ref}(\mathcal{G})$ are regular with uniform index μ and characteristics $r_0 \le \cdots \le r_{\mu-1} < r_\mu = m$.*
(3) *If all linearized DAEs (292) along reference functions $x_* \in \mathcal{C}^m_{ref}(\mathcal{G})$ are regular, then they have uniform index and characteristics, and the nonlinear DAE (283) is also regular on $\mathcal{G}$, with the same index and characteristics.*

Corollary 6.1. *Let the DAE (283) have a properly involved derivative and let f be sufficiently smooth. Let $\mathcal{D} \subseteq \mathcal{D}_f$ be an open connected set and $\mathcal{I} \subset \mathcal{I}_f$ be a compact interval. Then the following statements are valid:*

(1) *The DAE (283) is regular on $\mathcal{D} \times \mathcal{I}$ if the linearized DAE (292) along each arbitrary reference function $x_* \in \mathcal{C}^m(\mathcal{I}, \mathbb{R}^m)$ with values in $\mathcal{D}$ is regular, and vice versa.*
(2) *If the DAE (283) is regular on $\mathcal{D} \times \mathcal{I}$ with tractability index μ and characteristic values $r_0 \le \cdots \le r_{\mu-1} < r_\mu = m$, then all linearized DAEs (292) along reference functions $x_* \in \mathcal{C}^m(\mathcal{I}, \mathbb{R}^m)$ with values in $\mathcal{D}$ are regular with uniform index μ and characteristics $r_0 \le \cdots \le r_{\mu-1} < r_\mu = m$.*
(3) *If all linearized DAEs (292) along reference functions $x_* \in \mathcal{C}^m(\mathcal{I}, \mathbb{R}^m)$ with values in $\mathcal{D}$ are regular, then they have uniform index and characteristics, and the nonlinear DAE (283) is also regular on $\mathcal{D} \times \mathcal{I}$, with the same index and characteristics.*

Proof. Statement (1) is a consequence of the Statements (2) and (3).
Statement (2) follows from the construction of the admissible matrix function sequences. Namely, for each $x_* \in \mathcal{C}^m(\mathcal{I}, \mathbb{R}^m)$, with values in $\mathcal{D}$, we have

$$\begin{aligned} G_0(x_*'(t), x_*(t), t) &=: G_{*\,0}(t), \\ B_{i-1}(x_*^{(i+1)}(t), \cdots, x_*'(t), x_*(t), t) &=: B_{*\,i-1}(t), \\ G_i(x_*^{(i+1)}(t), \cdots, x_*'(t), x_*(t), t) &=: G_{*\,i}(t),\ t \in \mathcal{I}, \quad i = 1, \ldots, \mu, \end{aligned}$$

which represents an admissible matrix function sequence for the linearized along x_* DAE.

Statement (3) proves along the lines of [83, Theorem 3.33] by means of so-called widely orthogonal projector functions. The prove given in [83] also works, if one supposes solely compact individual intervals $\mathcal{I}_{x_*}$.

By Lemma 6.1 below, each reference function given on an individual compact interval can be extended to belong to $x_* \in \mathcal{C}^m(\mathcal{I}, \mathbb{R}^m)$, with values in $\mathcal{D}$. □

The next assertion is proved in [96].

Lemma 6.1. *Let $\mathcal{D} \subseteq \mathbb{R}^m$ be an open set and $\mathcal{I} \subset \mathbb{R}$ be a compact interval. Let $\mathcal{I}_* \subset \mathcal{I}$ be a compact subinterval and $s \in \mathbb{N}$.*
Then, for each function $x_* \in \mathcal{C}^s(\mathcal{I}_*, \mathbb{R}^m)$, with values in $\mathcal{D}$, there is an extension $\hat{x}_* \in \mathcal{C}^s(\mathcal{I}, \mathbb{R}^m)$, with values in $\mathcal{D}$.

6.1.4 Linear differential-algebraic operators

Let the linear DAE (289) be regular with tractability index $\mu \in \mathbb{N}$ on the interval $\mathcal{I} = [a,b]$. The function space

$$\mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m) = \{x \in \mathcal{C}(\mathcal{I},\mathbb{R}^m) : Dx \in \mathcal{C}^1(\mathcal{I},\mathbb{R}^n)\}$$

equipped with the norm $\|x\|_{\mathcal{C}_D^1} := \|x\|_\infty + \|(Dx)'\|_\infty$ is a Banach space. We consider the regular linear differential-algebraic operator (cf. [96])

$$Tx := A(Dx)' + Bx, \quad x \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m),$$

and, supposing accurately stated boundary conditions in the sense of Definition 2.3, the composed operator

$$\mathcal{T}x := (\,Tx,\, G_a x(a) + G_b x(b)\,), \quad x \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m),$$

so that the equations $Tx = q$ and $\mathcal{T}x = (q,\gamma)$ represent the DAE and the BVP, respectively.

We consider different image spaces Y and $Y \times \mathbb{R}^l$ for the operators T and $\mathcal{T}$. The natural one is

$$Y = \mathcal{C}(\mathcal{I},\mathbb{R}^m).$$

T and $\mathcal{T}$ are bounded in this setting:

$$\|Tx\|_\infty \le (\|A\|_\infty \|(Dx)'\|_\infty + \|b\|_\infty \|x\|_\infty) \le k\|x\|_{\mathcal{C}_D^1}, \quad x \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m).$$

The operator T is surjective exactly if the index μ equals one. Otherwise $\operatorname{im} T$ is a proper nonclosed subset in $\mathcal{C}(\mathcal{I},\mathbb{R}^m)$, see [83, Subsection 3.9.1], also Appendix 6.1.2. More precisely, one obtains

$$\operatorname{im} T = \{q \in \mathcal{C}(\mathcal{I},\mathbb{R}^m) : v_{\mu-1} := \mathcal{L}_{\mu-1}q,\ Dv_{\mu-1} \in \mathcal{C}^1(\mathcal{I},\mathbb{R}^n),\ \text{for } j = \mu-2,\dots,1 :$$
$$v_j := \mathcal{L}_j q + \sum_{i=j+1}^{\mu-1} \mathcal{M}_{j,i} v_i + \sum_{i=j+1}^{\mu-1} \mathcal{N}_{j,i}(Dv_i)',\ Dv_j \in \mathcal{C}^1(\mathcal{I},\mathbb{R}^n)\} =: \mathcal{C}^{\operatorname{ind}\mu}(\mathcal{I},\mathbb{R}^m).$$

If $\mu = 1$, then $\mathcal{T}$ acts bijectively between Banach spaces so that the inverse $\mathcal{T}^{-1}$ is also bounded and the BVP $\mathcal{T}x = (q,\gamma)$ is well-posed.

If $\mu > 1$, then the BVP $\mathcal{T}x = (q,\gamma)$ is essentially ill-posed in this natural setting because of the nonclosed image of T.

Let be $\mu > 1$. In an advanced setting we put

$$Y = \mathcal{C}^{\operatorname{ind}\mu}(\mathcal{I},\mathbb{R}^m)$$

and, by introducing the norm $\|q\|_{\operatorname{ind}\mu} := \|q\|_\infty + \|(Dv_{\mu-1})'\|_\infty + \cdots + \|(Dv_1)'\|_\infty$ we obtain again a Banach space. Regarding the structure of the DAE (cf. Subsubsection 6.1.2) one knows the operators t and $\mathcal{T}$ to be bounded again. Namely, we derive for

each arbitrary $x \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$ that

$$\|Tx\|_{\text{ind}\,\mu} := \|Tx\|_\infty + \|(D\Pi_{\mu-2}Q_{\mu-1}x)'\|_\infty + \cdots + \|(D\Pi_0 Q_1 x)'\|_\infty.$$

Taking into account that

$$(D\Pi_{\mu-2}Q_{\mu-1}x)' = (D\Pi_{\mu-2}Q_{\mu-1}D^-)'Dx + D\Pi_{\mu-2}Q_{\mu-1}D^-(Dx)'$$

etc. one achieves the wanted inequality $\|Tx\|_{\text{ind}\,\mu} \le k_{\text{ind}\,\mu}\|x\|_{\mathcal{C}_D^1}$ in fact.

In this advanced setting, as a bounded bijection acting in Banach spaces, $\mathcal{T}$ has a bounded inverse and the BVP is well-posed. This sounds fine, but it is quite illusory. The advanced image space $\mathcal{C}^{\text{ind}\,\mu}(\mathcal{I},\mathbb{R}^m)$ as well as its norm $\|.\|_{\text{ind}\,\mu}$ strongly depend on the special coefficients A, D, B. To describe them, one has to be aware of the full special structure of the given DAE. Except for the index-2 case, there seems to be no way to practice this formal well-posedness.

Furthermore, the higher the index the stronger the topology given by the norm $\|.\|_{\text{ind}\,\mu}$, see [83, Subsubsection 3.9.1], [96, Section 2]. It seems to be impossible to capture errors in practical computational procedures by those norms.

6.2 List of symbols and abbreviations

$\mathcal{L}(X,Y)$	set of linear operators from X to Y
$\mathcal{L}(X)$	$= \mathcal{L}(X,X)$
$\mathcal{L}(\mathbb{R}^m,\mathbb{R}^n)$	is identified with $\mathbb{R}^{n\times m}$
K^*	transposed matrix
K^-	generalized inverses
K^+	orthogonal generalized (Moore-Penrose) inverse
$\text{dom}\,K$	definition domain of the map K
$\ker K$	nullspace (kernel) of the operator K
$\text{im}\,K$	image (range) of the operator K
$\text{ind}\{E,F\}$	Kronecker index of the matrix pair $\{E,F\}$
$\langle\cdot,\cdot\rangle$	scalar product in $\mathbb{R}^m$
$(\cdot,\cdot)$	scalar product in function spaces
$\lvert\cdot\rvert$	vector and matrix norms
$\|\cdot\|$	norms on function spaces, operator norms
DAE	differential-algebraic equation
ODE	ordinary differential equation
IVP	initial value problem
BVP	boundary value problem
IERODE	inherent explicit ODE
LSS	least squares solution
TPBVP	two-point BVP

References

1. A. A. Abramov. On transfer of boundary conditions for systems of linear ordinary differential equations (a variant of transfer method). *U.S.S.R. Comput. Math. and Math. Phys.*, 1(3):542–544, 1961.
2. P. Amodio and F. Mazzia. Numerical solution of differential algebraic equations and computation of consistent initial/boundary conditions. *J. of Comp. Appl. Math.*, 87:135–146, 1997.
3. P. Amodio and F. Mazzia. An algorithm for the computation of consistent initial values for differential-algebraic equations. *Numerical Algorithms*, 19:13–23, 1998.
4. P. K. Anh. Multipoint boundary-value problems for transferable differential-algebraic equations. I–linear case. *Vietnam J. of Mathematics*, 25(4):347–358, 1997.
5. P. K. Anh. Multipoint boundary-value problems for transferable differential-algebraic equations. II–quasilinear case. *Vietnam J. of Mathematics*, 26(4):337–349, 1998.
6. P. K. Anh and N. V. Nghi. On linear regular multipoint boundary-value problems for differential algebraic equations. *Vietnam J. of Mathematics*, 28(2):183–188, 2000.
7. U. M. Ascher and L. R. Petzold. Projected collocation for higher-order higher-index differential-algebraic equations. *J. of Comp. Appl. Math.*, 43:243–259, 1992.
8. U. M. Ascher and L. R. Petzold. *Recent Developments in Numerical Methods and Software for ODEs/DAEs/PDEs*, chapter Numerical Methods for Boundary Value Problems in Differential-Algebraic Equations, pages 125–135. World Scientific Publishing Co. London Singapure, 1992. ed. by G. D. Byrne and W. E. Schiesser.
9. U. M. Ascher and L. R. Petzold. *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. SIAM Philadelphia, 1998.
10. U.M. Ascher, J. Christiansen, and R. Russell. Collocation software for boundary value ODEs. *ACM Trans. Math. Software*, 7(209-222), 1981.
11. U.M. Ascher and P. Lin. Sequential regularization methods for nonlinear higher index DAEs. *SIAM J. Sci. Comput.*, 18:160–181, 1997.
12. U.M. Ascher, R.M.M. Mattheij, and R.D. Russell. *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*. Prentice Hall, Englewood Cliffs, New Jersey, 1988.
13. U.M. Ascher and R. Spiteri. Collocation software for boundary value differential-algebraic equations. *SIAM J. Sci. Comput.*, 15:938–952, 1994.
14. W. Auzinger, G. Kneisl, O. Koch, and E. Weinmüller. SBVP 1.0 – A MATLAB solver for singular boundary value problems. ANUM Preprint 2/02, Vienna University of Technology, 2002.
15. W. Auzinger, G. Kneisl, O. Koch, and E. Weinmüller. A collocation code for boundary value problems in ordinary differential equations. *Numer. Algorithms*, 33:27–39, 2003.
16. W. Auzinger, O. Koch, and E. Weinmüller. Efficient collocation schemes for singular boundary value problems. *Numer. Algorithms*, 31:5–25, 2002.
17. W. Auzinger, O. Koch, and E. Weinmüller. Analysis of a new error estimate for collocation methods applied to singular boundary value problems. *SIAM J. Numer. Anal.*, 42:2366–2386, 2005.
18. W. Auzinger, H. Lehner, and E. Weinmüller. Defect-based A-posteriori Error Estimation for Index-1 DAEs. ASC Technical Report 20, Vienna University of Technology, 2007.
19. W. Auzinger, H. Lehner, and E. Weinmüller. An efficient asymptotically correct error estimator for collocation solution to singular index-1 DAEs. *BIT Numer. Math*, 51:43–65, 2011.
20. A. Backes. *Extremalbedingungen für Optimierungs-Probleme mit Algebro-Differentialgleichungen*. Logos Verlag Berlin, 2006. Dissertation, Humboldt-University Berlin, October 2005/January 2006.
21. G. Bader and U.M. Ascher. A new basis implementation for a mixed order boundary value ODE solver. *SIAM J. Scient. Stat. Comput.*, 8, 1987.
22. Y. Bai. *Modified collocation methods for boundary value problems in differential-algebraic equations*. PhD thesis, Philipps-Universität Marburg/Lahn, Fachbereich Mathematik, 1991.

23. Y. Bai. A perturbated collocation method for boundary value problems in differential-algebraic equations. *Appl. Math. Comput.*, 45:269–291, 1991.
24. Y. Bai. A modified Lobatto collocation for linear boundary value problems of differential-algebraic equations. *Computing*, 49:139–150, 1992.
25. A. Baiz. *Effiziente Lösung periodischer differential-algebraischer Gleichungssysteme in der Schaltungssimulation*. PhD thesis, Technische Universität Darmstadt, Fachbereich Informatik, 2003. Shaker Verlag Aachen.
26. K. Balla. Differential-algebraic equations and their adjoints. Dissertation, Doctor of the Hungarian Academy of Sciences, Hungarian Academy of Sciences, Budapest, 2004.
27. K. Balla and R. März. Transfer of boundary conditions for DAEs of index 1. *SIAM J. Numer. Anal.*, 33(6):2318–2332, 1996.
28. K. Balla and R. März. Linear differential-algebraic equations of index 1 and their adjoints. *Results in Mathematics*, 37:13–35, 2000.
29. K. Balla and R. März. A unified approach to linear differential-algebraic equations and their adjoints. *Journal for Analysis and its Applications*, 21(3):783–802, 2002.
30. K. Balla and R. März. Linear boundary value problems for differential-algebraic equations. *Miskolc Mathematical Notes*, 5(1):3–18, 2004.
31. B. Barz and E. Suschke. Numerische Behandlung eines Algebro-Differentialgleichungssystems. RZ-Mitteilungen, Humboldt-Universität, Berlin, 1994.
32. M.L. Bell and R.W.H. Sargent. Optimal control of inequality constrained DAE systems. *Computers and Chemical Engineering*, 24:2385–2404, 2000.
33. L. Biegler, S.L. Campbell, and V. Mehrmann. *Control and optimization with differential-algebraic constraints*. SIAM, 2011.
34. H.G. Bock, E. Eich, and J.P. Schlöder. Numerical solution of constrained least squares boundary value problems in differential-algebraic equations. In K. Strehmel, editor, *Numerical treatment of differential equations, NUMDIFF-4*, volume 104 of *Teubner Texte zur Mathematik*. Teubner, 1987.
35. K.E. Brenan, S.L. Campbell, and L.R. Petzold. *Numerical solution of initial-value problems in differential-algebraic equations*. North Holland, New York, 1989.
36. P.N. Brown, A.C. Hindmarsh, and L.R. Petzold. Consistent initial condition calculation for differential-algebraic systems. *SIAM J. Sci. Comput.*, 19(5):1495–1512, 1998.
37. R. Callies. *Entwurfsoptimierung und optimale Steuerung. Differential-algebraische Systeme, Mehrgitter-Mehrzielansätze und numerische Realisierung*. Habilitation, Technische Universität München, 2000.
38. K. D. Clark and L. R. Petzold. Numerical solution of boundary value problems in differential-algebraic systems. *SIAM J. Sci. Stat. Comput.*, 10:915–936, 1989.
39. C. de Boor and B. Swartz. Collocation at Gaussian Points. *SIAM J. Numer. Anal.*, 10:582–606, 1973.
40. F.R. de Hoog and R. Weiss. Difference methods for boundary value problems with a singularity for the first kind. *SIAM J. Numer. Anal.*, 13:775–813, 1976.
41. A. Degenhardt. A collocation method for boundary value problems of transferable differential-algebraic equations. Preprint (Neue Folge) 182, Humboldt-Universität zu Berlin, Sektion Mathematik, 1988.
42. A. Degenhardt. Collocation for transferable differential-algebraic equations. In E. Griepentrog, M. Hanke, and R. März, editors, *Berlin Seminar on Differential-Algebraic Equations*, volume 92-1 of *Seminarberichte*, pages 83–104. Humboldt-Universität zu Berlin, Fachbereich Mathematik, 1992.
43. A. Dick, O. Koch, R. März, and E. Weinmüller. Convergence of collocation schemes for boundary value problems in nonlinear index-1 DAEs with a singular point. *Mathematics of Computation*, 82(282):893–918, 2013.
44. R. Dokchan. *Numerical integration of differential-algebraic equations with harmless critical points*. PhD thesis, Humboldt-University of Berlin, Institute of Mathematics, 2011.
45. E. Eich-Soellner and C. Führer. *Numerical Methods in Multibody Dynamics*. B. G. Teubner Stuttgart, 1998.

46. H. W. Engl, M. Hanke, and A. Neubauer. Tikhonov regularization of nonlinear differential-algebraic equations. In P. C. Sabatier, editor, *Inverse Methods in Action*, pages 92–105. Springer Berlin Heidelberg, 1990.
47. R. England, R. Lamour, and J. Lopez-Estrada. Multiple shooting using a dichotomically stable integrator for solving DAEs. *APNUM*, 42:117–131, 2002.
48. D. Estévez Schwarz and R. Lamour. The computation of consistent initial values for nonlinear index-2 differential-algebraic equations. *Numerical Algorithms*, 26(1):49–75, 2001.
49. D. Estévez Schwarz and R. Lamour. *Progress in Differential-Algebraic Equations. Descriptor 2013*, chapter Monitoring singularities while integrating DAEs, pages 73–96. Differential-Algebraic Equations Forum. Springer Heidelberg, 2014.
50. D. Estévez Schwarz and R. Lamour. Diagnosis of singular points of properly stated DAEs using automatic differentiation. *Numerical Algorithms*, 2015. to appear.
51. C. Franke. *Numerical methods for the investigation of periodic motions in multibody dynamics. A collocation approach*. PhD thesis, Universität Ulm, 1998. Shaker Verlag Aachen 1998.
52. C. W. Gear. Maintaining solution invariants in the numerical solution of ODEs. *SIAM J. Sci. Stat. Comp.*, 7:734–743.
53. M. Gerdts. Direct shooting method for the numerical solution of higher-index DAE optimal control problems. *Journal of Optimization Theory and Applications*, 117(2):267–294, 2003.
54. M. Gerdts. *Surveys in Differential-Algebraic Equations II*, chapter A survey on optimal control problems with differential-algebraic equations. Springer Heidelberg, 2015. ed. by A. Ilchmann and T. Reis.
55. E. Griepentrog and R. März. *Differential-Algebraic Equations and Their Numerical Treatment*. Teubner-Texte zur Mathematik No. 88. BSB B.G. Teubner Verlagsgesellschaft, Leipzig, 1986.
56. M. Hanke. On a least-squares collocation method for linear differential-algebraic equations. *Numer. Math.*, 54:79–90, 1988.
57. M. Hanke. *Beiträge zur Regularisierung von Randwertaufgaben für Algebro-Differentialgleichungen mit höherem Index*. Dissertation(B), Habilitation, Humboldt-Universität zu Berlin, Institut für Mathematik, 1989.
58. M. Hanke. On the regularization of index 2 differential-algebraic equations. *Journal of Mathematical Analysis and Application*, 151:236–253, 1990.
59. M. Hanke. Asymptotic expansions for regularization methods of linear fully implicit differential-algebraic equations. *Zeitschrift für Analysis und ihre Anwendungen*, 13:513–535, 1994.
60. I. Higueras and R. März. Differential algebraic equations with properly stated leading term. *Comp. Math. Appl.*, 48:215–235, 2004.
61. I. Higueras, R. März, and C. Tischendorf. Stability preserving integration of index-1 DAEs. *Appl. Numer. Math.*, 45(2-3):175–200, 2003.
62. M. D. Ho. A collocation solver for systems of boundary-value differential/algebraic equations. *Computers and Chem. Eng.*, 7:735–737, 1983.
63. B. Houska and M. Diehl. A quadratically convergent inexact SQP method for optimal control of differential algebraic equations. *Optimal Control Applications and Methods*, 34:396–414, 2013.
64. L. V. Kalachev and R. E. O'Malley. Boundary value problems for differential-algebraic equations. *Numerical Functional Analysis and Optimization*, 16:363–378, 1995.
65. H. B. Keller and jr. A. B. White. Difference methods for boundary value problems in ordinary differential equations. *SINUM*, 12(5):791–802, 1975.
66. H.B. Keller. Approximation Methods for Nonlinear Problems with Application to Two-Point Boundary Value Problems. *Math. Comp.*, 29:464–474, 1975.
67. M. Kiehl. Sensitivity analysis of ODEs and DAEs – Theory and implementation guide. *Optimization Methods and Software*, 10:803–821, 1999.
68. O. Koch. Asymptotically correct error estimation for collocation methods applied to singular boundary value problems. *Numer. Math.*, 101:143–164, 2005.

69. O. Koch, P. Kofler, and E. Weinmüller. Initial value problems for systems of ordinary first and second order differential equations with a singularity of the first kind. *Analysis*, 21:373–389, 2001.
70. O. Koch, R. März, D. Praetorius, and E. Weinmüller. Collocation for solving DAEs with singularities. ASC Report 32/2007, Vienna University of Technology, Institute for Analysis and Scientific Computing, 2007.
71. O. Koch, R. März, D. Praetorius, and E. Weinmüller. Collocation methods for index-1 DAEs with a singularity of the first kind. *Mathematics of Computation*, 79(269):281–304, 2010.
72. O. Koch and E. Weinmüller. The convergence of shooting methods for singular boundary value problems. *Math. Comp.*, 72(241):289–305, 2003.
73. A. Kopelmann. Ein Kollokationsverfahren für überführbare Algebro-Differentialgleichungen. Preprint (Neue Folge) 151, Humboldt-Universität zu Berlin, Sektion Mathematik, 1987.
74. P. Kunkel and V. Mehrmann. *Differential-Algebraic Equations - Analysis and Numerical Solution*. EMS Publishing House, Zürich, Switzerland, 2006.
75. P. Kunkel, V. Mehrmann, and R. Stöver. Symmetric collocation methods for unstructured nonlinear differential-algebraic equations of arbitrary index. *Numer. Math.*, 98:277–304, 2004.
76. P. Kunkel and R. Stöver. Symmetric collocation methods for linear differential-algebraic boundary value problems. *Numer. Math.*, 91:475–501, 2002.
77. R. Lamour. A shooting method for fully implicit index-2 differential-algebraic equations. *SIAM J. Sci. Comput.*, 18(1):94–114.
78. R. Lamour. Bestimmung optimaler Integrationsrichtungen beim Mehrfachschießverfahren zur Lösung von Zwei-Punkt- Randwertproblemen. Wiss. Beitr., Martin-Luther-Univ. Halle Wittenberg 1984/24(M 33), 66-70 (1984)., 1984.
79. R. Lamour. A well–posed shooting method for transferable DAEs. *Numerische Mathematik*, 59, 1991.
80. R. Lamour. Oscillations in differential–algebraic equations. In *Seminarbericht Nr. 92–1*. Fachbereich Mathematik der Humboldt–Universität zu Berlin, 1992.
81. R. Lamour. Index determination and calculation of consistent initial values for DAEs. *Comp. Math. Appl.*, 50(2):1125–1140, 2005.
82. R. Lamour and R. März. Detecting structures in differential-algebraic equations: Computational aspects. *Journal of Computational and Applied mathematics*, 236(16):4055–4066, 2012. Special Issue: 40 years of Numerical Math.
83. R. Lamour, R. März, and C. Tischendorf. *Differential-Algebraic Equations: A Projector Based Analysis*. Differential-Algebraic Equations Forum. Springer-Verlag Berlin Heidelberg New York Dordrecht London, 2013. Series Editors: A. Ilchman, T. Reis.
84. R. Lamour, R. März, and R. Winkler. How floquet theory applies to index-1 differential-algebraic equations. *J. Appl. Math.*, 217(2):372–394, 1998.
85. R. Lamour, R. März, and R. Winkler. Stability of periodic solutions of index-2 differential algebraic systems. *J. Math. Anal. Appl.*, 279:475–494, 2003.
86. R. Lamour and F. Mazzia. Computation of consistent initial values for properly stated index-3 DAEs. *BIT Numer Math*, 49:161–175, 2009.
87. M. Lentini and R. März. The condition of boundary value problems in transferable differential-algebraic equations. *SIAM J. Numer. Anal.*, 27(4):1001–1015, 1990.
88. M. Lentini and R. März. Conditioning and dichotomy in differential-algebraic equations. *SIAM J. Numer. Anal.*, 27(6):1519–1526, 1990.
89. R. März. On difference and shooting methods for boundary value problems in differential-algebraic equations. *ZAMM*, 64(11):463–473, 1984.
90. R. März. On correctness and numerical treatment of boundary value problems in DAEs. *Zhurnal Vychisl. Matem. i Matem. Fiziki*, 26(1):50–64, 1986.
91. R. März. Numerical methods for differential-algebraic equations. *Acta Numerica*, pages 141–198, 1992.
92. R. März. On linear differential-algebraic equations and linearizations. *Applied Numerical Mathematics*, 18:267–292, 1995.

93. R. März. Managing the drift-off in numerical index-2 differential algebraic equations by projected defect corrections. Technical Report 96-32, Humboldt University, Institute of Mathematics, 1996.
94. R. März. Notes on linearization of differential-algebraic equations and on optimization with differential-algebraic constraints. Technical Report 2011-16, Humboldt-Universität zu Berlin, Institut für Mathematik, 2011.
95. R. März. Notes on linearization of DAEs and on optimization with differential-algebraic constraints. In L. T. Biegler, S. L. Campbell, and V. Mehrmann, editors, *Control and optimization with differential-algebraic constraints*, Advances in Design and Control, pages 37–58. SIAM, 2012.
96. R. März. *Surveys in Differential-Algebraic Equations II*, chapter Differential-Algebraic Equations from a Functional-Analytic Viewpoint: A Survey. Springer Heidelberg, 2015. ed. by A. Ilchmann and T. Reis.
97. R. März and R. Riaza. Linear differential-algebraic equations with properly leading term: A-critical points. *Math. Comp. Model. Dyn. Sus.*, 13:291–314, 2007.
98. R. März and E. B. Weinmüller. Solvability of boundary value problems for systems of singular differential-algebraic equations. *SIAM J. Math. Anal.*, 24(1):200–215, 1993.
99. K. Moszyński. A method of solving the boundary value problem for a system of linear ordinary differential equations. *Algorithmy*, 11(3):25–43, 1964.
100. T. Petry. On the stability of the Abramov transfer for differential-algebraic equations of index 1. *SIAM J. Numer. Anal.*, 35(1):201–216, 1998.
101. T. Petry. *Realisierung des Newton-Kantorovich-Verfahrens für nichtlineare Algebro-Differentialgleichungen mittels Abramov-Transfer*. PhD thesis, Humboldt-Universität zu Berlin, 1998. Logos Verlag Berlin.
102. P.J. Rabier and W.C. Rheinboldt. Theoretical and numerical analysis of differential-algebraic equations. In Ciarlet, P. G. et al., editor, *Handbook of numerical analysis*, volume VIII, Techniques of scientific computing (Part 4), pages 183–540. Amsterdam: North Holland/ Elsevier, 2002.
103. R. Riaza. *Differential-Algebraic Systems. Analytical Aspects and Circuit Applications*. World Scientific Singapure, 2008.
104. R. Riaza. *Surveys in Differential-Algebraic Equations I*, chapter DAEs in Circuit Modelling: A Survey. Differential-Algebraic Equations Forum. Springer Heidelberg, 2013. Eds. A. Ilchmann, T. Reis.
105. R. Riaza and R. März. Linear index-1 daes: regular and singular problems. *Acta Appl. Math.*, 84:29–53, 2004.
106. V. H. Schulz, H. G. Bock, and M. C. Steinbach. Exploiting invariants in the numerical solution of multipoint boundary value problems for DAE. *SIAM J. Sci. Comp.*, 19:440–467, 1998.
107. P. Selting and Q. Zheng. Numerical stability analysis of oscillating integrated circuits. *Journal of Computational and Applied Mathematics*, 82:367–378, 1997.
108. L.F. Shampine. Conservative laws and the numerical solution of ODEs. *Comp.and Maths. with Applications*, 12:1287–1296, 1986.
109. B. Simeon. *Computational Flexible Multibody Dynamics. A Differential-Algebraic Approach*. Differential-Algebraic Equations Forum. Springer Heidelberg, 2013.
110. H.J. Stetter. The defect correction principle and discretization methods. *Numer. Math.*, 29:425–443, 1978.
111. R. Stöver. *Numerische Lösung von linearen differential-algebraischen Randwertproblemen*. PhD thesis, Universität Bremen, January 1999. Doctoral thesis, Logos Verlag Berlin.
112. S. Trenn. *Surveys in Differential-Algebraic Equations I*, chapter Solution concepts for linear DAEs: A Survey, pages 137–172. Springer Heidelberg, 2013. ed. by A. Ilchmann and T. Reis.
113. B. Wernsdorf. *Ein Kollokationsverfahren zur numerischen Bestimmung periodischer Lösungen von nichtlinearen Algebro-Differentialgleichungen*. PhD thesis, Humboldt-Universität zu Berlin, Sektion Mathematik, 1984.

114. P. M. E. J. Wijckmans. *Conditioning of differential-algebraic equations and numerical solutions of multibody dynamics*. PhD thesis, Technische Universiteit Eindhoven, 1996.
115. P.E. Zadunaisky. On the estimation of errors propagated in the numerical integration of ODEs. *Numer. Math.*, 27, 1976.

Index