Differential-Algebraic Equations Forum

DAE-F

Achim Ilchmann
Timo Reis *Editors*

Surveys in Differential-Algebraic Equations III

Springer

Differential-Algebraic Equations Forum

Differential-Algebraic Equations Forum

The series “Differential-Algebraic Equations Forum” is concerned with analytical, algebraic, control theoretic and numerical aspects of differential algebraic equations (DAEs) as well as their applications in science and engineering. It is aimed to contain survey and mathematically rigorous articles, research monographs and textbooks. Proposals are assigned to an Associate Editor, who recommends publication on the basis of a detailed and careful evaluation by at least two referees. The appraisals will be based on the substance and quality of the exposition.

More information about this series at http://www.springer.com/series/11221

Achim Ilchmann • Timo Reis
Editors

Surveys in Differential-Algebraic Equations III

Springer

Editors
Achim Ilchmann
Institut für Mathematik
Technische Universität Ilmenau
Ilmenau, Germany

Timo Reis
Fachbereich Mathematik
Universität Hamburg
Hamburg, Germany

ISSN 2199-7497 ISSN 2199-840X (electronic)
Differential-Algebraic Equations Forum
ISBN 978-3-319-22427-5 ISBN 978-3-319-22428-2 (eBook)
DOI 10.1007/978-3-319-22428-2

Library of Congress Control Number: 2015954642

Mathematics Subject Classification (2010): 34A09, 65L80, 93A15 65L10, 34B09, 34B60, 65F15, 15A24, 49J15

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland Springer is part of Springer Science+Business Media (www.springer.com)

Preface of Surveys in Differential Algebraic Equations III

We are pleased to present the third volume of survey articles in various fields of differential-algebraic equations (DAEs), and we stress that a fourth volume will appear within the series "Differential-Algebraic Equations Forum".

In this volume, we again extend the list of survey articles in the sense that they are of theoretical interest and equally relevant to applications.

The chapter "The Flexibility of DAE Formulations" shows that DAEs are not only the outcome of modeling; they may further lead to more elegant formulations in control and observer design problems, their numerical solution, and simulation. In the chapter "Reachability Analysis and Deterministic Global Optimization of DAE Models", an overview on (optimal) control of parameterized DAEs is given. The chapter "Numerical Linear Algebra Methods for Differential-Algebraic Equations" is about numerical treatment of controller design and optimal control problems for large-scale differential-algebraic systems. The final chapter "Boundary-Value Problems for Differential-Algebraic Equations: A Survey" is a survey about boundary value problems for DAEs. Problems of this kind occur, for instance, in optimal control.

We hope that this issue will contribute to complete the picture of the latest developments in DAEs. The collection of survey articles may also indicate that DAEs are now an established field in applied mathematics.

Ilmenau, Germany — Achim Ilchmann
Hamburg, Germany — Timo Reis
May 2015

Contents

The Flexibility of DAE Formulations

Stephen L. Campbell

Abstract There has been extensive research on DAEs and their applications. One major reason given for the usefulness of DAEs is that they are the initial way that many complex systems are most naturally modeled. But there are other ways that DAE formulations are useful. This survey focuses on a number of problems where the extra flexibility of a DAE formulation permits the solution of a problem that would be hard to solve otherwise.

Keywords Delays • Differential-algebraic equation • Numerical methods • Observer • Optimal control

MSC: 34A09, 65L80, 93B07, 49J15, 34A40

1 Introduction

There has been extensive research on differential algebraic equations (DAEs) and their applications. Note the books [1, 22, 33, 61–63, 84, 86] and such survey papers as [5, 24]. One major reason given for the usefulness of DAEs is that they are the initial way that many complex systems are most naturally modeled. This is especially true in chemical, electrical, and mechanical engineering and with models formed by interconnecting various submodels. But there are other ways that DAE formulations are useful. This survey focuses on a number of problems where the extra flexibility of a DAE formulation permits the solution of a problem that would be hard to solve otherwise. This survey takes the form of carefully chosen case studies where the DAE formulation has been found useful. Our examples are taken from work on control problems, their numerical solution, and simulation. Some examples are from our work and some is from the work of others. No attempt is

S.L. Campbell (✉)
Department of Mathematics, North Carolina State University, Raleigh, NC, USA
e-mail: slc@ncsu.edu

A. Ilchmann, T. Reis (eds.), *Surveys in Differential-Algebraic Equations III*,
Differential-Algebraic Equations Forum, DOI 10.1007/978-3-319-22428-2_1

made at completeness. So perhaps "essay" is a more accurate word to describe this paper than "survey."

We shall assume that the reader is familiar with what a DAE is. In particular, we will assume that readers have heard of the index of a DAE. There are several definitions of index. We take the differential index based off the derivative array as discussed just before (2.40). See also [22, 23, 62]. However, we do not assume that they are familiar with the different control and numerical topics we discuss. For material that has appeared in journal articles we omit some of the proofs and technical detail unless they are relevant to the point being made. For material that has only appeared in conference papers, especially if the proceedings are not immediately accessible, more details are provided. In sections with material that has not appeared anywhere full details in establishing the statements are given.

Section 2 will give some examples from control theory and in particular observer design. We will not discuss the design of observers for DAEs, this is done, for example, in [24]. Rather, we will present two different examples where the flexibility of a DAE formulation when designing observers can be exploited. In Sect. 2.1 the use of a DAE observer allows us to get linear error dynamics which is very useful in observer design. Section 2.2 discusses the estimation of disturbances. Section 3 turns to the examination of optimal control problems. It turns out that the advantages and disadvantages of a DAE formulation are highly dependent on the type of numerical methods used. We will focus on direct transcription both because it is widely used and because the computational theory is not always what one first thinks it is. Section 3 starts by describing what direct transcription is. Then two illustrations of the advantages of a DAE formulation are given in Sects. 3.1 and 3.2. Two distinct examples are given in Sects. 3.1.1 and 3.1.2. Section 4 discusses the optimal control of delayed systems and the advantages of DAE formulations of them. Finally some conclusions are in Sect. 5. Enough citations are given to enable the reader to follow up on a given comment, example, or application, but citing all relevant work would make the bibliography as long as the text and so many relevant citations are omitted.

2 Observer Design

Observers play a fundamental role in control theory and applications. There is an extensive literature on observers. Observers and the system being observed can be continuous or discrete time, deterministic or stochastic. We focus here on the continuous time deterministic case. The basic idea is that there is a dynamical system

$$F(\dot{x}, x, u, \psi, t) = 0, \tag{2.1a}$$

and an output equation

$$y = h(x, u, \psi, t). \tag{2.1b}$$

Here x is the state, u is the control or input, y is the output or measurements. ψ if present represents noise or uncertainty or faults. The particular assumptions on ψ will depend on which application is being discussed, but in general it is at least piecewise smooth. Both u and y are considered known. Unless necessary for clarity we delete the "(t)" from functions such as $x, y, u, \psi, \hat{x}, \hat{y}, z$. The goal is to get estimates $\hat{x}$ of x. Later we include ψ in Sect. 2.2.

An observer is another dynamical system for $\hat{x}$,

$$\hat{F}(\dot{\hat{x}}, \hat{x}, \hat{y}, y, u, t) = 0 \tag{2.2a}$$

and an output equation

$$\hat{y} = \hat{h}(\hat{x}, u, t). \tag{2.2b}$$

If $x(0) = \hat{x}(0)$ and $\psi = 0$, then $x = \hat{x}$ for all $t > 0$. If $\psi = 0$ and $x(0) - \hat{x}(0) \neq 0$, then we want $x - \hat{x} \to 0$ as $t \to \infty$. This convergence of $x - \hat{x}$ can be global or local if $x(0) - \hat{x}(0)$ has to be small to begin with where small is determined by the amount of nonlinearity. If $\psi \neq 0$, then we either want $x(0) - \hat{x}(0)$ to go to zero if ψ goes to zero fast enough or for $x(0) - \hat{x}(0)$ to become small if ψ is small.

System (2.1) is the actual physical system with variables to be estimated. All that is known is u and y. On the other hand (2.2) exists in software or hardware and $\hat{x}, u, \hat{y}, y, t$ are all known and available.

For the linear time invariant system

$$\dot{x} = Ax + Bu + R\psi \tag{2.3a}$$

$$y = Cx + Du + S\psi, \tag{2.3b}$$

the Luenberger observer takes the form

$$\dot{\hat{x}} = A\hat{x} + Bu + L(y - \hat{y}) \tag{2.4a}$$

$$\hat{y} = C\hat{x} + Du \tag{2.4b}$$

and the error equation for the estimation error $e = x - \hat{x}$ is

$$\dot{e} = (A - LC)e + R\psi - LS\psi. \tag{2.5}$$

L is chosen to make $A - LC$ asymptotically stable if possible.

If A is $n \times n$, then

$$\Theta = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}$$

is called the observability matrix. The pair $\{A, C\}$ is observable if $rank(\Theta) = n$. If $\{A, C\}$ is not observable, then the nullspace of Θ, $N(\Theta)$, is an A invariant subspace called the unobservable subspace. Eigenvalues of A restricted to $N(\Theta)$ are called the unobservable eigenvalues.

If $\{A, C\}$ is an observable pair, then L can be chosen to place the eigenvalues of $A - LC$ arbitrarily. $\{A, C\}$ is called detectable if any unobservable eigenvalues have negative real part. If $\{A, C\}$ is detectable, then L can be chosen so that $A - LC$ is asymptotically stable and the observable eigenvalues can be placed arbitrarily.

Observers for DAEs and observers which are DAEs have been discussed extensively in the literature. Notes [17, 24, 30, 35–40, 102] and the bibliography of [24]. Our emphasis here is different. We focus on how using a DAE, or a higher index DAE, or a higher dimensional DAE, can provide advantages over a more standard observer.

2.1 Nonlinear Observers

If the dynamics of the system being observed is nonlinear, then either the observer dynamics, or the error equation, or both are nonlinear. This makes the design of the observer so that the error equation is asymptotically stable more difficult. There are several approaches to trying to design the observer. One is to try to reformulate the problem so that the error equation becomes linear. If the error equation becomes linear, then it is much easier to stabilize the error equation by choosing L appropriately. This is the approach of this section.

Observers are usually formulated as explicit systems of differential equations and implemented using standard ODE solvers. In this section, we show that there can be advantages in formulating the observer as a DAE even if the system is originally an ODE. We first review the general idea of DAE observer design of Nikoukhah [78]. We then give two special normal forms for which DAE observer design yields an observer with linear error dynamics. The idea to use DAE observer normal forms is introduced on index one DAE observers and then extended to index two Hessenberg DAEs. This allows us to enlarge the class of nonlinear systems for which linear observer error dynamics can be achieved. This section is based on the work of Von Wissel [96] and von Wissel et al. [97]. Note also [56].

Consider the nonlinear systems

$$\dot{x} = f(x, u) \tag{2.6a}$$

$$y = h(x). \tag{2.6b}$$

where f and h are smooth vector fields on $\mathbb{R}^n$ and $\mathbb{R}^p$, and the p measurements y in (2.6b) are independent. The problem of observer design consists in finding a nonlinear system

$$0 = \hat{f}(\dot{\omega}, \omega, u, y) \tag{2.7a}$$

$$\hat{x} = \hat{g}(\omega, u, y) \tag{2.7b}$$

that generates an estimate $\hat{x}(t)$ of the true value $x(t)$.

There are essentially three approaches that have been used in the past for nonlinear observer design with a number of variations on each approach. The first approach is a natural extension of linear observers and is very commonly adopted, for example see the techniques presented in the comparative study of [100]. The other approach to observer design is to work directly with system equations (2.6), either formulating the estimation problem as a nonlinear algebraic system of equations which must be solved periodically using for example Newton's method, see for example [76], or formulating it as an optimization problem over some sliding finite horizon which is again solved periodically [75]. The third approach is to use specially designed Lyapunov equations to stabilize the dynamics of the nonlinear error equation.

We present an alternative to these three approaches. We show that there can be advantages in formulating the observer as a DAE which can then be solved using a numerical DAE solver. For index one DAEs the numerical integration can, for example, be done by the DAE solver DASSL [22, 82]. More importantly, if (2.6) has a special form and verifies some algebraic conditions, we can easily construct a DAE index one observer that has linear time invariant observer error dynamics.

The class of nonlinear systems is even larger if we allow (2.7) to be an index two Hessenberg DAE [22]. Index two Hessenberg DAEs are of particular interest since this type of DAE can also be safely solved by differential algebraic system solvers, for instance DASSL with fixed stepsize [22] or Radau5 [47]. The usefulness of this approach will be shown on a simple example. For a more detailed analysis of DAE index two design and its application to mechanical type problems see [96].

2.1.1 Index One DAE Observer

System (2.6) is a DAE in x since u and y are supposed to be known. This DAE is over-determined with n unknowns and $n + p$ equations. This DAE describes all the constraints that we have for constructing $\hat{x}$. To make this DAE numerically integrable, we can do relaxation by introducing a p-dimensional vector function

λ into this DAE. One way to introduce λ is in the algebraic part which is the observation (2.6b). Then we have

$$\dot{\hat{x}} = f(\hat{x}, u) + g(\lambda) \tag{2.8a}$$

$$0 = y - h(\hat{x}) + \lambda, \tag{2.8b}$$

which means that we relax the entire estimate $\hat{x}$. Note that (2.8) is now a DAE in $\hat{x}, \lambda$ and the number of unknowns equals the number of equations. It is easy to see that introducing λ as in (2.8) leads to the usual explicit formulation of the observer. In particular, for the linear system (2.3a) with $\psi = 0$, solving (2.8b) for λ and having $g(\lambda) = L\lambda$ gives the Luenberger observer in (2.4).

The other way to regularize (2.6) is to introduce λ in the differential part (2.6a). In this case we constrain partially the estimate $\hat{x}$ through $h(\hat{x})$ by the observation y and relax only the remaining part. This is the way index one DAE observer design is introduced in [78].

System (2.13) which follows can be motivated by the linear time invariant case. Given

$$\dot{x} = Ax + Bu \tag{2.9a}$$

$$y = Cx \tag{2.9b}$$

with C full row rank, the system

$$\dot{x} = Ax + C^T\lambda + Bu \tag{2.10a}$$

$$y = Cx \tag{2.10b}$$

with u, y known is index two in x, λ. However,

$$\dot{x} = Ax + C^T(\dot{\lambda} + G\lambda) + Bu \tag{2.11a}$$

$$y = Cx \tag{2.11b}$$

is index one in x, λ. To see that (2.11) is index one, note that if we do one differentiation of the constraint (2.11b) the coefficient matrix of the derivatives $\{\dot{x}, \dot{\lambda}\}$ becomes invertible since C has full row rank. In fact,

$$\begin{bmatrix} I & -C^T \\ C & 0 \end{bmatrix}^{-1} = \begin{bmatrix} I - C^\dagger C & C^\dagger \\ (C^\dagger)^T & (CC^T)^{-1} \end{bmatrix},$$

where $C^\dagger$ is the Moore–Penrose generalized inverse of C [28].

Note that (2.11) can be written as

$$\dot{x} - C^T\dot{\lambda} = Ax + Bu + C^T G\lambda \tag{2.12a}$$

$$0 = y - Cx, \tag{2.12b}$$

which is in the form of (2.13) with $\Phi = -C^T, f = Ax + Bu, \Gamma = C^T G$.

Throughout the remainder of this section we say the DAE of the form

$$\dot{\hat{x}} + \Phi(\hat{x})\dot{\lambda} = f(\hat{x}, u) + \Gamma(\hat{x}, u)\lambda \tag{2.13a}$$

$$0 = y - h(\hat{x}) \tag{2.13b}$$

is the canonical observer DAE for system (2.6). For $\Phi = -h_x(x) = -\partial h(x)/\partial x$, Γ any matrix of appropriate dimensions depending on $\hat{x}$ and u, we recover the canonical index one DAE observer form of [78].

For the remainder of this paper we will sometimes use the MATLAB notation for stacking matrices or vectors. Thus if matrices H, K have the same number of columns, then $[H; K] = [H^T, K^T]^T$. Under appropriate observability conditions, $\lambda \to 0$ implies that $\hat{x} - x \to 0$ for $\hat{x}(0) \neq x(0)$. Γ has to be chosen such that $\lambda \to 0$. To see what type of conditions suffice, observe that for sufficiently small observation error $e = \hat{x} - x$ we have

$$\begin{bmatrix} I & -\Phi(\hat{x}) \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{e} \\ \dot{\lambda} \end{bmatrix} = \begin{bmatrix} f_x(\hat{x}) & \Gamma(\hat{x}, u) \\ h_x(\hat{x}) & 0 \end{bmatrix} \begin{bmatrix} e \\ \lambda \end{bmatrix} + O(\|e\|^2). \tag{2.14}$$

System (2.14) has an equilibrium point at $[e; \lambda] = 0$. Then from [78] we have Theorem 2.1.

Theorem 2.1 *Suppose that f, f_x, h, and h_x are bounded in a neighborhood of $(x(t), u(t))$. If*

$$\begin{bmatrix} I & -\Phi(\hat{x}) \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{e} \\ \dot{\lambda} \end{bmatrix} = \begin{bmatrix} f_x(\hat{x}) & \Gamma(\hat{x}, u) \\ h_x(\hat{x}) & 0 \end{bmatrix} \begin{bmatrix} e \\ \lambda \end{bmatrix} \tag{2.15}$$

is exponentially asymptotic stable at the origin, then the estimation error $\tilde{x} = \hat{x} - x$ associated with the canonical index one DAE observer with $\Phi(\hat{x}) = -h_x(\hat{s})$ converges exponentially to zero provided $\tilde{x}(0)$ and $\lambda(0)$ are sufficiently small.

For fixed $[\hat{x}; u]$ we may chose Γ such that this equilibrium point is stable for all fixed $[\hat{x}; u]$. If $[\hat{x}; u]$ varies slowly against $[\lambda; e]$, stability of (2.14) for fixed $[\hat{x}; u]$ implies stability for varying $[\hat{x}; u]$. See the discussion of extended linearization in [2].

2.1.2 Exact Linearization of the Error Equation

If system (2.6) is in the special form

$$\dot{x}_1 = f_1(x_1, u) + F_1(x_1, u)x_2 \tag{2.16a}$$

$$\dot{x}_2 = f_2(x_1, u) + F_2(x_1, u)x_2 \tag{2.16b}$$

$$y = x_1, \tag{2.16c}$$

we will refer to (2.16) as the index one DAE observer normal form. If we chose

$$\Phi = \begin{bmatrix} \Omega_1^{-1}(y) \\ \Omega_1^{-1}(y)\Omega_2(y) \end{bmatrix}, \quad \Gamma = \begin{bmatrix} \Gamma_1(y) \\ \Gamma_2(y) \end{bmatrix},$$

then (2.13) is an index one DAE which will be referred to as the index one canonical DAE observer form and (2.14) becomes the linear time varying system

$$\begin{bmatrix} \dot{\lambda} \\ \dot{\tilde{x}}_2 \end{bmatrix} = \begin{bmatrix} \Omega_1\Gamma_1 & \Omega_1 F_1 \\ \Gamma_2 - \Omega_2\Gamma_1 & F_2 - \Omega_2 F_1 \end{bmatrix} \begin{bmatrix} \lambda \\ \tilde{x}_2 \end{bmatrix}, \tag{2.17}$$

where $\tilde{x}$ is an approximation of e. There exist a number of methodologies to stabilize (2.17) by use of Ω and Γ such as extended linearization and Lyapunov design, for example. Here we consider the case where (2.17) can be made linear time invariant. For that we need that the matrices F_1 and F_2 have a particular structure.

Theorem 2.2 *Let $F_2(\hat{x}_1, u) = \overline{F}_2 + \tilde{F}_2(\hat{x}_1, u)$ where $\overline{F}_2$ is a constant matrix. If*

1. *there exists an invertible matrix Ω_1 such that $\Omega_1 F_1$ is constant,*
2. *$\overline{F}_2$ can be chosen such that the matrix pair $\{F_1, \overline{F}_2\}$ is observable and there exists a matrix F_2^* such that $\tilde{F}_2 = F_2^* F_1$,*

then the approximate error equation (2.17) can be made linear time invariant and its modes can arbitrarily be placed by proper choice of Γ_1 and Γ_2 for $\Omega_2 = F_2^*$.

Proof If (2.16) has Properties 1 and 2 of Theorem 2.2, the error equation (2.17) is

$$\begin{bmatrix} \dot{\lambda} \\ \dot{\tilde{x}}_2 \end{bmatrix} = \begin{bmatrix} \Omega_1\Gamma_1 & \Omega_1 F_1 \\ \Gamma_2 - F_2^*\Gamma_1 & \overline{F}_2 \end{bmatrix} \begin{bmatrix} \lambda \\ \tilde{x}_2 \end{bmatrix}. \tag{2.18}$$

Since $\Omega_1 F_1$ is constant and Ω_1 is invertible, (2.18) is a linear time invariant system for $\Gamma_1 = \Omega_1^{-1} K_1$ and $\Gamma_2 = K_2 + F_2^*\Gamma_1$, where K_1 and K_2 are constant matrices. The modes of (2.18) can arbitrarily be set by a proper choice of K_1 and K_2 if the matrix pair

$$\left\{ \begin{bmatrix} I & 0 \end{bmatrix}, \begin{bmatrix} 0 & \Omega_1 F_1 \\ 0 & \overline{F}_2 \end{bmatrix} \right\}$$

is observable which is the case if the matrix pair $\{\Omega_1 F_1, \overline{F}_2\}$ is observable.

2.1.3 Index Two DAE Observer

The idea of using DAEs as observers for (2.6) can be extended to the case where (2.13) is in Hessenberg semi-explicit index two form. Recall that a DAE is in Hessenberg semi-explicit index two form if it has the structure

$$\dot{\omega}_1 = \hat{f}(t, \omega_1, \omega_2) \tag{2.19a}$$

$$0 = \hat{g}(t, \omega_1), \tag{2.19b}$$

where $(\partial \hat{g}/\partial \omega_1)(\partial \hat{f}/\partial \omega_2)$ is nonsingular.

Hessenberg forms are of interest for several reasons. For one, many applications such as constrained mechanical system are index two or index three Hessenberg systems. Also there are numerical integrators for index two Hessenberg DAEs such as Radau5 [47]. Fixed step k-step BDF will converge after $k + 1$ steps [22] for Hessenberg index two DAEs.

The generalization of DAE index one observer design is, for example, of interest in the case where (2.6) can be put in the form

$$\dot{x}_1 = x_2 \tag{2.20a}$$

$$\dot{x}_2 = f_2(x_1, x_2, u) + F_2(x_1, u)x_3 \tag{2.20b}$$

$$\dot{x}_3 = f_3(x_1, x_2, u) + F_3(x_1, u)x_3 \tag{2.20c}$$

$$y = x_1. \tag{2.20d}$$

This particular form is frequently encountered in mechanical type systems where x_2 is velocity. Flexible joint robots are, for example, in this form [96]. System (2.20) will be referred to as the index two DAE observer normal form. The observer is the canonical observer DAE (2.13) with

$$\Phi = \begin{bmatrix} 0 \\ \Omega_1^{-1}(y) \\ \Omega_1^{-1}(y)\Omega_2(y) \end{bmatrix}, \qquad \Gamma = \begin{bmatrix} 0 \\ \Gamma_1(y, \hat{x}_2, u) \\ \Gamma_2(y, \hat{x}_2, u) \end{bmatrix},$$

which we will refer to as index two canonical DAE observer form.

One way to insure numerical integrability of the observer equations is to show that the index two canonical DAE observer form can be transformed into Hessenberg form. Since the transformation involves a number of equations, we will summarize it in Lemma 2.3.

Lemma 2.3 *The change of coordinates*

$$\omega_1 = \hat{x}_1 \tag{2.21a}$$

$$\omega_2 = \hat{x}_2 \tag{2.21b}$$

$$\omega_3 = \hat{x}_3 - \Omega_2(\hat{x}_1)\hat{x}_2 \tag{2.21c}$$

$$\omega_4 = \lambda + \Omega_1(\hat{x}_1)\hat{x}_2 \tag{2.21d}$$

transforms (2.20) *into the Hessenberg semi-explicit index two form*

$$\dot{\omega}_1 = \omega_2 \tag{2.22a}$$

$$\dot{\omega}_3 = \hat{f}_2(\omega_1, \omega_2, \omega_4, u) + \hat{F}_2(\omega_1, \omega_2, u)\omega_3 \tag{2.22b}$$

$$\dot{\omega}_4 = \hat{f}_3(\omega_1, \omega_2, \omega_4, u) + \hat{F}_3(\omega_1, \omega_2, u)\omega_3 \tag{2.22c}$$

$$0 = y - \omega_1, \tag{2.22d}$$

where

$$\hat{f}_2 = f_3 - \Omega_2 f_2 - \left(\dot{\Omega}_2 - F_3\Omega_2 + \Omega_2 F_2 \Omega_2\right)\omega_2 + (\Gamma_2 - \Omega_2\Gamma_1)\lambda \tag{2.23a}$$

$$\hat{f}_3 = \Omega_1 f_2 + \Omega_1\left(\dot{\Omega}_1 + F_2\Omega_2\right)\omega_2 + \Omega_1\Gamma_1\lambda \tag{2.23b}$$

$$\hat{F}_2 = F_3 - \Omega_2 F_2, \tag{2.23c}$$

$$\hat{F}_3 = \Omega_1 F_2 \tag{2.23d}$$

and $\lambda = \omega_4 - \Omega_1\omega_1$.

The system (2.21) is not a pure Hessenberg system but rather a cascade of two systems. Equations (2.22a) and (2.22d) form a Hessenberg index two system in ω_1, ω_2 for a given y. This system is then cascaded into (2.22b) and (2.22c) which is an ODE in ω_3, ω_4. Then (2.22) is index two in $\{\omega_1, \omega_2, \omega_3, \omega_4\}$ if y and u are known. We shall prove this directly.

Proof We have

$$\dot{\omega}_3 = \dot{\hat{x}}_3 - \dot{\Omega}_2\hat{x}_2 - \Omega_2\dot{\hat{x}}_2 \tag{2.24a}$$

$$\dot{\omega}_4 = \dot{\lambda} + \dot{\Omega}_1\hat{x}_2 + \Omega_1\dot{\hat{x}}_2. \tag{2.24b}$$

Elimination of $\Omega_2\dot{\hat{x}}_2$ in (2.24a) by use of (2.24b) and the left multiplication of (2.24b) by Ω_1^{-1} yields

$$\dot{\omega}_3 + \Omega_2\Omega_1^{-1}\dot{\omega}_4 = \left\{\dot{\hat{x}}_3 + \Omega_2\Omega_1^{-1}\dot{\lambda}\right\} - \left(\dot{\Omega}_2 - \Omega_2\Omega_1^{-1}\dot{\Omega}_1\right)\hat{x}_2$$

$$\Omega_1^{-1}\dot{\omega}_4 = \left\{\dot{\hat{x}}_2 + \Omega_1^{-1}\dot{\lambda}\right\} + \Omega_1^{-1}\dot{\Omega}_1\hat{x}_2.$$

The sums in the curly brackets on the right-hand side of the equations are the left-hand side of the third and second block row of the index two DAE observer

form. Together with (2.21) and (2.22) we get

$$\begin{aligned} 0 &= \hat{f}_2 + \Omega_2\Omega_1^{-1}\hat{f}_3 + \left(\hat{F}_2 + \Omega_2\Omega_1^{-1}\hat{F}_3\right)(\hat{x}_3 - \Omega_2\hat{x}_2) - \\ &\quad f_3 - F_3\hat{x}_3 - \Gamma_2\lambda + \left(\dot{\Omega}_2 - \Omega_2\Omega_1^{-1}\dot{\Omega}_1\right)\hat{x}_2 \\ 0 &= \Omega_1^{-1}\hat{f}_3 + \Omega_1^{-1}\hat{F}_3(\hat{x}_3 - \Omega_2\hat{x}_2) - f_2 - F_2\hat{x}_3 - \Gamma_1\lambda - \Omega_1^{-1}\dot{\Omega}_1\hat{x}_2, \end{aligned}$$

which yield Eq. (2.23).

2.1.4 Exact Linearization of the Error Equation

We can proceed as in the index one case due to the assumption that (2.6a) is in index two DAE observer normal form (2.20). The error equation is the linear time varying system,

$$\begin{bmatrix} \dot{\lambda} \\ \dot{\tilde{x}}_3 \end{bmatrix} = \begin{bmatrix} \Omega_1\Gamma_1 & \Omega_1 F_2 \\ \Gamma_2 - \Omega_2\Gamma_1 & F_3 - \Omega_2 F_2 \end{bmatrix} \begin{bmatrix} \lambda \\ \tilde{x}_3 \end{bmatrix}. \tag{2.25}$$

To get a linear time invariant error equation we need that F_2 and F_3 have a particular structure.

Theorem 2.4 *Let $F_3(\hat{x}_1, \hat{x}_2, u) = \overline{F}_3 + \tilde{F}_3(\hat{x}_1, \hat{x}_2, u)$ where $\overline{F}_3$ is a constant matrix. If*

1. *there exists an invertible matrix Ω_1 such that $\Omega_1 F_2$ is constant,*
2. *$\overline{F}_3$ can be chosen such that the matrix pair $\{F_2, \overline{F}_3\}$ is observable and there exists a matrix F_3^* such that $\tilde{F}_3 = F_3^* F_2$,*

then the error equation (2.25) can be made time invariant and its modes can arbitrarily be placed by proper choice of Γ_1 and Γ_2 for $\Omega_2 = F_3^*$.

2.1.5 Example

To illustrate the previous discussion we use as an example the model (2.26) of a three-phase current motor [18, 78]. We show on this model how to apply index one DAE observer design if (2.6) is in index one DAE normal form. Furthermore, we show that DAE index two observer design may yield linear error dynamics even if the DAE index one observer design does not. This shows that the extension of DAE observer design to index two DAEs allows us to enlarge the class of nonlinear systems with linearizable error dynamics.

Example 2.1 (Three Phase Current Motor) The model equations are

$$\dot{x}_1 = x_2 \tag{2.26a}$$

$$\dot{x}_2 = B_1 - A_1 x_2 - A_2 x_3 \sin(x_1) + \frac{1}{2}\sin(2x_1) \tag{2.26b}$$

$$\dot{x}_3 = = u - D_1 x_3 + D_2 \cos(x_1), \tag{2.26c}$$

where $x = [x_1, x_2, x_3]^T$ is the state, u a control input, and B_1, A_1, A_2, D_1, and D_2 are constants. We will consider several outputs in the discussion that follows.

Index One Observer

For the index one DAE observer design we use the output

$$y = [x_1,\ x_2]^T. \tag{2.27}$$

It can easily be seen that (2.26) and (2.27) is in index one DAE normal form (2.16). We have

$$f_1(y,u) = \begin{bmatrix} x_2 \\ B_1 - A_1 x_2 + \frac{1}{2}\sin(2x_1) \end{bmatrix}$$

$$F_1(y,u) = \begin{bmatrix} 0 \\ -A_2 \sin(x_1) \end{bmatrix}$$

$$f_2(y,u) = u + D_2 \cos(x_1)$$

$$F_2(y,u) = -D_1.$$

To see that (2.26) has for the output (2.27) a linear time invariant error equation we need to show that F_1 and F_2 satisfy Theorem 2.2 which is the case since $\overline{F}_2 = -D_1$ is a constant scalar.

This result is consistent with [18] where it is shown that (2.26) admits a linear time invariant error equation for the output (2.27).

For

$$\Gamma_1 = \begin{bmatrix} k_1 & 0 \\ 0 & -\frac{k_2}{k_3} A_2 \sin(x_1) \end{bmatrix},\ \Gamma_2 = \begin{bmatrix} k_4\ k_5 \end{bmatrix},\ \Omega_2 = \begin{bmatrix} 0\ 0 \end{bmatrix},$$

and (2.27) we obtain the DAE index one observer

$$\dot{\hat{x}}_1 = \hat{x}_2 - \dot{\lambda}_1 + k_1 \lambda_1$$

$$\dot{\hat{x}}_2 = B_1 - A_1 \hat{x}_2 - A_2 \hat{x}_3 \sin(\hat{x}_1) + \frac{1}{2}\sin(2\hat{x}_1) + A_2 \sin(\hat{x}_1)\frac{1}{k_3}(\dot{\lambda}_2 - k_2 \lambda_2)$$

$$\dot{\hat{x}}_3 = u - D_1\hat{x}_3 + D_2\cos(\hat{x}_1) + k_4\lambda_1 + k_5\lambda_2$$
$$y_1 = \hat{x}_1$$
$$y_2 = \hat{x}_2,$$

which has the linear time invariant error equation

$$\begin{bmatrix} \dot{\lambda}_1 \\ \dot{\lambda}_2 \\ \dot{\tilde{x}}_3 \end{bmatrix} = \begin{bmatrix} k_1 & 0 & 0 \\ 0 & k_2 & k_3 \\ k_4 & k_5 & -D_3 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \tilde{x}_3 \end{bmatrix}.$$

Now, instead of two observations, we take just one,

$$y = x_1. \tag{2.28}$$

It can easily be seen that for this observation, system (2.26) is no longer in index one DAE observer normal form (2.16). More importantly, the Lie bracket conditions of nonlinear observer design of [18, 59], or [58] for the single output case, applied to (2.26) and (2.28) show that (2.26) and (2.28) cannot be transformed into nonlinear observer form and, consequently, that it has no linear error equation for the output (2.28).

Index Two Observer

If we use DAE index two observer design for (2.26) with the output (2.28), we obtain a linear time invariant observer error equation. In fact, (2.26) is in index two DAE observer normal form (2.21), where

$$f_2(y, \dot{y}, u) = B_1 - A_1x_2 + \frac{1}{2}\sin(2x_1) \tag{2.29a}$$
$$F_2(y, \dot{y}, u) = -A_2\sin(x_1) \tag{2.29b}$$
$$f_3(y, \dot{y}, u) = D_2\sin(x_1) + u \tag{2.29c}$$
$$F_3(y, \dot{y}, u) = -D_3. \tag{2.29d}$$

Furthermore, F_3 and F_2 satisfy Theorem 2.4. Since we have $F_3 = \overline{F}_3$, the choice

$$\Omega_1 = -\frac{\kappa_2}{A_2\sin(x_1)} \tag{2.30}$$

yields $\Omega_1 F_2 = \kappa_2$ and the pair $\{F_2, \overline{F}_3\}$ is observable as $\overline{F}_3 = D_1$ is constant. For $\Gamma_1 = -\kappa_1 A_2 \sin(x_1)/\kappa_2$, $\Gamma_2 = \kappa_3$, and $\Omega_2 = 0$, the DAE index two observer is

$$\begin{aligned}
\dot{\hat{x}}_1 &= \hat{x}_2 \\
\dot{\hat{x}}_2 &= B_1 - A_1\hat{x}_2 - A_2\hat{x}_3\sin(\hat{x}_1) + \frac{1}{2}\sin(2\hat{x}_1) + A_2\sin(x_2)\frac{1}{\kappa_2}(\dot{\lambda} - \kappa_1\lambda) \\
\dot{\hat{x}}_3 &= u - D_1\hat{x}_3 + D_2\cos(\hat{x}_1) + \kappa_3\lambda \\
y &= \hat{x}_1,
\end{aligned}$$

which is an index two DAE provided $\hat{x}_1 \neq \pi k$, $k = 0, \pm 1, \pm 2, \cdots$. The error equation is

$$\begin{bmatrix} \dot{\lambda} \\ \dot{\tilde{x}}_3 \end{bmatrix} = \begin{bmatrix} \kappa_1 & \kappa_2 \\ \kappa_3 & -D_1 \end{bmatrix} \begin{bmatrix} \lambda \\ \tilde{x}_3 \end{bmatrix}.$$

To insure numerical integrability we transform the index two DAE into Hessenberg form, that is we need that the observer (2.7) be in the special form (2.22). We have for the observer (2.7)

$$0 = \begin{bmatrix} \dot{\omega}_1 - \omega_2 \\ \dot{\omega}_3 - \hat{f}_2(\omega_1, \omega_2, \omega_4, u) - \hat{F}_2(\omega_1, \omega_2, u)\omega_3 \\ \dot{\omega}_4 - \hat{f}_3(\omega_1, \omega_2, \omega_4, u) - \hat{F}_3(\omega_1, \omega_2, u)\omega_3 \\ y - \omega_1 \end{bmatrix},$$

where $[\hat{x}_1, \hat{x}_2, \hat{x}_3] = [\omega_1, \omega_2, \omega_3]$, and

$$\hat{f}_2 = D_2\sin(\omega_1) + u + \kappa_3\lambda \tag{2.31a}$$

$$\hat{f}_3 = \frac{-\kappa_2}{A_2\sin(\omega_1)}\left(B_1 - A_1\omega_2 + \frac{1}{2}\sin(2\omega_1) + \frac{\kappa\omega_2^2\cos\omega_1}{A_2\sin^2(\omega_1)}\right) + \kappa_1\lambda \tag{2.31b}$$

$$\hat{F}_2 = -A_2\sin(\omega_1) \tag{2.31c}$$

$$\hat{F}_3 = \kappa_2 \tag{2.31d}$$

$$\lambda = \omega_4 + \kappa_2\omega_1/(A_2\sin(\omega_1)). \tag{2.31e}$$

The resulting DAE is in Hessenberg semi-explicit index two form.

Numerical Simulation

For the following computer simulation we use the same values for the model parameters as in [18, 78]. That is, $A_1 = 0.2703$, $A_2 = 12.01$, $B_1 = 39.19$, $B_2 = -48.04$, $D_1 = 0.3222$, and $D_2 = 1.9$. The index two DAE observer includes the effect of numerical differentiation. To show the impact of a perturbation in the observation, we have perturbed the constraint by $v(t) = 0.001\cos(10t)$, i.e., the perturbed constraint is $0 = \hat{x}_1 - y + v$. The perturbation on $\hat{x}_1$ is obviously $v(t)$ and that on $\hat{x}_2$ is $\dot{v}(t) = -0.01\cos(10t)$. If the observation is not perturbed $\hat{x}_2$ jumps immediately to its true value x_2. If $\hat{x}_1(0) \neq y(0)$, then observation $\hat{x}_2$ is impulsive at $t = 0$.

Some of the simulation results are shown in the next four figures. Figure 1 shows x_3 and the estimate $\hat{x}_3$ while Fig. 2 gives the estimation error. Without any perturbation, Fig. 1 shows the estimate $\hat{x}_3$ (dotted line) converges to the true value x_3 (solid line). If the observation is perturbed, then $\hat{x}_3$ converges to a neighborhood of x_3 as shown by the solid line in Fig. 2.

λ is a measurement of the violation of the hidden constraint (2.32b) and the observation error $\tilde{x}_3$. Without perturbation λ goes to zero for growing time as shown

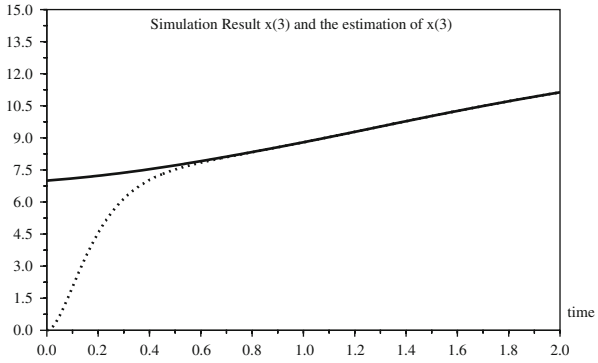

Fig. 1 Estimate $\hat{x}_3$ (*dotted line*) and the true value x_3 (*solid line*)

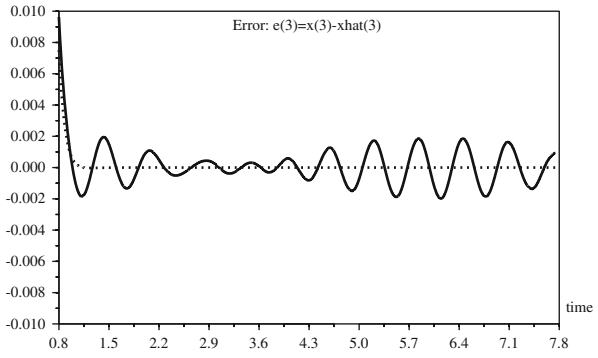

Fig. 2 Estimation error on x_3 with (*solid line*) and without perturbation (*dotted line*)

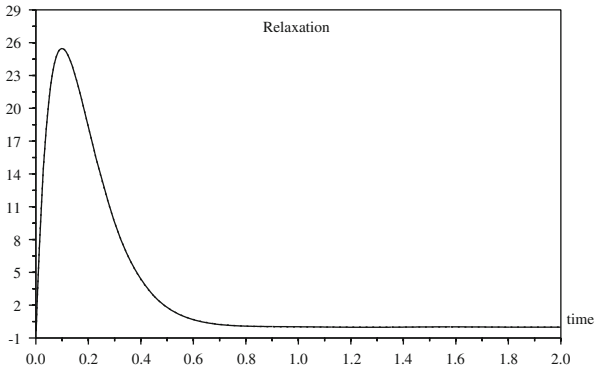

Fig. 3 Additional state λ with no perturbation

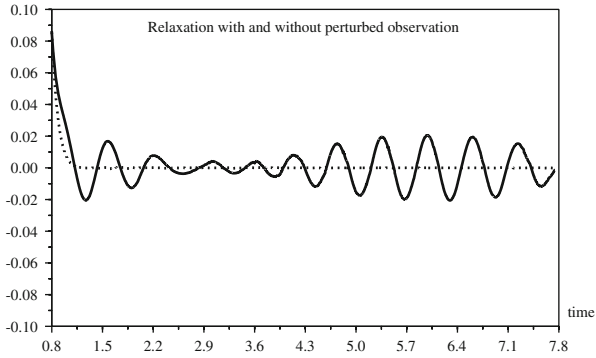

Fig. 4 Additional state λ with and without perturbation

in Fig. 3. However, if the observation is perturbed, λ stays in the neighborhood of zero as shown in Fig. 4.

The simulation shows that the additional state λ in the observer is a relaxation for hidden constraints. The constraint $0 = y - \hat{x}_1$ includes two hidden constraints:

$$0 = \dot{y} - \hat{x}_2 \tag{2.32a}$$

$$0 = \ddot{y} - \hat{f}_2(\hat{x}_1, \hat{x}_2, \lambda), \tag{2.32b}$$

where $\hat{f}_2(\hat{x}_1, \hat{x}_2, \lambda) = B_1 - A_1\hat{x}_2 - A_2\hat{x}_3 \sin(\hat{x}_1) + \frac{1}{2}\sin(2\hat{x}_1) + \Omega_1^{-1}\dot{\lambda} + \Gamma_1\lambda$. The integration by BDF integration schemes assures that the hidden constraint (2.32a) is verified for all $t > 0$. The hidden constraint (2.32b) is relaxed by the supplementary state λ. The constraint error $\ddot{y} - \hat{f}_2$ goes asymptotically to zero if $\lambda \to 0$.

2.2 *Estimation of Disturbances*

A different use of the DAE formulation in designing observers can be found in [44], note also [57]. Here the utilization of a DAE formulation assists in estimating some

disturbances. In Sect. 2.2.1 we will briefly summarize the idea of Gao and Wang [44] when the observer is a DAE but the system being estimated is an ODE. Then in Sect. 2.2.2 we will present an extension to the case where the original model is also a DAE.

2.2.1 Original Model Is an ODE

The original idea of [44] has been extended by Gao and others to cover a variety of types of disturbances. It suffices here to present part of the original idea. Both actuator and sensor faults can be handled by this approach given appropriate assumptions but we shall consider just sensor faults here.

The starting point is the usual linear time invariant model.

$$\dot{x} = Ax + Bu \tag{2.33a}$$

$$y = Cx + \omega, \tag{2.33b}$$

which is a special case of (2.3). As usual, u is the input or known control. y is the output and ω is the sensor noise.

The usual Luenberger observer (2.4) for estimating x has error dynamics

$$\dot{e} = (A - KC)e - K\omega. \tag{2.34}$$

If $\{A, C\}$ is detectable, then K can be found so that $A - KC$ has eigenvalues with negative real part. In the absence of ω, then e goes to zero as t goes to infinity. That is, the error in the estimate goes to zero.

Since y, u are considered known, we can try to set up an observer for both x and ω. Adding the equation

$$x_\omega = \omega \tag{2.35}$$

to the ODE system (2.33) we get the DAE with output

$$\bar{E}\dot{\bar{x}} = \bar{A}\bar{x} + \bar{B}u + \bar{N}\omega \tag{2.36a}$$

$$y = \bar{C}\bar{x}, \tag{2.36b}$$

where

$$\bar{x} = \begin{bmatrix} x \\ x_\omega \end{bmatrix}, \ \bar{B} = \begin{bmatrix} B \\ 0 \end{bmatrix}, \ \bar{N} = \begin{bmatrix} 0 \\ I \end{bmatrix},$$

$$\bar{E} = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \ \bar{A} = \begin{bmatrix} A & 0 \\ 0 & -I \end{bmatrix}, \ \bar{C} = \begin{bmatrix} C & I \end{bmatrix}, \ \bar{C}_0 = \begin{bmatrix} C & 0 \end{bmatrix}.$$

The Luenberger observer is based on using just a matrix multiple of the difference between the observed output y and the estimated output $\hat{y}$. That is, it uses proportionality. In the case being considered here we wish to integrate some of the terms so we get a PD (proportional derivative) observer. Thus the estimate of $\hat{x}$ is made up both of a term that depends on y and a term that is the solution of differential equation. This extra term is the η in (2.37).

Then in [44] it is shown that given (2.33), suppose that A is a stable matrix. Then there are matrices $\bar{K}, \bar{L}$ so that the observer

$$(\bar{E} + \bar{L}\bar{C})\dot{\zeta} = (\bar{A} - \bar{N}\bar{C}_0 - \bar{K}\bar{C})\zeta + \bar{B}u \tag{2.37a}$$

$$\hat{\bar{x}} = \zeta + (\bar{E} + \bar{L}\bar{C})^{-1}\bar{L}y \tag{2.37b}$$

provides an asymptotic estimate of $\bar{x}$. Note that $\hat{\bar{x}}$ is estimating both x and the disturbance ω.

If A is not stable, then one can still construct the observer with an additional assumption. Suppose that A, C is detectable and the noise ω is bounded. Then there are matrices $\bar{K}, \bar{L}$ so that the observer

$$(\bar{E} + \bar{L}\bar{C})\dot{\zeta} = (\bar{A} - \bar{K}\bar{C})\zeta + \bar{B}u + \bar{A}(\bar{E} + \bar{L}\bar{C})^{-1}\bar{L}y \tag{2.38a}$$

$$\hat{\bar{x}} = \zeta + (\bar{E} + \bar{L}\bar{C})^{-1}\bar{L}y \tag{2.38b}$$

provides an asymptotic estimate of $\bar{x}$. In (2.37) and (2.38) the $\bar{L}$ is to make $\bar{E} + \bar{L}\bar{C}$ invertible. The stabilization is then done by $\bar{K}$. Algorithms for computing $\bar{L}$ and $\bar{K}$ are given in [44].

2.2.2 Extension to When Original Model Is a DAE

The previous idea of expanding the observer to a larger DAE can be exploited in several ways. Here we consider the case when the original system is itself modeled by a DAE rather than an ODE. This section is from [87] and uses the idea of a completion of a DAE. A completion of a DAE is an ODE that includes the solutions of the DAE. Completions can be computed numerically for many DAEs of interest. At the same time that the completion is computed, a set of constraint equations characterizing the solution manifold can also be computed. Completions and their computation are studied in [23–26, 62, 79, 80] among other places. Completions are used extensively in [24].

Consider the descriptor system

$$E\dot{x} = Fx + Bu + D_1 f \tag{2.39a}$$

$$y = Hx + D_2 g, \tag{2.39b}$$

where $x \in \mathbb{R}^n$ is the state vector, $u \in \mathbb{R}^m$ is the control, $y \in \mathbb{R}^p$ is the measurement output vector, $g \in \mathbb{R}^p$ is a sensor fault, and $f \in \mathbb{R}^q$ is a process fault. To formally get the completion we form the derivative array by differentiating (2.39a) k times. Formally this gives us the system

$$\mathcal{E}\breve{\dot{x}} = \mathcal{F}\breve{x} + \mathcal{B}\breve{u} + \mathcal{D}_1\breve{f}, \tag{2.40}$$

where $\breve{z}$ indicates z and several of its derivatives. The key assumption is that there is a value of k so that $\begin{bmatrix}\mathcal{E} & \mathcal{F}\end{bmatrix}$ has full row rank and the first n columns of $\mathcal{E}$ are linearly independent of the rest of the columns. This is called 1-fullness in the literature. If the system (2.39) is linear time varying, then these assumptions are to hold for all t. The smallest value of k for which these assumptions hold is called the differentiation index of the DAE. From the derivative array (2.40) there are a number of completions that can be computed. For example, one could solve (2.40) for $\breve{\dot{x}}$ in the least squares sense and use the first n components for x'. This is called the least squares completion.

However, no matter which way you do it you get a system

$$\dot{x} = \hat{A}x + \hat{B}\bar{u} + \hat{G}\breve{f} \tag{2.41a}$$

$$0 = Gx + \tilde{B}\bar{u} + \tilde{G}\breve{f} \tag{2.41b}$$

$$y = Hx + D_2\, g. \tag{2.41c}$$

Equation (2.41b) characterizes all solutions of the DAE (2.39a). That is, it characterizes the solution manifold and provides all constraints. Equation (2.41a) is an ODE whose solutions include all the solutions of the DAE. The extra solutions of (2.41a) are called the extra dynamics. The extra dynamics vary depending on how the completion is computed. See [24] for example or [25].

For a simple purely algebraic system $x = f$, (2.41a) and (2.41b) become just

$$\dot{x} = \dot{f} \tag{2.42a}$$

$$0 = x - f. \tag{2.42b}$$

Gao and Wang [44] considered (2.41a) and (2.41c) and estimated constant additive faults, using a modified proportional, derivative, and integral (PID) observer.

However, faults are often not constant but can be generated by another system. The problem considered here is estimating a time varying fault $\dot{f} = Mf$, where $M \in \mathbb{R}^{q\times q}$. Then $\breve{f} = \begin{bmatrix} I & M & (M^T)^2 & \dots & (M^T)^k \end{bmatrix}^T f$, where k is the index of (2.39). Letting $\bar{\bar{G}} = \hat{G}\begin{bmatrix} I & M^T & (M^T)^2 & \dots & (M^T)^k \end{bmatrix}^T$, the system under consideration is thus

$$\dot{x} = \hat{A}x + \hat{B}\bar{u} + \bar{\bar{G}}f \tag{2.43a}$$

$$y = Hx + D_2\, g, \tag{2.43b}$$

if (2.41b) is ignored. Note that we are not saying the original system is an ODE. We are working with an ODE that has been computed from the DAE with some of the ODEs solutions from the DAE and some from our construction. The behavior of the extra dynamics must be taken into account. For this discussion, we also assume g is a constant fault. An augmented descriptor system

$$\bar{E}\dot{\bar{x}} = \bar{A}\bar{x} + \bar{B}u + \bar{N}g + \bar{G}f \tag{2.44a}$$

$$y = \bar{H}\bar{x} \tag{2.44b}$$

can be formed if we denote

$$\bar{x} = \begin{bmatrix} x \\ g \end{bmatrix},\ \bar{H} = \begin{bmatrix} H\ I_p \end{bmatrix},\ \bar{H}_0 = \begin{bmatrix} H\ 0 \end{bmatrix},\ \bar{G} = \begin{bmatrix} \bar{\bar{G}} \\ 0 \end{bmatrix},\ \bar{B} = \begin{bmatrix} \hat{B} \\ 0 \end{bmatrix}$$

$$\bar{N} = \begin{bmatrix} 0 \\ I_p \end{bmatrix},\ \bar{E} = \begin{bmatrix} I_n & 0 \\ 0 & 0 \end{bmatrix},\ \bar{A} = \begin{bmatrix} \hat{A} & 0 \\ 0 & -I_p \end{bmatrix}.$$

Modified PID Observer

We will include a full proof of the material in this section.

Corollary 2.5 *If the pair* $\{\hat{A}, H\}$ *is detectable,* $\operatorname{rank}\begin{bmatrix} -\bar{G} \\ sI_q - M \end{bmatrix} = q$, *and* $\operatorname{rank}\begin{bmatrix} \bar{A} & \bar{G} \\ 0 & M \\ \bar{H} & 0 \end{bmatrix} = n + p + q$, *and* g *is bounded, then there exist gain matrices* $\bar{L}, \bar{K}$ *and* K_1 *for the following observer*

$$(\bar{E} + \bar{L}\bar{H})\dot{\bar{\xi}} = (\bar{A} - \bar{K}\bar{H})\xi + \bar{B}u + \bar{A}(\bar{E} + \bar{L}\bar{H})^{-1}\bar{L}y + \bar{G}f \tag{2.45a}$$

$$\dot{\hat{f}} = -K_1\bar{H}\xi + M\hat{f} \tag{2.45b}$$

such that $\hat{x} = \xi + (\bar{E} + \bar{L}\bar{H})^{-1}\bar{L}y$ *is an asymptotic estimate of* $\bar{x}$ *and* $\hat{f}$ *is an asymptotic estimate of* f.

Proof Let $\bar{e} = \bar{x} - \bar{\hat{x}}$ and $\dot{e}_d = f - \hat{f}$. The proof follows along the same lines as in the ODE case, but (32) in [44] becomes

$$\dot{e}_d = -K_1\bar{H}\bar{e} + Me_d \tag{2.46}$$

so that (33) in [44] is

$$\begin{bmatrix} \bar{E} + \bar{L}\bar{H} & 0 \\ 0 & I_q \end{bmatrix} \begin{bmatrix} \dot{\bar{e}} \\ \dot{e}_d \end{bmatrix} = \left(\begin{bmatrix} \bar{A} & \bar{G} \\ 0 & M \end{bmatrix} - \begin{bmatrix} \bar{K} \\ K_I \end{bmatrix} \begin{bmatrix} \bar{H} & 0 \end{bmatrix} \right) \begin{bmatrix} \bar{e} \\ e_d \end{bmatrix} + \begin{bmatrix} \bar{N} \\ 0 \end{bmatrix} g \tag{2.47a}$$

or equivalently,

$$E_L \dot{\tilde{e}} = \left(\tilde{A} - \tilde{K}\tilde{H} \right) \tilde{e} + \begin{bmatrix} -L_1 \\ I + HL_1 \\ 0 \end{bmatrix} (L_2)^{-1} g. \tag{2.48}$$

As shown in [44], a $\bar{L}$ exists to make $\bar{E} + \bar{L}\bar{H}$ invertible and to reduce the amplification of the bounded fault g. A stabilizing $\tilde{K}$ exists if for all $s \in \mathcal{C}_+$,

$$n + p + q = \operatorname{rank} \begin{bmatrix} sI - (\tilde{E}_L)^{-1}\tilde{A} \\ \tilde{H} \end{bmatrix} \tag{2.49a}$$

$$= \operatorname{rank} \begin{bmatrix} s(\bar{E} + \bar{L}\bar{H}) - \bar{A} & -\bar{G} \\ 0 & sI_q - G \\ \bar{H} & 0 \end{bmatrix} \tag{2.49b}$$

which is equivalent to the following conditions:

1. For all $\operatorname{Re}(s) \geq 0$, $s \neq 0$,

$$\operatorname{rank} \begin{bmatrix} s(\bar{E} + \bar{L}\bar{H}) - \bar{A} \\ \bar{H} \end{bmatrix} = n + p \tag{2.50a}$$

$$\operatorname{rank} \begin{bmatrix} -\bar{G} \\ sI_q - M \end{bmatrix} = q. \tag{2.50b}$$

2. For $s = 0$,

$$\operatorname{rank} \begin{bmatrix} \bar{A} & \bar{G} \\ 0 & M \\ \bar{H} & 0 \end{bmatrix} = n + p + q. \tag{2.51}$$

Note that (2.50a) holds if $(\hat{A}, H)$ is detectable.

The eigenvalues of $\hat{A}$ are the finite eigenvalues of (2.39) and the additional dynamics generated by the completion. The additional eigenvalues generated by the stabilized least squares completion are $-\lambda$ of multiplicity k if the differential operator used for the derivative array is $\mathcal{D} = d/dt + \lambda$, for real $\lambda > 0$, and k is the index of the DAE [79].

If the finite eigenvalues of (2.39) are observable or stable, then $\{\hat{A}, H\}$ can always be made detectable when $\hat{A}$ is calculated using the stabilized least squares completion [20, 79].

A major disadvantage of using the differential operator $\mathcal{D} = d/dt + \lambda$ is limited observability. Specifically, the additional eigenvalues will be repeated k times; thus, they may have unobservable subspaces [65]. One way to overcome this is to choose the differential operator to be $\mathcal{D} = d/dt + \Delta$, where $Re(s) > 0$ for all eigenvalues s of Δ. The advantages to this approach is that if the DAE is index one, then there exists a Δ stabilized completion with distinct stable eigenvalues [25].

Suppose the DAE is in Hessenberg index two form,

$$J\dot{x}_1 = Ax_1 + Bx_2 \tag{2.52a}$$

$$0 = Cx_1, \tag{2.52b}$$

where $CJ^{-1}B$ is invertible, B is full column rank, and C is full row rank. Let Δ be a diagonal matrix with positive entries on the diagonal. Then the least squares completion has stabilized dynamics, but the eigenvalues may not be distinct. Analogous ideas relating the eigenvalues of Δ to those of the completion do not hold for Hessenberg index three DAEs [25].

The second assumption implies that $\hat{G}v \neq 0$ for any eigenvector v of M with corresponding eigenvalue $Re(s) \geq 0$. For if $\hat{G}v = 0$, then

$$\begin{bmatrix} -\bar{G} \\ sI_q - M \end{bmatrix} v = \begin{bmatrix} & I & \\ -\hat{G} & \begin{bmatrix} M \\ \vdots \\ M^k \end{bmatrix} & \\ & 0 & \\ & sI_q - M & \end{bmatrix} v = 0.$$

But this is a contradiction because, by assumption, $\begin{bmatrix} -\bar{G} \\ sI_q - M \end{bmatrix}$ has full column rank. Recall the time varying fault obeys $\dot{f} = Mf$. Hence, the fault is of the form

$$f(t) = \sum c_i v_i e^{s_i t}, \tag{2.53}$$

where (v_i, s_i) are eigenvector/eigenvalue pairs and c_i are scalars. Repeated eigenvalues may create terms with powers of t in them. $\hat{G}v \neq 0$ ensures that faults of this form will be visible in the completion as they will not lie in the null space of $\hat{G}$.

If M is invertible, then the third assumption implies

$$n + p = \operatorname{rank} \begin{bmatrix} \bar{A} \\ \bar{H} \end{bmatrix} = \operatorname{rank} \begin{bmatrix} \hat{A} & 0 \\ 0 & -I_p \\ H & I_p \end{bmatrix}.$$

The second column of this block matrix has rank p, so the assumption simplifies to rank $\begin{bmatrix} \hat{A} \\ H \end{bmatrix} = n$. If $n = q$, note that this is equivalent to $\{\hat{A}, H\}$ is an observable pair.

3 Optimization by Direct Transcription

Suppose that we are given a DAE with a control u of the form

$$F(\dot{x}, x, t, u) = 0. \tag{3.1}$$

If we have a given choice of u and wish to perform a simulation, then the index of the DAE plays its usual role in the theory of numerical DAE integrators. For example, for the linear system

$$A\dot{x} + Bx = Du, \tag{3.2}$$

if the index of the matrix pencil $\{A, B\}$ is greater than 1, then differentiation of u may be present depending on D. In general we have to deal with the usual numerical issues of integrating DAEs [22].

Suppose now that we have an optimal control problem

$$\min \mathcal{L}(x, u) \tag{3.3a}$$

$$F(\dot{x}, x, u, t) = 0 \tag{3.3b}$$

$$g(x, u, t) = 0. \tag{3.3c}$$

Here (3.3b) are the dynamics. Equation (3.3c) are any constraints on the time interval, and (3.3a) is some cost functional of x, u and possibly some parameters which is to be optimized. There may be additional inequality constraints not shown in (3.3) which we will discuss later in Sect. 3.2. Various initial and terminal equalities and inequalities may also hold.

One way to try to solve such problems is to parametrize the set of u, run integrations of the DAE given by (3.3b), (3.3c) to evaluate the cost (3.3a) for a given control u and parameter values and then feed this as a cost function to a general purpose optimizer. This is sometimes referred to as a control parameterization technique. This method can often be successful. However, it does mean that the usual DAE integration theory must be used and also that working with (3.3c) can be problematic if in addition there are inequality constraints since inequality constraints may go active and inactive causing changes in the DAE being integrated. Satisfying terminal conditions can also be an issue.

One popular alternative approach for solving industrial grade optimal control problems is direct transcription. The philosophy of direct transcription is completely different. One starts on a coarse time grid and discretizes the entire problem. This

nonlinear programming problem (NLP) is then fed to an optimization package. The solution is then evaluated and if necessary the grid is refined and the problem is solved again using the previous solution as an initial guess. The grid refinement is sometimes done by a process that tries to focus on where the computational difficulty is. The result is usually highly nonuniform grids. Some direct transcription softwares use uniform grids.

The necessary conditions of the original optimal control problem are never formulated and the approach has been very successful in solving many hard problems. Note that these direct transcription programs often do not actually worry too much about verifying optimality. They do worry about making sure the dynamics are integrated well. Thus the user can be sure that if the computed control is used, and the models are good, then the computed control will produce close to the computed cost.

It turns out that when a direct transcription approach is used that the usual DAE theory does not always apply [8, 31] and the software is able to solve some problems that at first glance should not be solvable. In this section we will discuss why that is the case and give examples in Sects. 3.1.1, 3.1.2, and 3.2.

Two popular examples of direct transcription codes that can work with some DAEs are GPOPS II and SOCX.

There are several optimal control problem solvers currently available for use. Pseudospectral Optimal Control Solver ($\mathcal{PSOPT}$) is an open source optimal control package written in C++ that uses direct collocation methods, which include pseudospectral as well as local discretizations [3]. General Pseudospectral OPtimal Control Software (GPOPS) is an open source MATLAB based optimal control software that implements the Gauss and Radau hp-adaptive pseudospectral methods [85]. The latest GPOPS II requires a license. When using either of these optimal control packages, one must also incorporate NLP solvers such as Sparse Nonlinear OPTimizer (SNOPT) [45] or Interior Point Optimizer (IPOPT) [98]. As noted, both software packages employ pseudospectral schemes which solve optimal control problems by approximating the time-dependent variables using orthogonal polynomials. A basic pseudospectral method is typically employed as a p-method where a single segment is used, and convergence is achieved by increasing the degree p of the polynomial [34]. In nice enough situations these methods display fast convergence in the states, controls, and costates. However, when solving problems with rapidly changing solutions, applying a very large-degree polynomial may not even guarantee a respectable solution. hp methods use several intervals with a polynomial on each subinterval.

Another direct transcription package is SOCX (Sparse Optimal Control Extended), a general purpose industrial grade software package capable of solving optimal control problems with both state and control delays and state and control constraints and is available from Applied Mathematical Analysis. SOCX is part of the Sparse Optimization Software (SOS) and is written in FORTRAN. SOCX can be used also with problems that have delays. This is discussed later in Sect. 4. As described later SOCX uses either the trapezoid or the Hermite-Simpson Runge-Kutta methods for discretization. It has available either a sequential

quadratic program (SQP) general optimizer or an interior point optimizer to solve the nonlinear programming problem formed by the discretization.

3.1 Virtual Index

For the remainder of this section we shall assume that the dynamics and path constraints take the form of

$$\dot{x}_1 = f_1(x_1, x_2, u, t) \tag{3.4a}$$

$$0 = f_2(x_1, x_2, u, t). \tag{3.4b}$$

Thus (3.4b) includes both any algebraic constraints in a DAE model and also any constraints for the optimization. x_1 is often called a differential variable since it is differentiated. The variables x_2, u are called algebraic variables since they only occur algebraically.

It does not make sense to talk of the index of (3.4) without additional information about how (3.4) is being used and interpreted. For example, in a simulation, or if using control parameterization, then u becomes known and we have a DAE in terms of x_1, x_2.

However, there are situations where u is not known such as when solving optimal control problems by direct transcription. Then the DAE is in terms of $\{x_1, x_2, u\}$.

There has been considerable discussion of this situation in terms of trying to determine a good control. This research is part of what is known as the behavioral approach. That is, the behavior is the set of all $\{x_1, x_2, u\}$ that satisfy (3.4). Properties like index are then determined after the choice of the control. The behavioral approach was promoted by Willems [70, 83] and others. We note [50] for a more DAE oriented behavioral approach. Generally these works seek to make choices of state and control that produce an index one system. For a nonlinear problem this may be difficult.

We shall do something a little different. We shall give two illustrations in Sects. 3.1.1 and 3.1.2 where these ideas are exploited but not necessarily by the user in the obvious way.

We shall say that (3.4) has virtual index one, if there is a differentiable and invertible transformation of $\{x_2, u\}$ into $\{\hat{x}_2, \hat{u}\}$ so that the DAE is index one in terms of $x_1, \hat{x}_2$ [41]. See also [27]. The variables $\hat{u}$ are called the virtual control. The variables $\{\hat{x}_2, \hat{u}\}$ may be in a different coordinate system than $\{x_2, u\}$ is written in. However, in many cases, including Sects. 3.1.1 and 3.1.2, the new coordinate system is just a reordering of the old one. The existence of this different coordinate system has major consequences. In general the concepts of differential index and tractability index [63] are different. However in our particular context, for a system having virtual index one, the two types of index are the same. Note that there are more general definitions of virtual and tractability index, see Chap. 10 of [63], but here we are talking about specific applications which are then solved by direct

transcription codes. These codes require, perhaps after a constant coordinate change, a semi-explicit formulation. Since this suffices to make our observations, we do not discuss the more general definitions here.

3.1.1 Utilized by the Software

Normally when faced with a DAE optimal control problem that does not have bounds on the controls, one wants the dynamics to be index one and for the control to appear in a nonlinear way in the cost. The exact statement is a bit more technical than this in particular in terms of what a good nonlinearity in the control is, but thinking of locally convex suffices for now.

In the solution of an optimal control problem by direct transcription, when there are not bounds on the controls or virtual controls, what is needed is that the virtual index is one and that the virtual controls appear in the cost in a good nonlinear manner. However, it is not necessary that the user see how this is done nor what the virtual controls are. What counts is only that such a choice of variables exists. This is automatically exploited in the optimization. If the control appears linearly in the cost or is absent from the cost, then control bounds are usually needed.

In optimal control problems the cost is part of the design process with some terms in the cost for performance reasons and some terms in the cost to help the numerics. Thus instead of the usual practice of making sure the control is included in the cost, when presented with higher index DAE dynamics, the correct thing to do is often to make sure that all the algebraic variables are included in the cost thereby making sure that the virtual controls, whatever they are, are included.

To illustrate how this can naturally occur consider a constrained mechanical system with a control u in the form of

$$\ddot{x} = f(x, \dot{x}, u, t) + g_x(x, u, t)^T \lambda \tag{3.5a}$$

$$0 = g(x, u, t), \tag{3.5b}$$

or equivalently,

$$\dot{x} = v \tag{3.6a}$$

$$\dot{v} = f(x, v, u, t) + g_x(x, u, t)^T \lambda \tag{3.6b}$$

$$0 = g(x, u, t). \tag{3.6c}$$

Assume that g_x is full row rank. Here (3.6c) is a state constraint that perhaps can be changed by the control u. The term $g_x(x, u, t)^T \lambda$ can be interpreted as the force that is exerted by this constraint.

If u is treated as a known function, then (3.6) is an index three DAE in the state variables x, v, λ as is often the case with constrained mechanics problems. Mechanical systems can have index higher than 3 [19]. However, in a direct

transcription setting x, v are the differential variables and u, λ are the algebraic variables. If (3.6) is a lower index DAE in x, v, λ, for some virtual control, then the direct transcription numerics will reflect this. The user can still use u as their control but in the solution of the numerical optimization, the software can automatically exploit the fact that there exists a lower index choice of control. Rigorous mathematical development of these statements can be found in [41]

A specific example from [41] is Example 3.1.

Example 3.1 (Hockey Puck Problem) A simple instance of the problem (3.6) can be stated as follows:

$$\min \quad x_1(1)^2 + x_2(1)^2 + \int_0^1 q_1 L^2 + q_2 b^2 + q_3 c^2 \, dt \tag{3.7a}$$

$$x_1' = v_1 \tag{3.7b}$$

$$x_2' = v_2 \tag{3.7c}$$

$$v_1' = -v_1 + L \tag{3.7d}$$

$$v_2' = -v_2 - bL \tag{3.7e}$$

$$0 = x_1 - bx_2 - c \tag{3.7f}$$

$$-1 = x_1(0), \quad 1 = x_2(0) \tag{3.7g}$$

$$-2 = v_1(0), \quad 2 = v_2(0). \tag{3.7h}$$

The dynamics and constraint (3.7f) can be thought of as a flat surface pushing an object at (x_1, x_2) across a smooth flat surface which has the usual type of friction model proportional to the velocity. The friction coefficient is -1 in (3.7d), (3.7e). The controls L, b, c determine the location and slope of the pushing surface. The initial conditions specify that the object starts at point $(-1, 1)$, moving away from the origin with initial velocity $(-2, 2)$. The objective is to push the object close to the origin ($x_1(1)^2 + x_2(1)^2$ penalizes not ending near the origin), while keeping the algebraic variables bounded, or small, depending on the weights q_1, q_2, q_3.

In [41] it is shown that for some classes of problems we can determine whether a given choice of virtual coordinates are properly weighted in the cost. This problem can be put in that form. If we take

$$\left[U_1 \middle| U_2 \right] = \left[\begin{array}{c|cc} 0 & 1 & 0 \\ 0 & 0 & -1 \\ -1 & 0 & x_2 \end{array}\right],$$

then the nonnegativity of

$$U_2^T \nabla^2_{L,b,c} \theta_0 U_2 = \begin{bmatrix} q_1 & 0 \\ 0 & q_2 + x_2^2 q_3 \end{bmatrix}, \tag{3.8}$$

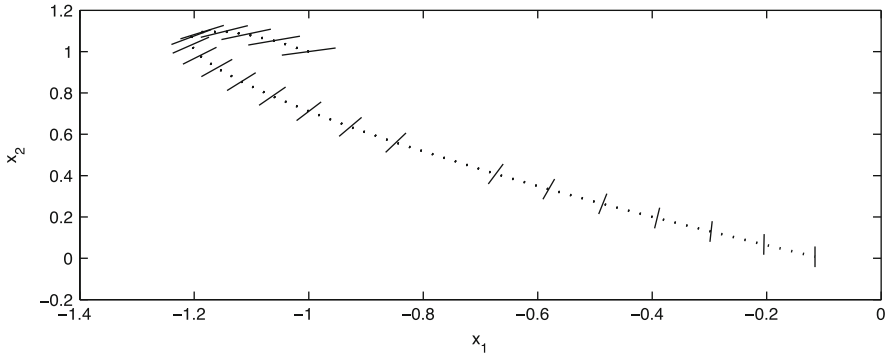

Fig. 5 Calculated trajectory of the object in $x_1 - x_2$ space with time-lapsed view of the pushing surface for Example 3.1, $q_i = 0.01$

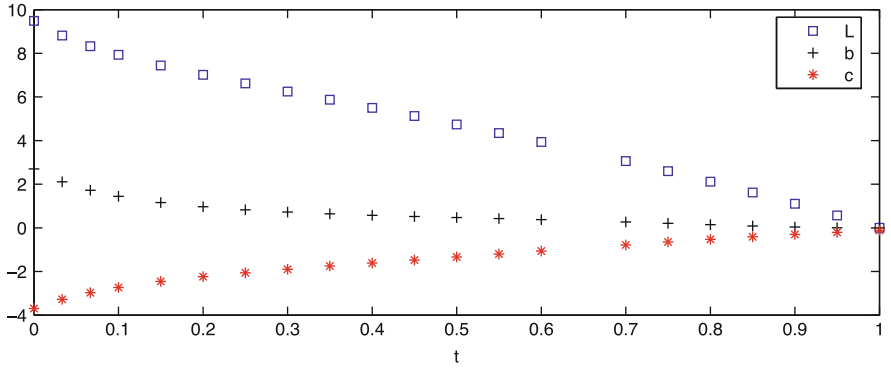

Fig. 6 L, b, c vs time for Example 3.1, $q_i = 0.01$

where θ_0 is the integrand of the cost (3.7a), determines whether we have a proper choice of virtual control.

We solve this optimal control problem using SOCX. The calculated trajectory of the object in $x_1 - x_2$ space and the values of L, b, c versus time for $q_1 = q_2 = q_3 = 0.01$ are shown in Figs. 5 and 6. This trajectory achieves $x_1 = -0.1155, x_2 = 0.0074$ at time $t = 1$, so that $\|x(1)\| = 0.0134$. Figure 5 also shows the orientation of the pushing surface at each temporal grid point. All values are rounded to four decimal places. The trajectories look the same when q_3 is changed to zero since (3.8) is still nonsingular.

As expected, SOCX fails to come up with a solution when just q_1 is changed to zero and (3.8) is always singular. When both q_2 and q_3 are set to zero, the resulting b and c trajectories oscillate wildly as shown in Figs. 7 and 8. This solution gives $x(1) = [-0.1271, 0]^T$, $\|x(1)\| = 0.0162$. If we consider getting the object close to the origin to be our main objective, and the second part of the cost function as merely there for regularization, then this is worse by 21 % than the first solution. When $q_2 = 0$ and q_1, q_3 are positive, the computed trajectories in Fig. 9 look similar

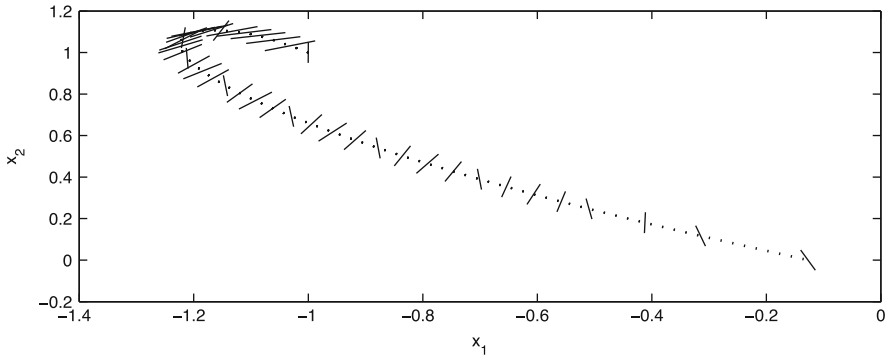

Fig. 7 Calculated trajectory of the object in $x_1 - x_2$ space for Example 3.1, $q_1 = 0.01$, $q_2 = q_3 = 0$

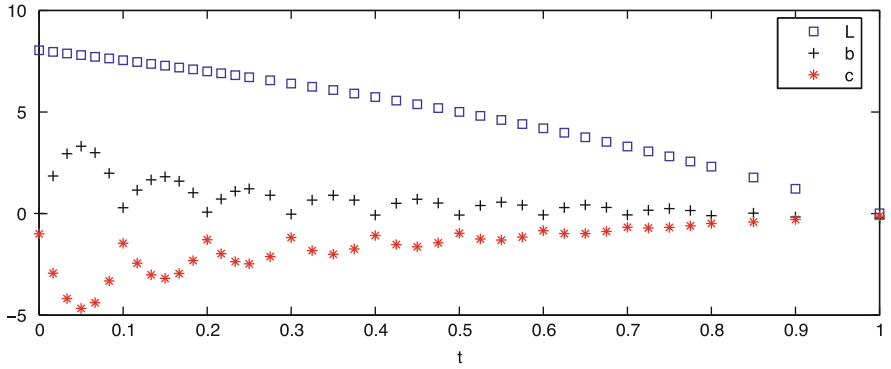

Fig. 8 L, b, c vs time for Example 3.1, $q_1 = 0.01, q_2 = q_3 = 0$

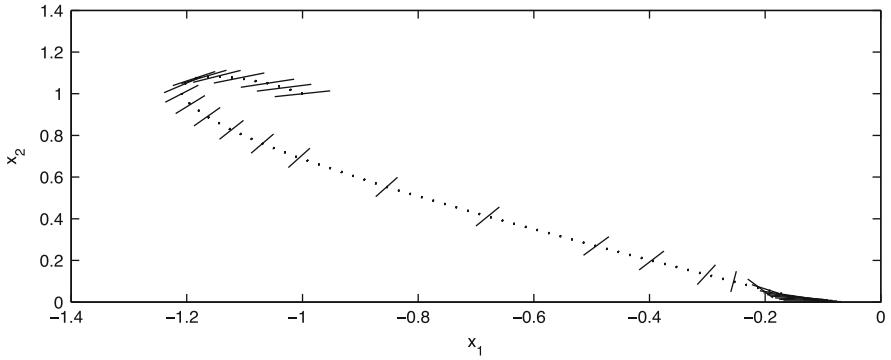

Fig. 9 Calculated trajectory of the object in $x_1 - x_2$ space for Example 3.1, $q_1 = q_3 = 0.01$, $q_2 = 0$

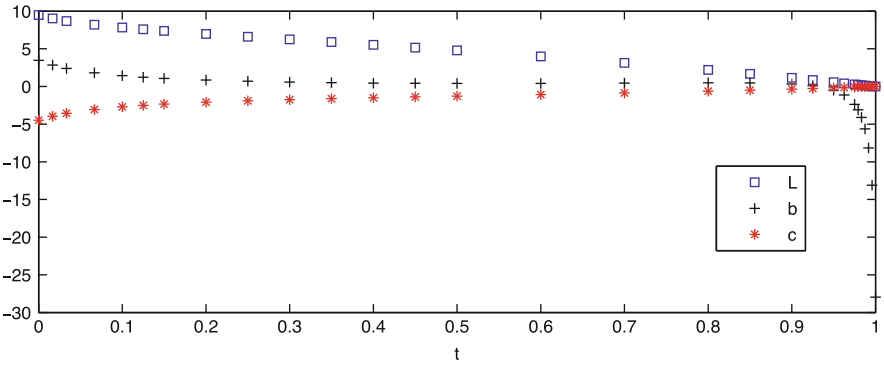

Fig. 10 L, b, c vs time for Example 3.1, $q_1 = q_3 = 0.01, q_2 = 0$

to Fig. 6 except at the last node, where b takes a sharp dive and x_2 is pushed a little closer to 0. Controls for this case are in Fig. 10. This is due to the $x_2^2 q_3$ term being near singular when x_2 is close to 0 resulting in a near singularity of (3.8). In this case $x(1) = [-0.1183, 0.0042]^T$ and $\|x(1)\| = 0.014$, which is 4 % worse than the first solution. The fact that the main objective is not satisfied as well does not necessarily mean that the solutions we obtained are suboptimal for their respective problems, but our theory allows us to conclude that the first solution is definitely optimal in the case $q_1 = q_2 = q_3 = 0.01$.

We conclude our discussion of this example by noting that other variations of this problem, such as more realistic spatially dependent friction models, curved pushing surfaces, and motor models for the actuators, naturally lead to examples with greater nonlinearity, and in the case of included actuator dynamics, higher index.

3.1.2 Utilized by the User

Our next example is one where the user has a control u they want to determine but the extra flexibility of a DAE formulation is exploited in a different manner.

Fault detection is an important part of most industrial processes and devices. Extensive work has been done on fault detection. We note only [44, 49, 52, 77, 81]. Fault detection approaches can be loosely placed into two types. One is passive and the other is active. In a passive approach, outputs from the system are monitored and used to try to tell when a fault has occurred. The system is not acted on. Usually design decisions have to make the trade off between having false positives and missing a fault. Passive approaches have been used on many applications. However, a passive approach may not be able to detect faults before they become serious. Sometimes developing faults are masked by the action of controllers or exist in subsystems that are only activated in critical conditions.

In an active approach, a test signal is used over a short time horizon to try to detect a fault if it is there, but to also not disturb the system performance any more than necessary. The goal is to detect the fault before it becomes too serious.

For both active and passive approaches the system and models may be deterministic or stochastic and continuous or discrete time.

We shall focus here on one model based active approach for guaranteed fault detection. Our discussion follows that in [88] which is an outgrowth of that begun in [29] where the models are ordinary differential equations. For our purposes here, it suffices to consider the additive uncertainty case. Model uncertainty is discussed in [88].

We assume that we have two DAE models of the form (3.9). Model $i = 0$ models the situation with no fault while $i = 1$ models the fault. We have a test horizon $[0,\ T]$. The models are both

$$E_i\dot{x}_i + F_i x_i = B_i u + M_i \mu_i \tag{3.9a}$$

$$y_i = C_i x_i + N_i \nu_i. \tag{3.9b}$$

Here μ_i, ν_i represent such things as model error, sensor error, and disturbances. The initial values $x_i(0)$ are also unknown. The coefficient matrices are assumed constant. y_i is the output from model i that is available to determine if the fault is present, that is, if model i is the correct model. States x_1 and x_0 do not have to be the same size. However, y_0 and y_1 have to have the same dimension.

The general approach is as follows. A bound is assumed on the total amount of uncertainty. Given a test signal u, if $y_0 = y_1$ implies the uncertainty bound is violated, then the signal u is called proper. That is, from a proper signal we cannot get the same output from both models. We then seek the smallest u that are proper. We assume that M_i, N_i are invertible. That is, we allow uncertainty into all of the equations. A proper u found this way will also be proper if any of the M_i, N_i are not invertible.

Unlike most of the work in this area, we do not assume that the (3.9a) are index one. We allow them to have any index but take the index at least one, otherwise the results are known [29]. There are several ways to measure the uncertainty. The uncertainty consists of $\{\mu_0, \mu_1, \nu_0, \nu_1, x_0(0), x_1(0)\}$. Here we take the uncertainty bound as

$$G(x(0), \mu, \nu) = x_0(0)^T P_0 x_0(0) + x_1(0)^T P_1 x_1(0) + \int_0^T \|\mu\|^2 + \|\nu\|^2 < 1, \tag{3.10}$$

where μ has components μ_0, μ_1 and the same holds for $x, \nu, x(0)$. Bounds other than one are treated with the same analysis. Other ways to measure the size of the test signal are discussed in [29].

For a given u, let $\phi(u)$ be the size of the smallest amount of uncertainty for which $y_0 = y_1$. That is, the inner minimization problem is

$$\phi(u) = \min_{x(0),\mu,\nu} G(x(0), \mu, \nu) \tag{3.11a}$$

$$E_0\dot{x}_0 + F_0x_0 = B_0u + M_0\mu_0 \tag{3.11b}$$

$$E_1\dot{x}_1 + F_1x_1 = B_1u + M_1\mu_1 \tag{3.11c}$$

$$0 = C_0x_0 + N_0\nu_0 - C_1x_1 - N_1\nu_1. \tag{3.11d}$$

Note that (3.11b)–(3.11d) is a DAE even if E_0 and E_1 are invertible.

If the size of u is measured by the L^2 norm, we get the final outer minimization problem is

$$\min_u \quad \int_0^T \|u\|^2 dt \tag{3.12a}$$

$$\phi(u) \geq 1. \tag{3.12b}$$

It is the constraint (3.12b) that makes the outer problem challenging. The cost of u in (3.12a) can be modified to reduce the effect of the test signal at the end of the testing period.

Direct transcription software, because they iterate on grids, often do not have the ability to call themselves in problems that involve multiple optimizations some of which occur in constraints like the one discussed here, or in function evaluations. In order to put our problem which minimizes subject to a minimization constraint in the form that can be quickly solved with optimization software we replace the inner minimization problem with equations that characterize that minimum. If the E_i are invertible, as in [29], we just solve (3.11d) for one of the noise variables and the inner problem reduces to a linear quadratic regulator (LQR) problem with known necessary conditions. So suppose that the E_i are not invertible and the models are DAEs. Because of how the noise is measured we are best restricted to orthogonal changes of coordinates. Using a singular value decomposition (SVD) of each E_i,

$$E_i = W_i \begin{bmatrix} E_{i1} & 0 \\ 0 & 0 \end{bmatrix} S_i,$$

where the E_{i1} are nonsingular, in fact positive definite, and performing coordinate changes based on the orthogonal matrices W_i, S_i for each model, we have

$$F_i = W_i \begin{bmatrix} F_{i11} & F_{i12} \\ F_{i21} & F_{i22} \end{bmatrix} S_i, \quad W_i^T M_i = \begin{bmatrix} M_{i1} \\ M_{i2} \end{bmatrix}, \quad W_i^T B_i = \begin{bmatrix} B_{i1} \\ B_{i2} \end{bmatrix},$$

$$C_i S_i^T = \begin{bmatrix} C_{i1} & C_{i2} \end{bmatrix}.$$

Then (3.11) becomes

$$\phi(u) = \min\ G(x(0), \mu, \nu) \tag{3.13a}$$

$$E_{01}\dot{x}_{01} + F_{011}x_{01} + F_{012}x_{02} = B_{01}u + M_{01}\mu_0 \tag{3.13b}$$

$$E_{11}\dot{x}_{11} + F_{111}x_{11} + F_{112}x_{12} = B_{11}u + M_{11}\mu_1 \tag{3.13c}$$

$$F_{021}x_{01} + F_{022}x_{02} = B_{02}u + M_{02}\mu_0 \tag{3.13d}$$

$$F_{111}x_{11} + F_{112}x_{12} = B_{11}u + M_{12}\mu_1 \tag{3.13e}$$

$$C_{01}x_{01} + C_{02}x_{02} + N_0\nu_0 = C_{11}x_{11} + C_{12}x_{12} + N_1\nu_1. \tag{3.13f}$$

The differential variables are x_{01}, x_{11}. The remaining variables are algebraic. Using the flexibility of the DAE formulation we want to have new control variables appear in the cost and the remaining DAE be index one since then the analysis can proceed as before. The coefficient matrix of the algebraic variables $\{x_{02}, x_{12}, \mu_0, \mu_1, \nu_0, \nu_1\}$ in the algebraic constraints is

$$\mathcal{A} = \begin{bmatrix} F_{022} & 0 & -M_{02} & 0 & 0 & 0 \\ 0 & F_{122} & 0 & -M_{112} & 0 & 0 \\ C_{02} & -C_{12} & 0 & 0 & N_0 & -N_1 \end{bmatrix}. \tag{3.14}$$

If

$$\begin{bmatrix} F_{022} & 0 \\ 0 & F_{122} \\ C_{02} & -C_{12} \end{bmatrix}$$

is full column rank and $\mathcal{A}$ is full row rank, then there is a partition of the $\{\mu_0, \mu_1, \nu_0, \nu_1\}$ which when considered part of the state gives an index one DAE and one may use the previously developed analysis. Note that by our assumptions $\mathcal{A}$ is always full row rank.

Uncertainty on the initial condition of any of the algebraic variables is a situation that can lead to numerical problems during optimization. Let

$$\tilde{P}_i = \begin{bmatrix} \hat{P}_i & 0 \\ 0 & 0 \end{bmatrix} \tag{3.15a}$$

$$\text{such that} \quad \tilde{P}_i \leq V_i^T P_{0_i} V_i \tag{3.15b}$$

$$\hat{P}_i > 0 \tag{3.15c}$$

for each model. Then we get a more conservative noise measure for each model,

$$\Gamma_i^2(x_{i1}(0), x_{i2}(0), \eta_i, \mu_i) = \frac{1}{2}x_{i1}^T(0)\hat{P}_i x_{i1}^T(0) + \frac{1}{2}\int_0^T \|\eta_i\|^2 + \|\mu_i\|^2\, dt, \tag{3.16}$$

and the total noise measure is $\Gamma_0^2 + \Gamma_1^2$. The u found using this measure will still be proper for the original measure.

To illustrate we give one example from [88].

Example 3.2 Let the normal and faulty model take the form of (3.9) where the fault under consideration occurs in E_i and F_i. The model coefficients are

$$E_0 = \begin{bmatrix} 1&2&3&0&0&0\\ 0&2&3&0&5&0\\ 0&0&3&4&0&6\\ 0&0&0&0&0&0\\ 0&0&0&0&0&0\\ 0&0&0&0&0&0 \end{bmatrix}, \quad E_1 = \begin{bmatrix} 1&2&3&0&0&0\\ 0&-2&3&0&5&0\\ 0&0&3&4&0&6\\ 0&0&0&0&0&0\\ 0&0&0&0&0&0\\ 0&0&0&0&0&0 \end{bmatrix},$$

$$F_0 = \begin{bmatrix} 0&0&0&-1&1&-1\\ 0&0&1&1&0&0\\ 0&-1&0&0&0&0\\ -1&1&0&-4&0&0\\ 1&0&0&0&-2&0\\ 1&0&0&0&0&0 \end{bmatrix}, \quad F_1 = \begin{bmatrix} 0.5&1&1.5&-1&1&-1\\ 0&-1&2.5&1&2.5&0\\ 0&-1&1.5&2&0&3\\ -1&1&0&-4&0&0\\ 1&0&0&0&-2&0\\ 1&0&0&0&0&0 \end{bmatrix}.$$

The rest of the parameters are

$$B_0 = B_1 = \begin{bmatrix} 1&1&-1&0&0&0 \end{bmatrix}^T, \quad D_0 = D_1 = \begin{bmatrix} 0\\1\\0 \end{bmatrix},$$

$$C_0 = C_1 = \begin{bmatrix} 0&0&0&0&1&0\\ 0&0&0&0&0&1\\ 0&0&0&1&0&0 \end{bmatrix},$$

$$P_{0_0} = \begin{bmatrix} 0.1116 & 0.1522 & 0.1946 & -0.0449 & -0.1777 & -0.0674\\ 0.1522 & 0.3166 & 0.4049 & -0.0934 & 0.0304 & -0.1401\\ 0.1946 & 0.4049 & 0.6652 & 0.0773 & 0.0389 & 0.1159\\ -0.0449 & -0.0934 & 0.0773 & 0.2899 & -0.0090 & 0.4348\\ -0.1777 & 0.0304 & 0.0389 & -0.0090 & 0.9645 & -0.0135\\ -0.0674 & -0.1401 & 0.1159 & 0.4348 & -0.0135 & 0.6522 \end{bmatrix},$$

$$P_{0_1} = \begin{bmatrix} 0.0813 & 0.1788 & 0.1870 & -0.0432 & -0.0407 & -0.0647\\ 0.1788 & 0.5025 & 0.2720 & -0.0628 & -0.3622 & -0.0942\\ 0.1870 & 0.2720 & 0.7563 & 0.0562 & 0.2550 & 0.0844\\ -0.0432 & -0.0628 & 0.0562 & 0.2947 & -0.0589 & 0.4421\\ -0.0407 & -0.3622 & 0.2550 & -0.0589 & 0.7021 & -0.0883\\ -0.0647 & -0.0942 & 0.0844 & 0.4421 & -0.0883 & 0.66311 \end{bmatrix},$$

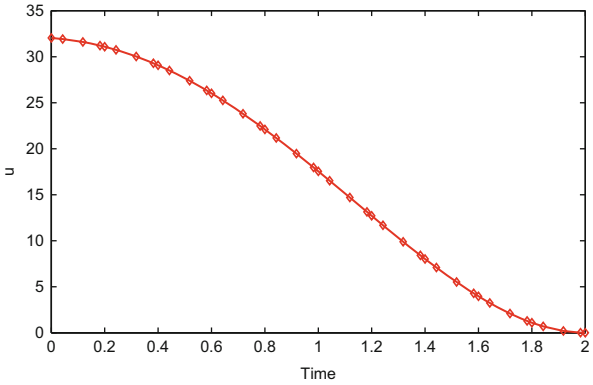

Fig. 11 Minimal proper u for the L^2 bound—Example 3.2

and $M_0 = M_1 = I_{6\times 6}$. $N_0 = N_1 = I_{3\times 3}$. Hence, model 0 is index 3 and model 1 is index 1. The noise for model i is given by (3.10) and we assume the total noise in the L^2 measure is bounded by $\gamma = 1$. With the choices for P_{0_0} and P_{0_1}, $\hat{P}_0 = \hat{P}_1 = I$ are selected to satisfy (3.15). Then, $P_0 = I$ in (3.10).

The software package chosen for implementation is GPOPS-II. Figure 11 shows the minimal proper u for this example. The test signal being zero at the end of the test interval is typical since the closer to the end of the test the less effect a signal can have. The exception is when the interval the test signal is applied on is shorter than the interval the output is observed on [90].

Even if our assumptions are not all met for particular higher index DAEs, it is sometimes possible to get a useful auxiliary signal. This is discussed in [87] and [89].

3.2 Differential Algebraic Inequalities

The next example is a case where the combination of a direct transcription approach along with a DAE formulation leads to good, but surprising results. This section is based on the work of Biehn [8].

Suppose that we are solving an optimal control problem of the form

$$\min \quad \int_0^T L(x,u,t)\,dt + \phi(x(T),T) \tag{3.17a}$$

$$\dot{x} = f(x,u,t) \tag{3.17b}$$

$$0 \le g(x,u,t) \tag{3.17c}$$

$$x(0) = x_0. \tag{3.17d}$$

Equation (3.17b) could be a DAE but that is not important for what we wish to discuss in this particular section.

It is important to distinguish the case when $g_x = 0$ from the $g_x \neq 0$ case. If g only depends on u and t, then it is called a control constraint and can often be dealt with using standard software although solutions can be bang-bang if u is not in the cost in a nice enough nonlinear manner.

However, if g depends on x and g_u is not full column rank, which includes when g depends only on x, t, then we have state inequality constraints present although the constraints may be implicit. That is the case that interests us most here. For the remainder of this section we assume that (3.17c) is a state inequality constraint. To simplify the discussion we will take $g_u = 0$ so that the state inequality constraint is explicit.

Inequality constraints such as (3.17c) often occur for safety and design or physical reasons. For example, a robot must stay within a certain workspace or a reaction must be kept below a certain temperature. If the constraint is never active, that is $g > 0$ at the optimum, then for this discussion, we have an unconstrained problem. If the constraint is not active at the minimum, but the minimum is very near where $g = 0$, then the comments made later about touch points and direct transcription apply.

If the constraint is always active, then (3.17b) and (3.17c) form a DAE which could easily be high index. If the dynamics were originally in the form $\ddot{x} = \hat{f}(x, \dot{x}, u, t)$, and g depends on just x, t, then the resulting DAE will be at least index three in x, u.

However, often the constraint will be active part of the time and inactive part of the time. In [6] there is an example with over ten state constraints that are frequently going active and inactive. If the optimal control software being used is based on integrators, numerous challenges are presented. For one, the integrator needs to figure out when the constraint is active in order to know what DAE it is integrating at any given time. This is often difficult. Even if the integrator knows when the constraint is active, the integration can involve a higher index DAE and integration can fail for stability or other reasons.

If direct transcription is being used, then the first difficulty is not so important. Feasible solutions can often still be found.

However, something else happens that is surprising. On a number of problems, where the optimal solution has subintervals where the constraints were active, the resulting DAE was one where the usual numerical DAE theory said the discretization used by the optimizer was not convergent and yet the software found a good solution.

The usual DAE numerical theory said that the error equation was unstable and thus a small error at the start would lead to a very large error later. However, a careful examination of the solutions obtained showed that in a DAE direct transcription solution, the optimizer has a choice in how close to solve the inequality. To the optimizer the instability of the error equation meant that a small perturbation at the start of the active interval could cancel out the large error later. Several examples and analysis verifying this can be found in [8, 31].

Another place where DAE theory and the numerical solution by direct transcription is different is with touch points [9, 54]. A touch point is when $g = 0$ at isolated values of t along the optimal solution. If integrating a DAE, then touch points can be a problem since the touch point means that the DAE has a singularity of some type.

However, with direct transcription the optimal solution is only asked to satisfy the inequality constraint up to the requested accuracy which is usually much larger than machine precision. We have seen examples where mathematically there were touch points in the theoretical solution and small bumps in the solution connecting the touch points in the mathematical solution, but the optimization would regularize the problem and avoid the bumps by small perturbations off the inequality. Since given the control found, the state is found to high accuracy this regularized optimal control could be used for the optimal control for all practical purposes. Also a formulation of the necessary conditions would be difficult and if formulated it would be extremely hard to solve numerically. We give one illustration from [9] where this is discussed much more carefully. It is a simplified problem that captures the behavior of more practical problems in [9].

Consider

$$\bar{J} = \min_{v} \frac{1}{2} \int_0^1 \rho(x_1 - 1)^2 + v^2 \, dt \tag{3.18a}$$

$$x_1' = x_2, \quad x_1(0) = x_1(1) = 0 \tag{3.18b}$$

$$x_2' = x_3, \quad x_2(0) = 1 = -x_2(1) \tag{3.18c}$$

$$x_3' = v, \quad x_3(0) = 2 = x_3(1) \tag{3.18d}$$

$$x_1(t) \leq 0.134, \tag{3.18e}$$

where $\rho \geq 0$. Note that (3.18e) is a state inequality constraint. For the choice of $\rho = 0$, problem (3.18) is the problem studied in [54]. Our problem is motivated in part by the observation that in a number of applications there can be very different weightings on the control and the state. For example, if we have a control weighting of 10^{-3} and a state weighting of 10^4, which amounts to 10^2 on the norm of x, then the corresponding value of ρ is 10^7. It is known from [54] that the constraint will be active for a range of L values when $\rho = 0$. We take one of these values, $L = 0.134$, throughout this section. The left side of Fig. 12 shows the computed state x_1 and the right side shows the computed control. Figure 13 expands the view of state x_1 near the constraint $x = 0.134$ by several orders of magnitude.

Figure 12 seems to show the state variable x_1 riding the constraint over the middle of the interval. However, a classical result in the literature [54] which says that under certain mild appearing technical assumptions the solution of an odd order state constrained problem of order 3 or higher cannot ride a constraint over a nonzero length interval. The only possibility is one or more touch and goes.

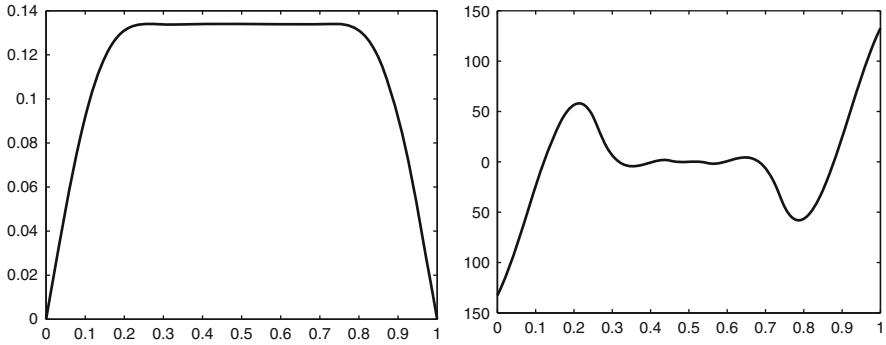

Fig. 12 Computed x_1 and control v for $\rho = 1.5 \times 10^5$, $\overline{J} = 60{,}148$

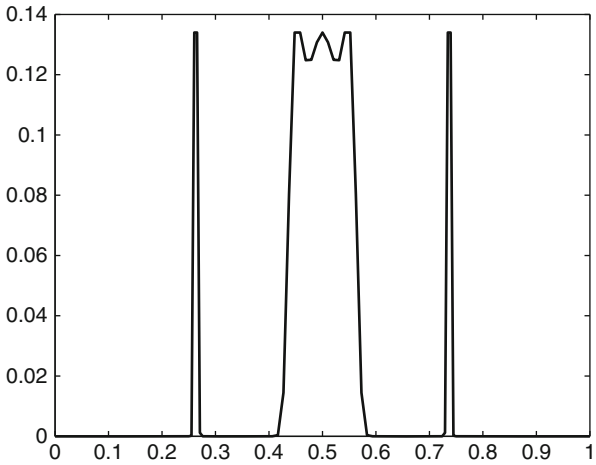

Fig. 13 Five touch points for x_1. Large nonlinear magnification near the state constraint

That is, the constraint intervals have zero length. The actual result from [54] (for the single constraint case) assumes that there is a pth order arc, the dynamics and constraint are $p + 1$ times continuously differentiable, the control is p times continuously differentiable along the boundary arc if there is a boundary arc, and the pth derivative of the inequality constraint along the optimal solution has a nonzero derivative with respect to the control. In addition, the Hamiltonian is assumed to have unique minimums in u for a given $x(t), \lambda(t)$ from the variational equations. Also, the optimal control is $p - 1$ times differentiable except at the junction points. Then the only way that there can be a boundary arc if p is odd is if the control has a continuous pth derivative at the junction points. This can be shown to not occur for the problems considered here, like for most smooth problems, since this condition at the end of a boundary arc would overdetermine the solution in an inconsistent way.

Thus what is observed in Fig. 12 seems to contradict the mathematical theory. Figure 13 greatly blows up the graph of x_1 by several orders of magnitude near the constraints in a nonlinear manner. We now see touch points emerging but on a very small scale. This is an example where the correct necessary conditions are very difficult to solve because they involve phenomena that are extremely ill conditioned. Yet this fine scale phenomena is not practically important. The direct transcription solution in Fig. 12 has regularized the numerical solution to only find those aspects of the solution that are computationally relevant.

Direct transcription's philosophy and theory is much closer to that of a boundary value solver rather than an initial value integrator. The equations are all solved at once rather than sequentially. Also any equation error tends to be distributed across the equations rather than accumulated sequentially. Inside the software the cost is also given dynamically so that doing a good job of finding the dynamics includes producing a solution whose actual cost is close to the computed cost.

4 Delayed DAEs

Many physical systems are naturally modeled as differential algebraic equations or DAEs. Many physical systems also possess delays either in the dynamics or in the application of the control. As noted earlier, direct transcription is a popular approach in industry for numerically solving nondelayed optimal control problems because of its ability to handle problems with constraints. In this section, we focus on how the use of the DAE formalism allows for the consideration of a much greater variety of delays.

Direct transcription has the advantage that it does not require forming the necessary conditions of an optimal control problem [7]. As just noted, this is especially useful when there are various operational constraints that go active and inactive. Some direct transcription software has been extended to handle delay problems [12]. The solution of delayed DAE optimal control problems is not a standard feature of optimal control codes. It also turns out that computational behavior is implementation dependent. The next section quickly describes the algorithm we are using. Our focus here is on how the capability to work with delayed DAEs (DDAEs) allows the solution of a much wider class of delayed systems.

For simpler problems or problems which are not DAEs and for which there are no state constraints, there are a number of ways to approach an optimal control problem particularly if the cost does not have endpoint conditions. For example, control parameterization can sometimes be used.

Time varying delays are of interest in a number of applications [32, 48, 51]. In [11, 12] examples are given to show that our direct transcription approach works well for problems with time varying state delays and constant control delay problems on uniform grids. We will not discuss time varying delays any further in this paper. In this section the way a particular software package implements direct

transcription may impact on the observations. All our computations were done in this section with SOCX.

In [10] it is pointed out that there can be computational problems with our approach when nonuniform grids are used on control delayed problems. This is not part of the main theme of this paper but will be briefly discussed at the end of Sect. 4.2.4. Recent research provides one way to work around this issue [92]. This section focuses on the power of the DAE formulation and its ability to describe a wide range of different appearing state and state derivative delay problems and then showing that our type of direct transcription algorithms can solve these challenging problems.

4.1 Direct Transcription Algorithm

Our particular direct transcription implementation is called SOCX (Sparse Optimal Control Extended) and is part of the SOS (Sparse Optimization Suite) package of software which is available from Applied Mathematical Analysis. Research oriented academic users can obtain a copy from Applied Mathematical Analysis. SOCX is a direct transcription software package that is being designed to solve nonlinear optimization, optimal control, parameter estimation, and delay problems. However, the information in this section is not specific to SOCX. This information can be used to aid or improve code for both nondelay and delay optimal control systems when the software has a similar philosophy to ours. The software begins by rewriting the dynamics of the properly formulated optimal control problem as a DAE or DDAE. This step aids in simplifying the problem, yet adds constraints to the formulation. Multiple varying delays and state and control delays can be accommodated. Determining the best way to formulate and solve a delayed optimal control problem when using direct transcription is still a research question.

Here we consider the optimal control delay problem formulation to minimize (4.1a) with constraints (4.1b)–(4.1e). To simplify the presentation, we continue to drop the dependence on just t from the variables in our notation, but keep the dependence on delayed or advanced variables. Thus $x(t), \dot{x}(t), u(t)$ and similar terms will be usually be written as $x, \dot{x}, u$ while $x(t-r)$ will remain as $x(t-r)$.

$$J = \phi(t_f) + \int_{t0}^{tf} L(x, u, x(\omega(t)), u(\eta(t)), t, p)dt \tag{4.1a}$$

$$\dot{x} = f(x, t, x(\omega(t)), u, u(\eta(t)), p), \quad t_0 \le t \le t_f \tag{4.1b}$$

$$0 = g(x, t, x(\omega(t)), u, u(\eta(t)), p), \quad t_0 \le t \le t_f \tag{4.1c}$$

$$x = \alpha(t), \quad -r \le t < 0, \; x_0 = q, \tag{4.1d}$$

$$u = \beta(t), \quad -s \le t < 0, \tag{4.1e}$$

with $\phi(t_f) = \phi(x(t_f), u(t_f), t_f, p)$, non-dynamic parameter vector p, and time delay functions $\omega(t)$, $\eta(t)$. Parameters r, s always denote the length of the prehistory and are constant. In the case of a single constant state and control delays, $\omega(t) = t - r$ and $\eta(t) = t - s$. However, we also consider time varying delays in which case we just write $\omega(t)$ and $\eta(t)$. The problem features cost (4.1a), DDAE (4.1b), (4.1c), and prehistory functions (4.1d) and (4.1e). Our implementation allows for several delays, state and control inequality constraints, and inequality constraints on initial and terminal conditions. In particular, there can be another vector equation like (4.1c) which gives any state or control inequality constraints. Note that even if the original process is an ODE, that a DAE can result if inequality state constraints become active. However, to simplify the presentation in this section we will use the simpler formulation (4.1) which has only equality constraints and one delay in the state and one delay in the algebraic or control variables. This suffices to make the desired points.

We will see that having DAEs gives much greater flexibility in the handling of different types of problems. Ideally the DAE would be index one, but SOCX can sometimes work with higher index DAEs depending on the cost function [41]. This is because in (4.1) the variable u denotes all algebraic variables. That is, those variables that do not appear differentiated in the equations. In a particular application where the process is a DAE, x would be the dynamic state variables while u is both the algebraic state variables and the control variables. This has positive consequences when solving numerically with direct transcription as discussed in [41] and Sect. 3.1.

Optimal control problems with delays require start-up functions on the delayed intervals (4.1d), (4.1e). Here r and s are taken to be positive. Systems of differential equations sometimes model phenomena requiring knowledge of the future state and/or control variables. In that case, $r, s < 0$ and (4.1d) and (4.1e) are replaced by post-history functions. Some of these types of problems with advances and other combinations of time shifted variables will be discussed later.

For simplicity, we consider (4.1) with no parameter vector p. When delays are present, it becomes necessary to relate variables and their delayed counterparts in an automatic way. The algorithm does this by reformulating the optimal control problem by enforcing consistency relationships between $x(\omega(t))$, $u(\eta(t))$ and pseudo variables, $v(t)$ and $w(t)$. Then the DAE (4.1b) and (4.1c) can be rewritten as a DDAE of the form

$$\dot{x} = f(x, u, v, w, t), \tag{4.2a}$$

$$0 = g(x, u, v, w, t), \tag{4.2b}$$

$$0 = v - x(\omega(t)), \tag{4.2c}$$

$$0 = w - u(\eta(t)). \tag{4.2d}$$

System (4.2) is written as if v is the same size as x and w is the same size as u. Actually (4.2c) only holds for those components of x that are delayed. In the same manner (4.2d) only holds for those components of u that are delayed.

Note that (4.2) is a DDAE even if (4.1c) is missing and the original model was a delayed ordinary differential equation. Also it is important to note that this reformulation is done within the software and is invisible to the user. Later we shall consider some reformulations that are done by the user so we will try to make it clear which reformulations are done by the user and which are done by the software.

Let us consider a DDAE system of the form (4.2), and show how the software approaches the problem. To simplify the presentation temporarily assume there are no parameters and a single state delay function ω. Define new algebraic variables $v(t)$, and require they be consistent with the delayed state variables. The original DDAE system can now be written as the larger DDAE

$$\dot{x} = f(x, u, v, t) \tag{4.3a}$$

$$0 = g(x, u, v, t) \tag{4.3b}$$

$$0 = v - z(\omega(t)), \tag{4.3c}$$

where z are just those entries in x that are delayed.

It is worth noting that in (4.3) that z is still a state variable since it consists of parts of x. The new variable v is a new algebraic variable. Thus in the case of state delays the reformulation keeps the delays to the state variables. This is important since computationally we observe more reliable solution behavior on state delay problems than on control delay problems. This will be commented on more later.

The basic idea is to first discretize the problem thereby creating a finite dimensional approximation. Large scale optimization methods can then be used to adjust the variables that define the discretization in order to find a minimum of the approximate problem. Then this solution is evaluated and if necessary used as the starting value for the solution of a finer approximation. The user can decide on the initial grid but further grid refinement is done automatically by the algorithm.

The direct transcription approach introduces a discretization of the problem by subdividing the time domain into M segments or intervals

$$0 = t_1 < t_2 < \cdots < t_M = t_f, \tag{4.4}$$

where the points are referred to as node, mesh, or grid points. For the remainder of this section, and only in this section, we use y_k as the variable for the estimate of variable y at time t_k for $y = x, u, v$. Thus one treats

$$m = (x_1, u_1, v_1, \ldots, x_M, u_M, v_M) \tag{4.5}$$

from the discretization as optimization variables in a nonlinear programming problem. We then approximate the differential equation using a nonlinear algebraic constraint. When the discretization is based on an implicit Runge-Kutta (IRK)

scheme the control problem is transcribed into a finite dimensional nonlinear program.

The algorithm uses the trapezoid (TR) and Hermite Simpson (HS) discretizations which are known to be second order and fourth order as ODE integrators. The user can specify either discretization but the default is to start with TR and then switch to HS after a couple of grid refinements.

For the trapezoidal method, we approximate the differential equations (4.3a) and algebraic constraints (4.3b) by a set of usually nonlinear algebraic constraints given by

$$0 = x_{k+1} - x_k - \frac{h_k}{2}(f_k + f_{k+1}) = \zeta_k, \tag{4.6a}$$

for $k = 1, \ldots, M-1$ and

$$0 = g_k, \text{ for } k = 1, \ldots, M, \tag{4.6b}$$

where (4.6a) are referred to as the *defect* constraints which are made small when solving the NLP. This discretization is implicit because the optimization variables (x_k, u_k, v_k) appear as arguments in the nonlinear functions $f_k \equiv f(x_k, u_k, v_k)$ and $g_k \equiv g(x_k, u_k, v_k)$.

Now in order to evaluate the right-hand side functions in (4.3a) and (4.3b) we must express the consistency relationship (4.3c) in terms of the NLP optimization variables. When the delay argument is exterior to the phase, that is when $\omega_k < 0$, or $\omega_k > t_M$, the value of the delayed state is given by the user defined function $\alpha(p, \omega_k)$. When the delay argument is interior to the phase, $0 \le \omega_k \le t_f$, let us define the interval J_k such that

$$t_{J_k} \le \omega(t_k) \le t_{J_k+1}. \tag{4.7}$$

Then for $\delta_k = (\omega_k - t_J)/h_{J_k} = (\omega_k - t_{J_k})/(t_{J_k+1} - t_{J_k})$ the interpolated value for the delayed state is just

$$z(\omega(t_k)) \doteq z(\omega_k) = c_1 z_{J_k} + c_2 z_{J_k+1} + c_3 \dot{z}_{J_k} + c_4 \dot{z}_{J_k+1}, \tag{4.8}$$

where the Hermite interpolation coefficients are

$$\begin{aligned} c_1 &= (1-\delta_k)^2(1+2\delta_k), \; c_2 = \delta_k^2(3-2\delta_k), \\ c_3 &= \delta_k(1-\delta_k)^2, \qquad c_4 = -\delta_k^2(1-\delta_k). \end{aligned} \tag{4.9}$$

Thus to enforce consistency at the grid points we can impose the NLP constraints

$$v(t_k) = z(\omega(t_k)) = \begin{cases} z(\omega_k) & \text{if } 0 \le \omega_k \le t_M, \\ \alpha(p, \omega_k) & \text{otherwise.} \end{cases} \tag{4.10}$$

Of course these consistency constraints are also implicit, since they involve the derivatives $\dot{z}_J$ and $\dot{z}_{J+1}$ which are given by the right-hand sides of (4.3a).

Time delay and time advance variables are permitted in a single problem. However, such variables are restricted from changing orientations, i.e., a time delayed variable cannot become a time advanced variable. To date, we cannot handle state dependent state delays with direct transcription. If no control is present the solution becomes a boundary value problem simulation. More information on the general philosophy of direct transcription can be found in [7] and technical details on how we are discretizing state delay problems are in [12].

4.2 DAEs and Delays

A key in dealing with some types of the delays is to be able to work with reformulations which are DAE models. This section will illustrate how to do this reformulation and what some of the advantages are. That a DAE formulation allows one to alter the appearance of a delay system is not a new observation [4]. However, those discussions were theoretical and we are interested in computational capabilities here. Sections 4.2.1–4.2.4 are based on [14, 92].

4.2.1 Advances

Problems with advances at first glance would seem to be less common than problems with delays. However, problems with advances do occur and, in fact, are common in some application areas. Some examples of applications with advances can be found in [53, 66]. For example, some control systems can look ahead. Pursuit and evasion problems can also involve advance information.

The existing methods for solving optimal control problems which rely on numerical integrators have trouble dealing with advances in the states. Depending on the problem, they can also have trouble with advances in the control. Since direct transcription discretizes the entire problem without a preference for the direction, and then uses sparse solvers, it does not matter to the software whether there is a delay or an advance or both. Advances are solved as easily as with delays as shown in [14].

The mixed-type problems discussed later in Sect. 4.2.3 also involve advances.

4.2.2 Neutral Systems

Neutral delay systems have both a delayed state and delayed state derivatives. Such systems appear in a variety of biological models [46] and in problems with networks

with lossless transmission lines [42]. While neutral equations are not immediately in the form of (4.1b), (4.1c), in practice many of them can be written in the required form. For example, given a neutral equation which is linear in the derivatives,

$$\dot{x} + B\dot{x}(t-r) = F_1(t, x, x(t-s), u, u(t-\tau)) \tag{4.11a}$$

$$0 = F_2(t, x, x(t-s), u, u(t-\tau)), \tag{4.11b}$$

we could let $z = x + Bx(t-r)$ and get the larger DDAE

$$\dot{z} = F_1(t, x, x(t-s), u, u(t-\tau)) \tag{4.12a}$$

$$0 = F_2(t, x, x(t-s), u, u(t-\tau)) \tag{4.12b}$$

$$0 = z - x - Bx(t-r), \tag{4.12c}$$

which is in a form that we can handle. Again note that the DAE (4.12a), (4.12c) results even if the original problem was not a DAE and only the ODE (4.11a).

We illustrate the reformulation of a neutral system (4.13) with a nonlinear biological model from [101]. This particular example is not a control problem so just a solution of the delayed equation is sought,

$$\frac{d}{dt}[x - \frac{1}{2}x(t-r)] = \cos(4t)x + 2\sin(4t)x(t-r) - 4[x - \frac{1}{2}x(t-r)]^2. \tag{4.13}$$

We rewrite (4.13) as the nonlinear DDAE in x, z

$$\dot{z} = \cos(4t)x + 2\sin(4t)x(t-r) - 4z^2 \tag{4.14a}$$

$$0 = z - x + \frac{1}{2}x(t-r). \tag{4.14b}$$

As in [101] we take $r = 0.3$, the interval of interest to be $[0,\ 4]$, and the prehistory in x to be $x = 1$ for $t \in [-0.3, 1]$ and $z(0) = \frac{1}{2}$. The solution found is given in Figs. 14 and 15.

4.2.3 Mixed-Type

Mixed-type or forward-backward systems have both delays and advances in the equations. These types of systems arise as the necessary conditions for optimal control problems with state delays and as models in a number of applications [43].

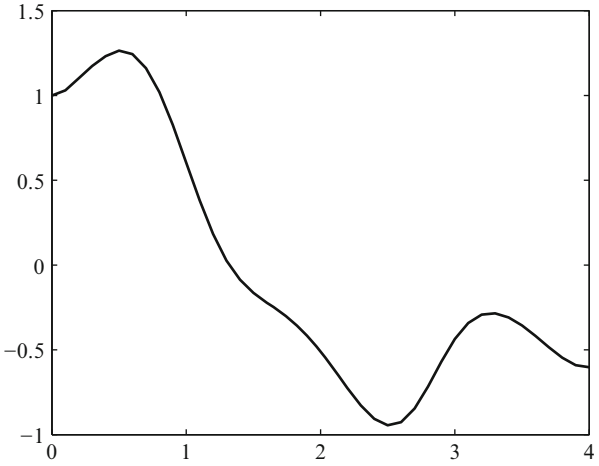

Fig. 14 State trajectory for neutral problem (4.13) using DDAE formulation (4.12)

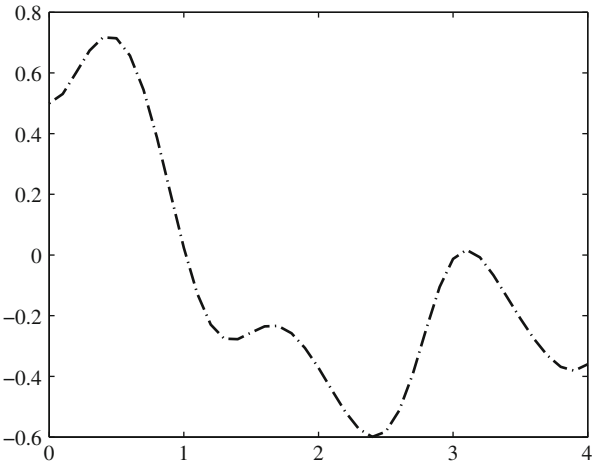

Fig. 15 z for the neutral delay problem (4.13) using DDAE formulation (4.14)

As an illustration of a mixed-type problem we use one example from [43] which is

Example 4.1

$$\dot{x} = cx + 2x(t-1) + 3x(t+1),\ 0 \le t \le 3 \tag{4.15a}$$

$$x = e^{3t}, \quad -1 \le t \le 0 \tag{4.15b}$$

$$x = e^{3t}, \quad 3 \le t \le 4, \tag{4.15c}$$

with $c = 3 - 2e^{-3} - 3e^{3}$.

This example was constructed so that the solution is e^{3t}. The software reformulates (4.15a) as a DDAE in x, z, w.

$$\dot{x} = cx + 2z + 3w, \tag{4.16a}$$

$$0 = z - x(t-1), \tag{4.16b}$$

$$0 = w - x(t+1). \tag{4.16c}$$

To illustrate how a mixed-type problem arises when using the necessary conditions consider the state delay optimal control problem

$$\dot{x} = Ax + Cx(t-r) + Bu, \tag{4.17}$$

with cost

$$\frac{1}{2}\int_a^b x^T Q x + u^T R u \, dt.$$

Then the necessary conditions which hold almost everywhere in $[a, b]$ are

$$\dot{x} = Ax + Cx(t-r) + Bu, \tag{4.18a}$$

$$-\dot{\lambda} = Qx + A^T\lambda + \chi_{[a,b-r]}(t)C^T\lambda(t+r) \tag{4.18b}$$

$$0 = Ru + B^T\lambda \tag{4.18c}$$

$$\lambda(b) = 0 \tag{4.18d}$$

$$u = \psi, \quad t \in [a-s, a) \tag{4.18e}$$

$$x = \phi, \quad t \in [a-r, a], \tag{4.18f}$$

where $\chi_{[a,b-r]}(t)$ is the characteristic function of the interval $[a,\ b-r]$. That is, $\chi_{[a,b-r]}(t) = 1$ if t is inside the interval and zero otherwise. If R is invertible, which frequently occurs when there are no control constraints, system (4.18) can be written more succinctly as

$$\dot{x} = Ax + Cx(t-r) - BR^{-1}B^T\lambda, \tag{4.19a}$$

$$-\dot{\lambda} = Qx + A^T\lambda + C^T\lambda(t+r) \tag{4.19b}$$

$$\lambda = 0, \quad t \in [b, b+r] \tag{4.19c}$$

$$x = \phi, \quad t \in [a-r, a]. \tag{4.19d}$$

If instead of the linear problem given here, the nonlinear problem equivalent of (4.18c) is nonlinear in u, then it might be preferable to consider the original DDAE (4.18a)–(4.18c) instead of (4.19).

Note that the system composed of (4.18a) through (4.18c) forms a mixed-type delay.

Another way that mixed-type problems arise is from problems involving integral operators with varying integration range. Consider for example, the operator

$$\theta = \int_{t-a}^{t+b} g(t, x, u)dt, \tag{4.20}$$

which looks like an averaging operator. This operator can be replaced by the system

$$\dot{z} = g(t, x, u) \tag{4.21a}$$

$$\theta = z(t + b) - z(t - a) \tag{4.21b}$$

and appropriate initial and terminal conditions.

4.2.4 Mixed-Type Neutral

By repetitively utilizing the above types of transformations a user can transform more complicated systems into ones that fit our formulation. Consider a system in which there are both past and future derivatives of state variables of the form

$$A\dot{x}(t - r) + B\dot{x} + C\dot{x}(t + s) = F, \tag{4.22}$$

where F depends on t and present, past, and future values of x, u. Then as before with the neutral equations, let $z = Ax(t - r) + Bx$ and (4.22) becomes

$$\dot{z} + C\dot{x}(t + s) = F \tag{4.23a}$$

$$z = Ax(t - r) + Bx. \tag{4.23b}$$

Repeating the process on (4.23) with $w = z + Cx(t + s)$ we get the delayed DDAE

$$\dot{w} = F \tag{4.24a}$$

$$z = Ax(t - r) + Bx \tag{4.24b}$$

$$w = z + Cx(t + s), \tag{4.24c}$$

which is now in the correct form for our software.

This example also illustrates how the rewriting to get a delayed DAE in the correct form is often not unique. For example, instead of (4.24) we could have gotten

$$\dot{w} = F \tag{4.25a}$$

$$z = Bx + Cx(t + s) \tag{4.25b}$$

$$w = z + Ax(t - r) \tag{4.25c}$$

or

$$\dot{w} = F \tag{4.26a}$$

$$w = Ax(t - r) + Bx + Cx(t + s). \tag{4.26b}$$

The examples given all had continuous solutions. Smoothness was not required. If there is a loss of smoothness (discontinuity of the derivative), then the software just refines the grid near the discontinuity of the derivative. This is seen in some of the computational grid plots. If there is a discontinuity in the solution, and it is a jump discontinuity, then we often get a good idea of what the optimal solution looks like but there will be some computational noise near the jump. This is like what happens with bang-bang controls. Bang-bang controls can often be best solved for using phases.

A numerical difficulty arises with control delays and nonuniform grids which are needed for some problems. A shown in (4.2c), (4.2d) both delayed differential and algebraic variables are handled in the same way. Additional variables denoted v, w in (4.2c), (4.2d) are introduced. This produces no observed difficulty in the solution of state delayed problems in the numerous examples solved to date. However, with the delayed controls the relationship (4.2d) can lead to some free variables in the optimization of the discretized NLP problem. This is carefully examined and illustrated in [15, 92].

An approach called EIC (Exogenous Input Parameterization) for dealing with this problem has been presented in [13]. Simply put EIC turns the algebraic variable of interest into a differential variable by adding dynamics $w' = z$ and lightly weighting z in the cost. Computational examples suggest that EIC can be very useful. Mathematical analysis and algorithm development of the EIC method is given in [15, 92]. Some issues remain. While this explains why there are problems with control delays and gives a way to address them, we are also in the process of explaining precisely why state variable delays are handled so much better than control delays. The key appears to be in the different types of interpolation that are involved in determining delayed state values as opposed to delayed control values.

In working with non-delayed optimal control problems the ability to formulate the problem as having several phases is often very helpful. We are looking at the problem of implementing phases within the delay setting. However, doing so in a naive way can destroy the sparsity which is so important for efficiency in implementing direct transcription approaches.

Often optimal control problems with delays have places of reduced smoothness. An optimal control model that describes the immune response of a pathogenic disease process is developed in Stengel, Ghigliazza, Kulkarni, and Laplace and illustrates a loss of smoothness [91]. The goal is to minimize the therapeutic treatment cost quantified by

$$F = \frac{1}{2}\left(x_1^2(t_f) + x_4^2(t_f)\right) + \frac{1}{2}\int_0^{t_f} x_1^2 + x_4^2 + \|u\|^2 \, dt \tag{4.27}$$

subject to the nonlinear delay equations

$$\dot{x}_1 = (a_{11} - a_{12}x_3)x_1 + b_1u_1, \tag{4.28a}$$

$$\dot{x}_2 = a_{21}(x_4)a_{22}x_1(t-r)x_3(t-r) - a_{23}(x_2 - x_2^*) + b_2u_2, \tag{4.28b}$$

$$\dot{x}_3 = a_{31}x_2 - (a_{32} + a_{33}x_1)x_3 + b_3u_3, \tag{4.28c}$$

$$\dot{x}_4 = a_{41}x_1 - a_{42}x_4 + b_4u_4. \tag{4.28d}$$

For $0 \le t \le t_F = 10$ with state delay $r = 1$, and startup functions given by

$$x_1(t) = 0 \quad -r \le t < 0 \tag{4.28e}$$

$$x_3(t) = 3 \quad -r \le t < 0. \tag{4.28f}$$

Define

$$a_{21}(x_4) = \begin{cases} \cos \pi x_4 & \text{if } 0 \le x_4 \le \frac{1}{2}, \\ 0 & \text{if } \frac{1}{2} \le x_4 \end{cases} \tag{4.28g}$$

$$x(0) = [3\ ,2,\ 4/3,\ 0]^T. \tag{4.28h}$$

The problem coefficients are defined as $a_{11} = 1, a_{12} = 1, a_{22} = 3, a_{23} = 1, a_{31} = 1, a_{32} = 1.5, a_{33} = 0.5, a_{41} = 1, a_{42} = 1, b_1 = -1, b_2 = 1, b_3 = 1, b_4 = -1, x_2^* = 2$. Note that a_{21} is not differentiable at $x_4 = 0.5$.

The graphs of the optimal control and state trajectories are in Figs. 16 and 17 respectively.

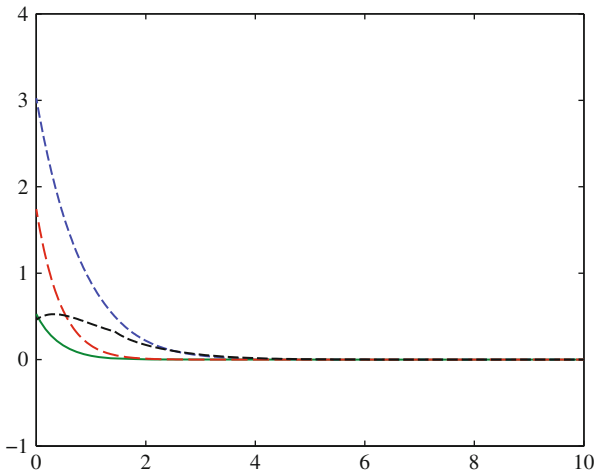

Fig. 16 Optimal control for the state delay problem (4.28)

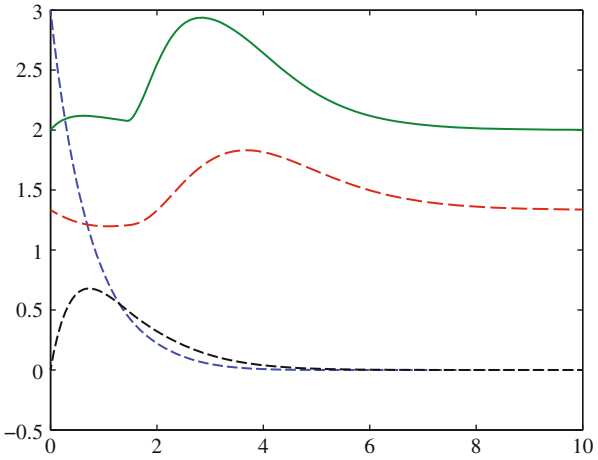

Fig. 17 Optimal state trajectories for the state delay problem (4.28)

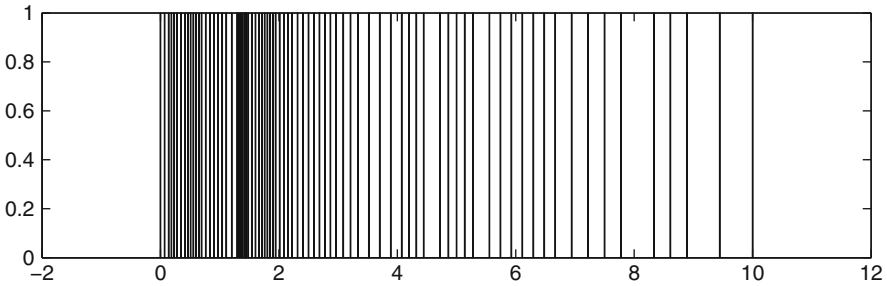

Fig. 18 Second to last grid for (4.28)

The state trajectory x_4 starts at zero and rises above 0.5. The corners in the other two state trajectories correspond to when x_4 crosses 0.5 again. The solution was verified by setting up a method of steps formulation with no delays and solving that problem. Solution of the method of steps formulation took 8.5 s. Solution of the delay problem by SOCX took 0.95 s. This and all other computations in this paper were done on a server consisting of dual 3 GHz quad core Intel Xeon processors (8 total cores) with 8 GB RAM. The final automatically generated grid had 192 points. The default of two iterations with TR and then switching to HS was followed.

The grid refinement strategy results in highly nonuniform grids. The initial grid for (4.28) was 10 uniformly spaced points. The final automatically generated grid was 192 points. Since it is easier to visualize, the second to last grid which had 83 points is plotted in Fig. 18. Note that the grid plotted in Fig. 18 has many more points near the beginning when the solutions are changing rapidly. In particular, the grid is very fine near the corners on the solution graphs.

Another example from [16] but with piecewise states is based on the hydraulic-transients model Example 2.1 from [95] except that we take specific parameter values.

$$Ex' = Mx + Nx(t-1) \tag{4.29a}$$
$$x_1(0) = 3 \tag{4.29b}$$
$$x_2(t) = 1, \quad -1 < t \le 0 \tag{4.29c}$$
$$x_3(t) = 2, \quad -1 \le t \le 0 \tag{4.29d}$$
$$E = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \; M = \begin{bmatrix} -2 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}, \; N = \begin{bmatrix} 0 & 0 & 2 \\ 0 & 0 & -\frac{1}{4} \\ 0 & -1 & 0 \end{bmatrix}. \tag{4.29e}$$

This problem is solved with SOCX. Figure 19 shows the solution using the Hermite Simpson discretization and a step of $h = 0.22$. This value of h was chosen to not divide the delay evenly. Figure 20 shows the solution with $h = 0.1$

With direct transcription the problem of points of reduced order at points of reduced continuity is handled by the grid getting to be very fine near the reduced continuity point and thus its contribution is very marginal. The sophisticated grid refinement strategy of SOCX estimates the error along the whole solution and then

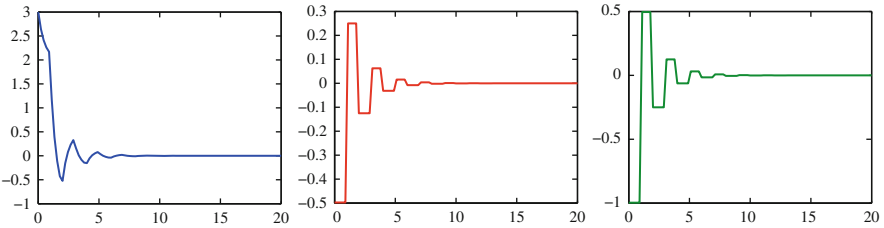

Fig. 19 States x_1, x_2, x_3 of (4.29) using HS with constant stepsize $h = 0.22$

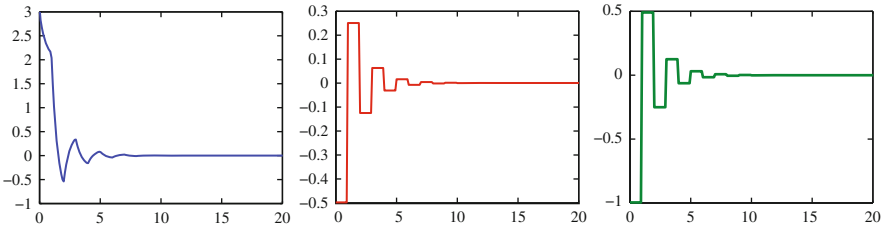

Fig. 20 States x_1, x_2, x_3 of (4.29) using HS with constant stepsize $h = 0.1$

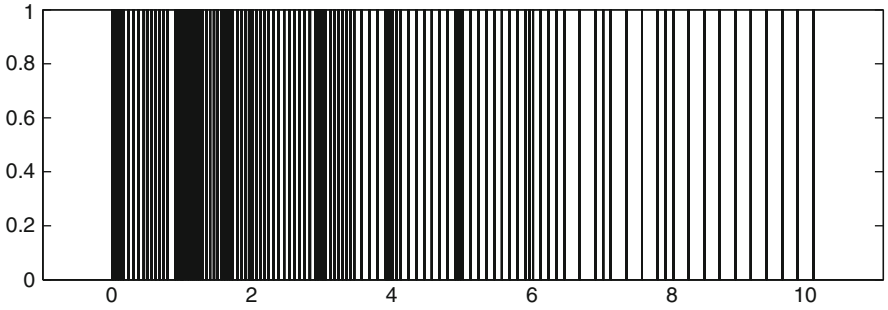

Fig. 21 Iteration 3 grid when solving (4.29) with TR and initial uniform stepsize $h = 0.22$ on [0, 10]

iteratively adds additional grid points where they are needed. Figure 21 shows the computational grid on [0; 10] for (4.29). This is the grid from iteration 3 with Trapezoid when the error was down to 10^{-5}. Note the highly nonuniform nature of the grid. The grid becomes much sparser to the right since although the some of the state variables continue to have discontinues these discontinuities have magnitude below the computational accuracy and thus can be ignored.

4.2.5 Delayed DAEs and RK Method Stability

In concluding, we give one more example where the combination of direct transcription and DAEs gives greater flexibility in the choice of discretization than one would expect based on the numerical initial value DAE theory. This discussion is from [16].

Consider the neutral delay DAE, or NDDAE,

$$E\dot{x} = Lx + Mx(t-\tau) + N\dot{x}(t-\tau), t \geq 0 \tag{4.30a}$$

$$x = \phi, \quad -\tau \leq t \leq 0, \tag{4.30b}$$

where $\{E, L\}$ is a regular pencil, that is $\det(\eta E + L) \neq 0$ for some scalar η. Also the delay variable $\tau > 0$.

Numerical methods and preservation of stability of delay problems are studied in [60, 93–95]. Theorem 2.4 of [95] gives three conditions that jointly imply delay-independently asymptotically stable. They then give formulas for Lagrange interpolation for delayed quantities based on three parameters α, β, m. A numerical method is then defined to be NAGP-stable if for systems meeting the assumptions of their Theorem 2.4, the coefficient matrix of the Runge-Kutta method is invertible, and the numerical solutions of the homogenous system goes to zero for consistent initial conditions whenever $\beta \geq m + 1$.

Then for our purposes the key result from [93] is that if the Lagrange interpolation satisfies $\alpha \leq \beta \leq \alpha + 2$ and $m \geq \beta + 1$, then the numerical processes generated by the Gauss, Lobatto IIIA, and Lobatto IIIB methods combined with the Lagrange interpolation are not NAGP-stable. The numerical processes generated by the Radau IA, Radau IIA, and Lobatto IIIC methods combined with the Lagrange interpolation are NAGP-stable. An example is given for a delayed DAE for which the instability holds.

The trapezoid and Hermite Simpson methods we have discussed earlier are Lobatto IIIA methods. The cited results would seem to suggest that these two methods would not be good choices for some delay problems. However, when the same examples from [93] are solved with SOCX asymptotically stable solutions are computed. There are two reasons for this. One is that our interpolation is a bit different that of [93] in how we compute prior values. However, experimentation shows that is not the primary reason. The difference lies in using the boundary value like direct transcription as opposed to an initial value approach. This is another example where one cannot just apply initial value theory to understanding the numerical behavior of direct transcription codes.

5 Conclusion

We have seen that allowing for the consideration of DAEs provides greater flexibility in solving a number of problems especially those in control and simulation. This is true for both ODE and DAE modeled processes. This flexibility has even more uses and advantages when it is combined with a direct transcription approach. Examples have included design of observers with linear error dynamics, estimation of disturbances, optimal control of high index dynamics, fault detection with DAE models, inequality constrained optimal control problems, and solving a variety of delayed optimal control problems.

Partial differential equations (PDEs) often have a boundary value type behavior in part to their solution. There has been some work on PDAEs, that is partial differential algebraic equations. The flexibility illustrated for delay DAEs carries over to PDAEs and many types of PDEs and PDAEs can be incorporated once a PDAE framework is included. A discussion of this is outside the scope of this paper. Such a survey would be of great interest, but the author is not the best person to write it. We note only [21, 55, 64, 67–69, 71–74, 99].

Acknowledgements The writing of this paper was supported in part by NSF Grants DMS-0907832 and DMS-1209251. Some of the results discussed were supported by earlier grants from ONR, AFOSR, ARO, and NSF. The number of people, both colleagues and students who have had an impact on the author's understanding of DAEs and their applications over the years, is too numerous to mention. This is especially true of the author's introduction to DAEs and their applications. But in terms of the specific topics discussed in this survey, the author would like to especially thank his colleagues John Betts, Ramine Nikoukhah, Peter Kunkel, Volker

Mehrmann, Roswitha Maerz and PhD students Neil Biehn, Jason Scott, Karmethia Thompson, Angela Engelsone, and Dirk von Wissel.

References

1. Aplevich, J.D.: Implicit Linear Systems. Lecture Notes in Control and Information Sciences, vol. 152. Springer, Berlin (1991)
2. Baumann, W.T.: Feedback control of multi-input nonlinear systems by extended linearization. IEEE Trans. Autom. Control **33**, 40–46 (1988)
3. Becerra, V.M.: $\mathcal{PSOPT}$ Optimal Control Solver User Manual Release 2 build, Reading RG6 6AY, University of Reading School of Systems Engineering United Kingdom (2010)
4. Bellena, A., Zennaro, M.: Numerical Methods for Delay Differential Equations. Oxford University Press, Oxford (2013)
5. Berger, T., Reis, T.: Controllability of linear differential-algebraic systems - a survey. In: Ilchmann, A., Reis, T. (eds.) Surveys in Differential-Algebraic Equations I, DAE Forum, pp. 1–62. Springer, Heidelberg (2013)
6. Betts, J.T.: Parametric Tool Path Trajectory Optimization, Technical Report SSGTECH-98-006, The Boeing Company (1998)
7. Betts, J.T.: Practical Methods for Optimal Control and Estimation Using Nonlinear Programming, 2nd edn. SIAM, Philadelphia (2010)
8. Betts, J.T., Biehn, N., Campbell, S.L.: Convergence of nonconvergent IRK discretizations of optimal control problems with state inequality constraints. SIAM J. Sci. Comput. **23**, 1981–2007 (2002)
9. Betts, J.T., Campbell, S.L., Engelsone, A.: Direct transcription solution of optimal control problems with higher order state constraints: theory vs practice. Optim. Eng. **8**, 1–19 (2007)
10. Betts, J.T., Campbell, S.L., Thompson, K.C.: Direct transcription solution of optimal control problems with control delays. In: Numerical Analysis and Applied Mathematics ICNAAM2011. AIP Conference Proceedings, Halkidiki, vol. 138, pp. 38–41 (2011)
11. Betts, J.T., Campbell, S.L., Thompson, K.C.: Optimal control software for constrained nonlinear systems with delays. In: Proceedings of 2011 IEEE Multi Conference on Systems and Control, Denver, pp. 444–449 (2011)
12. Betts, J.T., Campbell, S.L., Thompson, K.C.: Optimal control of a delay partial differential equation. In: Biegler, L., Campbell, S., Mehrmann, V. (eds.) Control and Optimization with Differential-Algebraic Constraints, pp. 213–232. SIAM, Philadelphia (2012)
13. Betts, J.T., Campbell, S.L., Thompson, K.C.: Simulation and optimization in systems with delays. In: Society for Modeling and Simulation Series 2013 Proceedings, San Diego, pp. 1084–1085 (2013)
14. Betts, J.T., Campbell, S.L., Thompson, K.C.: Direct transcription solution of optimal control problems with differential algebraic equations with delays. In: Proceedings of 14th IASTED International Symposium on Intelligent Systems and Control (ISC 2013), Marina del Rey, pp. 166–173 (2013)
15. Betts, J.T., Campbell, S.L., Thompson, K.C.: Solving optimal control problems with control delays using direct transcription, preprint (2015)
16. Betts, J.T., Campbell, S.L., Thompson, K.C.: LobattoIIIA methods, direct transcription, and DAEs with delays. Numer. Algorithms **69**, 291–300 (2015)
17. Biehn, N., Campbell, S.L., Nikoukhah, R., Delebecque, F.: Numerically constructible observers for linear time-varying descriptor systems. Automatica **37**, 445–452 (2001)
18. Birk, J., Zeitz, M.: Extended Luenberger observer for non-linear multivariable systems. Int. J. Control **47**, 1823–1836 (1988)
19. Blajer, W.: Index of differential-algebraic equations governing the dynamics of constrained mechanical systems. Appl. Math. Model. **16**, 70–77 (1992)

20. Bobinyec, K., Campbell, S.L., Kunkel, P.: Maximally reduced observers for linear time varying DAEs. In: Proceedings of IEEE Multiconference on Systems and Control, Denver, pp. 1373–1378 (2011)
21. Bodestedt, M., Tischendorf, C.: PDAE models of integrated circuits and perturbation analysis. Math. Comput. Model. Dyn. Syst. **13**, 1–17 (2007)
22. Brenan, K.E., Campbell, S.L., Petzold, L.R.: Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations. SIAM, Philadelphia (1996)
23. Campbell, S.L.: Least squares completions for nonlinear differential algebraic equations. Numer. Math. **65**, 77–94 (1993)
24. Campbell, S.L., Bobinyec, K.: Surveys in Differential-Algebraic Equations II, Differential-Algebraic Equations Forum, Springer, 1–67 (2014)
25. Campbell, S.L., Holte, L.E.: Eigenvalue placement in completions of DAEs. Electron. J. Linear Algebra **26**, 520–534 (2013)
26. Campbell, S.L., Kunkel, P.: Completions of nonlinear DAE flows based on index reduction techniques and their stabilization. J. Comput. Appl. Math. **233**, 1021–1034 (2009)
27. Campbell, S.L., März, R.: Direct transcription solution of high index optimal control problems and regular Euler-Lagrange equations. J. Comput. Appl. Math. **202**, 186–202 (2007)
28. Campbell, S.L., Meyer Jr., C.D.: Generalized Inverses of Linear Transformations. SIAM, Philadelphia (2009)
29. Campbell, S.L., Nikoukhah, R.: Auxiliary Signal Design for Failure Detection. Princeton University Press, Princeton (2004)
30. Campbell, S.L., Delebecque, F., Nikoukhah, R.: Observer design for linear time varying descriptor systems. In: Proceedings of Control Industrial Systems (CIS97), Belfort, pp. 507–512 (1997)
31. Campbell, S.L., Biehn, N., Jay, L., Westbrook, T.: Some comments on DAE theory for IRK methods and trajectory optimization. J. Comput. Appl. Math. **120**, 109–131 (2000)
32. Cloosterman, M., van de Wouw, N., Heemels, W., Nijmeijer, H.: Stability of networked control systems with uncertain time-varying delays. IEEE Trans. Autom. Control **54**, 1575–1580 (2009)
33. Dai, L.: Singular Control Systems. Lecture Notes in Control and Information Science, vol. 118. Springer, Berlin (1989)
34. Darby, C.L., Hager, W.W., Rao, A.V.: An hp-adaptive pseudospectral method for solving optimal control problems. Optim. Control Appl. Methods **32**, 476–502 (2011)
35. Darouach, M.: Functional observers for linear descriptor systems. In: Proceedings of 17th Mediterranean Conference on Control & Automation, Thessaloniki, pp. 1535–1539 (2009)
36. Darouach, M., Benzaouia, A.: Constrained observer based control for linear singular systems. In: Proceedings of 18th Mediterranean Conference on Control & Automation, Marrakech, pp. 29–33 (2010)
37. Darouach, M., Boutat-Baddas, L.: Observers for a class of nonlinear singular systems. IEEE Trans. Autom. Control **53**, 2627–2633 (2008)
38. Darouach, M., Boutayeb, M.: Design of observers for descriptor systems. IEEE Trans. Autom. Control **40**, 1323–1327 (1995)
39. Darouach, M., Zasadzinski, M., Hayar, M.: Reduced-order observer design for descriptor systems with unknown inputs. IEEE Trans. Autom. Control **41**, 1068–1072 (1996)
40. El-Tohami, M., Lovass-Nagy, V., Mukundan, R.: On the design of observers for generalized state space systems using singular value decomposition. Int. J. Control **38**, 673–683 (1983)
41. Engelsone, A., Campbell, S.L., Betts, J.T.: Direct transcription solution of higher-index optimal control problems and the virtual index. Appl. Numer. Math. **57**, 281–296 (2007)
42. Farrell, K., Grove, E.A., Ladas, G.: Neutral delay differential equations with positive and negative coefficients. Appl. Anal. **27**, 182–197 (1988)
43. Ford, N.J., Lumb, P.M.: Mixed-type functional differential equations: a numerical approach. J. Comput. Appl. Math. **229**, 471–479 (2009)

44. Gao, Z., Wang, H.: Descriptor observer approaches for multivariable systems with measurement noises and application in fault detection and diagnosis. Syst. Control Lett. **55**, 304–313 (2006)
45. Gill, P.E., Murray, W., Saunders, M.A.: User's Guide for SNOPT 7: A Fortran Package for Large-Scale Nonlinear Programming. Systems Optimization Laboratory, Stanford University, CA, 9, pp. 4305–4023 (2007)
46. Hadeler, K.P.: Neutral delay equations from and for population dynamics. Electron. J. Qual. Theory Differ. Equ. **11**, 1–18 (2008)
47. Hairer, E., Wanner, G.: Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems. Springer, Heidelberg (1996)
48. He, Y., Liu, G., Rees, D., Wu, M.: Stability analysis for neural networks with time-varying interval delay. IEEE Trans. Neural Netw. **18**, 1850–1854 (2007)
49. Hou, M., Müller, P.C.: Observer design for descriptor systems. IEEE Trans. Autom. Control **44**, 164–168 (1999)
50. Ilchmann, A., Mehrmann, V.: A behavioral approach to time-varying linear systems. Part 2: descriptor systems. SIAM J. Control Optim. **44**, 1748–1765 (2005)
51. Ionescu, C., Hodrea, R., De Keyser, R.: Variable time-delay estimation for anesthesia control during intensive care. IEEE Trans. Biomed. Eng. **58**, 363–369 (2011)
52. Isermann, R.: Fault Diagnosis Systems: An Introduction from Fault Detection to Fault Tolerance. Springer, Berlin (2006)
53. Ismail, G.A.F.: Modified technique for solving advance-delay differential system. Math. Comput. Model. **41**, 287–289 (2005)
54. Jacobsen, D.H., Lele, M.M., Speyer, J.L.: New necessary conditions of optimality for control problems with state variable inequality constraints. J. Math. Anal. Appl. **35**, 255–284 (1971)
55. Jansen, L., Tischendorf, C.: A unified (P)DAE modeling approach for flow networks. In: DAE-Forum Progress in Differential-Algebraic Equations - Deskriptor 2013. Springer, Heidelberg (2014)
56. Kidane, N., Yamashita, Y., Nishitani, H.: Observer based I/O-linearizing control of high index DAE systems. In: Proceedings of American Control Conference, Denver, CO, pp. 3537–3542 (2003)
57. Koenig, D., Mammar, S.: Design of proportional-integral observer for unknown input descriptor systems. IEEE Trans. Autom. Control **47**, 2057–2062 (2002)
58. Krener, A.J., Isidori, A.: Linearization by output injection and nonlinear observers. Syst. Control Lett. **3**, 47–52 (1983)
59. Krener, A.J., Respondek, W.: Nonlinear observers with linearizable error dynamics. SIAM J. Control Optim. **23**, 197–216 (1985)
60. Kuang, J., Tian, H., Shan, K.: Asymptotic stability of neutral differential systems with many delays. Appl. Math. Comput. **217**(24), 10087–10094 (2011)
61. Kumar, A., Daoutidis, P.: Control of Nonlinear Differential Algebraic Equation Systems. Chapman and Hall/CRC, New York (1999)
62. Kunkel, P., Mehrmann, V.: Differential-Algebraic Equations: Analysis and Numerical Solution. European Mathematical Society, Zürich (2006)
63. Lamour, R., März, R., Tischendorf, C.: Differential-Algebraic Equations: A Projector Based Analysis. Differential Algebraic Equations Forum. Springer, Heidelberg (2012)
64. Leugering, G., Engell, S., Griewank, A., Hinze, M., Rannacher, R., Schulz, V., Ulbrich, M., Ulbrich, S.: Constrained Optimization and Optimal Control for Partial Differential Equations. International Series of Numerical Mathematics, vol. 160. Birkhauser, Basel (2012)
65. Lewis, F.L., Syrmos, V.: Optimal Control, 2nd edn. Wiley, New York (1995)
66. Lucero, J.C.: Advanced-delay differential equation for aeroelastic oscillations in physiology. Biophys. Rev. Lett. **3**, 125–133 (2008)
67. Lucht, W., Strehmel, K.: Discretization based indices for semilinear partial differential algebraic equations. Appl. Numer. Math. **28**, 371–386 (1998)

68. Lucht, W., Strehmel, K., Eichler-liebenow, C.: Linear Partial Differential Algebraic Equations - Part I: Indexes, Consistent Boundary/Initial Conditions, Report 17. Martin-Luther-Universitat Halle, Fachbereich Mathematik und Informatik (1997)
69. Lucht, W., Strehmel, K., Eichler-Liebenow, C.: Indexes and special discretization methods for linear partial differential algebraic equations. BIT Numer. Math. **39**, 484–512 (1999)
70. Markovsky, I., Willems, J.C., Van Huffel, S., De Moor, B.L.M.: Exact and Approximate Modeling of Linear Systems: A Behavioral Approach. SIAM, Philadelphia (2006)
71. Marszalek, W., Campbell, S.L.: DAEs arising from traveling wave solutions of PDEs. J. Comput. Appl. Math. **82**, 41–58 (1997)
72. Marszalek, W., Campbell, S.L.: DAEs arising from traveling wave solutions of PDEs II. Comput. Math. Appl. **37**, 15–34 (1999)
73. Martinson, W.S., Barton, P.I.: Index and characteristic analysis of linear PDAE systems. SIAM J. Sci. Comput. **24**, 905–923 (2002)
74. Matthes, M., Tischendorf, C.: Convergence analysis of a partial differential algebraic system from coupling a semiconductor model to a circuit model. Appl. Numer. Math. **61**, 382–394 (2011)
75. Michalska, H., Mayne, D.Q.: Moving horizon observers and observer-based control. IEEE Trans. Autom. Control **40**, 995–1006 (1995)
76. Moraal, P.E., Grizzle, J.W.: Observer design for nonlinear systems with discrete-time measurements. IEEE Trans. Autom. Control **40**, 395–404 (1995)
77. Niemann, H.H.: A setup for active fault diagnosis. IEEE Trans. Autom. Control **51**, 1572–1578 (2006)
78. Nikoukhah, R.: A new methodology for observer design and implementation. IEEE Trans. Autom. Control **43**, 229–234 (1998)
79. Okay, I., Campbell, S.L., Kunkel, P.: Completions of implicitly defined vector fields and their applications. In: Proceedings of 18th International Symposium on Mathematical Theory of Networks and Systems, MTNS 08, Blacksburg, VA (2008)
80. Okay, I., Campbell, S.L., Kunkel, P.: Completions of implicitly defined linear time varying vector fields. Linear Algebra Appl. **431**, 1422–1438 (2009)
81. Patton, R.J., Frank, P.M., Clark, R.N.: Issues of Fault Diagnosis for Dynamic Systems. Springer, Berlin (2006)
82. Petzold, L.R.: A description of DASSL: a differential/algebraic system solver. In: Stepleman, R.S., et al. (eds.) Scientific Computing. North-Holland, Amsterdam (1983)
83. Polderman, J.W., Willems, J.C.: Introduction to Mathematical Systems Theory: A Behavioral Approach. Springer, New York (1998)
84. Rabier, P.J., Rheinboldt, W.C.: Nonholonomic Motion of Rigid Mechanical Systems from a DAE Viewpoint. SIAM, Philadelphia (2000)
85. Rao, A.V., Benson, D.A., Darby, C., Patterson, M.A., Francolin, C., Sanders, I., Huntington, G.T.: Algorithm 902: Gpops, a matlab software for solving multiple-phase optimal control problems using the Gauss pseudospectral method. ACM Trans. Math. Softw. **37**, 22:1–22:39 (2010)
86. Riaza, R.: Differential-Algebraic Systems. Analytical Aspects and Circuit Applications. Word Scientific, Hackensack, NJ (2008)
87. Scott, J.R.: Fault detection in differential algebraic equations. Ph.D. thesis, North Carolina State University (2015)
88. Scott, J.R., Campbell, S.L.: Auxiliary signal design for failure detection in differential-algebraic equations. Numer. Algebra Control Optim. **4**, 151–179 (2014)
89. Scott, J.R., Campbell, S.L.: Auxiliary signal design for failure detection in high index differential-algebraic equations. In: Proceedings of IEEE Conference on Decision and Control, Los Angeles (2014)
90. Scott, J.R., Campbell, S.L.: Asynchronous auxiliary signal design for failure detection. In: IEEE International Conference on Systems, Man, and Cybernetics (SMC2014), San Diego, CA, pp. 2785–2790 (2014)

91. Stengel, R.F., Ghigliazza, R., Kulkarni, N., Laplace, O.: Optimal control of innate immune response. Optim. Control Appl. Methods **23**, 91–104 (2002)
92. Thompson, K.C.: Solving Nonlinear Constrained Optimization Time Delay Systems with a Direct Transcription Approach. Ph.D. thesis, North Carolina State University (2014)
93. Tian, H., Kuang, J., Qiu, L.: The stability of linear multistep methods for linear systems of neutral differential equations. J. Comput. Math. **19**, 125–130 (2001)
94. Tian, H., Yu, Q., Kuang, J.: Asymptotic stability of linear neutral delay differential-algebraic equations and linear multistep methods. SIAM J. Numer. Anal. **49**, 608–618 (2011)
95. Tian, H., Yu, Q., Kuang, J.: Asymptotic stability of linear neutral delay differential-algebraic equations and Runge-Kutta methods. SIAM J. Numer. Anal. **52**, 68–82 (2014)
96. Von Wissel, D.: DAE Control of Dynamical Systems: Example of a Riderless Bicycle. Ph.D. thesis, INRIA (1996)
97. von Wissel, D., Nikoukhah, R., Campbell, S.L., Delebecque, F.: Nonlinear observer design using implicit system descriptions. In: Proceedings of Computational Engineering in Systems Applications, Lille, pp. 404–409 (1996)
98. Wächter, A., Biegler, L.T.: On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming. Math. Program. **106**, 25–57 (2006)
99. Wagner, Y.: A further index concept for linear PDAEs of hyperbolic type. Math. Comput. Simul. **53**, 287–291 (2000)
100. Walcott, B.L., Corless, M.J., Zak, S.H.: Comparative study of non-linear state-observation techniques. Int. J. Control **45**, 2109–2132 (1987)
101. Wang, L., Wang, Z., Zou, X.: Periodic solutions of neutral functional differential equations. J. Lond. Math. Soc. **65**, 439–452 (2002)
102. Wu, A.G., Duan, G.R.: Design of generalized PI observers for descriptor linear systems. IEEE Trans. Circuits Syst. I: Regul. Pap. **53**, 2828–2837 (2006)

Reachability Analysis and Deterministic Global Optimization of DAE Models

Joseph K. Scott and Paul I. Barton

Abstract This article provides a tutorial overview of recent progress in numerical methods for the reachability analysis and deterministic global optimization of DAE models. These problems are highly interrelated, and are global problems in the sense that they concern the parametric solutions of a DAE model over a potentially large range of model parameters, rather than locally about a single value. Many techniques are available for computing such global information for functions known in closed-form (i.e., factorable functions). Two of the simplest and most flexible techniques are interval arithmetic and McCormick's relaxation technique. The methods reviewed herein extend these techniques to functions defined as the solutions of DAE models, which are notably non-factorable. In doing so, we repeatedly exploit the idea that the factorable representations of the DAE governing equations, combined with insights from dynamical systems theory, can be used to infer global information about the DAE solutions. This concept is first used to derive methods for computing interval bounds and more general convex enclosures of the solutions of DAE models over a range of model parameters. Subsequently, these enclosures are employed in a branch-and-bound algorithm for deterministic global dynamic optimization with DAE's embedded. The article closes with an illustrative case study in parameter estimation and some prospects for future work.

Keywords Convex relaxations • Dynamic optimization • Global optimization • Interval methods • Reachability analysis

MSC: 93B03, 90C26, 34A09, 34A40, 34A60, 49J15, 49M37

J.K. Scott (✉)
Department of Chemical and Biomolecular Engineering, Clemson University, Clemson, SC 29609, USA
e-mail: jks9@clemson.edu

P.I. Barton
Department of Chemical Engineering, MIT, Cambridge, MA 02139, USA
e-mail: pib@mit.edu

A. Ilchmann, T. Reis (eds.), *Surveys in Differential-Algebraic Equations III*,
Differential-Algebraic Equations Forum, DOI 10.1007/978-3-319-22428-2_2

1 Introduction

Systems of differential-algebraic equations (DAEs) are used to model an incredible variety of dynamic phenomena. In the chemical process industry in particular, the numerical simulation of detailed DAE models has become a cornerstone of many core activities including process development, economic optimization, control system design, and safety analysis.

The purpose of this article is to review recent progress in numerical methods for the reachability analysis and deterministic global optimization of DAE models. These problems are highly interrelated, and are *global* problems in the sense that they concern the parametric solutions of a DAE model over a potentially large range of model parameters, rather than locally about a single value. This global information makes it possible to address a number of challenging problems in the design and control of chemical processes, thus motivating significant development of these techniques within the chemical engineering community over the past decade.

Given a DAE model, the *reachable set* at time t is defined to be the set of all states that can be reached by a solution of the DAEs at time t given initial conditions and model parameters in some specified sets. Reachability analysis generally refers to the characterization of this set. In modern process control, the computation of approximations or enclosures of reachable sets is an active area of research and finds quite extensive application. Such computations have been used, for example, for state estimation from online measurements in chemical and biological processes [54, 63, 64], feedback controller synthesis [46, 61], robust model predictive control [38, 39], and fault detection for chemical processes [33, 43]. Reachable sets also provide a means to quantify the effects of uncertainties in model parameters or inputs. A particular example of interest comes from models of chemical reaction kinetics, where the rate parameters are often only known to within an order of magnitude or worse [72, 88]. Since these models are nearly always nonlinear, the effects of such uncertainty on the model solution can be extremely difficult to infer. Reachable set enclosures have been applied in this context for uncertain chemical kinetics models [76, 88], ecology models [26, 42], and biological systems [54, 64].

A large variety of methods have been developed for computing enclosures of the reachable sets of dynamic systems. However, the vast majority of these methods apply only to systems of explicit ordinary differential equations (ODEs), rather than to DAEs, and often under further simplifying assumptions. For linear ODEs, enclosures are typically computed in the form of ellipsoids [37, 73] or polytopes [3, 14]. Many of these methods have been extended to treat nonlinear ODEs using local linearizations with a rigorous bound on the approximation error [4, 13]. However, a more efficient and widely used approach for nonlinear systems is to compute time-varying interval enclosures of the reachable set, either through interval Taylor series methods [57, 58] and their refinements [9, 31, 42], or through the solution of differential inequalities [26, 64, 76, 80, 87].

Two methods have been proposed for extending interval bounding methods for ODEs to the case of semi-explicit index-one DAEs [65, 78, 79]. In both methods, the addition of implicit algebraic equations is addressed through the use of interval Newton methods [59]. The methods differ in the treatment of the differential states, which is done using interval Taylor series methods in [65] and using differential inequalities in [78, 79]. In this article, we review the essential concepts and results from [78, 79], keeping in mind that the treatment of the algebraic equations is conceptually similar to that in [65]. Two further approaches that we do not review here can be found in [16, 28]. The method in [28] applies to implicit ODEs and could potentially be extended to treat semi-explicit DAEs. The method in [16] extends so-called level set methods for ODEs [51] to the case of semi-explicit index-one DAEs. However, methods of this type are designed to provide an accurate approximation of the reachable set, rather than a rigorous enclosure of it, and are therefore inappropriate for the applications of primary interest here.

Recently, it has been shown that generic convex enclosures of the reachable set can be characterized by convex and concave relaxations of the state variables with respect to the model parameters [75]. These relaxations are customarily used for solving global dynamic optimization problems, as discussed below. In the case of nonlinear ODEs, numerous methods for computing state relaxations have been developed over the past decade, exclusively within the chemical engineering community [62, 70, 71, 77, 82, 84, 87]. These methods require interval enclosures as input, and can therefore be thought of as refinements of these enclosures. Empirically, convex and concave state relaxations are known to provide tighter enclosures than the underlying interval methods, and have superior convergence behavior. Of the methods above, those in [82, 84] have been extended to semi-explicit index-one DAEs in [74, 81]. The main concepts of these approaches are reviewed herein.

The second problem addressed in this article is the global solution of dynamic optimization problems constrained by DAE models. Specifically, these are optimization problems in which the decision variables appear as parameters or control inputs in a DAE model, and the solution of this model in turn appears in the objective and constraints of the optimization problem. If control inputs are present among the decision variables, these problems are also commonly called optimal control problems. Dynamic optimization problems arise in a wide variety of applications. In the chemical process industry, dynamic optimization techniques are routinely used to locate optimal process designs, operating conditions, and control actions. For example, open-loop control of batch processes can be formulated as a dynamic optimization problem and has been widely studied in this context, particularly with application to high-value added industries such as specialty chemicals, pharmaceuticals, and bioprocessing [10, 47, 90]. Dynamic optimization problems also arise when considering processes with periodic dynamic behavior, such as pressure swing adsorption and simulated moving bed chromatography [19, 34]. Even for processes that are nominally operated at steady-state, several important problems require dynamic optimization, including the determination of optimal start-up and shut-down procedures [8, 12] and optimal policies for changeover from one product to

another [25]. A more fundamental application is the problem of estimating unknown parameters in a dynamic model from a given set of data [40, 55, 88]. Here, the model parameters are the decision variables, and the optimization algorithm finds those parameters which minimize the deviation of the model prediction from the measured data. This problem is extremely important, for example, for the determination of chemical reaction mechanisms from kinetic data [55, 88].

Solving dynamic optimization problems to local optimality is a very mature technology and can be done efficiently even for large and complex DAE models. The methods used in most modern codes can be classified as either *sequential* or *simultaneous*. In both approaches, any control inputs among the decision variables are first discretized through a procedure called control parameterization [93]. In the simultaneous approach, the DAEs themselves are also discretized, typically by collocation on finite elements [11, 17, 94]. This provides a representation of the state and control functions in terms of finitely many real parameters, so that the resulting optimization problem is a standard nonlinear program (NLP) on a Euclidean space with a large system of equality constraints approximating the original DAEs. This procedure makes standard methods in nonlinear programming applicable in principle, but typically generates very large-scale NLPs. On the other hand, the resulting NLPs are also highly structured, and the development of specialized interior point algorithms over the past several years has made this approach attractive for many problem classes [11].

In contrast, the sequential approach makes use of state-of-the-art dynamic simulation software to embed the DAE solution in the evaluation of the objective and constraint functions. This again leads to a standard NLP, with the caveat that the objective and constraint functions are not known explicitly as functions of the decision variables, but rather are evaluated by numerical solution of the embedded DAEs. The primary advantage of this scheme is that the resulting NLP is potentially much smaller than the NLP generated through the simultaneous approach because no decisions are introduced through discretization. On the other hand, the objective and constraint functions of this NLP, as well as their derivatives, can be costly to evaluate and have limited accuracy due to the embedded simulation. Nonetheless, this approach has become quite powerful due to the development of efficient and robust methods for numerical integration and sensitivity analysis of DAE systems [6, 24, 27, 49].

In general, local dynamic optimization algorithms require much less computational time than global algorithms. Unfortunately, local algorithms can only guarantee globally optimal solutions under restrictive convexity assumptions which are often violated in practical applications. For example, it has been shown that dynamic optimization problems arising in chemical engineering applications are very commonly nonconvex and exhibit multiple suboptimal local minima, especially when nonlinear models of chemical reaction kinetics are involved [45, 55]. The search for global solutions is well motivated in many such applications. One need only consider the problem of maximizing the profitability of a process. Clearly, a significant economic penalty may be incurred by designing and operating such a process according to a suboptimal local solution [85]. However, other

applications pose more serious problems. In parameter estimation problems, one is often interested in determining whether a model, equipped with its best fit parameter estimates, is consistent with measured data according to a statistical significance test. However, if only locally optimal parameter estimates are available, any conclusions drawn from such an analysis are dubious [52, 88].

One approach for solving dynamic optimization problems globally is the so-called dynamic programming approach, based on Bellman's principle of optimality [7]. However, this requires the solution of a boundary value problem in PDEs that becomes intractable for systems with many states [93]. Thus, more versatile algorithms have been developed as global extensions of the sequential and simultaneous approaches discussed above. Since both of these methods involve a reformulation of the original dynamic optimization problem to a standard NLP on a Euclidean space, the main idea is to combine these reformulations with the spatial branch-and-bound (B&B) global optimization framework for NLPs [67]. In the case of the simultaneous approach, this is conceptually straightforward [15, 21]. However, spatial B&B has worst-case exponential scaling in the number of decisions, which is problematic for the large-scale NLPs generated by the simultaneous approach. As a result, recent efforts in global dynamic optimization have primarily focused on the sequential approach. However, for this approach the application of spatial B&B is challenging. In particular, the objective and constraint functions in the resulting NLP are not known explicitly, but rather are defined implicitly through the solution of the embedded dynamic system, and this fact precludes the use of standard lower bounding procedures in the spatial B&B algorithm (see Sect. 7.1). In fact, establishing a valid lower bounding procedure for the sequential formulation requires a method for computing an enclosure of the reachable set, thus establishing the fundamental connection between global dynamic optimization and reachability analysis.

The first method for overcoming this problem was proposed in [20] using a lower bounding procedure based on a dynamic extension of the αBB method [1] known as βBB. However, the validity of the procedure depends on a user specified parameter, which must exceed a threshold value that is not known in general. This procedure was first made generally valid in [62] using an interval-based reachable set enclosure method. Notably, this method applies only to embedded ODE systems. In [86, 89], an alternative method was proposed based on the use of convex and concave relaxations of the state variables in the underlying reachability computation. Lin and Stadtherr proposed a further method in [41] using a sophisticated variant of the interval Taylor series reachable set enclosure methods discussed above [42]. Developments in this area are ongoing and have been primarily focused on improved state relaxation techniques for use in the spatial B&B algorithm [30, 70, 71, 82, 84]. Recently, the state relaxation methods in [82, 84] have been extended to treat semi-explicit index-one DAEs in [74, 81], making it possible to solve dynamic optimization problems with DAEs embedded to guaranteed global optimality for the first time. The essential features of this optimization algorithm are reviewed herein.

The purpose of this article is to give a self-contained review of the major concepts and tools necessary to address reachability analysis and global optimization problems for DAE systems. The problems considered here are stated formally in Sect. 2. In Sect. 3, the basic tools for computing global information about a given function, such as interval bounds and convex and concave relaxations, are presented for functions known explicitly in closed-form. The essential concepts here are the *factorable representation* of a function and its *natural interval and McCormick extensions*. These concepts are applied in Sect. 4 to obtain methods for bounding and relaxing the parametric solutions of systems of nonlinear algebraic equations. This is a direct prerequisite for treating differential-algebraic systems in subsequent sections, and is also the first example of a reoccurring theme: global information can be obtained for non-factorable functions that are defined as the solutions of equation systems with factorable governing equations. In Sects. 5 and 6, this theme is taken further to provide bounds and relaxations of the parametric solutions of DAEs. With these reachability algorithms as the enabling technology, global dynamic optimization is considered in Sect. 7. A numerical case study is presented in Sect. 8, and concluding remarks are given in Sect. 9.

Finally, we note that this review is intended as a tutorial overview of the specific results of several recent papers by the authors, rather than a high level survey of a large body of literature. This is primarily because the field is nascent, and with few exceptions noted above, the material reviewed here is the only available approach. At the same time, alternative approaches for problems with ODEs that seem likely to have fruitful extensions to DAEs are based on such similar foundational concepts that this review should serve as a useful introductory reference for them as well. In their original published form, the results and methods reviewed herein are presented in a highly technical manner, and are distributed among several papers in such a way that their synthesis into complete algorithms is difficult to appreciate. Thus, our intent is to provide a clear but thorough introduction that will encourage and direct further research into reachability and global optimization problems with nonlinear DAE models.

2 Problem Formulation

2.1 Notation

In the remainder of this article, vector quantities are denoted in bold, while scalar quantities are written without emphasis. For any $\mathbf{v} \in \mathbb{R}^n$, the standard p-norms are denoted by $\|\mathbf{v}\|_p = \left(\sum_{i=1}^n |v_i|^p\right)^{1/p}$, $1 \leq p < \infty$, and $\|\mathbf{v}\|_\infty = \max_i |v_i|$. For $\mathbf{v}, \mathbf{w}, \mathbf{u} \in \mathbb{R}^n$, the order relations $\mathbf{v} \leq \mathbf{w}$ and $\mathbf{v} < \mathbf{w}$ denote that these relations hold component-wise. Similarly, $\min(\mathbf{v}, \mathbf{w})$ and $\max(\mathbf{v}, \mathbf{w})$ denote the vectors with components $\min(v_i, w_i)$ and $\max(v_i, w_i)$, respectively, and $\mathrm{mid}(\mathbf{v}, \mathbf{w}, \mathbf{u})$ denotes the vector where each component is the middle value of v_i, w_i, and u_i. Let $C^k(D, \mathbb{R}^m)$

denote the set of k-times continuously differentiable functions from D into $\mathbb{R}^m$. For $D_s \subset \mathbb{R}^{n_s}$, $D_r \subset \mathbb{R}^{n_r}$, and $\boldsymbol{\ell} \in C^k(D_s \times D_r, \mathbb{R}^m)$, let $\frac{\partial \boldsymbol{\ell}}{\partial \mathbf{r}}(\hat{\mathbf{s}}, \hat{\mathbf{r}})$ denote the Jacobian matrix of $\boldsymbol{\ell}(\hat{\mathbf{s}}, \cdot)$ at $\hat{\mathbf{r}} \in D_r$.

2.2 Semi-explicit Index-One DAEs

This article considers exclusively the initial value problem in semi-explicit index-one DAEs

$$\left.\begin{aligned} \dot{\mathbf{x}}(t, \mathbf{p}) &= \mathbf{f}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p})) \\ \mathbf{0} &= \mathbf{g}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p})) \end{aligned}\right\}, \tag{2.1a}$$

$$\mathbf{x}(t_0, \mathbf{p}) = \mathbf{x}_0(\mathbf{p}), \tag{2.1b}$$

where t is the independent variable, $\mathbf{p}$ is the vector of problem parameters, $\dot{\mathbf{x}}(t, \mathbf{p})$ denotes the derivative of $\mathbf{x}(\cdot, \mathbf{p})$ at t, t_0 is the initial time, and $\mathbf{x}_0$ specifies the parametric initial conditions. It is assumed that $\mathbf{f} \in C^1(D_t \times D_p \times D_x \times D_y, \mathbb{R}^{n_x})$, $\mathbf{g} \in C^1(D_t \times D_p \times D_x \times D_y, \mathbb{R}^{n_y})$, and $\mathbf{x}_0 \in C^1(D_p, D_x)$, where $D_t \subset \mathbb{R}$, $D_p \subset \mathbb{R}^{n_p}$, $D_x \subset \mathbb{R}^{n_x}$, and $D_y \subset \mathbb{R}^{n_y}$ are open sets. A function $(\mathbf{x}, \mathbf{y}) \in C^1(I \times P, D_x) \times C^1(I \times P, D_y)$ is called a *solution of* (2.1) *on* $I \times P \subset D_t \times D_p$ if (2.1) holds for all $(t, \mathbf{p}) \in I \times P$.

The special form of the DAEs (2.1) is called the *semi-explicit* form. In addition to considering this special form, we will further restrict our attention to index-one, or *regular* solutions (although it is customary to refer to a DAE *system* as index-one or high-index, the differential index is actually a property of solutions). A solution $(\mathbf{x}, \mathbf{y}) \in C^1(I \times P, D_x) \times C^1(I \times P, D_y)$ is called an index-one solution, or a regular solution, if

$$\det \frac{\partial \mathbf{g}}{\partial \mathbf{y}}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p})) \neq 0, \quad \forall (t, \mathbf{p}) \in I \times P. \tag{2.2}$$

Note that, for any regular solution of (2.1) on $I \times P$, a single differentiation of the algebraic equations $\mathbf{g}$ gives the underlying ODEs

$$\dot{\mathbf{x}}(t, \mathbf{p}) = \mathbf{f}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p})), \tag{2.3}$$

$$\dot{\mathbf{y}}(t, \mathbf{p}) = -\left(\frac{\partial \mathbf{g}}{\partial \mathbf{y}}\right)^{-1} \left(\frac{\partial \mathbf{g}}{\partial \mathbf{x}} \mathbf{f}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p})) + \frac{\partial \mathbf{g}}{\partial t}\right), \tag{2.4}$$

for all $(t, \mathbf{p}) \in I \times P$, where all derivatives of $\mathbf{g}$ are evaluated at $(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p}))$.

The DAEs (2.1) have the following existence and uniqueness properties (see Theorems 4.13 and 4.18 in [36]):

- Given any point $(t_0, \hat{\mathbf{p}}, \hat{\mathbf{x}}_0, \hat{\mathbf{y}}_0)$ such that $\mathbf{x}_0(\hat{\mathbf{p}}) = \hat{\mathbf{x}}_0$, $\mathbf{g}(t_0, \hat{\mathbf{p}}, \hat{\mathbf{x}}_0, \hat{\mathbf{y}}_0) = \mathbf{0}$, and $\det \frac{\partial \mathbf{g}}{\partial \mathbf{y}}(t_0, \hat{\mathbf{p}}, \hat{\mathbf{x}}_0, \hat{\mathbf{y}}_0) \neq 0$, there exists a regular solution $(\mathbf{x}, \mathbf{y})$ of (2.1) defined on some sufficiently small open ball around $(t_0, \hat{\mathbf{p}})$ and satisfying $(\mathbf{x}(t_0, \hat{\mathbf{p}}), \mathbf{y}(t_0, \hat{\mathbf{p}})) = (\hat{\mathbf{x}}_0, \hat{\mathbf{y}}_0)$.
- If $(\mathbf{x}, \mathbf{y})$ and $(\mathbf{x}^*, \mathbf{y}^*)$ are solutions of (2.1) on $I \times P$, $(\mathbf{x}, \mathbf{y})$ is regular, P is connected, and there exists $\hat{\mathbf{p}} \in P$ such that $\mathbf{y}(t_0, \hat{\mathbf{p}}) = \mathbf{y}^*(t_0, \hat{\mathbf{p}})$, then $\mathbf{x}(t, \mathbf{p}) = \mathbf{x}^*(t, \mathbf{p})$ and $\mathbf{y}(t, \mathbf{p}) = \mathbf{y}^*(t, \mathbf{p})$, $\forall (t, \mathbf{p}) \in I \times P$.

In order to guarantee that (2.1) has a unique regular solution, the conditions above suggest that we must add a further specification on the initial condition of the form $\mathbf{y}(t_0, \hat{\mathbf{p}}) = \hat{\mathbf{y}}_0$. Indeed, without this condition, it is possible for (2.1) to have multiple solutions, even if both solutions are regular. The regularity only implies that any two such solutions must be completely isolated from one another (i.e., they cannot both satisfy $\mathbf{y}(t_0, \hat{\mathbf{p}}) = \hat{\mathbf{y}}_0$). The following example illustrates the need for this condition.

Example 2.1 Let $I \equiv [0, \delta] \subset D_t = \mathbb{R}$, $D_p = \emptyset$, $D_x = D_y = \mathbb{R}$, and define $g(t, z_x, z_y) = z_y^2 - z_x$. With fixed initial condition $x_0 = 1$ at $t_0 = 0$, there are two possible values for $y(t_0)$ satisfying $g(t_0, x(t_0), y(t_0)) = 0$; $y(t_0) = 1$ and $y(t_0) = -1$. Letting $f(t, z_x, z_y) = 1$, clearly $x(t) = 1 + t$ satisfies $\dot{x}(t) = 1 = f(t, x(t), y(t))$ for any $y : I \to \mathbb{R}$. However, both $y(t) = \sqrt{1+t}$ and $y(t) = -\sqrt{1+t}$ result in $g(t, x(t), y(t)) = (y(t))^2 - x(t) = 0$. In particular, $y(t) = \sqrt{1+t}$ satisfies (2.1) with the additional specification $y(t_0) = 1$, while $y(t) = -\sqrt{1+t}$ satisfies (2.1) along with the additional specification $y(t_0) = -1$.

2.3 Reachable Set Enclosures

This article considers two types of enclosures of the reachable set of (2.1). The first type of enclosure, discussed in Sect. 5, is an interval enclosure described by component-wise upper and lower bounds on each state. In particular, let $I = [t_0, t_f]$, let P be a compact, convex set, and let $(\mathbf{x}, \mathbf{y})$ be a regular solution of (2.1) on $I \times P$. Then, our objective is to compute functions $\mathbf{x}^L, \mathbf{x}^U : I \to \mathbb{R}^{n_x}$ and $\mathbf{y}^L, \mathbf{y}^U : I \to \mathbb{R}^{n_y}$ such that

$$\mathbf{x}^L(t) \leq \mathbf{x}(t, \mathbf{p}) \leq \mathbf{x}^U(t) \quad \text{and} \quad \mathbf{y}^L(t) \leq \mathbf{y}(t, \mathbf{p}) \leq \mathbf{y}^U(t), \quad \forall (t, \mathbf{p}) \in I \times P. \tag{2.5}$$

These functions are referred to as *state bounds* for the solution $(\mathbf{x}, \mathbf{y})$. An interesting feature of the methods discussed is that the existence of a unique regular solution satisfying the given initial data need not be assumed; it can be verified computationally by the bounding method.

The second type of reachable set enclosure, discussed in Sect. 6, is characterized by convex and concave relaxations. Recall that, for $D \subset \mathbb{R}^n$ convex, a vector function $\mathbf{w} : D \to \mathbb{R}^m$ is called convex if each component is convex; i.e.,

$$\mathbf{w}(\lambda \mathbf{z}_1 + (1-\lambda)\mathbf{z}_2) \leq \lambda \mathbf{w}(\mathbf{z}_1) + (1-\lambda)\mathbf{w}(\mathbf{z}_2), \quad \forall (\lambda, \mathbf{z}_1, \mathbf{z}_2) \in [0,1] \times D \times D,$$

and it is called concave if the opposite (weak) inequality holds. For arbitrary $\mathbf{w} : D \to \mathbb{R}^m$ with D convex, the functions $\mathbf{w}^{cv}, \mathbf{w}^{cc} : D \to \mathbb{R}^m$ are called *convex and concave relaxations for* $\mathbf{w}$ *on* D, respectively, if $\mathbf{w}^{cv}$ is convex on D, $\mathbf{w}^{cc}$ is concave on D, and

$$\mathbf{w}^{cv}(\mathbf{z}) \leq \mathbf{w}(\mathbf{z}) \leq \mathbf{w}^{cc}(\mathbf{z}), \quad \forall \mathbf{z} \in D. \tag{2.6}$$

Returning to (2.1), given a regular solution $(\mathbf{x}, \mathbf{y})$ on $I \times P$, our objective is to compute functions $\mathbf{x}^{cv}, \mathbf{x}^{cc} : I \times P \to \mathbb{R}^{n_x}$ and $\mathbf{y}^{cv}, \mathbf{y}^{cc} : I \times P \to \mathbb{R}^{n_y}$ such that

$$\mathbf{x}^{cv}(t,\mathbf{p}) \leq \mathbf{x}(t,\mathbf{p}) \leq \mathbf{x}^{cc}(t,\mathbf{p}) \quad \text{and} \quad \mathbf{y}^{cv}(t,\mathbf{p}) \leq \mathbf{y}(t,\mathbf{p}) \leq \mathbf{y}^{cc}(t,\mathbf{p}), \tag{2.7}$$

for all $(t,\mathbf{p}) \in I \times P$, and, for each $t \in I$, $\mathbf{x}^{cv}(t,\cdot)$ and $\mathbf{y}^{cv}(t,\cdot)$ are convex on P, and $\mathbf{x}^{cc}(t,\cdot)$ and $\mathbf{y}^{cc}(t,\cdot)$ are concave on P. These functions are called *state relaxations for* $(\mathbf{x},\mathbf{y})$ *on* $I \times P$. In Sect. 6, it will be shown that these relaxations describe a convex enclosure of the reachable set that can be outer-approximated by a polytope of any desired complexity [75]. State relaxations often provide a tighter enclosure of the reachable set than do state bounds. Moreover, state relaxations are better suited for use in a global dynamic optimization algorithm, as will be shown in Sect. 7.

2.4 Global Dynamic Optimization

The second problem considered in this article is the global solution of the dynamic optimization problem

$$\begin{aligned} \min_{\mathbf{p} \in P} \quad & \phi(\mathbf{p}, \mathbf{x}(t_f,\mathbf{p}), \mathbf{y}(t_f,\mathbf{p})) \\ \text{s.t.} \quad & \boldsymbol{\eta}(\mathbf{p}, \mathbf{x}(t_f,\mathbf{p}), \mathbf{y}(t_f,\mathbf{p})) \leq \mathbf{0}. \end{aligned} \tag{2.8}$$

Above, ϕ and $\boldsymbol{\eta}$ are continuous functions of the form $(\phi, \boldsymbol{\eta}) : D_p \times D_x \times D_y \to \mathbb{R} \times \mathbb{R}^{n_c}$, and $(\mathbf{x},\mathbf{y})$ satisfies (2.1) on $I \times P$, along with the additional initial condition

$$\mathbf{y}(t_0, \hat{\mathbf{p}}) = \hat{\mathbf{y}}_0, \tag{2.9}$$

where $\hat{\mathbf{p}} \in P$ and $\hat{\mathbf{y}}_0 \in D_y$ satisfy $\mathbf{g}(t_0, \hat{\mathbf{p}}, \mathbf{x}_0(\hat{\mathbf{p}}), \hat{\mathbf{y}}_0) = \mathbf{0}$ and $\det \frac{\partial \mathbf{g}}{\partial \mathbf{y}}(t_0, \hat{\mathbf{p}}, \mathbf{x}_0(\hat{\mathbf{p}}), \hat{\mathbf{y}}_0) \neq 0$. Note that $\mathbf{x}$ and $\mathbf{y}$ are well-defined functions of $(t, \mathbf{p}) \in I \times P$ only if the solution of (2.1) with (2.9) exists and is unique for all $(t, \mathbf{p}) \in I \times P$. As discussed in Sect. 2.2, the condition (2.9) ensures local existence. For (2.8) to be well-posed, it is assumed that a solution exists on all of $I \times P$. Given this assumption, (2.9) ensures uniqueness on $I \times P$. In most applications, there is a consistent initial condition of interest, so that the specification (2.9) is easily made. On the other hand, if one is interested in an optimization problem that considers all possible solutions of (2.1), then some additional method will be required for exhaustively enumerating such solutions. No such method has yet been proposed in the literature.

As discussed in Sect. 1, there are many algorithms available for solving (2.8) to local optimality. In contrast, our concern here is with algorithms that are guaranteed to terminate finitely with an ϵ-global optimum $\mathbf{p}^*$. In particular, for a user specified $\epsilon > 0$, $\mathbf{p}^*$ satisfies $\boldsymbol{\eta}(\mathbf{p}^*, \mathbf{x}(t_f, \mathbf{p}^*), \mathbf{y}(t_f, \mathbf{p}^*)) \leq \mathbf{0}$ and

$$\phi(\mathbf{p}^*, \mathbf{x}(t_f, \mathbf{p}^*), \mathbf{y}(t_f, \mathbf{p}^*)) \leq \phi(\mathbf{p}, \mathbf{x}(t_f, \mathbf{p}), \mathbf{y}(t_f, \mathbf{p})) + \epsilon, \tag{2.10}$$

for all $\mathbf{p} \in P$ such that $\boldsymbol{\eta}(\mathbf{p}, \mathbf{x}(t_f, \mathbf{p}), \mathbf{y}(t_f, \mathbf{p})) \leq \mathbf{0}$.

Note that (2.8) considers only optimization problems with a real vector of decision variables $\mathbf{p}$. Although many dynamic optimization problems take this form (e.g., parameter estimation), there are also many problems of practical interest involving continuous control inputs $u(t)$ as decisions. For such problems, the methods discussed here can be used only after applying *control parameterization*. Control parameterization refers to the approximation of the control inputs by functions that can be characterized by a finite number of real parameters (e.g., polynomials, piece-wise affine controls, etc.). In most cases, control parameterization can be done with minimal error, and is common practice in state-of-the-art dynamic optimization algorithms [93]. Nonetheless, the reader should note that the guarantee of global optimality provided by the method discussed herein applies to the parameterized problem (2.8), and any inferences about the original problem are subject to the parameterization error. A method for overcoming this limitation in the case of ODE embedded problems has recently been proposed in [30].

Without yet considering the details of solving (2.8) globally, it is possible to intuitively appreciate the relationship between (2.8) and reachability analysis. In short, a global solver must provide a guarantee that the located solution $\mathbf{p}^*$ minimizes the function $\phi(\cdot, \mathbf{x}(t_f, \cdot), \mathbf{y}(t_f, \cdot))$ among all feasible $\mathbf{p} \in P$. Clearly such a guarantee cannot be made unless one has characterized all possible values of $\mathbf{x}(t_f, \cdot)$ and $\mathbf{y}(t_f, \cdot)$ that can be obtained with $\mathbf{p} \in P$, which is exactly the reachability problem discussed in the previous section.

The problem (2.8) admits several generalizations that we have omitted for notational convenience:

1. Integral terms can be included in the objective and constraints; i.e.,

$$\begin{aligned} \min_{\mathbf{p}\in P} \quad & \phi(\mathbf{p}, \mathbf{x}(t_f, \mathbf{p}), \mathbf{y}(t_f, \mathbf{p})) + \int_{t_0}^{t_f} \psi(s, \mathbf{p}, \mathbf{x}(s, \mathbf{p}), \mathbf{y}(s, \mathbf{p}))ds \\ \text{s.t.} \quad & \boldsymbol{\eta}(\mathbf{p}, \mathbf{x}(t_f, \mathbf{p}), \mathbf{y}(t_f, \mathbf{p})) + \int_{t_0}^{t_f} \boldsymbol{\ell}(s, \mathbf{p}, \mathbf{x}(s, \mathbf{p}), \mathbf{y}(s, \mathbf{p}))ds \leq \mathbf{0}. \end{aligned} \tag{2.11}$$

   Problems of this type can always be recast in the form of (2.8) by introducing quadrature variables as described in [74].
2. The objective and constraints may contain values of the states at multiple time points $t_0, \ldots, t_m \in I$:

$$\begin{aligned} \min_{\mathbf{p}\in P} \quad & \phi(\mathbf{p}, \mathbf{x}(t_0, \mathbf{p}), \ldots, \mathbf{x}(t_m, \mathbf{p}), \mathbf{y}(t_0, \mathbf{p}), \ldots, \mathbf{y}(t_m, \mathbf{p})) \\ \text{s.t.} \quad & \boldsymbol{\eta}(\mathbf{p}, \mathbf{x}(t_0, \mathbf{p}), \ldots, \mathbf{x}(t_m, \mathbf{p}), \mathbf{y}(t_0, \mathbf{p}), \ldots, \mathbf{y}(t_m, \mathbf{p})) \leq \mathbf{0}. \end{aligned} \tag{2.12}$$

   The algorithm for solving (2.8) presented in Sect. 7 is easily extended to this case. The restriction to final time terms only simplifies the notation.

3 Factorable Functions, Interval Arithmetic, and McCormick Relaxations

Computing interval bounds and/or convex and concave relaxations of a given function requires global information about that function. In general, local characterizations of a function, such as its value or its derivative at a point, are not enough. Rather, one requires information about the behavior of the function on the entire domain of interest. An essential tool in this regard is the so-called *factorable representation* of a function [50, 56]. Indeed, the primary complication in computing the state bounds and state relaxations defined in Sect. 2.3 is that the parametric solutions $\mathbf{x}(t, \cdot)$ and $\mathbf{y}(t, \cdot)$ are not known in closed-form, but rather are evaluated by numerical integration. Because of this, they have no known factorable representation to work from. Nonetheless, factorable representations, as well as the methods for computing bounds and relaxations of them, will be central to the state bounding and relaxation methods for DAEs presented in the following sections. Therefore, these concepts are developed in detail in this section.

Informally, a function is called *factorable* if it can be written as a finite sequence of simple operations, including basic arithmetic operations as well as intrinsic functions available on a computer. For example, the function

$$h(s_1, s_2) = 10s_1 + s_2 e^{s_2} \tag{3.1}$$

is factorable because it can be evaluated for any $(s_1, s_2) \in \mathbb{R}^2$ by executing the following sequence of simple computations:

$$\begin{aligned} v_1(s_1, s_2) &= s_1, \\ v_2(s_1, s_2) &= s_2, \\ v_3(s_1, s_2) &= 10 v_1(s_1, s_2), \\ v_4(s_1, s_2) &= \exp(v_2(s_1, s_2)), \\ v_5(s_1, s_2) &= v_2(s_1, s_2) \times v_4(s_1, s_2), \\ v_6(s_1, s_2) &= v_3(s_1, s_2) + v_5(s_1, s_2), \\ h(s_1, s_2) &= v_6(s_1, s_2). \end{aligned}$$

Each of the intermediates v_i is called a factor, and the factorable representation is the sequence $v_1, \ldots, v_6$. In essence, any function written explicitly in computer code is factorable.

To formalize the notion of a factorable function, we must first define the set of operations that will be permissible in the sequence of computations defining such functions. This set will contain binary addition, binary multiplication, and elements of a *library of univariate functions* $u : B \subset \mathbb{R} \to \mathbb{R}$, denoted by $\mathscr{L}$. The elements of $\mathscr{L}$ will be used to represent functions such as $\sqrt{s}$, s^n, e^s, $\sin s$, etc. Furthermore, $\mathscr{L}$ should include the negative and reciprocal functions $-s$ and $1/s$, so that subtraction and division can be achieved by combination with binary addition and multiplication. Any univariate function of interest may be included in $\mathscr{L}$, provided that certain information related to bounding and relaxing u is available (see Assumptions 3.1–3.2). This information has been collected for a large number of common univariate operations in [83].

Definition 3.1 A function $\mathbf{h} : D_s \subset \mathbb{R}^n \to \mathbb{R}^m$ is *factorable* if it can be expressed in terms of a finite number of factors $v_1, \ldots, v_q : D_s \to \mathbb{R}$ such that $v_i(\mathbf{s}) = s_i$ for $i = 1, \ldots, n$, $v_k(\mathbf{s})$ is defined for each $n < k \leq q$ as either

(a) $v_k(\mathbf{s}) = v_i(\mathbf{s}) + v_j(\mathbf{s})$, with $i, j < k$, or
(b) $v_k(\mathbf{s}) = v_i(\mathbf{s}) v_j(\mathbf{s})$, with $i, j < k$, or
(c) $v_k(\mathbf{s}) = u_k(v_i(\mathbf{s}))$, with $i < k$, $u_k \in \mathscr{L}$,

and $\mathbf{h}(\mathbf{s}) = (v_{i(1)}(\mathbf{s}), \ldots, v_{i(m)}(\mathbf{s}))$ for some indices $i(1), \ldots, i(m) \in \{1, \ldots, q\}$.

3.1 Interval Arithmetic

For $a, b \in \mathbb{R}$, $a \leq b$, define the *interval* $[a, b]$ as the compact, connected set $\{x \in \mathbb{R} : a \leq x \leq b\}$. The set of all nonempty intervals is denoted $\mathbb{IR}$. Intervals are denoted by capital letters, $S \in \mathbb{IR}$. Since S is a subset of $\mathbb{R}$, the notation $s \in S$ is

well-defined. The set of n-dimensional interval vectors is denoted $\mathbb{IR}^n$. In particular, $S \in \mathbb{IR}^n$ has elements $S_i \in \mathbb{IR}$, $i = 1, \ldots, n$. Every $S \in \mathbb{IR}^n$ can be regarded as a subset of $\mathbb{R}^n$ defined by the Cartesian product $S_1 \times \ldots \times S_n$, so that $\mathbf{s} \in \mathbb{R}^n$ satisfies $\mathbf{s} \in S$ if $s_i \in S_i$, $i = 1, \ldots, n$. The set of $n \times m$ interval matrices is denoted $\mathbb{IR}^{n \times m}$ and defined analogously to $\mathbb{IR}^n$; $A \in \mathbb{IR}^{n \times m}$ has elements $A_{ij} \in \mathbb{IR}$, for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$, and, for any $\mathbf{A} \in \mathbb{R}^{n \times m}$ with elements a_{ij}, $\mathbf{A} \in A$ if $a_{ij} \in A_{ij}$ for all indices i and j. For any $D \subset \mathbb{R}^n$, let $\mathbb{I}D$ denote the set $\{S \in \mathbb{IR}^n : S \subset D\}$. This notation is also used for $D \subset \mathbb{R}^{n \times m}$.

If $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ and $\mathbf{v} \leq \mathbf{w}$, then $[\mathbf{v}, \mathbf{w}]$ denotes the n-dimensional interval $[v_1, w_1] \times \ldots \times [v_n, w_n]$. Moreover, for any $S \in \mathbb{IR}$, the notation $\mathbf{s}^L, \mathbf{s}^U \in \mathbb{R}^n$ will be commonly used to denote the vectors such that $S = [\mathbf{s}^L, \mathbf{s}^U]$. The notation $m(S)$ denotes the *midpoint* of S, $m(S) \equiv 0.5(\mathbf{s}^L + \mathbf{s}^U)$, and $w(S)$ denotes the *width* of S, $w(S) \equiv \mathbf{s}^U - \mathbf{s}^L$. For $A \in \mathbb{IR}^{n \times m}$, $m(A)$ and $w(A)$ are real-valued matrices defined analogously. For any $\mathbf{s} \in \mathbb{R}^n$, the singleton $[\mathbf{s}, \mathbf{s}]$ is called a *degenerate* interval.

The central task in interval analysis is to compute an interval which encloses the range of a given function [56].

Definition 3.2 Let $\mathbf{h} : D_s \subset \mathbb{R}^n \to \mathbb{R}^m$. An interval mapping $H : \mathbb{I}D_s \to \mathbb{IR}^m$ is an *inclusion function* for $\mathbf{h}$ if $\mathbf{h}(S) \subset H(S)$, $\forall S \in \mathbb{I}D_s$, where $\mathbf{h}(S)$ is the image of S under $\mathbf{h}$.

Typically, inclusion functions are derived from a simpler object known as an *interval extension*. The function H is an *interval extension of* $\mathbf{h}$ if, for every $\mathbf{s} \in D_s$, $H([\mathbf{s}, \mathbf{s}]) = [\mathbf{h}(\mathbf{s}), \mathbf{h}(\mathbf{s})]$. It is *inclusion monotonic* if

$$S_1 \subset S_2 \implies H(S_1) \subset H(S_2), \quad \forall S_1, S_2 \in \mathbb{I}D_s. \tag{3.2}$$

The following theorem is a central result in interval analysis:

Theorem 3.1 *Let* $\mathbf{h} : D_s \subset \mathbb{R}^n \to \mathbb{R}^m$. *If* $H : \mathbb{I}D_s \to \mathbb{IR}^m$ *is an inclusion monotonic interval extension of* $\mathbf{h}$, *then it is an inclusion function for* $\mathbf{h}$.

Inclusion monotonic interval extensions for binary addition and multiplication are given by the formulas

$$S + Q = [s^L + q^L, s^U + q^U], \tag{3.3}$$

$$SQ = [\min(s^L q^L, s^L q^U, s^U q^L, s^U q^U), \max(s^L q^L, s^L q^U, s^U q^L, s^U q^U)], \tag{3.4}$$

where $S = [s^L, s^U]$ and $Q = [q^L, q^U]$. Moreover, formulas are readily available for many common univariate functions [56, 59]. We make the following assumption throughout:

Assumption 3.1 *For every* $u \in \mathscr{L}$, $u : B \subset \mathbb{R} \to \mathbb{R}$, *an inclusion monotonic interval extension* $[u] : \mathbb{I}B \to \mathbb{IR}$ *is available and can be evaluated computationally.*

For any factorable function $\mathbf{h}$, one can compute a particular interval extension called the *natural interval extension* and denoted by $[\mathbf{h}] : \mathbb{I}D_s \to \mathbb{IR}^m$ (the notations

$[\mathbf{h}]^L(S)$ and $[\mathbf{h}]^U(S)$ are used to denote the lower and upper bounds of $[\mathbf{h}](S)$, respectively). The natural interval extension is computed by recursively applying the known interval extensions of the factors of $\mathbf{h}$. That is, each operation in the definition of $\mathbf{h}$ (Definition 3.1) is replaced by its interval counterpart. For example, if h is defined by (3.1), then

$$[h](S_1, S_2) = 10S_1 + S_2[\exp](S_2), \tag{3.5}$$

$$= [10s_1^L, 10s_1^U] + [s_2^L, s_2^U][e^{s_2^L}, e^{s_2^U}], \tag{3.6}$$

$$= [10s_1^L, 10s_1^U] + [\min(s_2^L e^{s_2^L}, s_2^L e^{s_2^U}, s_2^U e^{s_2^L}, s_2^U e^{s_2^U}), \tag{3.7}$$

$$\max(s_2^L e^{s_2^L}, s_2^L e^{s_2^U}, s_2^U e^{s_2^L}, s_2^U e^{s_2^U})],$$

$$= [10s_1^L + \min(s_2^L e^{s_2^L}, s_2^L e^{s_2^U}, s_2^U e^{s_2^L}, s_2^U e^{s_2^U}), \tag{3.8}$$

$$10s_1^U + \max(s_2^L e^{s_2^L}, s_2^L e^{s_2^U}, s_2^U e^{s_2^L}, s_2^U e^{s_2^U})]. \tag{3.9}$$

The natural interval extension of a factorable function is inclusion monotonic, and therefore defines an inclusion function. The reader is referred to [56, 59] for further details on interval analysis.

3.2 McCormick Relaxations

McCormick's relaxation technique [50] is a method for computing convex and concave relaxations of a given factorable function $\mathbf{h} : D_s \subset \mathbb{R}^n \to \mathbb{R}^m$. Given an interval $S = [\mathbf{s}^L, \mathbf{s}^U] \subset D_s$ and a point $\mathbf{s} \in S$, the method computes three quantities associated with each factor v_k in Definition 3.1: $V_k(S)$, $v_k^{cv}(S, \mathbf{s})$, and $v_k^{cc}(S, \mathbf{s})$. $V_k(S)$ is an interval bound for v_k on S computed using interval arithmetic. The remaining numbers $v_k^{cv}(S, \mathbf{s})$ and $v_k^{cc}(S, \mathbf{s})$ are the values of convex and concave relaxations of v_k on S, respectively, evaluated at $\mathbf{s}$. For every $k \leq n$, we have $v_k = s_k$, so these quantities can be trivially assigned as $(V_k(S), v_k^{cv}(S, \mathbf{s}), v_k^{cc}(S, \mathbf{s})) = (S_k, s_k, s_k)$. For every successive factor v_k, relaxations and bounds are computed recursively using known rules based on the definition of v_k in Definition 3.1. For example, if $v_k(\mathbf{s}) = v_i(\mathbf{s}) + v_j(\mathbf{s})$, we assign

$$V_k(S) = V_i(S) + V_j(S), \tag{3.10}$$

$$v_k^{cv}(S, \mathbf{s}) = \max(V_i^L(S), v_i^{cv}(S, \mathbf{s})) + \max(V_j^L(S), v_j^{cv}(S, \mathbf{s})), \tag{3.11}$$

$$v_k^{cc}(S, \mathbf{s}) = \min(V_i^U(S), v_i^{cc}(S, \mathbf{s})) + \min(V_j^U(S), v_j^{cc}(S, \mathbf{s})). \tag{3.12}$$

It is straightforward to see that $v_k(\mathbf{s}) \in [v_k^{cv}(S, \mathbf{s}), v_k^{cc}(S, \mathbf{s})]$, $\forall \mathbf{s} \in S$, and convexity of $v_k^{cv}(S, \cdot)$ (resp. concavity of $v_k^{cc}(S, \cdot)$) follows from the well-known results that

the minimum (resp. maximum) of two convex (resp. concave) functions is convex (resp. concave) and the sum of two convex (resp. concave) functions is convex (resp. concave). Rules for multiplication and composition with many common univariate functions are readily available [83]. Noting that $\mathbf{h}(\mathbf{s}) = (v_{i(1)}(\mathbf{s}), \ldots, v_{i(m)}(\mathbf{s}))$, this procedure provides convex and concave relaxations for $\mathbf{h}$ on S through the definitions $h_j^{cv}(\mathbf{s}) = v_{i(j)}^{cv}(S, \mathbf{s})$ and $h_j^{cc}(\mathbf{s}) = v_{i(j)}^{cc}(S, \mathbf{s}), j = 1, \ldots, m$.

On account of the initialization $(V_k(S), v_k^{cv}(S, \mathbf{s}), v_k^{cc}(S, \mathbf{s})) = (S_k, s_k, s_k)$, $k \leq n$, McCormick's relaxation procedure can be thought of as a mapping of the form $(S, \mathbf{s}) \mapsto (H(S), \mathbf{h}^{cv}(\mathbf{s}), \mathbf{h}^{cc}(\mathbf{s}))$. However, in [83], it was observed that generalizing this initialization to $(V_k(S), v_k^{cv}(S, \mathbf{s}), v_k^{cc}(S, \mathbf{s})) = (S_k, s_k^{cv}, s_k^{cc})$, $k \leq n$, where the inputs satisfy $S \cap [\mathbf{s}^{cv}, \mathbf{s}^{cc}] \neq \emptyset$, results in a generalized mapping $(S, \mathbf{s}^{cv}, \mathbf{s}^{cc}) \mapsto (H(S), \mathbf{u}(\mathbf{s}^{cv}, \mathbf{s}^{cc}), \mathbf{o}(\mathbf{s}^{cv}, \mathbf{s}^{cc}))$ with an interesting interpretation: $\mathbf{u}$ and $\mathbf{o}$ are *composite relaxations for* $\mathbf{h}$ *on* S.

Definition 3.3 Let $\mathbf{h} : D_s \subset \mathbb{R}^n \to \mathbb{R}^m$ and let $S \subset D_s$ be convex. Two functions $\mathbf{u}, \mathbf{o} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^m$ are called convex and concave composite relaxations for $\mathbf{h}$ on S, respectively, if the following condition holds: For any convex set $P \subset \mathbb{R}^q$ and any functions $\mathbf{s}, \mathbf{s}^{cv}, \mathbf{s}^{cc} : P \to \mathbb{R}^n$ such that $\mathbf{s}(\mathbf{p}) \in S$, $\forall \mathbf{p} \in P$, and $\mathbf{s}^{cv}$ and $\mathbf{s}^{cc}$ are, respectively, convex and concave relaxations of $\mathbf{s}$ on P, convex and concave relaxations of the composite mapping $\mathbf{h}(\mathbf{s}(\cdot))$ are given by $\mathbf{u}(\mathbf{s}^{cv}(\cdot), \mathbf{s}^{cc}(\cdot))$ and $\mathbf{o}(\mathbf{s}^{cv}(\cdot), \mathbf{s}^{cc}(\cdot))$, respectively.

Composite relaxations are central to the relaxation theory for DAEs presented in Sect. 6. Given a function $\boldsymbol{\psi} : P \to \mathbb{R}^m$ defined by $\boldsymbol{\psi}(\mathbf{p}) \equiv \mathbf{h}(\mathbf{s}(\mathbf{p}))$ with $\mathbf{s} : P \to D_s$, composite relaxations of $\mathbf{h}$ provide a means to compute relaxations for $\boldsymbol{\psi}$ on P given bounds and relaxations for $\mathbf{s}$ on P. Note that this requires factorability of $\mathbf{h}$, but not of $\boldsymbol{\psi}$. The notion of composite relaxations can also be used to treat the case where $\mathbf{h} : P \times D_s \to \mathbb{R}^m$ and $\boldsymbol{\psi}(\mathbf{p}) \equiv \mathbf{h}(\mathbf{p}, \mathbf{s}(\mathbf{p}))$. By initializing the factors corresponding to the arguments $\mathbf{p}$ as in the standard relaxation procedure and using the generalized initialization for the factors corresponding to $\mathbf{s}$, one obtains a mapping of the form $(S, \mathbf{p}, \mathbf{s}^{cv}, \mathbf{s}^{cc}) \mapsto (H(S), \mathbf{u}(\mathbf{p}, \mathbf{s}^{cv}, \mathbf{s}^{cc}), \mathbf{o}(\mathbf{p}, \mathbf{s}^{cv}, \mathbf{s}^{cc}))$ with the obvious properties.

Since McCormick's relaxation technique is based on the recursive application of relaxation rules for a set of basic operations, it is convenient to denote the procedure explicitly for a given function, e.g., for (3.1) as

$$\{h\}(\mathscr{S}_1, \mathscr{S}_2) = 10\mathscr{S}_1 + \mathscr{S}_2\{\exp\}(\mathscr{S}_2). \tag{3.13}$$

Formalizing this type of expression requires a more abstract development of McCormick's procedure and is intimately related to the concept of composite relaxations. To begin, we define the space of *McCormick objects*

$$\mathbb{MR}^n \equiv \{\mathscr{S} = (S^B, S^C) \in \mathbb{IR}^n \times \mathbb{IR}^n : S^B \cap S^C \neq \emptyset\}. \tag{3.14}$$

Elements of $\mathbb{MR}^n$ are denoted by script capitals. For any $\mathscr{S} \in \mathbb{MR}^n$, the notations $S^B, S^C \in \mathbb{IR}^n$, and $\mathbf{s}^L, \mathbf{s}^U, \mathbf{s}^{cv}, \mathbf{s}^{cc} \in \mathbb{R}^n$ will commonly be used to denote the intervals and vectors satisfying $\mathscr{S} = (S^B, S^C) = ([\mathbf{s}^L, \mathbf{s}^U], [\mathbf{s}^{cv}, \mathbf{s}^{cc}])$. As with intervals, the set

$\mathbb{MR}^{n\times m}$ can be defined analogously to $\mathbb{MR}^n$; $\mathscr{A} \in \mathbb{MR}^{n\times m}$ has elements $\mathscr{A}_{ij} \in \mathbb{MR}$, for all $i \in \{1,\ldots,n\}$ and $j \in \{1,\ldots,m\}$. For any $D \subset \mathbb{R}^n$, let $\mathbb{M}D$ denote the set $\{\mathscr{S} \in \mathbb{MR}^n : S^B \subset D\}$. This notation is also used for $D \subset \mathbb{R}^{n\times m}$.

Now, consider again a factorable function $\mathbf{h} : D_s \subset \mathbb{R}^n \to \mathbb{R}^m$. Using the notation above, the McCormick relaxation of $\mathbf{h}$ can be formalized as a mapping of the form $\{\mathbf{h}\} : \mathbb{M}D_s \to \mathbb{MR}^m$, as suggested by (3.13). This mapping is a *relaxation function* for $\mathbf{h}$.

Definition 3.4 Let $\mathbf{h} : D_s \subset \mathbb{R}^n \to \mathbb{R}^m$ and let $S \subset D_s$ be convex. A function $\mathscr{H} : \mathbb{M}D_s \to \mathbb{MR}^m$ is a relaxation function for $\mathbf{h}$ if, for every $\mathscr{S} = (S, [\mathbf{s}^{cv}, \mathbf{s}^{cc}])$, $\mathscr{H}(\mathscr{S}) = (H(S), [\mathbf{u}(S, \mathbf{s}^{cv}, \mathbf{s}^{cc}), \mathbf{o}(S, \mathbf{s}^{cv}, \mathbf{s}^{cc})])$, where H is an inclusion function for $\mathbf{h}$ and $\mathbf{u}(S,\cdot,\cdot)$ and $\mathbf{o}(S,\cdot,\cdot)$ are, respectively, convex and concave composite relaxations for $\mathbf{h}$ on S.

To construct $\{\mathbf{h}\}$, McCormick's procedure makes use of relaxation functions for the basic operations $+$, $\times$, and $u \in \mathscr{L}$ appearing in the factorable representation of $\mathbf{h}$. For example, the relaxation function for binary addition is defined as

$$\mathscr{X} + \mathscr{Y} = (X^B + Y^B, (X^C \cap X^B) + (Y^C \cap Y^B)), \tag{3.15}$$

which agrees with (3.10). A relaxation function for binary multiplication can be similarly defined [74], and relaxation functions for many common univariate functions and are compiled in [74, 83]. We make the following assumption throughout.

Assumption 3.2 *For every $u \in \mathscr{L}$, $u : B \subset \mathbb{R} \to \mathbb{R}$, a relaxation function $\{u\} : \mathbb{M}B \to \mathbb{MR}$ is available and can be evaluated computationally.*

Naturally, $\{\mathbf{h}\}$ is constructed by applying the relaxation functions of the basic operations $+$, $\times$, and $u \in \mathscr{L}$ sequentially according to the factorable representation of $\mathbf{h}$. Thus, $\{\mathbf{h}\}$ is a particular relaxation function for $\mathbf{h}$ called the *natural McCormick extension of* $\mathbf{h}$. For example, the natural McCormick extension of (3.1) is given by (3.13), which is now well-defined. The fact that this procedure defines a relaxation function for $\mathbf{h}$ is a consequence of the following basic composition result.

Theorem 3.2 *Let $\mathbf{v} : D_s \subset \mathbb{R}^n \to \mathbb{R}^m$ and $o : B \subset \mathbb{R}^m \to \mathbb{R}$ have relaxations functions $\mathscr{V} : \mathbb{M}D_s \to \mathbb{MR}^m$ and $\mathscr{O} : \mathbb{M}B \to \mathbb{MR}$, respectively. If $\mathbf{v}(D_s) \subset B$ and $\mathscr{V}(\mathbb{M}D_s) \subset \mathbb{M}B$, then $\mathscr{O} \circ \mathscr{V}$ is a relaxation function for $o \circ \mathbf{v}$.*

Remark 3.1 For simplicity of exposition, the definition of a relaxation function here is slightly different than the original definition in [74]. For a function $\mathscr{H} : \mathbb{M}D_s \to \mathbb{MR}^m$, it can be shown that if $\mathscr{H}$ is a relaxation function as defined here, then it is also a relaxation function as defined in [74]. The converse holds provided that $\mathscr{H}$ satisfies an inclusion monotonicity property (Definition 2.4.13, [74]). Accordingly, the composition result in [74] (Lemmas 2.4.15 and 2.4.17) require inclusion monotonicity, while Theorem 3.2 here is a direct consequence of Definitions 3.3 and 3.4 above.

For any $\mathscr{S} = (S, [\mathbf{s}^{cv}, \mathbf{s}^{cc}]) \in \mathbb{M}D_s$, the natural McCormick extension of $\mathbf{h}$ evaluates to $\{\mathbf{h}\}(\mathscr{S}) = ([\mathbf{h}](S), [\mathbf{u}(S, \mathbf{s}^{cv}, \mathbf{s}^{cc}), \mathbf{o}(S, \mathbf{s}^{cv}, \mathbf{s}^{cc})])$, where $[\mathbf{h}]$ is the natural interval extension of $\mathbf{h}$ and $\mathbf{u}(S, \cdot, \cdot)$ and $\mathbf{o}(S, \cdot, \cdot)$ are convex and concave composite relaxations for $\mathbf{h}$ on S, respectively. Convex and concave relaxations are recovered from $\{\mathbf{h}\}$ by simply considering arguments of the form $\mathscr{S} = (S, [\mathbf{s}, \mathbf{s}])$ for $\mathbf{s} \in S$. This is equivalent to the initialization $(V_k(S), v_k^{cv}(S_k, \mathbf{s}), v_k^{cc}(S, \mathbf{s})) = (S_k, s_k, s_k)$, $k \leq n$, discussed above. For any such $\mathscr{S}$, we have $\{\mathbf{h}\}(\mathscr{S}) = ([\mathbf{h}](S), [\mathbf{u}(S, \mathbf{s}, \mathbf{s}), \mathbf{o}(S, \mathbf{s}, \mathbf{s})])$, and convex and concave relaxations of $\mathbf{h}$ on S are given by the definitions $\mathbf{h}^{cv}(\mathbf{s}) = \mathbf{u}(S, \mathbf{s}, \mathbf{s})$ and $\mathbf{h}^{cc}(\mathbf{s}) = \mathbf{o}(S, \mathbf{s}, \mathbf{s})$, respectively. This follows from Definition 3.3 after choosing $P \equiv S$ and observing that the identity mapping $P = S \ni \mathbf{s} \mapsto \mathbf{s} \in S$ is both a convex and a concave relaxation of itself on P. The reader is referred to [53, 74, 83] for a comprehensive treatment of McCormick relaxations.

3.3 Implementation

Several computational tools are available for performing interval and McCormick arithmetic operations. Of course, any function written explicitly in computer code is factorable, and its factorable representation can be generated automatically by parsing this code. Thus, it is possible to use code generation to create additional subroutines that compute the natural interval and McCormick extensions of such a function by replacing the standard arithmetic operations with interval and McCormick operations, respectively. A more flexible way to achieve this is through so-called operator overloading, which is available in most object oriented programming languages (e.g., `C++`). In this scheme, one can create interval and McCormick objects as new variable types with defined rules for the standard arithmetic operations. Then, the existing code for a factorable function can be executed with these objects to obtain the natural interval and McCormick extensions. A number of libraries defining interval objects and their associated arithmetic are freely available. For example, the Boost interval library in `C++` (http://www.boost.org/doc/libs/1_55_0/libs/numeric/interval/doc/interval.htm) and the `MATLAB` toolbox `INTLAB` (http://www.ti3.tu-harburg.de/~rump/intlab/). Similarly, relaxation functions can be computed using the McCormick arithmetic library `MC++` (http://www3.imperial.ac.uk/people/b.chachuat/research).

4 Bounds and Relaxations for Implicit Functions

In this section, procedures are derived for computing interval bounds and convex and concave relaxations for functions defined implicitly as the solutions of systems of algebraic equations. This is a fundamental step required for treating DAEs in the

following two sections. In general, we consider the nonlinear algebraic equations

$$\boldsymbol{\ell}(\mathbf{s}, \mathbf{r}) = \mathbf{0}, \tag{4.1}$$

where $\boldsymbol{\ell} \in C^k(D_s \times D_r, \mathbb{R}^n)$, $k \geq 1$, and $D_s \subset \mathbb{R}^{n_s}$ and $D_r \subset \mathbb{R}^n$ are open sets. Given an interval $S \subset D_s$, our primary interest is in the case where (4.1) defines a unique implicit function $\mathbf{h} : S \to \mathbb{R}^n$ such that $\boldsymbol{\ell}(\mathbf{s}, \mathbf{h}(\mathbf{s})) = \mathbf{0}$, $\forall \mathbf{s} \in S$. In this case, bounds and relaxations for $\mathbf{h}$ on S are well-defined. More generally, we can define inclusion and relaxation functions for $\mathbf{h}$ on S of the form $H : \mathbb{I}S \to \mathbb{IR}^n$ and $\mathscr{H} : \mathbb{M}S \to \mathbb{MR}^n$, respectively.

In doing so, the key challenge that must be addressed is that $\mathbf{h}$ is implicitly defined via (4.1), and hence no factorable representation of $\mathbf{h}$ is available in general. Of course, this implies that the methods of Sect. 3 cannot be applied directly. Instead, the key idea here is to use the factorable representation of $\boldsymbol{\ell}$, and hence the ability to compute inclusion and relaxation functions for $\boldsymbol{\ell}$, to infer the required information about $\mathbf{h}$ through (4.1). This approach originated with the development of so-called *interval Newton methods* [59], and has recently been extended to compute relaxations in [91]. To follow this approach here, the following basic assumption is required.

Assumption 4.1 *The functions* $\boldsymbol{\ell}$ *and* $\frac{\partial \boldsymbol{\ell}}{\partial \mathbf{r}}$ *are factorable and have natural interval and McCormick extensions* $[\boldsymbol{\ell}] : \mathbb{I}D_s \times \mathbb{I}D_r \to \mathbb{IR}^n$, $[\frac{\partial \boldsymbol{\ell}}{\partial \mathbf{r}}] : \mathbb{I}D_s \times \mathbb{I}D_r \to \mathbb{IR}^{n \times n}$, $\{\boldsymbol{\ell}\} : \mathbb{M}D_s \times \mathbb{M}D_r \to \mathbb{MR}^n$, *and* $\{\frac{\partial \boldsymbol{\ell}}{\partial \mathbf{r}}\} : \mathbb{M}D_s \times \mathbb{M}D_r \to \mathbb{MR}^{n \times n}$.

Remark 4.1 Although any factorable function $\mathbf{w} : D \to \mathbb{R}^n$ has natural interval and McCormick extensions, the content of the above assumption is that these are defined on all of $\mathbb{I}D$ and $\mathbb{M}D$, respectively. For example, $w(s) = 1/(s - \sqrt{s})$ is well defined on $D = (1, +\infty]$. However, overestimation in the interval evaluation of the denominator leads to a division by zero error for $S = [4, 100] \in \mathbb{I}D$, so that the interval extension is not defined at this point: $[w](S) = 1/([4, 100] - [2, 10]) = 1/[-6, 98] = \text{NaN}$. Assuming well-defined interval and McCormick extensions on all of $\mathbb{M}D$ is not essential for the following developments, but simplifies the presentation.

This simplest way to infer information about $\mathbf{h}$ from the factorable representation of $\boldsymbol{\ell}$ is through a direct rearrangement of (4.1) of the form $\mathbf{r} = \mathbf{h}(\mathbf{s})$, $\forall \mathbf{s} \in D_s$. However, such a rearrangement is rarely possible. Instead, we pursue the more modest feat of deriving a factorable function $\boldsymbol{\psi}$ that satisfies $\mathbf{r} = \boldsymbol{\psi}(\mathbf{s}, \mathbf{r})$ for all $(\mathbf{s}, \mathbf{r}) \in D_s \times D_r$ such that $\boldsymbol{\ell}(\mathbf{s}, \mathbf{r}) = \mathbf{0}$. In general, even this cannot be accomplished precisely as written due to some technicalities discussed below. However, we will derive a similar expression and then provide a computational test capable of verifying that this expression is well-defined, and that a unique implicit function $\mathbf{h}$ exists, on some given intervals $S \times R$. This then implies that $\mathbf{h}(\mathbf{s}) = \boldsymbol{\psi}(\mathbf{s}, \mathbf{h}(\mathbf{s}))$, $\forall \mathbf{s} \in S$, which provides the opportunity to compute inclusion and relaxation functions for $\mathbf{h}$ through iterative procedures.

We begin by presenting the details of a particular interval Newton method called the interval Hansen-Sengupta method [59]. The existence of a unique implicit function to be bounded is not assumed a priori. Instead, the algorithm is centered around the following more general task: Given intervals $S \subset D_s$ and $R \subset D_r$, compute a refined interval $R' \subset R$ that contains every $\mathbf{r} \in R$ for which (4.1) holds for some $\mathbf{s} \in S$. If it is possible to derive a factorable function $\boldsymbol{\psi}$ satisfying the implication

$$(\mathbf{s}, \mathbf{r}) \in S \times R, \quad \boldsymbol{\ell}(\mathbf{s}, \mathbf{r}) = \mathbf{0} \quad \Longrightarrow \quad \mathbf{r} = \boldsymbol{\psi}(\mathbf{s}, \mathbf{r}), \tag{4.2}$$

then this task is accomplished by the update $R' = R \cap [\boldsymbol{\psi}](S, R)$, where $[\boldsymbol{\psi}]$ is the natural interval extension of $\boldsymbol{\psi}$.

The key step in deriving such a function is an application of the mean-value theorem. Choose a fixed reference point $\tilde{\mathbf{r}} \in R$. For each fixed $\mathbf{s} \in S$ and $i \in \{1, \ldots, n\}$, applying the mean-value theorem to the function $\ell_i(\mathbf{s}, \cdot)$ ensures that there exists $\lambda_i(\mathbf{s}) \in [0, 1]$ such that

$$\frac{\partial \ell_i}{\partial \mathbf{r}}(\mathbf{s}, \tilde{\mathbf{r}} + \lambda_i(\mathbf{s})(\mathbf{r} - \tilde{\mathbf{r}}))\,(\mathbf{r} - \tilde{\mathbf{r}}) = \ell_i(\mathbf{s}, \mathbf{r}) - \ell_i(\mathbf{s}, \tilde{\mathbf{r}}). \tag{4.3}$$

Define $\boldsymbol{\lambda} : S \to [\mathbf{0}, \mathbf{1}]$ by choosing $\lambda_i(\mathbf{s})$ in this way for all $\mathbf{s} \in S$ and every $i \in \{1, \ldots, n\}$. Additionally, define

$$\mathbf{M}(\mathbf{s}, \mathbf{r}, \boldsymbol{\lambda}') \equiv \begin{bmatrix} \frac{\partial \ell_1}{\partial \mathbf{r}}(\mathbf{s}, \tilde{\mathbf{r}} + \lambda_1'(\mathbf{r} - \tilde{\mathbf{r}})) \\ \vdots \\ \frac{\partial \ell_n}{\partial \mathbf{r}}(\mathbf{s}, \tilde{\mathbf{r}} + \lambda_n'(\mathbf{r} - \tilde{\mathbf{r}})) \end{bmatrix}, \quad \forall (\mathbf{s}, \mathbf{r}, \boldsymbol{\lambda}') \in D_s \times D_r \times \mathbb{R}^n. \tag{4.4}$$

Then, from the definition of $\boldsymbol{\lambda}$, $\mathbf{M}(\mathbf{s}, \mathbf{r}, \boldsymbol{\lambda}(\mathbf{s}))\,(\mathbf{r} - \tilde{\mathbf{r}}) = \boldsymbol{\ell}(\mathbf{s}, \mathbf{r}) - \boldsymbol{\ell}(\mathbf{s}, \tilde{\mathbf{r}})$, $\forall (\mathbf{s}, \mathbf{r}) \in S \times R$. In particular,

$$(\mathbf{s}, \mathbf{r}) \in S \times R, \quad \boldsymbol{\ell}(\mathbf{s}, \mathbf{r}) = \mathbf{0} \quad \Longrightarrow \quad \mathbf{M}(\mathbf{s}, \mathbf{r}, \boldsymbol{\lambda}(\mathbf{s}))\,(\mathbf{r} - \tilde{\mathbf{r}}) = -\boldsymbol{\ell}(\mathbf{s}, \tilde{\mathbf{r}}). \tag{4.5}$$

Thus, any $(\mathbf{s}, \mathbf{r})$ satisfying (4.1) must have $(\mathbf{r} - \tilde{\mathbf{r}})$ as a solution of an $n \times n$ linear system. To obtain (4.2) from (4.5), we consider rearrangements of these linear equations that isolate each r_i on the left-hand side. Assume for the moment that the diagonal elements of $\mathbf{M}(\mathbf{s}, \mathbf{r}, \boldsymbol{\lambda}(\mathbf{s}))$ are nonzero. Then, the ith linear relation above can be rearranged as

$$r_i = \tilde{r}_i + \frac{1}{m_{ii}(\mathbf{s}, \mathbf{r}, \boldsymbol{\lambda}(\mathbf{s}))} \left(-\ell_i(\mathbf{s}, \tilde{\mathbf{r}}) - \sum_{j \neq i} m_{ij}(\mathbf{s}, \mathbf{r}, \boldsymbol{\lambda}(\mathbf{s}))(r_j - \tilde{r}_j) \right). \tag{4.6}$$

With this in mind, we make the following definition for every $(\mathbf{s}, \mathbf{r}, \boldsymbol{\lambda}') \in D_s \times D_r \times \mathbb{R}^n$ such that $m_{ii}(\mathbf{s}, \mathbf{r}, \boldsymbol{\lambda}') \neq 0$ for all $i \in \{1, \ldots, n\}$:

$$\boldsymbol{\psi}(\mathbf{s}, \mathbf{r}, \boldsymbol{\lambda}') \equiv \mathbf{r}', \tag{4.7}$$

where, using the abbreviation $m_{ij} = m_{ij}(\mathbf{s}, \mathbf{r}, \boldsymbol{\lambda}')$, for $i = 1, \ldots, n$,

$$r_i' = \tilde{r}_i + \frac{1}{m_{ii}} \left(-\ell_i(\mathbf{s}, \tilde{\mathbf{r}}) - \sum_{j<i} m_{ij}(r_j' - \tilde{r}_j) - \sum_{j>i} m_{ij}(r_j - \tilde{r}_j) \right). \tag{4.8}$$

The reason for splitting the sum above will become clear shortly. In any case, it is easily shown from (4.6) that this definition satisfies

$$(\mathbf{s}, \mathbf{r}) \in S \times R, \quad \boldsymbol{\ell}(\mathbf{s}, \mathbf{r}) = \mathbf{0} \quad \Longrightarrow \quad \mathbf{r} = \boldsymbol{\psi}(\mathbf{s}, \mathbf{r}, \boldsymbol{\lambda}(\mathbf{s})). \tag{4.9}$$

In this last implication, $\boldsymbol{\psi}$ is not quite in the form we set out to derive due to the third argument, $\boldsymbol{\lambda}(\mathbf{s})$. However, in this form, $\boldsymbol{\psi}$ is a factorable function wherever it is defined. Indeed, by Assumption 4.1, both $\boldsymbol{\ell}$ and $\frac{\partial \boldsymbol{\ell}}{\partial \mathbf{r}}$ are factorable. Thus, $\mathbf{M}$ is factorable, and by (4.8), so is $\boldsymbol{\psi}$. On the other hand, $\boldsymbol{\lambda}$ is a theoretical construct that is known to exist on account of the mean-value theorem, but does not have a known factorable representation. Fortunately, we are assured that $\boldsymbol{\lambda}(\mathbf{s}) \in [\mathbf{0}, \mathbf{1}]$, $\forall \mathbf{s} \in S$, so that a factorable representation is not necessary in order to bound $\boldsymbol{\lambda}$. Thus, our task of computing a refined interval R' is now accomplished through the update

$$R' = R \cap [\boldsymbol{\psi}](S, R, [\mathbf{0}, \mathbf{1}]). \tag{4.10}$$

Here $[\boldsymbol{\psi}]$ is the natural interval extension of $\boldsymbol{\psi}$, derived by applying interval arithmetic to (4.7) and (4.8). In particular, $[\boldsymbol{\psi}](S, R, [\mathbf{0}, \mathbf{1}]) \equiv R'$, where, using the abbreviation $[m_{ij}] = [m_{ij}](S, R, [\mathbf{0}, \mathbf{1}])$,

$$R_i' = \tilde{r}_i + \frac{1}{[m_{ii}]} \left(-[\ell_i](S, \tilde{\mathbf{r}}) - \sum_{j<i} [m_{ij}](R_j' - \tilde{r}_j) - \sum_{j>i} [m_{ij}](R_j - \tilde{r}_j) \right). \tag{4.11}$$

For the moment, it is assumed that $0 \notin [m_{ii}](S, R, [\mathbf{0}, \mathbf{1}])$ for every i, so that interval division by $[m_{ii}](S, R, [\mathbf{0}, \mathbf{1}])$ is defined [56]. Note that, by splitting the sum in the definition of $\boldsymbol{\psi}$, we have arranged that $[\psi_i]$ is computed with the updated intervals R_j' in place of R_j for all $j < i$, which helps to tighten the refinement R' in (4.10). In fact, a further improvement can be achieved by carrying out the intersection

in (4.10) component-wise during the computation of $[\boldsymbol{\psi}]$. Thus, we further define $[\boldsymbol{\psi}_\cap](S,R,[\mathbf{0},\mathbf{1}]) \equiv R'$, where, using the abbreviation $[m_{ij}] = [m_{ij}](S,R,[\mathbf{0},\mathbf{1}])$,

$$R_i' = R_i \cap \left(\tilde{r}_i + \frac{1}{[m_{ii}]} \left(-[\ell_i](S,\tilde{\mathbf{r}}) - \sum_{j<i} [m_{ij}](R_j' - \tilde{r}_j) - \sum_{j>i} [m_{ij}](R_j - \tilde{r}_j) \right) \right). \tag{4.12}$$

This gives the modified update $R' = [\boldsymbol{\psi}_\cap](S,R,[\mathbf{0},\mathbf{1}])$.

In order to extend the definition of $[\boldsymbol{\psi}_\cap]$ to intervals S and R such that $0 \in [m_{ii}](S,R,[\mathbf{0},\mathbf{1}])$ for some i, it is necessary to define a special interval operator:

$$\Gamma(A,B,Z) \equiv \text{hull}\left(\{z \in Z : az = b,\ a \in A,\ b \in B\}\right), \tag{4.13}$$

where hull(Q) denotes the smallest interval containing the set Q. Given an a priori bound Z, $\Gamma(A,B,Z)$ is a bound on the solutions $z \in Z$ of a single interval linear equation, even when rearranging that equation for z would involve division by an interval containing zero. It has been shown that Γ can be evaluated computationally through the simple rules

$$\Gamma(A,B,Z) \equiv \begin{cases} (B/A) \cap Z & \text{if } 0 \notin A \\ \text{hull}\left(Z \backslash \text{int}([b^L/a^L, b^L/a^U])\right) & \text{if } 0 \in A \text{ and } b^L > 0 \\ \text{hull}\left(Z \backslash \text{int}([b^U/a^U, b^U/a^L])\right) & \text{if } 0 \in A \text{ and } b^U < 0 \\ Z & \text{if } 0 \in A \text{ and } 0 \in B \end{cases}, \tag{4.14}$$

where int(Q) is the interior of Q [59]. Using Γ, we now define $[\boldsymbol{\psi}_\Gamma](S,R,[\mathbf{0},\mathbf{1}]) \equiv R'$, where, using the abbreviation $[m_{ij}] = [m_{ij}](S,R,[\mathbf{0},\mathbf{1}])$,

$$R_i' = \tilde{r}_i + \Gamma\left([m_{ii}], -[\ell_i](S,\tilde{\mathbf{r}}) - \sum_{j<i} [m_{ij}](R_j' - \tilde{r}_j) - \sum_{j>i} [m_{ij}](R_j - \tilde{r}_j), (R_i - \tilde{r}_i) \right). \tag{4.15}$$

In contrast to $[\boldsymbol{\psi}_\cap]$, $[\boldsymbol{\psi}_\Gamma]$ is defined for all arguments with $(S,R) \in \mathbb{I}D_s \times \mathbb{I}D_r$. Moreover, we have the following variant of Theorem 5.1.8 in [59] proven in [78]:

Theorem 4.1 *Let $S \in \mathbb{I}D_s$, $R \in \mathbb{I}D_r$, $\tilde{\mathbf{r}} \in R$, and define $R' \equiv [\boldsymbol{\psi}_\Gamma](S,R,[\mathbf{0},\mathbf{1}])$. The following conclusions hold:*

1. *If $(\mathbf{s},\mathbf{r}) \in S \times R$ satisfies $\boldsymbol{\ell}(\mathbf{s},\mathbf{r}) = \mathbf{0}$, then $\mathbf{r} \in R'$.*
2. *If $R' = \emptyset$, then $\nexists (\mathbf{s},\mathbf{r}) \in S \times R$ such that $\boldsymbol{\ell}(\mathbf{s},\mathbf{r}) = \mathbf{0}$.*
3. *If $\tilde{\mathbf{r}} \in \text{int}(R)$ and $\emptyset \neq R' \subset \text{int}(R)$, then $\exists \mathbf{h} \in C^k(S,R')$ such that, for every $\mathbf{s} \in S$, $\mathbf{r} = \mathbf{h}(\mathbf{s})$ is the unique element of R satisfying $\boldsymbol{\ell}(\mathbf{s},\mathbf{r}) = \mathbf{0}$. Moreover, $[\mathbf{M}](S,R,[\mathbf{0},\mathbf{1}])$ does not contain a singular matrix and does not contain zero in any of its diagonal elements.*

Remark 4.2 Since $\tilde{\mathbf{r}} \in R$, it is straightforward to show that $\tilde{\mathbf{r}} + [\mathbf{0}, \mathbf{1}](R - \tilde{\mathbf{r}}) = R$, and hence

$$[\mathbf{M}](S, R, [\mathbf{0}, \mathbf{1}]) \equiv \begin{bmatrix} [\frac{\partial \ell_1}{\partial \mathbf{r}}](S, R) \\ \vdots \\ [\frac{\partial \ell_n}{\partial \mathbf{r}}](S, R) \end{bmatrix} = \left[\frac{\partial \boldsymbol{\ell}}{\partial \mathbf{r}}\right](S, R). \tag{4.16}$$

The statement of Theorem 4.1 in [78] uses this substitution. Conclusion 3 of Theorem 4.1 then states that if $\tilde{\mathbf{r}} \in \text{int}(R)$ and the inclusion $\emptyset \neq [\boldsymbol{\psi}_\Gamma](S, R, [\mathbf{0}, \mathbf{1}]) \subset \text{int}(R)$ holds, then $\left[\frac{\partial \boldsymbol{\ell}}{\partial \mathbf{r}}\right](S, R)$ contains no singular matrices and does not contain zero in any of its diagonal elements. In practice, it is almost always necessary to precondition (4.1) by a non-singular matrix $\mathbf{C} \in \mathbb{R}^{n \times n}$ for this to be true, and hence for the desired inclusion to hold. A common choice is the midpoint inverse, $\mathbf{C} \equiv \left(m(\left[\frac{\partial \boldsymbol{\ell}}{\partial \mathbf{r}}\right](S, R))\right)^{-1}$, in which case $[\mathbf{M}](S, R, [\mathbf{0}, \mathbf{1}]) = \mathbf{C}\left[\frac{\partial \boldsymbol{\ell}}{\partial \mathbf{r}}\right](S, R)$ is an interval enclosure of the identity matrix. For ease of notation, the preconditioner $\mathbf{C}$ is omitted from Theorem 4.1 and all subsequent developments.

Without further note, we will always define $[\boldsymbol{\psi}_\Gamma](S, R, [\mathbf{0}, \mathbf{1}])$ with $\tilde{\mathbf{r}} = m(R)$, so that $\tilde{\mathbf{r}} \in \text{int}(R)$ whenever R has nonempty interior. Now, suppose that intervals S^* and R^* have been found such that $\emptyset \neq [\boldsymbol{\psi}_\Gamma](S^*, R^*, [\mathbf{0}, \mathbf{1}]) \subset \text{int}(R^*)$. Then, Conclusion 3 of Theorem 4.1 ensures the existence of an implicit function satisfying $\boldsymbol{\ell}(\mathbf{s}, \mathbf{h}(\mathbf{s})) = \mathbf{0}$ on all of S^*, and further ensures that $\mathbf{h}$ is the only such function taking values in R^*. Thus, the inclusion test of Conclusion 3 isolates a single parametric solution branch of (4.1), and provides the bound $\mathbf{h}(S^*) \subset [\boldsymbol{\psi}_\Gamma](S^*, R^*, [\mathbf{0}, \mathbf{1}]) \subset \text{int}(R^*)$. Applying Conclusion 1 of Theorem 4.1, this bound can be iteratively refined by the scheme

$$R_{k+1} = [\boldsymbol{\psi}_\Gamma](S^*, R_k, [\mathbf{0}, \mathbf{1}]), \quad R_0 = R^*. \tag{4.17}$$

This is known as the interval Hansen-Sengupta method. More generally, this scheme defines a sequence of progressively tighter inclusion functions $H_k : \mathbb{I}S^* \to \mathbb{IR}^n$ for $\mathbf{h}$ via

$$H_{k+1}(S) = [\boldsymbol{\psi}_\Gamma](S, H_k(S), [\mathbf{0}, \mathbf{1}]), \quad H_0(S) = R^*, \quad \forall S \in \mathbb{I}S^*. \tag{4.18}$$

The fact that $[\mathbf{M}](S^*, R^*, [\mathbf{0}, \mathbf{1}])$ does not contain zero in any of its diagonal elements also has important consequences. Most importantly, it follows that the matrix $\mathbf{M}(\mathbf{s}, \mathbf{r}, \boldsymbol{\lambda}(\mathbf{s}))$ has nonzero diagonals for all $(\mathbf{s}, \mathbf{r}) \in S^* \times R^*$, so that $\boldsymbol{\psi}$ is defined on all of $(\mathbf{s}, \mathbf{r}) \in S^* \times R^*$ and, by (4.9),

$$\mathbf{h}(\mathbf{s}) = \boldsymbol{\psi}(\mathbf{s}, \mathbf{h}(\mathbf{s})), \quad \forall \mathbf{s} \in S^*. \tag{4.19}$$

Additionally, it can be shown that $[\boldsymbol{\psi}_{\Gamma}](R, S, [\mathbf{0}, \mathbf{1}]) = [\boldsymbol{\psi}_{\cap}](R, S, [\mathbf{0}, \mathbf{1}])$, $\forall (S, R) \in \mathbb{I}S^* \times \mathbb{I}R^*$, and, using the fact that $\emptyset \neq [\boldsymbol{\psi}_{\Gamma}](S^*, R^*, [\mathbf{0}, \mathbf{1}]) \subset \mathrm{int}(R^*)$,

$$[\boldsymbol{\psi}_{\Gamma}](R^*, S^*, [\mathbf{0}, \mathbf{1}]) = [\boldsymbol{\psi}_{\cap}](R^*, S^*, [\mathbf{0}, \mathbf{1}]) = [\boldsymbol{\psi}](R^*, S^*, [\mathbf{0}, \mathbf{1}]). \tag{4.20}$$

We are now prepared to show how the developments of Sect. 3.2 can be used to compute convex and concave relaxations of the implicit function $\mathbf{h}$ on $S \in \mathbb{I}S^*$. It is assumed that the inclusion test of Theorem 4.1 has been passed with (S^*, R^*), so that $\mathbf{h} : S^* \to R^*$ exists. Moreover, it is assumed that an inclusion function $H : \mathbb{I}S^* \to \mathbb{IR}^n$ is available, e.g., by truncation of (4.18). By analogy to (4.18), we will derive an iterative scheme that generates refined inclusion functions $\mathscr{H}_k : \mathbb{M}S^* \to \mathbb{MR}^n$. Recall that $\mathbb{M}S^*$ contains all $\mathscr{S} = (S, [\mathbf{s}^{cv}, \mathbf{s}^{cc}])$ with $S \subset S^*$ (although we will be primarily concerned with arguments of the form $\mathscr{S} = (S, [\mathbf{s}, \mathbf{s}])$ with $\mathbf{s} \in S$). To initialize this scheme, we note that one relaxation function is given by

$$\mathscr{H}_0(\mathscr{S}) = (H(S), H(S)), \quad \forall \mathscr{S} = (S, [\mathbf{s}^{cv}, \mathbf{s}^{cc}]) \in \mathbb{M}S^*. \tag{4.21}$$

Using the notation $[\mathbf{h}^L(S), \mathbf{h}^U(S)] = H(S)$, this is true because the constant functions $\mathbf{h}^{cv}(\mathbf{s}) = \mathbf{h}^L(S)$ and $\mathbf{h}^{cc}(\mathbf{s}) = \mathbf{h}^U(S)$ are trivially convex and concave relaxations of $\mathbf{h}$ on S.

In order to refine $\mathscr{H}_0$, we again use the semi-explicit characterization of $\mathbf{h}$ given in (4.19). To begin, note that $\boldsymbol{\lambda}(\mathbf{s}) \in [\mathbf{0}, \mathbf{1}]$, $\forall \mathbf{s} \in S$, so that the constant function $\mathbb{M}S^* \ni \mathscr{S} \mapsto \mathscr{L} \equiv ([\mathbf{0}, \mathbf{1}], [\mathbf{0}, \mathbf{1}]) \in \mathbb{MR}^n$ is a relaxation function for $\boldsymbol{\lambda}$ on $\mathbb{M}S^*$. Now, since $\boldsymbol{\psi}$ is factorable, the natural relaxation function $\{\boldsymbol{\psi}\}$ can be computed by simply applying McCormick arithmetic to (4.8). However, it is more useful to define $\{\boldsymbol{\psi}_{\cap}\}$ by analogy to $[\boldsymbol{\psi}_{\cap}]$ as

$$\{\boldsymbol{\psi}_{\cap}\}(\mathscr{S}, \mathscr{R}, \mathscr{L}) \equiv \mathscr{R}', \tag{4.22}$$

where, using the abbreviation $\{m_{ij}\} = \{m_{ij}\}(\mathscr{S}, \mathscr{R}, \mathscr{L})$,

$$\mathscr{R}_i' \equiv \mathscr{R}_i \cap \left(\tilde{r}_i + \frac{1}{\{m_{ii}\}} \left(\begin{array}{r} -\{\ell_i\}(\mathscr{S}, \tilde{\mathbf{r}}) - \sum_{j<i} \{m_{ij}\}(\mathscr{R}_j' - \tilde{r}_j) \\ - \sum_{j>i} \{m_{ij}\}(\mathscr{R}_j - \tilde{r}_j) \end{array} \right) \right). \tag{4.23}$$

The intersection above is defined for arbitrary $\mathscr{R}$ and $\overline{\mathscr{R}}$ in $\mathbb{MR}^n$ by $\mathscr{R} \cap \overline{\mathscr{R}} = (R^B \cap \overline{R}^B, R^C \cap \overline{R}^C)$. Division by a McCormick object is defined whenever its interval part does not contain zero. But, for any $\mathscr{S} \in \mathbb{M}S^*$, $\{m_{ii}\}(\mathscr{S}, \mathscr{R}, \mathscr{L})$ has interval part $[m_{ii}](S, R, [\mathbf{0}, \mathbf{1}]) \subset [m_{ii}](S^*, R^*, [\mathbf{0}, \mathbf{1}])$, which is guaranteed not to contain zero. Thus, $\{\boldsymbol{\psi}_{\cap}\}$ is defined for all $\mathscr{S} \in \mathbb{M}S^*$, and the same is easily seen to be true of $\{\boldsymbol{\psi}\}$.

Now, a sequence of progressively refined relaxation functions $\mathscr{H}_k : \mathbb{M}S^* \to \mathbb{MR}^n$ can be defined via

$$\mathscr{H}_{k+1}(\mathscr{S}) = \{\boldsymbol{\psi}_{\cap}\}(\mathscr{S}, \mathscr{H}_k(\mathscr{S}), \mathscr{L}), \quad \mathscr{H}_0(\mathscr{S}) = (H(S), H(S)), \tag{4.24}$$

for all $\mathscr{S} = (S, [\mathbf{s}^{cv}, \mathbf{s}^{cc}]) \in \mathbb{M}S^*$. Convex and concave relaxations of $\mathbf{h}$ on any $S \subset S^*$ are evaluated computationally at a particular point $\mathbf{s} \in S$ by simply setting $\mathscr{S} = (S, [\mathbf{s}, \mathbf{s}])$ and executing the iteration (4.24). For any $k \geq 0$, the result is a McCormick object $\mathscr{H}_k(\mathscr{S}) = (H_k(S), [\mathbf{h}_k^{cv}(\mathbf{s}), \mathbf{h}_k^{cc}(\mathbf{s})])$ with the obvious interpretation.

To see that this works in more detail, denote $\mathscr{H}_k(\mathscr{S}) = (H_k(S), [\mathbf{h}_k^{cv}(\mathbf{s}), \mathbf{h}_k^{cc}(\mathbf{s})])$ and assume that $\mathbf{h}_k^{cv}$ and $\mathbf{h}_k^{cc}$ are convex and concave relaxations of $\mathbf{h}$ on S, which is trivially true for $k = 0$. Now consider a single iteration of (4.24). Using the definition of a relaxation function in terms of composite relaxations (see Sect. 3.2), the computation of $\mathbf{h}_{k+1}^{cv}$ and $\mathbf{h}_{k+1}^{cc}$ in (4.24) can be expressed as

$$\mathbf{h}_{k+1}^{cv}(\mathbf{s}) = \max(\mathbf{h}_k^{cv}(\mathbf{s}), \mathbf{u}_{\psi}(\mathbf{s}, \mathbf{s}, \mathbf{h}_k^{cv}(\mathbf{s}), \mathbf{h}_k^{cc}(\mathbf{s}))), \tag{4.25}$$

$$\mathbf{h}_{k+1}^{cc}(\mathbf{s}) = \min(\mathbf{h}_k^{cc}(\mathbf{s}), \mathbf{o}_{\psi}(\mathbf{s}, \mathbf{s}, \mathbf{h}_k^{cv}(\mathbf{s}), \mathbf{h}_k^{cc}(\mathbf{s}))), \tag{4.26}$$

where $\mathbf{u}_{\psi}$ and $\mathbf{o}_{\psi}$ are convex and concave composite relaxations of $\boldsymbol{\psi}$ on $S \times H_k(S)$, respectively. Since $\mathbf{h}_k^{cv}$ and $\mathbf{h}_k^{cc}$ are convex and concave relaxations of $\mathbf{h}$ on S by hypothesis, the properties of composite relaxations ensure that the functions $\mathbf{s} \mapsto \mathbf{u}_{\psi}(\mathbf{s}, \mathbf{s}, \mathbf{h}_k^{cv}(\mathbf{s}), \mathbf{h}_k^{cc}(\mathbf{s}))$ and $\mathbf{s} \mapsto \mathbf{o}_{\psi}(\mathbf{s}, \mathbf{s}, \mathbf{h}_k^{cv}(\mathbf{s}), \mathbf{h}_k^{cc}(\mathbf{s}))$ are, respectively, convex and concave relaxations of $\mathbf{s} \mapsto \boldsymbol{\psi}(\mathbf{s}, \mathbf{h}(\mathbf{s}))$ (and hence of $\mathbf{h}$) on S. The same is true of $\mathbf{h}_{k+1}^{cv}$ and $\mathbf{h}_{k+1}^{cc}$ because the maximum (resp. minimum) of two convex (resp. concave) functions is convex (resp. concave).

4.1 Example

Section 8 presents a parameter estimation problem in semi-explicit DAEs with the following algebraic equation:

$$0 = y_1^3 + 2m_3 y_1^2 + K_b y_1 - (K_b + y_1^2)x_2, \tag{4.27}$$

where m_3 and K_b are known constants. As discussed in the following section, the methods of this section are applied to the algebraic equations in semi-explicit DAE systems by interpreting $\mathbf{r} = \mathbf{y}$ as the dependent variables and $\mathbf{s} = (t, \mathbf{p}, \mathbf{x})$ as the independent variables. Thus, we may write (4.27) in the notation of this section as

$$0 = r^3 + 2m_3 r^2 + K_b r - (K_b + r^2)s. \tag{4.28}$$

The derivative $\frac{\partial \ell}{\partial r}$ is given by

$$\frac{\partial \ell}{\partial r}(s, r) = r(3r + 2(2m_3 - s)) + K_b. \tag{4.29}$$

The equations above are written to represent the specific factorable representations used in the computation of bounds and relaxations of the implicit function $h(s)$ below. They are arranged so as to reduce the conservatism of the bounds and relaxations of these expressions when they are evaluated with interval- or McCormick-valued arguments. In general, this is done by minimizing the number of appearances of interval- or McCormick-valued variables. As a simple example, $a(b+c)$ is preferred to $ab+ac$ because a appears twice in the latter expression. Since these appearances of a will be considered independent when evaluated, e.g., in interval arithmetic, the latter expression gives a weaker enclosure. With $A = [2,3]$, $B = [-2,-1]$, and $C = [1,2]$, the first expression gives $[2,3]([-2,-1]+[1,2]) = [2,3][-1,1] = [-3,3]$, whereas the second gives $[2,3][-2,-1]+[2,3][1,2] = [-6,-2]+[2,6] = [-4,4]$. In arranging (4.28), we have used the observation that the evaluations of this function needed for computing $[\psi_\Gamma]$ and $\{\psi_\cap\}$ always use a real-valued reference value for r. Thus, multiple appearances of r are not problematic. On the other hand, (4.28) is evaluated with interval- and McCormick-valued arguments for s. Thus, (4.28) is arranged so that s appears only once, which would not be the case if, for example, the r^2 terms were collected. In contrast to the situation with (4.28), computing $[\psi_\Gamma]$ and $\{\psi_\cap\}$ does involve evaluating (4.29) with interval- and McCormick-valued arguments for r. Moreover, (4.29) cannot be arranged so that r appears only once. The preferred arrangement in this case is the so-called Horner's form [56].

Using (4.28) and (4.29), Eqs. (4.15) and (4.23) become, respectively,

$$R' = \tilde{r} + \left[(R-\tilde{r}) \cap \left(\frac{-\tilde{r}^3 - 2m_3\tilde{r}^2 - K_b\tilde{r} + (K_b+\tilde{r}^2)S}{R(3R+2(2m_3-S))+K_b} \right) \right], \quad \text{and} \tag{4.30}$$

$$\mathscr{R}' = \tilde{r} + \left[(\mathscr{R}-\tilde{r}) \cap \left(\frac{-\tilde{r}^3 - 2m_3\tilde{r}^2 - K_b\tilde{r} + (K_b+\tilde{r}^2)\mathscr{S}}{\mathscr{R}(3\mathscr{R}+2(2m_3-\mathscr{S}))+K_b} \right) \right], \tag{4.31}$$

provided that no division by an interval or McCormick object containing zero occurs. Choosing the values $m_3 = 100$, $K_b = \frac{36}{540}$, and $S = [4,8]\times 10^{-3}$, we find that $\emptyset \neq R' \subset \mathrm{int}(R)$ with $R = [0.96164, 1.5531]\times 10^{-3}$. This R interval was found using a nonsmooth Newton iteration as described in [79]. It follows from Theorem 4.1 that there exists a unique implicit function $r = h(s)$ satisfying (4.28) on S and bounded by R. This bound changes by less that 10^{-6} after ten iterations of (4.30). Relaxations of h can now be computed by iteration on (4.31) with $\mathscr{R}$ initially set to (R,R) and $\mathscr{S} = (S,[\mathbf{s},\mathbf{s}])$ for values of $\mathbf{s}$ in S. The resulting relaxations after five iterations are given in Fig. 1.

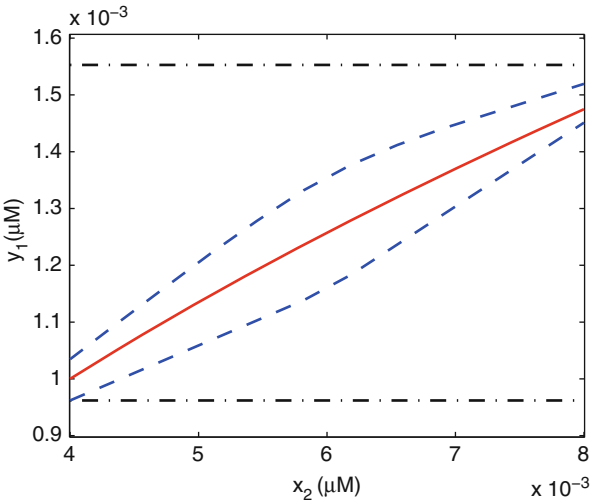

Fig. 1 Interval bounds (*dot-dashed*) and convex and concave relaxations (*dashed*) for the implicit solution $y_1 = h(x_2)$ of (4.27) (*solid*) computed via (4.30) and (4.31)

5 State Bounds for Semi-explicit Index-One DAEs

In this section, we summarize the main results in [78, 79] for the computation of interval bounds on the reachable set of (2.1), which we restate here for convenience:

$$\left.\begin{aligned}\dot{\mathbf{x}}(t,\mathbf{p}) &= \mathbf{f}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}),\mathbf{y}(t,\mathbf{p}))\\ \mathbf{0} &= \mathbf{g}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}),\mathbf{y}(t,\mathbf{p}))\end{aligned}\right\}, \tag{5.1a}$$

$$\mathbf{x}(t_0,\mathbf{p}) = \mathbf{x}_0(\mathbf{p}). \tag{5.1b}$$

In what follows, we consider a regular solution $(\mathbf{x},\mathbf{y})$ of (5.1) on $I\times P$ and compute functions $\mathbf{x}^L,\mathbf{x}^U : I \to \mathbb{R}^{n_x}$ and $\mathbf{y}^L,\mathbf{y}^U : I \to \mathbb{R}^{n_y}$ such that

$$\mathbf{x}^L(t) \le \mathbf{x}(t,\mathbf{p}) \le \mathbf{x}^U(t) \quad \text{and} \quad \mathbf{y}^L(t) \le \mathbf{y}(t,\mathbf{p}) \le \mathbf{y}^U(t), \quad \forall (t,\mathbf{p}) \in I\times P.$$

These functions are referred to as *state bounds* for the solution $(\mathbf{x},\mathbf{y})$. As mentioned in Sect. 2.3, the existence and uniqueness of $(\mathbf{x},\mathbf{y})$ on $I\times P$ is verified by the method and need not be assumed. The behavior of the method for DAEs with multiple solutions is discussed below, after some key results are in place.

Exactly as in Sect. 4, the difficulty in computing state bounds for $(\mathbf{x},\mathbf{y})$ is that these functions have no known factorable representation, so that the methods of Sect. 3 cannot be applied directly. This complication was overcome in Sect. 4 by instead exploiting the factorable representation of the algebraic system there (specifically, $\boldsymbol{\ell}$ and $\frac{\partial \boldsymbol{\ell}}{\partial \mathbf{r}}$) to obtain global information about the function of interest (the implicit function $\mathbf{h}$). Our approach in this section and the next is exactly analogous, although the functions of interest here are defined as the solutions of DAEs, so that some additional insights are required. In brief, the factorable

representation of the governing equations in (5.1) will be used to derive an auxiliary system of DAEs describing bounding trajectories $\mathbf{x}^L$, $\mathbf{x}^U$, $\mathbf{y}^L$, and $\mathbf{y}^U$ as its solutions. Accordingly, we make the following assumption on (5.1) throughout.

Assumption 5.1 *The functions* $\mathbf{f}$, $\mathbf{x}_0$, $\mathbf{g}$ *and* $\frac{\partial \mathbf{g}}{\partial \mathbf{y}}$ *are factorable with natural interval and McCormick extensions* $[\mathbf{x}_0] : \mathbb{I}D_p \to \mathbb{IR}^{n_x}$, $\{\mathbf{x}_0\} : \mathbb{M}D_p \to \mathbb{MR}^{n_x}$, $[\mathbf{f}] : \mathbb{I}D_t \times \mathbb{I}D_p \times \mathbb{I}D_x \times \mathbb{I}D_y \to \mathbb{IR}^{n_x}$, $\{\mathbf{f}\} : \mathbb{M}D_t \times \mathbb{M}D_p \times \mathbb{M}D_x \times \mathbb{M}D_y \to \mathbb{MR}^{n_x}$, $[\mathbf{g}] : \mathbb{I}D_t \times \mathbb{I}D_p \times \mathbb{I}D_x \times \mathbb{I}D_y \to \mathbb{IR}^{n_y}$, $\{\mathbf{g}\} : \mathbb{M}D_t \times \mathbb{M}D_p \times \mathbb{M}D_x \times \mathbb{M}D_y \to \mathbb{MR}^{n_y}$, $[\frac{\partial \mathbf{g}}{\partial \mathbf{y}}] : \mathbb{I}D_t \times \mathbb{I}D_p \times \mathbb{I}D_x \times \mathbb{I}D_y \to \mathbb{IR}^{n_y \times n_y}$, *and* $\{\frac{\partial \mathbf{g}}{\partial \mathbf{y}}\} : \mathbb{M}D_t \times \mathbb{M}D_p \times \mathbb{M}D_x \times \mathbb{M}D_y \to \mathbb{MR}^{n_y \times n_y}$.

5.1 Theoretical Considerations

The treatment of the algebraic equations in (5.1) follows exactly the developments in Sect. 4. The first step is to apply the main result Theorem 4.1 to the algebraic equations $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) = \mathbf{0}$. In the semi-explicit form (5.1), the role of the algebraic equations $\mathbf{g}$ is essentially to specify the algebraic variables $\mathbf{z}_y$ given fixed values of t, $\mathbf{p}$, and the differential variables $\mathbf{z}_x$. Thus, in applying Theorem 4.1 to these equations, we identify the dependent variable $\mathbf{r}$ with $\mathbf{z}_y$ and the independent variable $\mathbf{s}$ with $(t, \mathbf{p}, \mathbf{z}_x)$. With these substitutions, the functions $\mathbf{M}$, $\boldsymbol{\psi}$, $[\boldsymbol{\psi}]$, $[\boldsymbol{\psi}_\cap]$, and $[\boldsymbol{\psi}_\Gamma]$ can be defined exactly as in Sect. 4. For example, in this case (4.4) becomes

$$\mathbf{M}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y, \boldsymbol{\lambda}') \equiv \begin{bmatrix} \frac{\partial g_1}{\partial \mathbf{y}}(t, \mathbf{p}, \mathbf{z}_x, \tilde{\mathbf{z}}_y + \lambda_1'(\mathbf{z}_y - \tilde{\mathbf{z}}_y)) \\ \vdots \\ \frac{\partial g_{n_y}}{\partial \mathbf{y}}(t, \mathbf{p}, \mathbf{z}_x, \tilde{\mathbf{z}}_y + \lambda_{n_y}'(\mathbf{z}_y - \tilde{\mathbf{z}}_y)) \end{bmatrix}, \tag{5.2}$$

for all $(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y, \boldsymbol{\lambda}') \in \mathbb{R}^{1+n_p+n_x+n_y+n_y}$, where $\tilde{\mathbf{z}}_y \in \mathbb{R}^{n_y}$ is a fixed reference point. Theorem 4.1 yields the following Corollary:

Corollary 5.1 *Let* $(J, P, Z_x, Z_y) \in \mathbb{I}D_t \times \mathbb{I}D_p \times \mathbb{I}D_x \times \mathbb{I}D_y$, $\tilde{\mathbf{z}}_y \in Z_y$, *and define*

$$Z_y' \equiv [\boldsymbol{\psi}_\Gamma](J, P, Z_x, Z_y, [\mathbf{0}, \mathbf{1}]). \tag{5.3}$$

The following conclusions hold:

1. *If* $(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) \in J \times P \times Z_x \times Z_y$ *satisfies* $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) = \mathbf{0}$, *then* $\mathbf{z}_y \in Z_y'$.
2. *If* $Z_y' = \emptyset$, *then* $\nexists (t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) \in J \times P \times Z_x \times Z_y$ *such that* $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) = \mathbf{0}$.
3. *If* $\tilde{\mathbf{z}}_y \in \mathrm{int}(Z_y)$ *and* $\emptyset \neq Z_y' \subset \mathrm{int}(Z_y)$, *then* $\exists \mathbf{h} \in C^1(J \times P \times Z_x, Z_y')$ *such that, for every* $(t, \mathbf{p}, \mathbf{z}_x) \in J \times P \times Z_x$, $\mathbf{z}_y = \mathbf{h}(t, \mathbf{p}, \mathbf{z}_x)$ *is the unique element of* Z_y *satisfying* $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) = \mathbf{0}$. *Moreover,* $[\mathbf{M}](J, P, Z_x, Z_y, [\mathbf{0}, \mathbf{1}])$ *does not contain a singular matrix and does not contain zero in any of its diagonal elements.*

Remark 5.1 As in Remark 4.2, note that $\tilde{\mathbf{z}}_y \in Z_y$ implies $\tilde{\mathbf{z}}_y + [\mathbf{0}, \mathbf{1}](Z_y - \tilde{\mathbf{z}}_y) = Z_y$, and hence

$$[\mathbf{M}](J, P, Z_x, Z_y, [\mathbf{0}, \mathbf{1}]) = \left[\frac{\partial \mathbf{g}}{\partial \mathbf{y}}\right](J, P, Z_x, Z_y). \tag{5.4}$$

The statement of Corollary 5.1 in [78] uses this substitution. As mentioned in Sect. 4, it is almost always necessary to precondition the system $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) = \mathbf{0}$ in order to pass the inclusion test of Conclusion 3, but we omit the preconditioner for brevity.

Conclusion 3 of Corollary 5.1 is crucial to the state bounding method presented below. First, it provides a computational test for the existence and uniqueness of an implicit function $\mathbf{h} : J \times P \times Z_x \to \mathbb{R}^{n_y}$ satisfying $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{h}(t, \mathbf{p}, \mathbf{z}_x)) = \mathbf{0}$, $\forall (t, \mathbf{p}, \mathbf{z}_x) \in J \times P \times Z_x$. The existence of this implicit function makes (5.1) equivalent to an explicit system of ODEs on $I \times P$, which provides an inroad for the application of state bounding methods for ODEs. However, the complication remains that $\mathbf{h}$ is not factorable. Thus, the second essential piece of information gained from Conclusion 3 is the a priori bound $\mathbf{h}(J, P, Z_x) \subset Z_y' \subset \text{int}(Z_y)$.

As discussed in Sect. 4, the inclusion test in Conclusion 3 of Corollary 5.1 has some further useful consequences. Specifically, when it holds for some (J, P, Z_x, Z_y), it implies that $\boldsymbol{\psi}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y, \boldsymbol{\lambda}')$ is defined for all $(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y, \boldsymbol{\lambda}') \in J \times P \times Z_x \times Z_y \times [\mathbf{0}, \mathbf{1}]$, and satisfies

$$\mathbf{h}(t, \mathbf{p}, \mathbf{z}_x) = \boldsymbol{\psi}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{h}(t, \mathbf{p}, \mathbf{z}_x), \boldsymbol{\lambda}(t, \mathbf{p}, \mathbf{z}_x)), \quad \forall (t, \mathbf{p}, \mathbf{z}_x) \in J \times P \times Z_x. \tag{5.5}$$

Moreover, we have $[\boldsymbol{\psi}_\Gamma](J, P, Z_x, Z_y) = [\boldsymbol{\psi}_\cap](J, P, Z_x, Z_y) = [\boldsymbol{\psi}](J, P, Z_x, Z_y)$. We will use both of these facts below.

The following theorem is the main result from [78] that enables the computation of state bounds. One of the main hypotheses is that the inclusion test of Conclusion 3 in Corollary 5.1 is satisfied pointwise in time; i.e., with $J = [t, t]$ and time-varying bounds $X(t)$ and $Y(t)$ in place of Z_x and Z_y.

Theorem 5.2 *Let $(I, P) \in \mathbb{I}D_t \times \mathbb{I}D_p$. Suppose that $\mathbf{y}^L, \mathbf{y}^U : I \to \mathbb{R}^{n_y}$ are continuous, $\mathbf{x}^L, \mathbf{x}^U : I \to \mathbb{R}^{n_x}$ are absolutely continuous, and for all $t \in I$, $\mathbf{x}^L(t) \leq \mathbf{x}^U(t)$, $\mathbf{y}^L(t) \leq \mathbf{y}^U(t)$, $X(t) \equiv [\mathbf{x}^L(t), \mathbf{x}^U(t)] \in \mathbb{I}D_x$, and $Y(t) \equiv [\mathbf{y}^L(t), \mathbf{y}^U(t)] \in \mathbb{I}D_y$. Let the following hypotheses hold:*

(IC): $\mathbf{x}_0(\mathbf{p}) \in X(t_0)$, $\forall \mathbf{p} \in P$.
(ALG): *For every $t \in I$, $\emptyset \neq [\boldsymbol{\psi}_\Gamma]([t, t], P, X(t), Y(t), [\mathbf{0}, \mathbf{1}]) \subset \text{int}(Y(t))$.*
(RHS): *For a.e. $t \in I$ and each index i,*

1. $\dot{x}_i^L(t) \leq f_i(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y)$ *for all* $(\mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) \in P \times X(t) \times Y(t)$ *such that* $z_{x,i} = x_i^L(t)$ *and* $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) = \mathbf{0}$,
2. $\dot{x}_i^U(t) \geq f_i(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y)$ *for all* $(\mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) \in P \times X(t) \times Y(t)$ *such that* $z_{x,i} = x_i^U(t)$ *and* $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \mathbf{z}_y) = \mathbf{0}$.

Then every regular solution of (5.1) *on* $I \times P$ *with* $\mathbf{y}(t_0, \hat{\mathbf{p}}) \in Y(t_0)$ *for at least one* $\hat{\mathbf{p}} \in P$ *must satisfy* $(\mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p})) \in X(t) \times Y(t)$ *for all* $(t, \mathbf{p}) \in I \times P$.

The proof of Theorem 5.2 is quite technical and can be found in Theorem 5.5.6 of [78]. However, the main ideas are not difficult to appreciate. Consider a solution $(\mathbf{x}, \mathbf{y})$ as in the statement of the theorem. First, note that Hypothesis (ALG) and Conclusion 3 of Corollary 5.1 imply that, for every $t \in I$, there exists $\mathbf{h}(t, \cdot, \cdot) : P \times X(t) \to \text{int}(Y(t))$ such that $\mathbf{h}(t, \mathbf{p}, \mathbf{z}_x)$ solves $\mathbf{g}(t, \mathbf{p}, \mathbf{z}_x, \cdot) = \mathbf{0}$ uniquely among elements of $Y(t)$, provided that $(\mathbf{p}, \mathbf{z}_x) \in P \times X(t)$. With $\mathbf{z}_x = \mathbf{x}(t, \mathbf{p})$, $\mathbf{y}(t, \mathbf{p})$ satisfies this equation by definition. Thus, we have the following implication for all $(t, \mathbf{p}) \in I \times P$:

$$\mathbf{x}(t, \mathbf{p}) \in X(t), \quad \mathbf{y}(t, \mathbf{p}) \in Y(t) \quad \Longrightarrow \quad \mathbf{y}(t, \mathbf{p}) = \mathbf{h}(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p})) \in \text{int}(Y(t)). \tag{5.6}$$

Then, from Hypotheses (IC) and the assumption that $\mathbf{y}(t_0, \hat{\mathbf{p}}) \in Y(t_0)$, continuity of $\mathbf{y}(t_0, \cdot)$ implies that $\mathbf{y}(t_0, \cdot)$ is confined to the interior of $Y(t)$ on all of P. Thus, $(\mathbf{x}(t_0, \mathbf{p}), \mathbf{y}(t_0, \mathbf{p})) \in X(t_0) \times Y(t_0)$ for all $\mathbf{p} \in P$.

Now, in order for $(\mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p}))$ to leave the bounds $X(t) \times Y(t)$ at some $t > t_0$ and for some $\mathbf{p} \in P$, it must happen that $(\mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p}))$ first intersects the boundary of $X(t) \times Y(t)$. Here again, (5.6) implies that $\mathbf{y}(t, \mathbf{p})$ cannot reach the boundary of $Y(t)$ before $\mathbf{x}(t, \mathbf{p})$ leaves $X(t)$. Thus, if a violation is to occur, we must come to a point such that $(\mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p})) \in X(t) \times Y(t)$ and either $x_i(t, \mathbf{p}) = x_i^L(t)$ or $x_i(t, \mathbf{p}) = x_i^U(t)$ for at least one i. Supposing that the first case occurs, we now note that $(\mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p}))$ satisfies all of the conditions imposed on $(\mathbf{z}_x, \mathbf{z}_y)$ in Hypothesis (RHS).1 of the theorem, and it follows that

$$\dot{x}_i^L(t) \leq f_i(t, \mathbf{p}, \mathbf{x}(t, \mathbf{p}), \mathbf{y}(t, \mathbf{p})) = \dot{x}_i(t, \mathbf{p}). \tag{5.7}$$

From the inequality $\dot{x}_i^L(t) \leq \dot{x}_i(t, \mathbf{p})$ and some further technical conditions it can then be shown that it is impossible for $\dot{x}_i(\cdot, \mathbf{p})$ to cross $\dot{x}_i^L$ at t, completing the argument.

Note that Hypotheses (IC) and (RHS) in Theorem 5.2 involve global information about the functions $\mathbf{f}$ and $\mathbf{x}_0$ in the form of bounds on their images. The content of Theorem 5.2 is then to deduce global information about the solution $(\mathbf{x}, \mathbf{y})$ from global information about the system equations. In this sense, Theorem 5.2 is well aligned with our strategy so far. Not surprisingly, the implementation of these ideas in the next section uses the techniques of Sect. 3 to obtain the required bounds on the system equations, which have known factorable representations.

5.2 Implementation

Using the theoretical developments of the previous two sections, we now derive an auxiliary system of semi-explicit DAEs that describes valid state bounds as its

solutions. Let $\mathscr{B}_i^L, \mathscr{B}_i^U : \mathbb{IR}^{n_x} \to \mathbb{IR}^{n_x}$ be defined by $\mathscr{B}_i^L([\mathbf{v},\mathbf{w}]) = \{\mathbf{z} \in [\mathbf{v},\mathbf{w}] : z_i = v_i\}$ and $\mathscr{B}_i^U([\mathbf{v},\mathbf{w}]) = \{\mathbf{z} \in [\mathbf{v},\mathbf{w}] : z_i = w_i\}$, for every $i = 1,\ldots,n_x$. Moreover, let $[\mathbf{f}]^L$ and $[\mathbf{f}]^U$ denote the upper and lower bounds of $[\mathbf{f}]$, and defined $[\boldsymbol{\psi}]^L$ and $[\boldsymbol{\psi}]^U$ analogously. For any continuous and pointwise positive $\gamma : I \to \mathbb{R}$, consider the initial value problem in DAEs

$$\dot{x}_i^L(t) = [f_i]^L\left([t,t], P, \mathscr{B}_i^L(X(t)), Y(t)\right), \tag{5.8}$$

$$\dot{x}_i^U(t) = [f_i]^U\left([t,t], P, \mathscr{B}_i^U(X(t)), Y(t)\right), \tag{5.9}$$

$$\mathbf{0} = \mathbf{y}^L(t) - [\boldsymbol{\psi}]^L([t,t], P, X(t), Y(t), [\mathbf{0},\mathbf{1}]) + \mathbf{1}\gamma(t), \tag{5.10}$$

$$\mathbf{0} = -\mathbf{y}^U(t) + [\boldsymbol{\psi}]^U([t,t], P, X(t), Y(t), [\mathbf{0},\mathbf{1}]) + \mathbf{1}\gamma(t), \tag{5.11}$$

for all $i = 1,\ldots,n_x$, with initial conditions

$$[\mathbf{x}^L(t_0), \mathbf{x}^U(t_0)] = [\mathbf{x}_0](P). \tag{5.12}$$

A solution of (5.8)–(5.12) is any $\mathbf{x}^L, \mathbf{x}^U : I \to \mathbb{R}^{n_x}$ and $\mathbf{y}^L, \mathbf{y}^U : I \to \mathbb{R}^{n_y}$ satisfying (5.8)–(5.12) for all $t \in I$, with $\mathbf{x}^L$ and $\mathbf{x}^U$ continuously differentiable and $\mathbf{y}^L$ and $\mathbf{y}^U$ piecewise C^1 (see [22] for a definition of this class of functions).

The main result of [79] is that any solution of (5.8)–(5.12) provides state bounds for a unique regular solution of (2.1) on $I \times P$. Of course, this is a direct application of Theorem 5.2. Clearly, (5.12) verifies Hypothesis (IC) since natural interval extensions are inclusion functions (see Sect. 3). Furthermore, (5.10) and (5.11), along with positivity of γ, imply that $[\boldsymbol{\psi}]([t,t], P, X(t), Y(t), [\mathbf{0},\mathbf{1}])$ is strictly contained in $Y(t)$. Typically, $\gamma = 10^{-6}$ works well in practice. Observing that $[\boldsymbol{\psi}] = [\boldsymbol{\psi}_\Gamma]$ for all arguments such that (5.10) and (5.11) hold (see the discussion following Corollary 5.1), Hypothesis (ALG) is verified. The function $[\boldsymbol{\psi}]$ is used in place of $[\boldsymbol{\psi}_\Gamma]$ here for numerical reasons that are beyond the scope of this review [79]. Finally, (5.8) and (5.9) establish Hypothesis (RHS), again through the inclusion properties of the natural interval extension (note that the operators $\mathscr{B}_i^L$ and $\mathscr{B}_i^U$ are used to reflect the conditions $z_{x,i} = x_i^L(t)$ and $z_{x,i} = x_i^U(t)$ in Hypothesis (RHS), respectively).

With some technical modifications to Eqs. (5.8)–(5.12) that are beyond the scope of this article, it is possible to assert the existence of a regular solution within the computed bounds [79, Theorem 6.6.4]. Note that Theorem 5.2 does not assert existence.

Theorem 5.3 *Let* $(\mathbf{x}^L, \mathbf{x}^U, \mathbf{y}^L, \mathbf{y}^U)$ *be a solution of* (5.8)–(5.12) *on* I. *Then there exists a unique regular solution* $(\mathbf{x}, \mathbf{y})$ *of* (2.1) *on* $I \times P$ *satisfying* $\mathbf{x}(t,\mathbf{p}) \in X(t)$ *and* $\mathbf{y}(t,\mathbf{p}) \in Y(t)$, $\forall (t,\mathbf{p}) \in I \times P$.

In light of the discussion above, state bounds are given by solving the DAEs (5.8)–(5.12) for $\mathbf{x}^L$, $\mathbf{x}^U$, $\mathbf{y}^L$, and $\mathbf{y}^U$. This can be done using any state-of-the-art DAE solver, for example, `IDA` [27]. Moreover, the required natural interval extensions can be computed automatically using operator overloading as described in Sect. 3.

The only caveat is that the governing equations in (5.8)–(5.12) are nonsmooth and are undefined for certain arguments, e.g., $x_i^L(t) > x_i^U(t)$. Arguments of this type do not occur along solution trajectories, but may occur during numerical integration. For these reasons, some specialized methods are required, as discussed in [79].

Given (5.12), consistent initial conditions for $\mathbf{y}^L$ and $\mathbf{y}^U$ can be computed by solving (5.10) and (5.11) at t_0. If (5.1) has multiple regular solutions, then there may be multiple consistent initial conditions, and which one is chosen determines which solution branch will be bounded by the solution of (5.8)–(5.12). In particular, at t_0, (5.10) and (5.11) verifies the existence of an implicit function $\mathbf{h}(t_0,\cdot,\cdot) : P \times X(t_0) \to Y(t_0)$ providing the unique solution of $\mathbf{g}(t_0,\mathbf{p},\mathbf{x}(t_0,\mathbf{p}),\cdot) = \mathbf{0}$ within $Y(t_0)$, but says nothing about potential solutions outside of $Y(t_0)$. Thus, there may exist other solutions that lie outside $Y(t_0)$ for all $\mathbf{p} \in P$, and these solutions will not be bounded. Example 2.1 provides a simple DAE of this type.

The bounding equations (5.8)–(5.12) only apply to regular (i.e., index-one) solutions. In particular, the algebraic equations (5.10) and (5.11) cannot be satisfied by any $X(t)$ and $Y(t)$ containing a solution that is not regular at t. This is because satisfaction of (5.10) and (5.11) implies by Conclusion 3 of Corollary 5.1 that $\frac{\partial \mathbf{g}}{\partial \mathbf{y}}(t,\mathbf{p},\mathbf{z}_x,\mathbf{z}_y)$ is non-singular for every $(\mathbf{p},\mathbf{z}_x,\mathbf{z}_y) \in P \times X(t) \times Y(t)$. Thus, the solution of (5.8)–(5.12) will cease to exist if the bounded solution approaches a bifurcation point. In general, even in the case of a unique regular solution, there is no guarantee that (5.8)–(5.12) has a solution for a given I and P. However, (5.8)–(5.12) has been shown to have solutions providing tight state bounds for several examples of practical interest [79]. Moreover, whenever a solution of (5.8)–(5.12) does exist, it is guaranteed that a unique regular solution of (5.1) exists within the computed bounds.

It is possible to tighten the bounds given by the solution of (5.8)–(5.12) quite considerably by refining the argument $Y(t)$ in (5.8) and (5.9). In fact, this is the standard procedure as presented in [79] and was initially omitted here only for simplicity of exposition. To develop this refinement procedure, we note that from (5.6) we have

$$\mathbf{y}(t,\mathbf{p}) = \mathbf{h}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p})), \quad \forall (t,\mathbf{p}) \in I \times P. \tag{5.13}$$

Then, using (5.5) with $J = [t,t]$, $Z_x = X(t)$, and $Z_y = Y(t)$, it follows that

$$\mathbf{y}(t,\mathbf{p}) = \boldsymbol{\psi}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}),\mathbf{y}(t,\mathbf{p}),\boldsymbol{\lambda}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}))), \quad \forall (t,\mathbf{p}) \in I \times P. \tag{5.14}$$

Thus, we may refine $Y(t)$ at each t through the iteration

$$Y_{k+1}(t) = [\boldsymbol{\psi}_{\Gamma}]([t,t],P,X(t),Y_k(t),[\mathbf{0},\mathbf{1}]), \quad Y_0(t) = Y(t). \tag{5.15}$$

Further refinement is possible if we consider refining only the particular interval $Y(t)$ that will be used to evaluate $[f_i]^L$ in (5.8). Of course, analogous procedures also apply to the argument of $[f_i]^U$, and for $j \neq i$. Considering the condition that $[f_i]^L$ must satisfy according to Hypothesis (RHS) in Theorem 5.2, the argument $Y(t)$

in (5.8) can be refined by the iteration

$$Y_{k+1}(t) = [\boldsymbol{\psi}_\Gamma]([t,t], P, \mathscr{B}_i^L(X(t)), Y_k(t), [\mathbf{0},\mathbf{1}]), \quad Y_0(t) = Y(t). \tag{5.16}$$

In practice, these refinements have been found to lead to much sharper results than applying (5.8)–(5.12) directly.

5.3 *Alternative Approaches*

As mentioned in the Sect. 1, an alternative state bounding approach for semi-explicit DAEs has been proposed in [65]. The treatment of the algebraic equations in this method is very similar to that presented here, using an interval Newton method as described in Sect. 4. However, in [65] the *interval Krawczyk method* is used instead of the interval Hansen-Sengupta method. The interval Krawczyk method is based on a different rearrangement of (4.5) that leads to a weaker enclosure but avoids the issue of division by an interval containing zero. The reader is referred to [59] for a comprehensive description and comparison of the two methods.

In its treatment of the differential equations, the method proposed in [65] differs significantly from the developments here. Rather than deriving a *bounding system* which can be solved numerically to obtain state bounds [e.g., (5.8)–(5.12)], the method applies the interval Krawczyk method to the differential equations as well. In essence, the authors propose a time-stepping scheme in which, in each interval $[t_j, t_{j+1}]$, the interval Krawczyk method is applied to the nonlinear system of equations

$$\mathbf{0} = -\boldsymbol{\sigma}_x + \mathbf{f}(t, \mathbf{p}, \mathbf{x}_j + (t-t_j)\boldsymbol{\sigma}_x, \mathbf{z}_y), \tag{5.17}$$

$$\mathbf{0} = \mathbf{g}(t, \mathbf{p}, \mathbf{x}_j + (t-t_j)\boldsymbol{\sigma}_x, \mathbf{z}_y). \tag{5.18}$$

Above, $\boldsymbol{\sigma}_x$ and $\mathbf{z}_y$ are dummy variables for $\dot{\mathbf{x}}$ and $\mathbf{y}$, respectively, and the remaining variables are known to lie within interval bounds at the beginning of the time step (e.g., $t \in [t_j, t_{j+1}]$, $\mathbf{p} \in P$, $\mathbf{x}_j \in X_j$). The authors of [65] claim that, if an inclusion test similar to that in Conclusion 3 of Theorem 4.1 is passed with some intervals Z_y and Σ_x, then a unique solution of the DAE system is guaranteed to exist on $[t_j, t_{j+1}] \times P$, and $\mathbf{y}(t,\mathbf{p}) \in Z_y$ and $\dot{\mathbf{x}}(t,\mathbf{p}) \in \Sigma_x$, $\forall (t,\mathbf{p}) \in [t_j, t_{j+1}] \times P$. These intervals can then be refined by interval iteration on (5.17)–(5.18) as described in Sect. 4. Finally, the differential state is bounded by

$$\mathbf{x}(t,\mathbf{p}) = \mathbf{x}(t_j,\mathbf{p}) + \int_{t_j}^{t} \dot{\mathbf{x}}(s,\mathbf{p})ds \in X_j + [0, (t_{j+1}-t_j)]\Sigma_x, \tag{5.19}$$

for all $(t,\mathbf{p}) \in [t_j, t_{j+1}] \times P$, and the next time step can be initialized.

The existence and uniqueness claim is not proven in [65], and does not follow from a careful application of Theorem 4.1 to (5.17) and (5.18). Thus some skepticism is warranted. However, it is likely that the method can be shown to be valid using a fixed-point result for ODE solutions, such as the successive substitution method. A similar proof is used to establish an interval existence and uniqueness test for DAE solutions in Theorem 4.2 of [79].

Compared to the method presented in Sects. 5.1 and 5.2, the method of [65] has the apparent disadvantage that it requires an interval inclusion test to pass on Eqs. (5.17) and (5.18), which notably involve equations related to the differential equations in addition to the original algebraic equations. As discussed in [78], and briefly in Sect. 8, passing the interval inclusion test numerically can be quite difficult when P is large in width. Thus, applying such a test to an inflated system of equations may cause unnecessary failure of the method for some problems. Ultimately, a thorough computational comparison of these two methods will be required to clarify the possible advantages and disadvantages of each approach.

6 State Relaxations for Semi-explicit Index-One DAEs

In this section, we take up the computation of state relaxations for (5.1). It assumed throughout that state bounds have been computed via the procedure of Sect. 5. Thus, $X : I \to \mathbb{IR}^{n_x}$ and $Y : I \to \mathbb{IR}^{n_y}$ satisfying (5.8)–(5.12) are available, and hence we are ensured that a unique regular solution $(\mathbf{x}, \mathbf{y})$ of (5.1) exists satisfying

$$(\mathbf{x}(t,\mathbf{p}), \mathbf{y}(t,\mathbf{p})) \in X(t) \times Y(t), \qquad \forall (t,\mathbf{p}) \in I \times P. \tag{6.1}$$

Thus, the objective of this section is to compute functions $\mathbf{x}^{cv}, \mathbf{x}^{cc} : I \times P \to \mathbb{R}^{n_x}$ and $\mathbf{y}^{cv}, \mathbf{y}^{cc} : I \times P \to \mathbb{R}^{n_y}$ such that

$$\mathbf{x}^{cv}(t,\mathbf{p}) \le \mathbf{x}(t,\mathbf{p}) \le \mathbf{x}^{cc}(t,\mathbf{p}) \quad \text{and} \quad \mathbf{y}^{cv}(t,\mathbf{p}) \le \mathbf{y}(t,\mathbf{p}) \le \mathbf{y}^{cc}(t,\mathbf{p}), \tag{6.2}$$

for all $(t,\mathbf{p}) \in I \times P$, and, for each $t \in I$, $\mathbf{x}^{cv}(t,\cdot)$ and $\mathbf{y}^{cv}(t,\cdot)$ are convex on P, and $\mathbf{x}^{cc}(t,\cdot)$ and $\mathbf{y}^{cc}(t,\cdot)$ are concave on P. These functions are called *state relaxations for* $(\mathbf{x},\mathbf{y})$ *on* $I \times P$. More generally, functions $\mathscr{X} : I \times \mathbb{M}P \to \mathbb{MR}^{n_x}$ and $\mathscr{Y} : I \times \mathbb{M}P \to \mathbb{MR}^{n_y}$ will be computed such that, for each fixed $t \in I$, $\mathscr{X}(t,\cdot)$ and $\mathscr{Y}(t,\cdot)$ are relaxation functions for $\mathbf{x}(t,\cdot)$ and $\mathbf{y}(t,\cdot)$, respectively.

As in the previous section, we begin by treating the algebraic equations. However, matters here are much simpler since existence and uniqueness of the solution $(\mathbf{x},\mathbf{y})$ has already been established by the state bounding procedure. The remaining task is simply to derive a procedure that computes the inclusion function $\mathscr{Y}$ supposing that $\mathscr{X}$ is known. This will then be used during the computation of $\mathscr{X}$ described below to obtain $\mathscr{X}$ and $\mathscr{Y}$ simultaneously. The starting point for this derivation is the relation (5.14), established in the previous section. Mirroring the developments of Sect. 4, this relation immediately suggest that a sequence of

progressively refined relaxation functions for $\mathbf{y}(t,\cdot)$ can be computed through the iteration

$$\begin{aligned}\mathscr{Y}_{k+1}(t,\mathscr{P}) &= \{\boldsymbol{\psi}_\cap\}(t,\mathscr{P},\mathscr{X}(t,\mathscr{P}),\mathscr{Y}_k(t,\mathscr{P}),\mathscr{L}),\\ \mathscr{Y}_0(t,\mathscr{P}) &= (Y(t),Y(t)),\end{aligned} \tag{6.3}$$

for all $\mathscr{P}\in\mathbb{M}P$, where $\mathscr{L}=([\mathbf{0},\mathbf{1}],[\mathbf{0},\mathbf{1}])$ is a relaxation function for $\boldsymbol{\lambda}$ in (5.14). Of course, $\mathscr{Y}_0(t,\cdot)$ is trivially a relaxation function for $\mathbf{y}(t,\cdot)$ by (6.1). Assuming this is true of $\mathscr{Y}_k(t,\cdot)$, and further considering that $\mathscr{X}(t,\cdot)$ is assumed to be a relaxation function for $\mathbf{x}(t,\cdot)$, the properties of $\{\boldsymbol{\psi}_\cap\}$ imply that $\mathscr{Y}_{k+1}(t,\cdot)$ is also a relaxation function for $\mathbf{y}(t,\cdot)$, as discussed in Sect. 4. Finally, note that all divisions by McCormick objects in the evaluation of $\{\boldsymbol{\psi}_\cap\}$ above are guaranteed to be defined for all $\mathscr{P}\in\mathbb{M}P$ on account of Conclusion 3 of Corollary 5.1 and (5.10) and (5.11), provided that the interval part of $\mathscr{X}(t)$ is $X(t)$. Then, convex and concave relaxations of $\mathbf{y}(t,\cdot)$ on $P'\subset P$ are evaluated computationally at a particular point $\mathbf{p}\in P'$ by simply setting $\mathscr{P}=(P',[\mathbf{p},\mathbf{p}])$ and executing the iteration (6.3). For any $k\geq 0$, the result is a McCormick object $\mathscr{Y}_k(\mathscr{P})=(Y_k(t,P'),[\mathbf{y}_k^{cv}(t,\mathbf{p}),\mathbf{y}_k^{cc}(t,\mathbf{p})])$ with the obvious interpretation. Moving forward, we use the notation

$$\mathscr{Y}(t,\mathscr{P}) = \mathscr{H}_K(t,\mathscr{P},\mathscr{X}(t,\mathscr{P})),\quad \forall\mathscr{P}\in\mathbb{M}P, \tag{6.4}$$

to denote the relaxation function $\mathscr{Y}(t,\mathscr{P})$ computed from $\mathscr{X}(t,\mathscr{P})$ through the iteration (6.3) truncated at $K>0$.

We are now prepared to state the main result from [81] used to compute state bounds for $(\mathbf{x},\mathbf{y})$. By Assumption 5.1, $\mathbf{f}$ is factorable, so we can consider its natural McCormick extension, $\{\mathbf{f}\}$. Below, we use the notation $\{\mathbf{f}\}=([\mathbf{f}],[\{\mathbf{f}\}^{cv},\{\mathbf{f}\}^{cc}])$.

Theorem 6.1 *Let $K>0$ and, for every $\mathscr{P}\in\mathbb{M}P$, and let $\mathscr{X}^{cv}(\cdot,\mathscr{P}),\mathscr{X}^{cc}(\cdot,\mathscr{P}):I\to\mathbb{R}^{n_x}$ be solutions of the following initial value problem in ODEs*

$$\dot{\mathscr{X}}^{cv}(t,\mathscr{P}) = \{\mathbf{f}\}^{cv}(t,\mathscr{P},\mathscr{X}(t,\mathscr{P}),\mathscr{Y}(t,\mathscr{P})), \tag{6.5}$$

$$\dot{\mathscr{X}}^{cc}(t,\mathscr{P}) = \{\mathbf{f}\}^{cc}(t,\mathscr{P},\mathscr{X}(t,\mathscr{P}),\mathscr{Y}(t,\mathscr{P})), \tag{6.6}$$

$$\mathscr{X}(t,\mathscr{P}) = (X(t),X(t)\cap[\mathscr{X}^{cv}(t,\mathscr{P}),\mathscr{X}^{cc}(t,\mathscr{P})]), \tag{6.7}$$

$$\mathscr{Y}(t,\mathscr{P}) = \mathscr{H}_K(t,\mathscr{P},\mathscr{X}(t,\mathscr{P})), \tag{6.8}$$

$$\mathscr{X}(t_0,\mathscr{P}) = \{\mathbf{x}_0\}(\mathscr{P}). \tag{6.9}$$

Then, for each $t\in I$, $\mathscr{X}(t,\cdot)$ and $\mathscr{Y}(t,\cdot)$ are relaxation functions for $\mathbf{x}(t,\cdot)$ and $\mathbf{y}(t,\cdot)$, respectively. In particular, state relaxations for $(\mathbf{x},\mathbf{y})$ on $P'\subset P$ are given by the definitions,

$$\mathbf{x}^{cv}(t,\mathbf{p}) = \mathscr{X}^{cv}(t,(P',[\mathbf{p},\mathbf{p}])),\qquad \mathbf{y}^{cv}(t,\mathbf{p}) = \mathscr{Y}^{cv}(t,(P',[\mathbf{p},\mathbf{p}])), \tag{6.10}$$

$$\mathbf{x}^{cc}(t,\mathbf{p}) = \mathscr{X}^{cc}(t,(P',[\mathbf{p},\mathbf{p}])),\qquad \mathbf{y}^{cc}(t,\mathbf{p}) = \mathscr{Y}^{cc}(t,(P',[\mathbf{p},\mathbf{p}])), \tag{6.11}$$

for all $(t,\mathbf{p})\in I\times P$.

By Theorem 6.1, convex and concave relaxations of $\mathbf{x}(t,\cdot)$ and $\mathbf{y}(t,\cdot)$ on $P' \subset P$ are evaluated computationally at a particular point $\mathbf{p} \in P'$ by simply setting $\mathscr{P} = (P', [\mathbf{p},\mathbf{p}])$ and solving (6.5)–(6.9). Since $\mathscr{Y}(t,\mathscr{P})$ can be computed explicitly given $\mathscr{X}(t,\mathscr{P})$ [i.e., through a fixed number of iterations of (6.3)], this system can be treated as a system of explicit ODEs. Thus, (6.5)–(6.9) can be solved by any numerical integration code (e.g., `CVODES` [27]), using a McCormick arithmetic library (see Sect. 3).

The proof of Theorem 6.1 is based on the integral form of the differential equations in (5.1) ,

$$\mathbf{x}(t,\mathbf{p}) = \mathbf{x}_0(\mathbf{p}) + \int_{t_0}^{t} \mathbf{f}(s,\mathbf{p},\mathbf{x}(s,\mathbf{p}),\mathbf{y}(s,\mathbf{p}))ds, \quad \forall (t,\mathbf{p}) \in I \times P. \tag{6.12}$$

Based on a similar rearrangement of (6.5) and (6.6) we define a sequence of functions $\mathscr{X}^{cv,\ell}(\cdot,\mathscr{P}), \mathscr{X}^{cc,\ell}(\cdot,\mathscr{P}) : I \to \mathbb{R}^{n_x}$ with $\ell \in \mathbb{N}$ by

$$\mathscr{X}^{cv,\ell+1}(t,\mathscr{P}) = \{\mathbf{x}_0\}^{cv}(\mathscr{P}) + \int_{t_0}^{t} \{\mathbf{f}\}^{cv}(s,\mathscr{P},\mathscr{X}^{\ell}(s,\mathscr{P}),\mathscr{Y}^{\ell}(s,\mathscr{P}))ds, \tag{6.13}$$

$$\mathscr{X}^{cc,\ell+1}(t,\mathscr{P}) = \{\mathbf{x}_0\}^{cc}(\mathscr{P}) + \int_{t_0}^{t} \{\mathbf{f}\}^{cc}(s,\mathscr{P},\mathscr{X}^{\ell}(s,\mathscr{P}),\mathscr{Y}^{\ell}(s,\mathscr{P}))ds, \tag{6.14}$$

$$\mathscr{X}^{\ell}(t,\mathscr{P}) = (X(t), X(t) \cap [\mathscr{X}^{cv,\ell}(t,\mathscr{P}), \mathscr{X}^{cc,\ell}(t,\mathscr{P})]), \tag{6.15}$$

$$\mathscr{Y}^{\ell}(t,\mathscr{P}) = \mathscr{H}_K(t,\mathscr{P},\mathscr{X}^{\ell}(t,\mathscr{P})). \tag{6.16}$$

Using some regularity properties of $\{\mathbf{f}\}$, it can be shown that these sequences converge to the true solutions of (6.5)–(6.9), for every $\mathscr{P} \in \mathbb{M}P$ and every continuous choice of $\mathscr{X}^{cv,0}(\cdot,\mathscr{P})$ and $\mathscr{X}^{cc,0}(\cdot,\mathscr{P})$. Choosing these initial approximations such that $[\mathscr{X}^{cv,0}(t,\mathscr{P}), \mathscr{X}^{cc,0}(t,\mathscr{P})] = X(t)$ for all $t \in I$ and $\mathscr{P} \in \mathbb{M}P$, it follows that $\mathscr{X}^{0}(t,\cdot)$ is a relaxation function for $\mathbf{x}(t,\cdot)$ for each $t \in I$. Assuming that this is true of $\mathscr{X}^{\ell}(t,\cdot)$ it is shown for $\mathscr{X}^{\ell+1}(t,\cdot)$ as follows. First, from the definition of $\mathscr{H}_K$ above, it is guaranteed that $\mathscr{Y}^{\ell}(t,\cdot)$ is a relaxation function for $\mathbf{y}(t,\cdot)$ for each $t \in I$. Then, the definition of $\{\mathbf{f}\}$ and the composition theorem for relaxation functions (Theorem 3.2) imply that $\mathscr{P} \mapsto \{\mathbf{f}\}(t,\mathscr{P},\mathscr{X}^{\ell}(t,\mathscr{P}),\mathscr{Y}^{\ell}(t,\mathscr{P}))$ is a relaxation function for $\mathbf{p} \mapsto \mathbf{f}(t,\mathbf{p},\mathbf{x}(t,\mathbf{p}),\mathbf{y}(t,\mathbf{p}))$. Using basic properties of the integral, it can therefore be shown that the right-most terms in (6.13) and (6.14) together form a relaxation function for the right-most term of (6.12). In fact, this same relationship is also true of the first terms on the right-hand sides of (6.13), (6.14), and (6.12). It follows then that $\mathscr{X}^{\ell+1}(t,\cdot)$ is a relaxation function for $\mathbf{x}(t,\cdot)$ for each $t \in I$. This recovers the inductive hypothesis at $\ell+1$, and hence for all ℓ, and taking limits then guarantees that $\mathscr{X}(t,\cdot)$ is a relaxation function for $\mathbf{x}(t,\cdot)$ for each $t \in I$, which proves the theorem. For a formal mathematical argument, the reader is referred to [81].

The state relaxations described by (6.5)–(6.9) can be considerably tightened through a slightly more involved formulation [74]:

$$\dot{\mathscr{X}}_i^{cv}(t,\mathscr{P}) = \{f_i\}^{cv}(t,\mathscr{P},\underline{\mathscr{X}}(i,t,\mathscr{P}),\underline{\mathscr{Y}}(i,t,\mathscr{P})), \tag{6.17}$$

$$\dot{\mathscr{X}}_i^{cc}(t,\mathscr{P}) = \{f_i\}^{cc}(t,\mathscr{P},\overline{\mathscr{X}}(i,t,\mathscr{P}),\overline{\mathscr{Y}}(i,t,\mathscr{P})), \tag{6.18}$$

$$\underline{\mathscr{X}}(i,t,\mathscr{P}) = (X(t),\mathscr{B}_i^L([\mathscr{X}^{cv}(t,\mathscr{P}),\mathscr{X}^{cc}(t,\mathscr{P})])), \tag{6.19}$$

$$\overline{\mathscr{X}}(i,t,\mathscr{P}) = (X(t),\mathscr{B}_i^U([\mathscr{X}^{cv}(t,\mathscr{P}),\mathscr{X}^{cc}(t,\mathscr{P})])), \tag{6.20}$$

$$\underline{\mathscr{Y}}(i,t,\mathscr{P}) = \mathscr{H}_K(t,\mathscr{P},\underline{\mathscr{X}}(i,t,\mathscr{P})), \tag{6.21}$$

$$\overline{\mathscr{Y}}(i,t,\mathscr{P}) = \mathscr{H}_K(t,\mathscr{P},\overline{\mathscr{X}}(i,t,\mathscr{P})), \tag{6.22}$$

$$\mathscr{X}(t_0,\mathscr{P}) = \{\mathbf{x}_0\}(\mathscr{P}). \tag{6.23}$$

The argument that the solutions of this system provide state relaxations is much more involved than that for (6.5)–(6.9). The validity of the bounding relations (6.2) in this case results from an argument similar to that given after Theorem 5.2 with regard to the constraints $z_{x,i} = x_i^L(t)$ and $z_{x,i} = x_i^U(t)$ appearing in Hypothesis (RHS) of that theorem. For this reason, the interval operators $\mathscr{B}_i^{L/U}$ appear in (6.17)–(6.23), as they did in the bounding system (5.8)–(5.12). The convexity and concavity of the state bounds obtained via (6.17)–(6.23) again rely on the composition properties of relaxation functions (Theorem 3.2), but the argument is made difficult by the use of $\mathscr{B}_i^{L/U}$. In fact, due to these difficulties, (6.17)–(6.23) can only be guaranteed to provide valid state relaxations when the solutions satisfy $[\mathscr{X}^{cv}(t,\mathscr{P}),\mathscr{X}^{cc}(t,\mathscr{P})] \subset X(t)$, $\forall t \in I$. However, this can be ensured by a modification of (6.17)–(6.23) that yields a hybrid system with sliding modes. The reader is referred to [74] for these technical details (an analogous development for explicit ODEs can be found in [82]).

In the next section, it is shown that the convex and concave parameter dependence of state bounds is very useful for solving global dynamic optimization problems. However, for standard problems in reachability analysis, it is typically desirable to have an enclosure that is not parameter dependent. Of course, state bounds provide exactly this kind of enclosure. However, state relaxations can be used to provide a sharper enclosure $E(t) \subset \mathbb{R}^{n_x+n_y}$ through the definition

$$E(t) \equiv \bigcup_{\mathbf{p}\in P} ([\mathbf{x}^{cv}(t,\mathbf{p}),\mathbf{x}^{cc}(t,\mathbf{p})] \times [\mathbf{y}^{cv}(t,\mathbf{p}),\mathbf{y}^{cc}(t,\mathbf{p})]), \qquad \forall t \in I. \tag{6.24}$$

It is straightforward to argue that $(\mathbf{x}(t,\mathbf{p}),\mathbf{y}(t,\mathbf{p})) \in E(t)$, $\forall (t,\mathbf{p}) \in I \times P$, and moreover that $E(t)$ is convex for each fixed $t \in I$. From this last observation, it follows that $E(t)$ can be outer-approximated by a polytope of the form

$$E_{\text{poly}}(t) \equiv \{(\mathbf{z}_x,\mathbf{z}_y) : \boldsymbol{\lambda}_k^{\mathrm{T}}\mathbf{z}_x + \boldsymbol{\sigma}_k^{\mathrm{T}}\mathbf{z}_y \le c_k(t),\ k = 1,\ldots,K\}, \tag{6.25}$$

by defining

$$c_k(t) \equiv \max_{\mathbf{p}\in P} \Big[\sum_{i=1}^{n_x} \max(\lambda_{k,i} x_i^{cv}(t,\mathbf{p}), \lambda_{k,i} x_i^{cc}(t,\mathbf{p})) + \sum_{i=1}^{n_y} \max(\sigma_{k,i} y_i^{cv}(t,\mathbf{p}), \sigma_{k,i} y_i^{cc}(t,\mathbf{p}))\Big], \tag{6.26}$$

for all $t \in I$. Using the convexity/concavity properties of the state relaxations, it is straightforward to show that these maximizations are concave for any $\boldsymbol{\lambda}_k \in \mathbb{R}^{n_x}$ and $\boldsymbol{\sigma}_k \in \mathbb{R}^{n_y}$, and can therefore be solved efficiently [74, 75].

6.1 Alternative Approaches

At present, no alternative approaches have been proposed for computing state relaxations for DAEs. However, several methods are available for ODEs that can likely be extended to the DAE case without undue complications. Indeed, once state bounds have been computed, the issue of existence and uniqueness is resolved and the relaxation procedure can be seen as a refinement of the bounds. In this context, one promising direction is the extension of the state relaxation methods for ODEs in [70, 71]. These methods are fundamentally different from the state relaxation method proposed here in that they do not define a relaxed dynamic system to be solved numerically [e.g., as in (6.17)–(6.23)]. Rather, these methods first discretize the dynamics, and then compute state relaxations for an approximate solution that is essentially known in factorable form. These relaxations are then made valid by adding a bound on the approximation error. Despite this difference, these methods are quite similar to the method proposed here in their use of factorable representations and McCormick relaxations as fundamental building blocks (in addition to so-called *Taylor-model arithmetic*, a more complex enclosure method for factorable functions that was not discussed in Sect. 3). Thus, the extension of these methods to DAEs following the developments of Sect. 4 is a plausible and potentially fruitful area for future research.

7 Global Optimization with Semi-explicit Index-One DAEs Embedded

In this section, we review the method in [74] for deterministic global optimization problems with DAEs embedded. The specific problem under consideration is discussed in detail in Sect. 2.4. As mentioned in Sect. 1, all existing global optimization algorithms for dynamic problems are based on the spatial branch-and-bound (B&B)

framework originally developed for standard NLPs [29, 69]. Accordingly, we begin with a brief overview of this approach, and subsequently extend it to DAE embedded problems using the developments of the last several sections.

7.1 The Spatial Branch-and-Bound Global Optimization Algorithm

Consider the standard NLP

$$\begin{aligned} \min_{\mathbf{p}\in P} \quad & J(\mathbf{p}) \\ \text{s.t.} \quad & \mathbf{G}(\mathbf{p}) \leq \mathbf{0}, \end{aligned} \tag{7.1}$$

where $P \in \mathbb{IR}^{n_p}$, and J and $\mathbf{G}$ are continuous on P. To solve this problem to global optimality, the spatial B&B method considers a sequence of subproblems in which (7.1) is restricted to a subinterval $P^l \subset P$:

$$\begin{aligned} \min_{\mathbf{p}\in P^l} \quad & J(\mathbf{p}) \\ \text{s.t.} \quad & \mathbf{G}(\mathbf{p}) \leq \mathbf{0}, \end{aligned} \tag{7.2}$$

The basic requirement for applying spatial B&B is that, for any subinterval $P^l \subset P$ (which may be P itself), procedures are available that compute guaranteed upper and lower bounds on the optimal objective value of (7.2). These bounds are denoted by UBD^l and LBD^l, respectively. Since the value of the objective function at any feasible point provides an upper bound on the optimal objective value of (7.2), UBD^l can be computed by solving (7.1) to local optimality. Computing a lower bound is substantially more difficult and is the key step in the spatial B&B algorithm. Methods for accomplishing this are discussed below.

Supposing that upper and lower bounding procedures are available, the spatial B&B algorithm proceeds as follows. First, upper and lower bounds are computed for the optimal objective value of (7.1). Since these bounds apply to the original problem of interest, rather than to the subproblem (7.2), they are denoted by UBD and LBD, respectively. If it happens that $UBD - LBD$ is less than a specified tolerance ε, then the B&B algorithm terminates, having bracketed the optimal objective value of (7.1) within the given tolerance. An estimate of the solution value $\mathbf{p}^*$ is then given by the value which attained the upper bound UBD. If this termination test fails, then P is partitioned into two subintervals, termed *branching*, typically by bisection in its dimension of largest width. These subintervals inherit the bounds UBD and LBD, which are obviously valid for the corresponding subproblems (7.2) on account of being valid for (7.1). These two subintervals are then added to a stack Σ of subintervals, or *nodes*, to be processed that is maintained throughout the algorithm.

At the beginning of a generic iteration of the algorithm, UBD and LBD are the best known upper and lower bounds on the optimal objective value of (7.1), respectively, and the stack Σ contains a number of nodes P^l, each of which is equipped with upper and lower bounds UBD^l and LBD^l that have been inherited from the parent node from which it was generated through bisection. Collectively, the nodes P^l may not form a partition of P, but the complement of $\cup_l P^l$ in P will have been proven not to contain the optimal solution of (7.1) through the procedures below. The iteration proceeds by selecting from the stack a node P^l for which $LBD^l = LBD$. The upper and lower bounds UBD^l and LBD^l are then refined by computing bounds on the optimal objective value of (7.2) using the procedures that we have assumed to be available. If it is found that (7.2) is infeasible, then P^l is eliminated from further consideration and a new element is selected from the stack. In this case, we say that P^l is *fathomed by infeasibility*. Otherwise, upper and lower bounds on the optimal objective value of the original problem (7.1) are updated according to

$$UBD := \min_k UBD^k \quad \text{and} \quad LBD := \min_k LBD^k, \tag{7.3}$$

where the min is taken over all elements of Σ. These assignments are valid because the complement of $\cup_k P^k$ in P has been shown not to contain a global optimum of (7.1). Moreover, if P^l was the only element of Σ for which $LBD^l = LBD$ at the beginning of the iteration, and if LBD^l was improved by the application of the lower bounding procedure to (7.2), then LBD is improved by this assignment. If UBD is improved by this assignment, then there is an opportunity to fathom some nodes in the stack. This is done by checking the inequality $LBD^k > UBD$ for every $P^k \in \Sigma$. If this is true for some P^k, then the optimal solution cannot lie in P^k and P^k is eliminated from further consideration. In this case, P^k is said to be *fathomed by value dominance*. If P^l has not been fathomed either by infeasibility or by value dominance, then it is bisected and the two resulting nodes are added to the stack.

The iteration outlined above is repeated until either the stack becomes empty, indicating that (7.1) is infeasible, or it is found in some iteration that $UBD - LBD < \varepsilon$, indicating that a point $\mathbf{p}^*$ has been found which achieves an objective value within ε of the globally optimal objective value. Roughly, if the lower bounding procedure has the property that it provides sharper bounds on smaller intervals P^l and becomes exact in the limit as P^l tends toward a singleton, then it can be shown that one of these outcomes will occur after finitely many iterations [29]. A notable exception to this result can occur in the presence of inequality constraints if the algorithm is unable to find feasible points at which to compute upper bounds. In the following discussion, we assume that upper bounds can be computed without difficulty.

Due to the repeated partitioning of P, the spatial B&B algorithm exhibits worst-case exponential run-time with respect to the dimension of $\mathbf{p}$ and the magnitude of $1/\varepsilon$. In practice, the primary determinants of the run-time are the computational cost and the accuracy of the lower bounding procedure. In addition, a number of more

advanced techniques have been developed which can greatly accelerate convergence through the use of domain reduction techniques [67–69]. Thus, while it is true that the basic procedure outlined above can be prohibitively expensive, impressive results have been achieved for many challenging problems using advanced implementations of the method [66, 67, 85, 92].

Several methods are available for computing lower bounds on the optimal objective value of the subproblem (7.2). If J is factorable, we may use the lower bound of the natural interval extension $[J](P^l)$, as described in Sect. 3. Although some early implementations are based on this approach [32], the lower bounds computed in this way are relatively weak. Moreover, these bounds obey a first-order convergence rate property [56], while it has been demonstrated that at least second-order convergence is required to avoid serious convergence problems in spatial B&B algorithms [95]. In most modern implementations, lower bounds are computed by constructing and solving convex underestimating programs [5, 23, 50, 92]. A convex underestimating program for (7.2) is a program that is convex and has an optimal objective value that is guaranteed to underestimate that of (7.2). Although there are many ways to accomplish this, we consider the program

$$\begin{aligned} \min_{\mathbf{p}\in P^l} \quad & J_l^{cv}(\mathbf{p}) \\ \text{s.t.} \quad & \mathbf{G}_l^{cv}(\mathbf{p}) \leq \mathbf{0}, \end{aligned} \tag{7.4}$$

where J_l^{cv} and $\mathbf{G}_l^{cv}$ are convex relaxations of J and $\mathbf{G}$ on P^l, respectively. This program is clearly convex, and hence solvable to global optimality using standard local optimization techniques. Moreover, its optimal objective value is easily seen to underestimate that of (7.2). If J and $\mathbf{G}$ are factorable functions, J_l^{cv} and $\mathbf{G}_l^{cv}$ can be obtained from the natural McCormick extensions $\{J\}$ and $\{\mathbf{G}\}$, as described in Sect. 3. In fact, when J and $\mathbf{G}$ are factorable, a number of approaches are available for constructing a convex relaxations [1, 2, 5, 50], and more generally to construct convex underestimating programs for (7.2) [92]. This last method is used in the popular code `BARON`.

7.2 A Lower Bounding Procedure for Optimization with DAEs

We now consider the application of spatial B&B global optimization to the dynamic optimization problem (2.8), which we restate here for convenience:

$$\begin{aligned} \min_{\mathbf{p}\in P} \quad & \phi(\mathbf{p}, \mathbf{x}(t_f, \mathbf{p}), \mathbf{y}(t_f, \mathbf{p})) \\ \text{s.t.} \quad & \boldsymbol{\eta}(\mathbf{p}, \mathbf{x}(t_f, \mathbf{p}), \mathbf{y}(t_f, \mathbf{p})) \leq \mathbf{0}. \end{aligned} \tag{7.5}$$

Recall that $(\phi, \boldsymbol{\eta}) : D_p \times D_x \times D_y \to \mathbb{R} \times \mathbb{R}^{n_c}$ are continuous, and $(\mathbf{x}, \mathbf{y})$ is a regular solution of (5.1) on $I \times P$ that uniquely satisfies the additional initial condition

$$\mathbf{y}(t_0, \hat{\mathbf{p}}) = \hat{\mathbf{y}}_0, \tag{7.6}$$

where $\hat{\mathbf{p}} \in P$ and $\hat{\mathbf{y}}_0 \in D_y$ satisfy $\mathbf{g}(t_0, \hat{\mathbf{p}}, \mathbf{x}_0(\hat{\mathbf{p}}), \hat{\mathbf{y}}_0) = \mathbf{0}$ and $\det \frac{\partial \mathbf{g}}{\partial \mathbf{y}}(t_0, \hat{\mathbf{p}}, \mathbf{x}_0(\hat{\mathbf{p}}), \hat{\mathbf{y}}_0) \neq 0$. The existence of such a solution can be verified by the state bounding algorithm of Sect. 5, and in fact this will be done during the course of the optimization algorithm described here.

The optimization problem (7.5) can be written in the form of the standard optimization problem (7.1) with the definitions

$$J(\mathbf{p}) \equiv \phi(\mathbf{p}, \mathbf{x}(t_f, \mathbf{p}), \mathbf{y}(t_f, \mathbf{p})), \tag{7.7}$$

$$\mathbf{G}(\mathbf{p}) \equiv \boldsymbol{\eta}(\mathbf{p}, \mathbf{x}(t_f, \mathbf{p}), \mathbf{y}(t_f, \mathbf{p})). \tag{7.8}$$

Thus, spatial B&B is applicable to (7.5), provided that upper and lower bounds on the subproblems (7.2) can be computed, for any $P \subset P$, with the definitions (7.7) and (7.8). Any local dynamic optimization algorithm can be used to compute the required upper bounds. Moreover, a valid lower bound can be computed by solving the convex problem (7.4) locally, provided that convex relaxations J^{cv} and $\mathbf{G}^{cv}$ can be derived for (7.7) and (7.8). Although many methods for computing convex relaxations were discussed above, the key assumption in all of these techniques is that the objective and constraint functions are factorable functions of $\mathbf{p}$. Of course, this is not the case for (7.7) and (7.8), precisely because the parametric DAE solutions $\mathbf{x}(t_f, \cdot)$ and $\mathbf{y}(t_f, \cdot)$ are not factorable. Thus, we find ourselves in the now familiar situation of requiring global information about non-factorable functions.

The difficult work needed to address this problem has already been done in the last several sections, leading to state relaxations for $(\mathbf{x}, \mathbf{y})$. What remains is only propagate these relaxations through ϕ and $\boldsymbol{\eta}$ to obtain J^{cv} and $\mathbf{G}^{cv}$. To do this, we will use the factorable representations of ϕ and $\boldsymbol{\eta}$ to compute their natural McCormick extensions, and subsequently leverage the composition theorem for relaxation functions. The following assumption is required:

Assumption 7.1 *The functions ϕ and $\boldsymbol{\eta}$ are factorable with natural McCormick extensions $\{\phi\} : \mathbb{M}D_p \times \mathbb{M}D_x \times \mathbb{M}D_y \to \mathbb{MR}$ and $\{\boldsymbol{\eta}\} : \mathbb{M}D_p \times \mathbb{M}D_x \times \mathbb{M}D_y \to \mathbb{MR}^{n_c}$.*

Relaxation functions $(\mathscr{J}, \mathscr{G}) : \mathbb{M}P \to \mathbb{MR} \times \mathbb{MR}^{n_c}$ for $(J, \mathbf{G})$ can now be defined through the following procedure for each $\mathscr{P} = (P^l, [\mathbf{p}^{cv}, \mathbf{p}^{cc}]) \in \mathbb{M}P$:

1. Compute state bounds $X(t)$ and $Y(t)$ on P^l by solving (5.8)–(5.12) with P^l in place of P.
2. Evaluate the relaxation functions $\mathscr{X}(t_f, \cdot)$ and $\mathscr{Y}(t_f, \cdot)$ at $\mathscr{P}$ by solving (6.17)–(6.23).

3. Assign

$$\mathscr{J}(\mathscr{P}) \equiv \{\phi\}(\mathscr{P}, \mathscr{X}(t_f, \mathscr{P}), \mathscr{Y}(t_f, \mathscr{P})), \tag{7.9}$$

$$\mathscr{G}(\mathscr{P}) \equiv \{\eta\}(\mathscr{P}, \mathscr{X}(t_f, \mathscr{P}), \mathscr{Y}(t_f, \mathscr{P})). \tag{7.10}$$

Since $\mathscr{X}(t_f, \cdot)$ and $\mathscr{Y}(t_f, \cdot)$ have been shown to be relaxation functions for $\mathbf{x}(t_f, \cdot)$ and $\mathbf{y}(t_f, \cdot)$, respectively, and $\{\phi\}$ is a relaxation function for ϕ by definition, Theorem 3.2 implies that $\mathscr{J}$ is a relaxation function for J. An analogous argument holds for $\mathscr{G}$. As discussed previously, the required McCormick arithmetic in these computations can be done automatically using the McCormick arithmetic library `MC++` (http://www3.imperial.ac.uk/people/b.chachuat/research).

Now, using the notation $\mathscr{J} = ([\mathscr{J}^L, \mathscr{J}^U], [\mathscr{J}^{cv}, \mathscr{J}^{cc}])$, and similarly for $\mathscr{G}$, define

$$J_l^{cv}(\mathbf{p}) \equiv \mathscr{J}^{cv}((P^l, [\mathbf{p}, \mathbf{p}])), \quad \mathbf{G}_l^{cv}(\mathscr{P}) \equiv \mathscr{G}^{cv}((P^l, [\mathbf{p}, \mathbf{p}])). \tag{7.11}$$

By the properties of relaxation functions with arguments of the form $(P^l, [\mathbf{p}, \mathbf{p}])$, it follows that J_l^{cv} and $\mathbf{G}_l^{cv}$ are convex relaxations of J and $\mathbf{G}$ as defined by (7.7) and (7.8), respectively (see Sect. 3.2). Thus, a valid lower bounding problem for (7.5) on P^l is given by

$$\begin{aligned} \min_{\mathbf{p} \in P^l} \quad & J_l^{cv}(\mathbf{p}) \\ \text{s.t.} \quad & \mathbf{G}_l^{cv}(\mathbf{p}) \leq \mathbf{0}. \end{aligned} \tag{7.12}$$

Since the procedure for evaluating $J_l^{cv}(\mathbf{p})$ and $\mathbf{G}_l^{cv}(\mathbf{p})$ outlined above involves the solution of (6.17)–(6.23) for each $\mathbf{p}$, (7.12) is itself a dynamic optimization problem. However, it is guaranteed to be convex, so that it can be solved to global optimality using a local solver, thereby providing the lower bound required by the spatial B&B algorithm. Note that the bounding system (5.8)–(5.12) is solved only once during the solution of (7.12) because (5.8)–(5.12) depend only on P', and not on $\mathbf{p}$. Implementing this requires that one stores the interpolating polynomial used by the integrator in each time step of the state bounds integration, so that accurate bound values can be recovered during integration of the state relaxations. This information can be obtained as output from the DAE solver `IDA` [27].

Due to the use of McCormick's relaxation technique, it is possible that the objective and constraints in (7.12) are non-differentiable. However, subgradients can be computed using the subgradient propagation rules for McCormick relaxations developed in [53], along with the developments for state relaxations in [74]. These subgradients can be computed automatically using the McCormick arithmetic library `MC++` (http://www3.imperial.ac.uk/people/b.chachuat/research). Because (7.12) is a potentially nonsmooth convex optimization problem, it is best to solve it using a specialized nonsmooth solver, such as a bundle method [44, 48]. However, these methods are not as mature as those for differentiable problems,

and the available solvers of this type remain problematic. Numerical experiments using the sequential quadratic programming code SNOPT are presented in [74], without serious numerical difficulties. An alternative approach is to use the values $J_l^{cv}(\hat{\mathbf{p}})$ and $\mathbf{G}_l^{cv}(\hat{\mathbf{p}})$ for some fixed $\hat{\mathbf{p}} \in P^l$, along with subgradients at this point, to construct affine underestimators for J and $\mathbf{G}$ on P^l. This reduces the lower bounding problem to a linear program and thereby avoids the issue of nonsmoothness. The lower bounds generated by this scheme are more conservative. However, there is a significant advantage in terms of the computational cost-per-node because (6.17)–(6.23) are integrated only once during the solution of the lower bounding problem. This method is used in the case study in Sect. 8.

7.3 Alternative Approaches

As mentioned in Sect. 1, an alternative approach to global dynamic optimization is to combine the simultaneous formulation, in which the dynamics are discretized in time and reduced to a large system of algebraic equations, with state-of-the-art branch-and-bound codes for standard NLPs [15]. A serious drawback of this approach is that it generates a very large number of new decision variables representing the values of the differential and algebraic states at each time point in the discretization scheme. While the special structure of this large-scale NLP makes it tractable for the purposes of local optimization, it has not been demonstrated that this structure can mitigate the characteristic worst-case exponential dependence of branch-and-bound run-time on the number of decision variables. However, even if this could be done, there is a more subtle problem that we are now in a position to appreciate. For a branch-and-bound search to be exhaustive, upper and lower bounds on the permissible values of all decision variables must be provided by the user. Given the fact that the decisions in the simultaneous formulation include the state variables themselves, at numerous time points throughout the horizon of interest, it follows that a reachable set enclosure must be provided as input. Thus, the primary complication of the sequential approach appears again in the simultaneous formulation and apparently eliminates the advantages of discretization.

Looking toward future improvements in relaxation theory for dynamic optimization problems, it is useful to note that state-of-the-art branch-and-bound codes for standard NLPs do not typically apply McCormick's relaxation technique directly, nor do they use the simplistic relaxation of (7.2) by (7.4). In particular, it is widely appreciated that accounting for the interdependence of the objective and constraint functions, as well as the interdependence of individual terms within each of these equations, is essential to obtaining tight relaxations in general. In the context of dynamic optimization problems, these observations have not been applied principally because of the lack of factorable expressions for the states, objective, and constraints. Indeed, the generality of McCormick's relaxation technique, and specifically its composition properties, are central to overcoming this problem. Moreover, moving beyond this approach is not a straightforward application of

existing relaxation techniques for NLPs. This is because the interdependence of different equations and sub-expressions in dynamic optimization problems are most often a result of correlations between the state variables. Thus, these interdependencies cannot be detected by methods that rely on a factorable representation. Rather, they are properties of the differential-algebraic system that must be inferred from the factorable representations of the governing equations using insights from dynamical systems theory. This is an open challenge that needs to be addressed moving forward. As a specific example, it suggests an alternative approach in which the state variables are not relaxed component-wise, as is done in (6.17)–(6.23). Instead, one might try for a more general convex enclosure of the reachable set as a whole. At present, this idea has not been investigated in the state relaxation literature, either for ODEs or DAEs, and is a promising area for future investigations.

8 Numerical Results and Directions for Improvement

In this section, the methods developed in Sects. 5–7 are demonstrated on a parameter estimation problem from [35]. Shortcomings and directions for future research are also highlighted.

8.1 Problem Formulation

Consider the following two-component signal transduction network:

$$\mathrm{ATP} + \mathrm{KdpD} \rightleftharpoons \mathrm{ADP} + \mathrm{KdpD}^p \tag{8.1}$$

$$\mathrm{KdpD}^p + \mathrm{KdpE} \rightleftharpoons \mathrm{KdpD} + \mathrm{KdpE}_f^p \tag{8.2}$$

$$\mathrm{KdpE}_f^p + \mathrm{KdpD} \rightarrow \mathrm{KdpE} + \mathrm{KdpD} \tag{8.3}$$

$$2\mathrm{KdpE}_f^p + \mathrm{DNA}_f \rightleftharpoons \mathrm{KdpE\text{-}DNA} \tag{8.4}$$

This network provides a simplified description of the regulation of K^+ uptake in *E. coli*, and is mechanistically similar to a variety stimulus response mechanisms found in bacteria, archaea, and plants [35]. A superscript p above denotes a phosphorylated species. Thus, reaction (8.1) describes the autophosphorylation of KdpD, (8.2) is a phosphoryl transfer reaction, (8.3) describes the dephosphorylation of KdpE_f, and (8.4) describes the binding of DNA to form the transcription complex KdpE-DNA.

In [35], purified KdpD, KdpE, and DNA fragments were combined in a mixture of ATP/ADP, and the concentrations of the phosphorylated proteins were measured over time. Specifically, the measured concentrations were C_{KdpD^p} and $C_{\mathrm{KdpE}_f^p} + 2C_{\mathrm{KdpE\text{-}DNA}}$, where the latter figure is the total concentration of phosphorylated

KdpE in both its free and complexed forms. The concentration of ATP/ADP were essentially constant during the experiments.

To present the model equations for this system in the notation of previous sections, we write

$$\mathbf{x} \equiv (C_{\mathrm{KdpD}^p}, C_{\mathrm{KdpE}_f^p} + 2C_{\mathrm{KdpE\text{-}DNA}}), \tag{8.5}$$

$$\mathbf{y} \equiv (C_{\mathrm{KdpE}_f^p}, C_{\mathrm{DNA}_f}, C_{\mathrm{KdpD}}, C_{\mathrm{KdpE}}). \tag{8.6}$$

We further collect some known initial concentrations in the vector $\mathbf{m}$ defined as

$$\mathbf{m} \equiv (C_{\mathrm{ATP0}}, C_{\mathrm{ADP0}}, C_{\mathrm{DNA0}}, C_{\mathrm{KdpD0}}, C_{\mathrm{KdpE0}}) = (100, 8, 100, 1, 4)\,\mu\mathrm{M}. \tag{8.7}$$

Finally, the forward and reverse rate constants of reaction i are denoted as k_i and k_{-i}, the reactions (8.1)–(8.4) are numbered consecutively from 1 to 4, and the equilibrium constant for DNA binding in (8.4) is denoted by K_b. With these definitions, the model equations are

$$\dot{x}_1 = k_1 m_1 y_3 - k_{-1} m_2 x_1 - k_2 x_1 y_4 + k_{-2} y_3 y_1, \tag{8.8}$$

$$\dot{x}_2 = k_2 x_1 y_4 - k_{-2} y_3 y_1 - k_3 y_3 y_1, \tag{8.9}$$

$$0 = -x_2 + y_1 + 2\frac{y_1^2 y_2}{K_b}, \tag{8.10}$$

$$0 = -m_3 + y_2 + \frac{y_1^2 y_2}{K_b}, \tag{8.11}$$

$$0 = -m_4 + y_3 + x_1, \tag{8.12}$$

$$0 = -m_5 + y_4 + x_2. \tag{8.13}$$

To arrive at these equations, differential species balances are first derived for x_1, x_2, and $C_{\mathrm{KdpE\text{-}DNA}}$ assuming mass-action kinetics in (8.1)–(8.4). In this form, the system specification is completed by three algebraic equations that specify the concentrations of KdpD, KdpE, and DNA using the observation that the total concentration of each of these species in all its phosphorylated, unphosphorylated, and complexed forms must equal C_{KdpD0}, C_{KdpE0}, and C_{DNA0}, respectively. This model is then reduced by the observation that DNA binding is fast relative to the other reactions, and can therefore be assumed to be in quasi-equilibrium. This leads to the algebraic relation

$$y_1^2 y_2 = K_b C_{\mathrm{KdpE\text{-}DNA}}, \tag{8.14}$$

which is substituted throughout to arrive at the DAEs (8.8)–(8.13).

Table 1 Data used for the parameter estimation problem (8.15), based on experimental data from Fig. 3(B) in [35]

ℓ	t_ℓ (min)	$\hat{x}_1$ (μM)	$\hat{x}_2$ (μM)
0	0	0	0
1	1	0.0027	0.003
2	2.5	0.0068	0.0075
3	5	0.0068	0.0108
4	7.5	0.0063	0.0148
5	10	0.0069	0.019
6	12.5	0.0067	0.0205
7	15	0.0072	0.0230

In [35], parameter estimation was done using a local dynamic optimization method to find the unknown rate constants k_i and k_{-i} for all $i \in \{1, \ldots, 4\}$. Here, we consider solving this problem globally. To simplify the problem, we fit only k_1 and k_3, which were found to be the parameters with the highest sensitivities in [35]. Thus, we solve the optimization problem

$$\begin{aligned} \min_{k_1,k_3} \quad & \frac{1}{7}\sum_{\ell=1}^{7}\left[\left(\frac{x_1(t_\ell,k_1,k_3)-\hat{x}_1(t_\ell)}{\hat{x}_1(t_\ell)}\right)^2+\left(\frac{x_2(t_\ell,k_1,k_3)-\hat{x}_2(t_\ell)}{\hat{x}_2(t_\ell)}\right)^2\right] \\ \text{s.t.} \quad & k_1 \in [0.001, 0.01]\ (\text{h}\mu\text{M})^{-1}, \quad k_3 \in [10, 300]\ (\text{h}\mu\text{M})^{-1}, \\ & (8.8)\text{–}(8.13) \quad \text{hold} \quad \forall t \in [0, 0.25]\ \text{h}, \\ & \mathbf{x}(t_0, k_1, k_3) = \mathbf{0}\ \mu\text{M}, \\ & \mathbf{y}(t_0, k_1, k_3) = (0, m_3, m_4, m_5)\ \mu\text{M}. \end{aligned} \tag{8.15}$$

Above, $\hat{x}_1$ and $\hat{x}_2$ denote measured values, which are given in Table 1. These data are based on experimental values reported in [35]. Unfortunately, the measured values were not reported directly, so the data in Table 1 are estimated from Fig. 3(B) in [35]. Although inaccurate, this data is sufficient for the illustrative purposes of this example. The fixed rate constants in (8.15) are set to the optimal values found in [35]: $k_{-1} = 0.0029\ (\text{h}\mu\text{M})^{-1}$, $k_2 = 108\ (\text{h}\mu\text{M})^{-1}$, $k_{-2} = 1080\ (\text{h}\mu\text{M})^{-1}$, $K_b = \frac{36}{540}$ $(\mu\text{M})^2$.

8.2 Global Dynamic Optimization

For the purposes of solving (8.15) to global optimality, some further rearrangements of (8.8)–(8.13) are helpful. First, the final three algebraic equations are solved analytically for y_2, y_3, and y_4 respectively, and substituted into the remaining equations to arrive at a system of two differential equations and one nonlinear algebraic equation. Secondly, the resulting equations are rearranged to give factorable

representations that are favorable for interval and McCormick computations. The final form used for the computations below is

$$\dot{x}_1 = k_1 m_1 (m_4 - x_1) - k_{-1} m_2 x_1 - k_2 x_1 (m_5 - x_2) + k_{-2} (m_4 - x_1) y_1, \tag{8.16}$$

$$\dot{x}_2 = -[(k_{-2} + k_3) y_1 + k_2 (m_5 - x_2)](m_4 - x_1) + k_2 m_4 (m_5 - x_2), \tag{8.17}$$

$$0 = y_1^3 + 2m_3 y_1^2 + K_b y_1 - (K_b + y_1^2) x_2. \tag{8.18}$$

The global dynamic optimization algorithm is also provided with the analytical derivative of the algebraic equation with respect to y_1 with the factorable representation

$$\frac{\partial g}{\partial y_1} = y_1 (3y_1 + 2(2m_3 - x_2)) + K_b. \tag{8.19}$$

The purpose of these rearrangements is to reduce the conservatism of the enclosures obtained through the interval and McCormick methods described in the preceding sections. The analytical solution of (8.11)–(8.13) avoids potentially conservative interval and McCormick iterations on these equations using the methods of Sect. 4. These iterations are applicable to general nonlinear equations, and do not always reduce to the natural interval extension of the analytical solution when the latter is possible. The factorable representations used in (8.18) and (8.19) have already been justified in Sect. 4.1, and are designed to minimize the number of appearances of variables that will be interval- or McCormick-valued in the interval or McCormick evaluations of these expressions. The rearrangements to the right-hand side of (8.17) are similarly motivated. In this case, we note that the state bounding method of Sect. 5 uses exclusively interval evaluations of (8.17) in which x_2 is a degenerate interval, while x_1, y_1, and k_3 are potentially nondegenerate intervals. Thus, the rearranged equation reduces the appearances of x_1 from 2 to 1, without concern for the additional appearances of x_2 so generated. Rearrangements of the type discussed here and in Sect. 4.1 are standard in interval computations and are typically not difficult to perform. However, good arrangements depend on which variables will be interval- or McCormick-valued in a given expression, and can become cumbersome for large systems. Automating these rearrangements is an interesting topic for future research at the intersection of numerical and symbolic computations.

Using Eqs. (8.16)–(8.18), the least-squares estimation problem (8.15) was solved to global optimality using branch-and-bound as described in Sect. 7. In each node, upper bounds were determined by simply evaluating the objective at the midpoint $\mathbf{p}^l_{\text{mid}}$ of the parameter interval P^l. To compute a lower bound for the optimal objective value on P^l, state bounds were first computed on P^l by solving (5.8)–(5.11). The parameter γ in (5.10) and (5.11) was set to 10^{-6} and (5.8)–(5.11) were solved using `IDA` [27] with relative and absolute tolerances of 10^{-6}. Subsequently, state relaxations were computed at $\mathbf{p}^l_{\text{mid}}$ by solving (6.17)–(6.23) with $K = 5$ using `CVODES` [27] with relative and absolute tolerances of 10^{-6}. Additionally, sensitivity

equations for (6.17)–(6.23) were integrated in order to provide subgradients for the state relaxations as described in detail in [74]. The state relaxations and subgradients were then propagated through the least-squares objective in (8.15) using McCormick arithmetic along with the subgradient propagation rules in [53]. This gives the relaxation value $J_l^{cv}(\mathbf{p}_{\text{mid}}^l)$ along with the associated subgradient at $\mathbf{p}_{\text{mid}}^l$. Using this information, an affine relaxation of the objective was formulated and minimized on the box P^l to obtain the final lower bound. The termination condition for the branch-and-bound algorithm was specified as $|UBD - LBD| \leq 10^{-5}$. No domain reduction techniques were used. Branching was done using a weighted-width heuristic that bisects the ith parameter interval if $i \in \text{argmax}(W_i w(P_i^l))$. Here, w denotes the interval width and $W_i \in \mathbb{R}_+$ is a weighting factor. For the present problem, $W_1 = 10^5$ and $W_2 = 1$ were specified based on the observation that the sensitivity of the objective function to k_1 at $\mathbf{p}_{\text{mid}}$ of the root node is larger than that with respect to k_3 by a factor of 10^5.

The ϵ-global minimum objective value was found to be 0.029069, which is attained at $\mathbf{p}^* = (k_1^*, k_3^*) = (4.31348 \times 10^{-3}, 162.9297)$ $(\text{h}\mu\text{M})^{-1}$. This optimal solution is quite different from that found in [35]; $(k_1, k_3) = (2.9 \times 10^{-3}, 90)$ $(\text{h}\mu\text{M})^{-1}$. However, while it is likely that the value found in [35] is not globally optimal, the discrepancy here is probably primarily due to the error introduced in our estimation of the experimental data from Fig. 3(B) in [35].

The global solution for (8.15) was found in 52s. The optimization was done on a virtual machine with 4 GB of memory running Ubuntu 12.04 on a Macbook Pro OSX 10.9.2 host with a 2.6 GHz Intel Core i7 CPU. The branch-and-bound algorithm visited 1587 nodes in total, generating the partition of the search space shown in Fig. 2. The state bounds and the relaxation of the objective function computed in the 1100th node visited by the algorithm are shown in Figs. 3 and 4 as representative examples.

Although the state bounds in Fig. 3 are relatively sharp, the bounding technique can generate rapidly diverging bounds on the larger P^l intervals that occur early in the branch-and-bound tree. This is a common problem in reachability analysis and also occurs for problems with ODEs embedded. However, a unique feature of the DAE methods here is the need to pass an inclusion test to verify the existence of a solution of the algebraic equations as described in Sect. 4. This manifests itself in the state bounding method through the need to satisfy Eqs. (5.10) and (5.11) everywhere along the bounding trajectories, which can become difficult or impossible when the interval arguments become large in width. For example, the state bounds computed in the 900th node visited by the branch-and-bound algorithm are shown in Fig. 5. These are clearly divergent, and although the bounding procedure succeeds, slightly larger parameter intervals cause it to fail due to an inability to satisfy (5.10) and (5.11) for t near the end of the horizon. For such nodes, the lower bound is set to $-\infty$ and the node is branched. These failures make a significant contribution to the overall number of nodes visited by the branch-and-bound algorithm, as well as to the CPU time. Thus, an area for future research is in deriving sharper inclusion tests that can potentially succeed on larger parameter intervals.

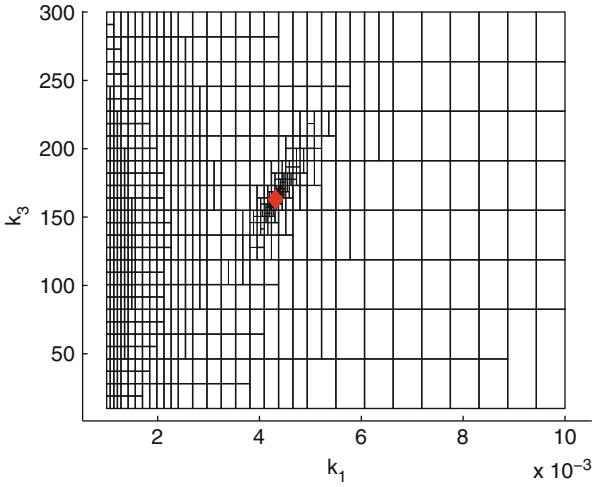

Fig. 2 Intervals in the search space $P = [0.001, 0.01] \times [10, 300]$ that were fathomed by value dominance during the branch-and-bound global solution of (8.15). The *diamond* indicates the global solution found

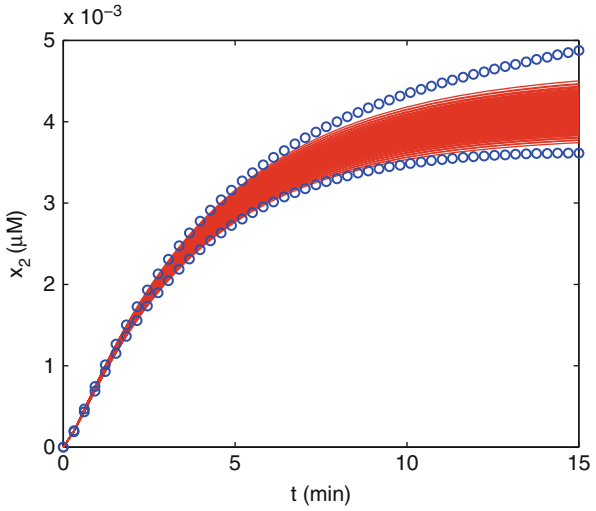

Fig. 3 Upper and lower state bounds for x_2 (*circles*) computed on $P^{1100} = [1.421875 \times 10^{-3}, 1.4921875 \times 10^{-3}] \times [136.875, 145.9375]$, as well as true solution trajectories of (8.16)–(8.18) (*solid*) computed on a uniform 12×12 grid of P^{1100}. P^{1100} is the interval corresponding to the 1100th node processed by the branch-and-bound algorithm while solving (8.15)

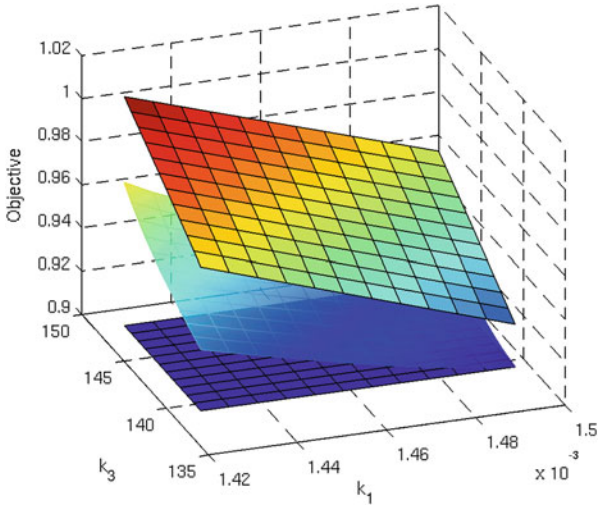

Fig. 4 Objective function of (8.15) on P^{1100} (*upper surface*), along with its convex relaxation (*middle surface*) and interval lower bound (*lower surface*). P^{1100} is the interval corresponding to the 1100th node processed by the branch-and-bound algorithm while solving (8.15)

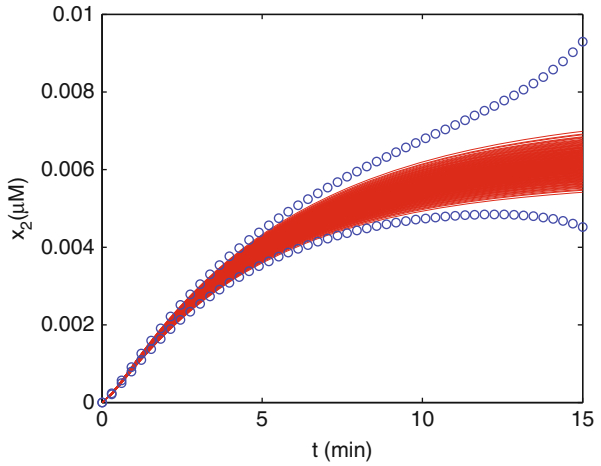

Fig. 5 Upper and lower state bounds for x_2 (*circles*) computed on $P^{900} = [1.5625 \times 10^{-3}, 1.703125 \times 10^{-3}] \times [118.75, 127.8125]$, as well as true solution trajectories of (8.16)–(8.18) (*solid*) computed on a uniform 12×12 grid of P^{900}. P^{900} is the interval corresponding to the 900th node processed by the branch-and-bound algorithm while solving (8.15)

A further area for research in DAE state bounding is to consider the use of a priori information that is known about the DAE reachable set, for example, through physical arguments. This has been studied in the case of ODE state bounding techniques in [76, 80]. It was shown there that many models of interest in chemical engineering applications satisfy affine solution invariants of the form $\mathbf{M}(\mathbf{x}(t,\mathbf{p}) - \mathbf{x}_0(\mathbf{p})) = \mathbf{0}$, $\forall (t,\mathbf{p}) \in I \times P$, and that these invariants can be exploited within the state bounding procedure to provide dramatically improved bounds in many cases. The same observation likely holds for DAE models. Indeed, the algebraic equations (8.12) and (8.13) are derived from species balances and are used to replace differential balances on y_3 and y_4. However, if these differential balances were included in the model, then (8.12) and (8.13) would be redundant and would describe affine solution invariants. Experience with ODEs suggest that these could be used in the bounding procedure to significantly improve the computed bounds on larger P^l intervals. What remains to be done is to establish methods for exploiting this information within the DAE state bounding theory in a manner analogous to the ODE methods in [76, 80].

Figure 2 clearly shows that there is a large accumulation of small intervals P^l surrounding the unconstrained global minimum, which adds significantly to the computational cost of global optimization. This phenomenon is known as the cluster problem and has been investigated in a number of articles [18, 95]. The severity of this problem is known to be related to the convergence order of the relaxation method used; i.e., the rate at which the convex relaxations converge to the original objective function as the parameter interval P^l tends toward degeneracy at a global minimizer, $P^l \to [\mathbf{p}^*, \mathbf{p}^*]$. The cluster problem is severe for first-order convergent methods, non-existent for third-order convergent methods, and second-order methods constitute a boundary case in which the behavior is determined by the size of the pre-factor in the convergence order estimate. In general, computing a third-order convergent relaxation is thought to be NP-hard [60]. At present, there are no published results on the convergence order of relaxation methods for dynamic optimization problems. However, preliminary research suggests that the ODE state relaxation methods in [82, 84] are both second-order convergent, with the pre-factor for the latter being potentially much better than for the former. However, the relaxation methods for implicit functions presented in Sect. 4 are thought to be only first-order convergent, which would then imply first-order convergence of the DAE relaxation methods presented here. Thus, a third area for future research is in developing second-order convergent relaxation methods for implicit functions, and hence for DAE solutions. It is expected that such a development would result in a very substantial reduction in CPU time for DAE embedded global optimization problems.

9 Conclusions

In this article, we have given a tutorial overview of the main concepts and tools used for reachability analysis and deterministic global optimization of nonlinear DAE models. These problems are highly interrelated, and are *global* problems in the sense that they concern the parametric solutions of a DAE model over a potentially large range of model parameters, rather than locally about a single value. The fundamental tool used to compute global information in the methods presented here, as well as similar methods in the literature, is the factorable representation of a function, along with its natural interval and McCormick extensions. The primary challenge when dealing with DAE systems is that the solutions are not factorable in general. To circumvent this, the methods detailed in this article repeatedly used the key idea that global information can be computed for a non-factorable function if that function is fully specified as the solution of a system of factorable equations. Using this theme, bounds and relaxations were computed for the solutions of implicit functions, and these methods were then extended to DAEs using some key theorems relating global properties of the DAEs to global properties of the solutions. In particular, these extensions led to computational methods for computing interval bounds, as well as convex and concave relaxations for the parametric solutions of semi-explicit index-one DAEs. Interestingly, these methods were also shown to be capable of computationally verifying the existence and uniqueness of a DAE solution within the computed bounds. Finally, we have highlighted the importance of reachability analysis in the context of global dynamic optimization, and argued that reachable set enclosure methods are a key enabling tool for the application of powerful global optimization algorithms to dynamic problems. In particular, we have illustrated the use of reachable set enclosures described in terms of convex and concave relaxations to extend the spatial branch-and-bound framework to dynamic optimization problems, leading a guaranteed global optimization algorithm for problems with semi-explicit DAE models embedded.

References

1. Adjiman, C.S., Dallwig, S., Floudas, C.A., Neumaier, A.: A global optimization method, αBB, for general twice-differentiable constrained NLPs - I. Theoretical advances. Comput. Chem. Eng. **22**(9), 1137–1158 (1998)
2. Adjiman, C.S., Androulakis, I.P., Floudas, C.A.: A global optimization method, αBB, for general twice-differentiable constrained NLPs - II. Implementation and computational results. Comput. Chem. Eng. **22**(9), 1159–1179 (1998)
3. Althoff, M., Stursberg, O., Buss, M.: Reachability analysis of linear systems with uncertain parameters and inputs. In: Proceedings of 46th IEEE Conference on Decision and Control, pp. 726–732 (2007)
4. Althoff, M., Stursberg, O., Buss, M.: Reachability analysis of nonlinear systems with uncertain parameters using conservative linearization. In: Proceedings of 47th IEEE Conference on Decision and Control, pp. 4042–4048 (2008)

5. Androulakis, I.P., Maranas, C.D., Floudas, C.A.: αBB: a global optimization method for general constrained nonconvex problems. J. Glob. Optim. **7**, 337–363 (1995)
6. Ascher, U.M., Petzold, L.R.: Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations. SIAM, Philadelphia (1998)
7. Bellman, R.: Dynamic Programming. Princeton University Press, New Jersey (1957)
8. Benyahia, B., Lakerveld, R., Barton, P.I.: A plant-wide dynamic model of a continuous pharmaceutical process. Ind. Eng. Chem. Res. **51**(47), 15393–15412 (2012)
9. Berz, M., Makino, K.: Verified integration of ODEs and flows using differential algebraic methods on high-order Taylor models. Reliab. Comput. **4**, 361–369 (1998)
10. Bhatia, T., Biegler, L.: Dynamic optimization in the design and scheduling of multiproduct batch plants. Ind. Eng. Chem. Res. **35**, 2234–2246 (1996)
11. Biegler, L.T., Zavala, V.M.: Large-scale nonlinear programming using IPOPT: an integrating framework for enterprise-wide dynamic optimization. Comput. Chem. Eng. **33**(3), 575–582 (2009)
12. Chachuat, B., Mitsos, A., Barton, P.I.: Optimal start-up of microfabricated power generation processes employing fuel cells. Optim. Control Appl. Methods **31**(5), 471–495 (2010)
13. Chernousko, F.L.: Ellipsoidal state estimation for dynamical systems. Nonlinear Anal. **63**, 872–879 (2005)
14. Chisci, L., Garulli, A., Zappa, G.: Recursive state bounding by parallelotopes. Automatica **32**(7), 1049–1055 (1996)
15. Cizniar, M., Podmajersky, M., Hirmajer, T., Fikar, M., Latifi, A.M.: Global optimization for parameter estimation of differential-algebraic systems. Chem. Pap. **63**(3), 274–283 (2009)
16. Cross, E.A., Mitchell, I.M.: Level set methods for computing reachable sets of systems with differential algebraic equation dynamics. In: Proceedings of 2008 American Control Conference, pp. 2260–2265 (2008)
17. Cuthrell, J.E., Biegler, L.T.: On the optimization of differential-algebraic process systems. AIChE J. **33**(8), 1257–1270 (1987)
18. Du, K.S., Kearfott, R.: The cluster problem in multivariate global optimization. J. Glob. Optim. **5**(3), 253–265 (1994)
19. Dunnebier, G., Fricke, J., Klatt, K.U.: Optimal design and operation of simulated moving bed chromatographic reactors. Ind. Eng. Chem. Res. **39**(7), 2290–2304 (2000)
20. Esposito, W.R., Floudas, C.A.: Deterministic global optimization in nonlinear optimal control problems. J. Glob. Optim. **17**, 97–126 (2000)
21. Esposito, W.R., Floudas, C.A.: Global optimization for the parameter estimation of differential-algebraic systems. Ind. Eng. Chem. Res. **39**, 1291–1310 (2000)
22. Facchinei, F., Pang, J.S.: Finite-Dimensional Variational Inequalities and Complementarity Problems, vol. 1. Springer, New York (2003)
23. Falk, J.E., Soland, R.M.: An algorithm for separable nonconvex programming problems. Manag. Sci. **15**(9), 550–569 (1969)
24. Feehery, W., Tolsma, J., Barton, P.: Efficient sensitivity analysis of large-scale differential-algebraic systems. Appl. Numer. Math. **25**(1), 41–54 (1997)
25. Flores-Tlacuahuac, A., Biegler, L.T., Saldívar-Guerra, E.: Optimal grade transitions in the high-impact polystyrene polymerization process. Ind. Eng. Chem. Res. **45**(18), 6175–6189 (2006)
26. Harrison, G.W.: Dynamic models with uncertain parameters. In: Avula, X. (ed.) Proceedings of the First International Conference on Mathematical Modeling, vol. 1, pp. 295–304 (1977)
27. Hindmarsh, A.C., Brown, P.N., Grant, K.E., Lee, S.L., Serban, R., Shumaker, D.E., Woodward, C.S.: SUNDIALS, suite of nonlinear and differential/algebraic equation solvers. ACM Trans. Math. Softw. **31**, 363–396 (2005)
28. Hoefkens, J., Berz, M., Makino, K.: Computing validated solutions of implicit differential equations. Adv. Comput. Math. **19**, 231–253 (2003)
29. Horst, R., Tuy, H.: Global Optimization: Deterministic Approaches, 3rd edn. Springer, New York (1996)
30. Houska, B., Chachuat, B.: Branch-and-lift algorithm for deterministic global optimization in nonlinear optimal control. J. Optim. Theor. Appl. (2014). doi:10.1007/s10957-013-0426-1

31. Houska, B., Villanueva, M., Chachuat, B.: A validated integration algorithm for nonlinear ODEs using Taylor models and ellipsoidal calculus. In: Proceedings of 2013 IEEE 52nd Annual Conference on Decision and Control, pp. 484–489 (2013). doi:10.1109/CDC.2013.6759928
32. Kearfott, R.: Rigorous Global Search: Continuous Problems. Kluwer Academic Publishers, Dordrecht (1996)
33. Kesavan, P., Lee, J.H.: A set based approach to detection and isolation of faults in multivariable systems. Comput. Chem. Eng. **25**, 925–940 (2001)
34. Ko, D., Siriwardane, R., Biegler, L.T.: Optimization of pressure-swing adsorption process using zeolite 13X for CO2 sequestration. Ind. Eng. Chem. Res. **42**(2), 339–348 (2003)
35. Kremling, A., Heermann, R., Centler, F., Jung, K., Gilles, E.: Analysis of two-component signal transduction by mathematical modeling using the KdpD/KdpE system of Escherichia coli. Biosystems **78**(1–3), 23–37 (2004)
36. Kunkel, P., Mehrmann, V.: Differential-Algebraic Equations: Analysis and Numerical Solution. European Mathematical Society, Zurich (2006)
37. Kurzhanski, A.B., Varaiya, P.: Ellipsoidal techniques for reachability analysis. In: Hybrid Systems: Computation and Control. Lecture Notes in Computer Science, vol. 1790, Sringer, Berlin, pp. 202–214 (2000)
38. Le, V.T.H., Stoica, C., Dumur, D., Alamo, T., Camacho, E.F.: Robust tube-based constrained predictive control via zonotopic set-membership estimation. In: Proceedings of 50th IEEE Conference on Decision and Control, pp. 4580–4585 (2011)
39. Limon, D., Bravo, J.M., Alamo, T., Camacho, E.F.: Robust MPC of constrained nonlinear systems based on interval arithmetic. IEEE Proc. Control Theory Appl. **152**(3), 325–332 (2005)
40. Lin, Y., Stadtherr, M.A.: Deterministic global optimization for parameter estimation of dynamic systems. Ind. Eng. Chem. Res. **45**, 8438–8448 (2006)
41. Lin, Y., Stadtherr, M.A.: Deterministic global optimization of nonlinear dynamic systems. AIChE J. **53**(4), 866–875 (2007)
42. Lin, Y., Stadtherr, M.A.: Validated solutions of initial value problems for parametric ODEs. Appl. Numer. Math. **57**, 1145–1162 (2007)
43. Lin, Y., Stadtherr, M.A.: Fault detection in nonlinear continuous-time systems with uncertain parameters. AIChE J. **54**(9), 2335–2345 (2008)
44. Luksan, L., Vlcek, J.: Algorithm 811: NDA: algorithms for nondifferentiable optimization. ACM Trans. Math. Softw. **27**(2), 193–213 (2001)
45. Luus, R., Dittrich, J., Keil, F.J.: Multiplicity of solutions in the optimization of a bifunctional catalyst blend in a tubular reactor. Can. J. Chem. Eng. **70**, 780–785 (1992)
46. Lygeros, J., Tomlin, C., Sastry, S.: Controllers for reachability specifications for hybrid systems. Automatica **35**, 349–370 (1999)
47. Ma, D.L., Chung, S.H., Braatz, R.D.: Worst-case performance analysis of optimal batch control trajectories. AIChE J. **45**(7), 1496–1476 (1999)
48. Makela, M.M.: Survey of bundle methods for nonsmooth optimization. Optim. Methods Softw. **17**(1), 1–29 (2002)
49. Maly, T., Petzold, L.R.: Numerical methods and software for sensitivity analysis of differential-algebraic systems. Appl. Numer. Math. **20**, 57–79 (1996)
50. McCormick, G.P.: Computability of global solutions to factorable nonconvex programs: Part I - convex underestimating problems. Math. Program. **10**, 147–175 (1976)
51. Mitchell, I., Bayen, A.M., Tomlin, C.: A time-dependent Hamilton-Jacobi formulation of reachable sets for continuous dynamic games. IEEE Trans. Autom. Control **50**(7), 947–957 (2005)
52. Mitsos, A., Bollas, G.M., Barton, P.I.: Bilevel optimization formulation for parameter estimation in liquid-liquid phase equilibrium problems. Chem. Eng. Sci. **64**(3), 548–559 (2009)
53. Mitsos, A., Chachuat, B., Barton, P.I.: McCormick-based relaxations of algorithms. SIAM J. Optim. **20**(2), 573–601 (2009)

54. Moisan, M., Bernard, O., Gouze, J.L.: Near optimal interval observers bundle for uncertain bioreactors. Automatica **45**(1), 291–295 (2009)
55. Moles, C.G., Mendes, P., Banga, J.R.: Parameter estimation in biochemical pathways: a comparison of global optimization methods. Genome Res. **13**(11), 2467–2474 (2003)
56. Moore, R.E.: Methods and Applications of Interval Analysis. SIAM, Philadelphia, PA (1979)
57. Nedialkov, N.S., Jackson, K.R., Corliss, G.F.: Validated solutions of initial value problems for ordinary differential equations. Appl. Math. Comput. **105**, 21–68 (1999)
58. Neher, M., Jackson, K.R., Nedialkov, N.S.: On Taylor model based integration of ODEs. SIAM J. Numer. Anal. **45**(1), 236–262 (2007)
59. Neumaier, A.: Interval Methods for Systems of Equations. Cambridge University Press, Cambridge (1990)
60. Neumaier, A.: Complete search in continuous global optimization. In: Iserles, A. (ed.) Acta Numerica. Cambridge University Press, Cambridge (2004)
61. Oishi, M., Mitchell, I., Tomlin, C., Saint-Pierre, P.: Computing viable sets and reachable sets to design feedback linearizing control laws under saturation. In: Proceedings of 45th IEEE Conference on Decision and Control, San Diego, CA, pp. 3801–3807 (2006)
62. Papamichail, I., Adjiman, C.S.: A rigorous global optimization algorithm for problems with ordinary differential equations. J. Glob. Optim. **24**(1), 1–33 (2002)
63. Raissi, T., Ramdani, N., Candau, Y.: Set membership state and parameter estimation for systems described by nonlinear differential equations. Automatica **40**, 1771–1777 (2004)
64. Rapaport, A., Dochain, D.: Interval observers for biochemical processes with uncertain kinetics and inputs. Math. Biosci. **193**, 235–253 (2005)
65. Rauh, A., Brill, M., Gunther, C.: A novel interval arithmetic approach for solving differential-algebraic equations with Valencia-IVP. Int. J. Appl. Math. Comput. Sci. **19**(3), 381–397 (2009)
66. Ryoo, H., Sahinidis, N.: Global optimization of nonconvex NLPs and MINLPs with application in process design. Comput. Chem. Eng. **19**(5), 551–566 (1995)
67. Ryoo, H., Sahinidis, N.: A branch-and-reduce approach to global optimization. J. Glob. Optim. **2**, 107–139 (1996)
68. Sahinidis, N., Tawarmalani, M.: Accelerating branch-and-bound through a modeling language construct for relaxation-specific constraints. J. Glob. Optim. **32**(2), 259–280 (2005)
69. Sahinidis, N., Tawarmalani, M.: A polyhedral branch-and-cut approach to global optimization. Math. Program. **130**(2), 225–249 (2005)
70. Sahlodin, A.M., Chachuat, B.: Convex/concave relaxations of parametric ODEs using Taylor models. Comput. Chem. Eng. **35**, 844–857 (2011)
71. Sahlodin, A.M., Chachuat, B.: Discretize-then-relax approach for convex/concave relaxations of the solutions of parametric ODEs. Appl. Numer. Math. **61**, 803–820 (2011)
72. Schaber, J., Liebermeister, W., Klipp, E.: Nested uncertainties in biochemical models. IET Syst. Biol. **3**(1), 1–9 (2009)
73. Schweppe, F.: Recursive state estimation: unknown but bounded errors and system inputs. IEEE Trans. Autom. Control **13**(1), 22–28 (1968)
74. Scott, J.K.: Reachability analysis and deterministic global optimization of differential-algebraic systems. Ph.D. thesis, Massachusetts Institute of Technology (2012)
75. Scott, J.K., Barton, P.I.: Convex enclosures for the reachable sets of nonlinear parametric ordinary differential equations. In: Proceedings of 49th IEEE Conference on Decision and Control, Atlanta, GA, pp. 5695–5700 (2010)
76. Scott, J.K., Barton, P.I.: Tight, efficient bounds on the solutions of chemical kinetics models. Comput. Chem. Eng. **34**, 717–731 (2010)
77. Scott, J.K., Barton, P.I.: Convex relaxations for nonconvex optimal control problems. In: Proceedings of 50th IEEE Conference on Decision and Control, Orlando, FL, pp. 1042–1047 (2011)
78. Scott, J.K., Barton, P.I.: Interval bounds on the solutions of semi-explicit index-one DAEs. Part 1: Analysis. Numer. Math. **125**(1), 1–25 (2011)
79. Scott, J.K., Barton, P.I.: Interval bounds on the solutions of semi-explicit index-one DAEs. Part 2: Computation. Numer. Math. **125**(1), 27–60 (2011)

80. Scott, J.K., Barton, P.I.: Bounds on the reachable sets of nonlinear control systems. Automatica **49**, 93–100 (2013)
81. Scott, J.K., Barton, P.I.: Convex and concave relaxations for the parametric solutions of semi-explicit index-one DAEs. J. Optim. Theory Appl. **156**(3), 617–649 (2013)
82. Scott, J.K., Barton, P.I.: Improved relaxations for the parametric solutions of odes using differential inequalities. J. Glob. Optim. **57**(1), 143–176 (2013)
83. Scott, J.K., Stuber, M.D., Barton, P.I.: Generalized McCormick relaxations. J. Glob. Optim. **51**, 569–606 (2011)
84. Scott, J.K., Chachuat, B., Barton, P.I.: Nonlinear convex and concave relaxations for the solutions of parametric ODEs. Optim. Control Appl. Methods **34**(2), 145–163 (2013)
85. Selot, A., Kuok, L.K., Robinson, M., Mason, T., Barton, P.I.: A short-term operational planning model for natural gas production systems. AIChE J. **54**(2), 495–515 (2007)
86. Singer, A.B., Barton, P.I.: Global solution of optimization problems with parameter-embedded linear dynamic systems. J. Optim. Theory Appl. **121**, 613–646 (2004)
87. Singer, A.B., Barton, P.I.: Bounding the solutions of parameter dependent nonlinear ordinary differential equations. SIAM J. Sci. Comput. **27**, 2167–2182 (2006)
88. Singer, A.B., Barton, P.I.: Global dynamic optimization for parameter estimation in chemical kinetics. J. Phys. Chem. A **110**(3), 971–976 (2006)
89. Singer, A.B., Barton, P.I.: Global optimization with nonlinear ordinary differential equations. J. Glob. Optim. **34**, 159–190 (2006)
90. Srinivasan, B., Palanki, S., Bonvin, D.: Dynamic optimization of batch processes - I. Characterization of the nominal solution. Comput. Chem. Eng. **27**(1), 1–26 (2003)
91. Stuber, M.D., Scott, J.K., Barton, P.I.: Convex and concave relaxations of implicit functions. Optim. Methods Softw. **30**(3), 424–460 (2015)
92. Tawarmalani, M., Sahinidis, N.V.: Convexification and Global Optimization in Continuous and Mixed-Integer Nonlinear Programming. Kluwer Academic Publishers, Dordrecht (2002)
93. Teo, K.L., Goh, G., Wong, K.: A Unified Computational Approach to Optimal Control Problems. Wiley, New York (1991)
94. Tsang, T., Himmelblau, D., Edgar, T.: Optimal control via collocation and nonlinear programming. Int. J. Control **21**, 763–768 (1975)
95. Wechsung, A., Schaber, S.D., Barton, P.I.: The cluster problem revisited. J. Glob. Optim. **58**(3), 429–438 (2014)

Numerical Linear Algebra Methods for Linear Differential-Algebraic Equations

Peter Benner, Philip Losse, Volker Mehrmann, and Matthias Voigt

Contents

P. Benner
Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstraße 1, 39106 Magdeburg, Germany
e-mail: benner@mpi-magdeburg.mpg.de

P. Losse
Pestalozzistraße 13B, 10625 Berlin, Germany
e-mail: philip.losse@gmail.com

V. Mehrmann • M. Voigt (✉)
Institut für Mathematik, Technische Universität Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany
e-mail: mehrmann@math.tu-berlin.de; mvoigt@math.tu-berlin.de

A. Ilchmann, T. Reis (eds.), *Surveys in Differential-Algebraic Equations III*,
Differential-Algebraic Equations Forum, DOI 10.1007/978-3-319-22428-2_3

Abstract A survey of methods from numerical linear algebra for linear constant coefficient differential-algebraic equations (DAEs) and descriptor control systems is presented. We discuss numerical methods to check the solvability properties of DAEs as well as index reduction and regularization techniques. For descriptor systems we discuss controllability and observability properties and how these can be checked numerically. These methods are based on staircase forms and derivative arrays, obtained by real orthogonal transformations that are discussed in detail. Then we use the reformulated problems in several control applications for differential-algebraic equations ranging from regular and singular linear-quadratic optimal and robust control to dissipativity checking. We discuss these applications and give a systematic overview of the theory and the numerical solution methods. In particular, we show that all these applications can be treated with a common approach that is based on the computation of eigenvalues and deflating subspaces of even matrix pencils. The unified approach allows us to generalize and improve several techniques that are currently in use in systems and control.

Keywords Canonical form • Controllability • Descriptor system • Differential-algebraic equation • Dissipativity • Even matrix pencil • $\mathscr{H}_\infty$ control • Kronecker index • $\mathscr{L}_\infty$-norm • Linear-quadratic optimal control • Observability

AMS Subject Classification (2010): 15A21, 15A22, 34A09, 65F15, 65L80, 93C05, 93D09

Notation

$\mathbb{N}$, $\mathbb{N}_0$	The set of natural numbers, $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$
$\mathbb{R}$, $\mathbb{C}$	The fields of real and complex numbers, resp.
$\mathbb{C}^-$, $\mathbb{C}^+$	The sets of complex numbers with negative and positive real parts, resp.
i	The imaginary unit
u	The roundoff unit
$\mathbb{R}[s]$, $\mathbb{C}[s]$	The rings of polynomials with real and complex coefficients in the indeterminate s, resp.
$\mathbb{R}(s)$	The field of real-rational functions in the indeterminate s
$\mathscr{R}^{m,n}$	The sets of $m \times n$ matrices with entries in a ring $\mathscr{R}$
A^T, A^H, A^{-1}	Transpose, conjugate transpose, and inverse of the matrix A
range A, ker A	Range and kernel of the matrix A, resp.

$$\operatorname{diag}(A_1, \ldots, A_k) := \begin{bmatrix} A_1 & & \\ & \ddots & \\ & & A_k \end{bmatrix}$$

$\Lambda(A)$ The spectrum of $A \in \mathbb{R}^{n,n}$

$\Lambda(E, A)$ The set of finite eigenvalues of $sE - A \in \mathbb{R}[s]^{m,n}$

1 Introduction

In modern modeling and simulation software packages such as MODELICA[1] or MATLAB/SIMULINK,[2] the mathematical models are generated via a network of standardized submodels. This network approach has become the industrial standard in many physical and engineering domains, see, e.g., [8, 58, 68, 83, 101, 103–106, 112], and leads to differential-algebraic equations (DAEs), or descriptor systems in the control setting. The models include differential equations that model the dynamical behavior and algebraic equations that model constraints, interface and boundary conditions, or balance equations.

In this survey we study linear constant coefficient DAEs and descriptor systems, which arise from general nonlinear DAEs or descriptor systems by linearizing around a stationary solution [46], or via realization procedures [3, 4]. Linear constant coefficient DAEs take the form

$$E\dot{x}(t) = Ax(t) + f(t), \quad x(0) = x_0, \tag{1.1}$$

and linear time-invariant descriptor systems have the form

$$E\dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = x_0, \tag{1.2a}$$

$$y(t) = Cx(t) + Du(t), \tag{1.2b}$$

with matrices $E, A \in \mathbb{R}^{k,n}, B \in \mathbb{R}^{k,m}, C \in \mathbb{R}^{p,n}$ and $D \in \mathbb{R}^{p,m}$. Here, $x : [0, \infty) \to \mathbb{R}^n$ represents the state, $u : [0, \infty) \to \mathbb{R}^m$ denotes a control input signal, $y : [0, \infty) \to \mathbb{R}^p$ is the output signal, and $f : [0, \infty) \to \mathbb{R}^k$ is a given inhomogeneity.

For a uniform presentation, we combine both DAE and descriptor system in the form

$$E\dot{x}(t) = Ax(t) + Bu(t) + f(t), \quad x(0) = x_0, \tag{1.3a}$$

$$y(t) = Cx(t) + Du(t), \tag{1.3b}$$

[1] https://www.modelica.org/.

[2] http://www.mathworks.com/.

where in the DAE case the term $Bu(\cdot)$ and the output equation are missing, whereas in the descriptor system case the inhomogeneity $f(\cdot)$ is dropped.

The survey is organized as follows. In Sect. 2 we briefly discuss the existence and uniqueness of solutions, as well the consistency of initial values. With a given DAE or descriptor system we can carry out numerical simulation, control, and optimization tasks. However, in the automatically generated models many difficulties arise which require a preliminary treatment, a reformulation, or a regularization, see [47]. In the case of linear constant coefficient DAEs or descriptor systems, this preliminary treatment is achieved using techniques from numerical linear algebra. In Sect. 3 the methods are based on derivative arrays and in Sect. 4 on staircase forms. These numerically stable methods allow us to check solvability and consistency of initial values for DAEs, as well as controllability and observability properties of descriptor systems.

After discussing the analysis and regularization techniques, we can proceed to more advanced control and optimization applications for descriptor systems. All these applications lead to generalized eigenvalue problems for even matrix pencils. Therefore, in Sect. 5 we discuss their structured condensed forms as well as the appropriate numerical methods. Afterward we consider the linear-quadratic regulator problem in Sect. 6 and the $\mathscr{H}_\infty$ optimal control problem in Sect. 7. In Sect. 8 we consider the computation of the $\mathscr{L}_\infty$-norm for continuous-time descriptor systems and finally, in Sect. 9 the dissipativity checking problem. Conclusions and comments on open problems complete the paper.

2 Solvability Theory

We begin our survey with the solvability theory of system (1.3a). This can done in terms of Kronecker canonical form (KCF) of the matrix pencil $sE-A \in \mathbb{R}[s]^{k,n}$, see, e.g., [43, 60].

Theorem 2.1 (Kronecker Canonical Form) *Let $sE - A \in \mathbb{R}[s]^{k,n}$ be given. Then there exist nonsingular matrices $P \in \mathbb{C}^{k,k}$ and $Q \in \mathbb{C}^{n,n}$ such that*

$$P(sE-A)Q = \operatorname{diag}\left(\mathscr{L}_{\varepsilon_1}(s), \ldots, \mathscr{L}_{\varepsilon_k}(s), \mathscr{L}_{\delta_1}(s)^T, \ldots, \mathscr{L}_{\delta_\ell}(s)^T, \mathscr{N}_{\sigma_1}(s), \ldots, \mathscr{N}_{\sigma_q}(s), \mathscr{J}_{\rho_1}(s), \ldots, \mathscr{J}_{\rho_r}(s)\right),$$

where

(i) each $\mathscr{L}_{\varepsilon_j}(s)$ is an $\varepsilon_j \times (\varepsilon_j + 1)$ right singular block with right minimal index $\varepsilon_j \in \mathbb{N}_0$ and form

$$s\begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix};$$

(ii) each $\mathscr{L}_{\delta_j}(s)^T$ *is a* $(\delta_j+1)\times\delta_j$ *left singular block with left minimal index* $\delta_j \in \mathbb{N}_0$ *and form*

$$s\begin{bmatrix} 0 & & \\ 1 & \ddots & \\ & \ddots & 0 \\ & & 1 \end{bmatrix} - \begin{bmatrix} 1 & & \\ 0 & \ddots & \\ & \ddots & 1 \\ & & 0 \end{bmatrix};$$

(iii) each $\mathscr{N}_{\sigma_j}(s)$ *is a* $\sigma_j \times \sigma_j$ *infinite eigenvalue block with index* $\sigma_j \in \mathbb{N}$ *and form*

$$s\begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix} - \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & 1 \end{bmatrix};$$

(iv) each $\mathscr{J}_{\rho_j}(s)$ *is a* $\rho_j \times \rho_j$ *Jordan block with index* $\rho_j \in \mathbb{N}$ *and finite eigenvalue* $\lambda_j \in \mathbb{C}$ *and form*

$$s\begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} - \begin{bmatrix} \lambda_j & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_j \end{bmatrix}.$$

The Kronecker canonical form is unique up to permutation of the blocks, i.e., the kind, size, and number of the blocks are invariants of the pencil $sE - A$.

In the real version of the KCF, the blocks $\mathscr{J}_{\rho_j}(s)$ are in real Jordan form [69] and the transformation matrices P, Q are real. Based on the KCF we have the following definition.

Definition 2.1

(i) A matrix pencil $sE-A \in \mathbb{R}[s]^{k,n}$ is called *regular*, if $k = n$ and $\det(\lambda E-A) \neq 0$ for some $\lambda \in \mathbb{C}$. Otherwise the pencil is called *singular*.

(ii) If $sE-A$ is regular, then a complex number λ_0 is a *finite eigenvalue* of $sE-A$, if $\det(\lambda_0 E - A) = 0$. The finite eigenvalues are associated with the $\mathscr{J}_{\rho_j}(s)$ blocks in the KCF whereas the $\mathscr{N}_{\sigma_j}(s)$ blocks are said to be corresponding the *infinite eigenvalues* of the pencil $sE - A$.

(iii) If $sE - A$ is a singular pencil, then its eigenvalues are the eigenvalues of the regular blocks in its Kronecker canonical form, i.e., the union of the eigenvalues of the $\mathscr{N}_{\sigma_j}(s)$ and $\mathscr{J}_{\rho_j}(s)$ blocks in Theorem 2.1.

(iv) The *Kronecker index* of a regular matrix pencil $sE-A$ is the size of the largest block $\mathscr{N}_{\sigma_j}(s)$ in Theorem 2.1. It is denoted by $\nu = \operatorname{ind}(E, A)$.

In the DAE case ($Bu \equiv 0$), it is clear from the KCF that for an arbitrary inhomogeneity $f(\cdot)$ and for arbitrary consistent initial conditions, to have a chance for a unique solution of (1.1), the pencil $sE - A$ has to be regular [44]. Nevertheless, if the pencil is singular, then for special $f(\cdot)$ and special initial conditions, a solution may exist and it even may be unique. Characterizations of existence and uniqueness of solutions can also be analyzed by different condensed forms, for instance the quasi-Kronecker form, see [24, 25]. For descriptor control systems (1.2a) the regularity of $sE - A$ is good to have, but not necessary.

In the regular case, the KCF specializes to the Weierstraß canonical form (WCF), see, e.g., [41, 60].

Theorem 2.2 (Weierstraß Canonical form (WCF)) *If $sE - A \in \mathbb{R}[s]^{n,n}$ is a regular pencil, then there exist nonsingular matrices $X = \begin{bmatrix} X_{\mathrm{f}} & X_\infty \end{bmatrix} \in \mathbb{C}^{n,n}$ and $Y = \begin{bmatrix} Y_{\mathrm{f}} & Y_\infty \end{bmatrix} \in \mathbb{C}^{n,n}$ for which*

$$Y^H(sE - A)X = \begin{bmatrix} Y_{\mathrm{f}}^H \\ Y_\infty^H \end{bmatrix} (sE - A) \begin{bmatrix} X_{\mathrm{f}} & X_\infty \end{bmatrix} = s \begin{bmatrix} I_r & 0 \\ 0 & N \end{bmatrix} - \begin{bmatrix} J & 0 \\ 0 & I_{n-r} \end{bmatrix}, \tag{2.1}$$

where $sI_r - J \in \mathbb{C}[s]^{r,r}$ with $J \in \mathbb{C}^{r,r}$ in Jordan canonical form contains the finite eigenvalues of $sE - A$, whereas the pencil $sN - I_{n-r} \in \mathbb{R}[s]^{n-r,n-r}$ with a nilpotent $N \in \mathbb{R}^{r,r}$ in Jordan canonical form corresponds to the infinite eigenvalues of $sE - A$.

Again there exists a real version of the WCF where J is in real Jordan canonical form and the transformation matrices X and Y are real. Since we prefer real-valued solutions we assume in the following considerations that the pencil $sE - A$ is transformed to real WCF and hence that X, Y as in Theorem 2.2 are real. With the notation of (2.1), classical continuously differentiable solutions of (1.3a) attain the form

$$x(t) = X_{\mathrm{f}} x_1(t) + X_\infty x_2(t),$$

where $x_1(\cdot)$, $x_2(\cdot)$ are solutions of

$$\begin{aligned} \dot{x}_1(t) &= Jx_1(t) + Y_{\mathrm{f}}^T(Bu(t) + f(t)) \\ N\dot{x}_2(t) &= x_2(t) + Y_\infty^T(Bu(t) + f(t)), \end{aligned}$$

respectively. With $\nu = \mathrm{ind}(E, A)$, there are the explicit solution formulas

$$\begin{aligned} x_1(t) &= e^{Jt}x_1(0) + \int_0^t e^{J(t-\tau)} Y_{\mathrm{f}}^T(Bu(\tau) + f(\tau))\,\mathrm{d}\tau, \\ x_2(t) &= -\sum_{i=0}^{\nu-1} \frac{\mathrm{d}^i}{\mathrm{d}t^i} \left(N^i Y_\infty^T(Bu(t) + f(t))\right). \end{aligned} \tag{2.2}$$

In the descriptor system case this shows that the input functions must belong to some suitable function space $\mathscr{U}_{\mathrm{ad}}$. In particular, they must be sufficiently smooth.

Equation (2.2) also shows that the possible values of the initial condition x_0 are restricted. The initial state must be an element of the set of *consistent* initial conditions

$$\mathscr{S} := \left\{ X_{\mathrm{f}} x_{0,1} + X_{\infty} x_{0,2} : x_{0,1} \in \mathbb{R}^r, \right.$$
$$\left. x_{0,2} = -\sum_{i=0}^{\nu-1} \frac{\mathrm{d}^i}{\mathrm{d}t^i} \left(N^i Y_{\infty}^T (Bu(0) + f(0)) \right), u(\cdot) \in \mathscr{U}_{\mathrm{ad}} \right\}.$$

To ensure a smooth response for every continuous input $u(\cdot)$ and every consistent initial value, it is necessary for the system to be regular and have index less than or equal to one.

The presented existence and uniqueness results are useful from a theoretical point of view, but it is well known that arbitrarily small perturbations can radically change the kind and number of the Kronecker blocks, and thus it is problematic to compute the KCF or WCF with a numerical algorithm in finite precision arithmetic [109].

A better way to obtain the full information about the characteristic invariants in the WCF and KCF is *staircase algorithms*, which use a sequence of rank decisions, orthogonal matrix multiplications, and small perturbations to transform a pencil into a generalized upper triangular (GUPTRI) form [52–54, 113], see Sect. 4. These staircase forms can be used to check solvability conditions and consistency of initial conditions. However, if the system violates the above-mentioned consistency and solvability conditions, or is close to such a system (in the sense that there exist small perturbations of the data that lead to a system that violates these conditions), then it is necessary to remodel or regularize the system such that further simulation and control methods are applicable. Again this should be done via numerically stable methods and this is the topic of the next section.

3 Regularization and Derivative Arrays

If not all the information about the characteristic quantities in the KCF is needed, then a very good alternative to the staircase form is to use a derivative array, see [45, 75]. This leads to a numerically stable method that allows us to check solvability and consistency of initial conditions [51]. Furthermore, if some of the conditions do not hold, then this approach can be used to obtain a regularization of the system. For general nonlinear DAEs and descriptor systems this general procedure has been presented in [47]. Here we briefly summarize this regularization procedure for the linear constant coefficient case.

In the following we assume that B and C^T have full column rank, otherwise we can just remove the kernels, by considering fewer inputs and outputs, respectively. This can be achieved by performing a singular value decomposition (SVD) or QR decomposition with column pivoting.

One first writes (1.3a) in behavior form, by combining input and state to a joint vector $z(\cdot) = \begin{bmatrix} x(\cdot)^T & u(\cdot)^T \end{bmatrix}^T$ as

$$\mathscr{E}\dot{z}(t) = \mathscr{A}z(t) + f(t) \tag{3.1}$$

with $\mathscr{E} = \begin{bmatrix} E & 0 \end{bmatrix}$, $\mathscr{A} = \begin{bmatrix} A & B \end{bmatrix}$ partitioned analogously. Then for given $\mu \in \mathbb{N}$, one forms an enlarged DAE, namely

$$\mathscr{M}_\mu \dot{z}_\mu(t) = \mathscr{N}_\mu z_\mu(t) + \phi_\mu(t),$$

where

$$\mathscr{M}_\mu = \begin{bmatrix} \mathscr{E} & & & \\ -\mathscr{A} & \mathscr{E} & & \\ & \ddots & \ddots & \\ & & -\mathscr{A} & \mathscr{E} \end{bmatrix} \in \mathbb{R}^{(\mu+1)k,(\mu+1)(n+m)},$$

$$\mathscr{N}_\mu = \begin{bmatrix} \mathscr{A} & 0 & \dots & 0 \\ 0 & 0 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \dots & \dots & 0 \end{bmatrix} \in \mathbb{R}^{(\mu+1)k,(\mu+1)(n+m)},$$

$$z_\mu(\cdot) = \begin{bmatrix} z(\cdot) \\ \dot{z}(\cdot) \\ \vdots \\ z^{(\mu)}(\cdot) \end{bmatrix}, \quad \phi_\mu(\cdot) = \begin{bmatrix} f(\cdot) \\ \dot{f}(\cdot) \\ \vdots \\ f^{(\mu)}(\cdot) \end{bmatrix}.$$

With the above notation, the pair $(\mathscr{M}_\mu, \mathscr{N}_\mu)$ is called derivative array [45]. One obtains the following theorem, see [74, 75, 78].

Theorem 3.1 *Let the system (3.1) be given. Then there exist integers μ, d, a, and v such that $(\mathscr{M}_\mu, \mathscr{N}_\mu)$ has the following properties:*

(i) $\operatorname{corank} \mathscr{M}_{\mu+1} - \operatorname{corank} \mathscr{M}_\mu = v$.

(ii) $\operatorname{rank} \mathscr{M}_\mu = (\mu+1)k - a - v$, *i.e., there exists a matrix $Z \in \mathbb{R}^{(\mu+1)k,a+v}$ with orthonormal columns and maximal rank, satisfying $Z^T \mathscr{M}_\mu = 0$.*

(iii) $\operatorname{rank} Z^T \mathscr{N}_\mu \begin{bmatrix} I_{n+m} & 0 & \dots & 0 \end{bmatrix}^T = a$, *i.e., Z can be partitioned as $Z = \begin{bmatrix} Z_2 & Z_3 \end{bmatrix}$, with $Z_2 \in \mathbb{R}^{(\mu+1)k,a}$ and $Z_3 \in \mathbb{R}^{(\mu+1)k,v}$ such that $\hat{A}_2 := Z_2^T \mathscr{N}_\mu \begin{bmatrix} I_{n+m} & 0 & \dots & 0 \end{bmatrix}^T$*

has full row rank a and $Z_3^T \mathscr{N}_\mu \begin{bmatrix} I_{n+m} & 0 & \dots & 0 \end{bmatrix}^T = 0$. *Furthermore, there exists a matrix* T_2 *with orthonormal columns of maximal rank satisfying* $\hat{A}_2 T_2 = 0$.

(iv) $\operatorname{rank} \mathscr{E} T_2 = d = k - a - v$, *i.e., there exists* $Z_1 \in \mathbb{R}^{k,d}$ *with orthonormal columns and maximal rank satisfying* $\hat{E}_1 := Z_1^T \mathscr{E}$ *with* $\operatorname{rank} \hat{E}_1 = d$.

Furthermore, system (3.1) *has the same solution set as the* strangeness-free *system*

$$\begin{bmatrix} \hat{E}_1 \\ 0 \\ 0 \end{bmatrix} \dot{z}(t) = \begin{bmatrix} \hat{A}_1 \\ \hat{A}_2 \\ 0 \end{bmatrix} z(t) + \begin{bmatrix} \hat{f}_1(t) \\ \hat{f}_2(t) \\ \hat{f}_3(t) \end{bmatrix}, \tag{3.2}$$

where $\hat{E}_1 = Z_1^T \mathscr{E}$, $\hat{A}_1 = Z_1^T \mathscr{A}$, $\hat{f}_1(\cdot) = Z_1^T f(\cdot)$, *and* $\hat{f}_i(\cdot) = Z_i^T \phi_\mu(\cdot)$, $i = 2, 3$.

The number μ is called the *strangeness-index* of the DAE. It is equal to the size of the largest block of types $\mathscr{L}_\varepsilon(s)$ or $\mathscr{N}_\sigma(s)$ and is equal to $\nu - 1$ with $\nu = \operatorname{ind}(E, A)$ if the pencil is regular with $\nu > 0$, see [75, 78]. It satisfies $\mu = 0$ if the system is regular and of index at most one. If μ is known, then the coefficients of the differential-algebraic system (3.2) can be computed by using three nullspace computations, which can be implemented via singular value decompositions or QR decompositions with column pivoting. If μ is not known, then one proceeds recursively, increasing μ until the form (3.2) can be numerically safely determined.

System (3.2) is a *reformulation* of (3.1) (using the original system and its derivatives), without changing the solution set, since no transformation of the vector $z(\cdot)$ has been made. The constructed submatrices $\hat{A}_1$ and $\hat{A}_2$ have been obtained from the block matrix

$$\begin{bmatrix} A & B \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{(\mu+1)k, n+m}$$

by transformations from the left, and this means that no derivatives of $u(\cdot)$ are needed, and hence, there are no additional smoothness requirements for $u(\cdot)$. Furthermore, we immediately obtain again a descriptor system of the form

$$E_1 \dot{x}(t) = A_1 x(t) + B_1 u(t) + \hat{f}_1(t), \quad x(0) = x_0 \tag{3.3a}$$

$$0 = A_2 x(t) + B_2 u(t) + \hat{f}_2(t), \tag{3.3b}$$

$$0 = \hat{f}_3(t), \tag{3.3c}$$

$$y(t) = Cx(t) + Du(t), \tag{3.3d}$$

where

$$E_1 = \hat{E}_1 \begin{bmatrix} I_n \\ 0 \end{bmatrix}, \quad A_i = \hat{A}_i \begin{bmatrix} I_n \\ 0 \end{bmatrix}, \quad B_i = \hat{A}_i \begin{bmatrix} 0 \\ I_m \end{bmatrix}, \quad i = 1, 2,$$

and $E_1, A_1 \in \mathbb{R}^{d,n}$, while $A_2 \in \mathbb{R}^{a,n}$.

The equations in (3.3c) characterize the solvability of (1.3a), which is given if $\hat{f}_3 \equiv 0$. If $\hat{f}_3 \not\equiv 0$, then the system does not have a classical solution. In this case either the model should be discarded or one can perform a regularization by just setting $\hat{f}_3 \equiv 0$ and release a warning that the system has been modified. In the latter case these equations can just be removed from the system and one continues with a modified state equation of $d + a$ equations in n state variables

$$E_1 \dot{x}(t) = A_1 x(t) + B_1 u(t) + \hat{f}_1(t), \quad x(0) = x_0 \tag{3.4a}$$

$$0 = A_2 x(t) + B_2 u(t) + \hat{f}_2(t), \tag{3.4b}$$

$$y(t) = Cx(t) + Du(t), \tag{3.4c}$$

together with the given initial conditions.

Consistency of initial values can be easily checked, they have to satisfy the equation

$$A_2 x_0 + B_2 u(0) + \hat{f}_2(0) = 0, \tag{3.5}$$

and this also restricts the set of admissible inputs $u(\cdot)$. Again if the given initial conditions do not satisfy (3.5), then a regularization would make them consistent.

In (3.4a) and (3.4b) we have $d + a$ equations and n variables in $x(\cdot)$ and m variables in $u(\cdot)$. In order for this system to be uniquely solvable for all sufficiently smooth inputs $u(\cdot)$, and all consistent initial conditions, as a necessary condition we would need that $d + a = n$ [44, 75]. In a reasonable model, this should be the case, but since automatically generated models typically have redundancies, and also there may be modeling errors, a mismatch may happen which can, however, be easily fixed. If $d + a < n$, then for given $u(\cdot)$ we cannot expect a unique solution, so we can just attach $n - d - a$ variables from $x(\cdot)$ to $u(\cdot)$ and if $d + a > n$, then we just attach $d + a - n$ of the input variables in $u(\cdot)$ to $x(\cdot)$. Note that we must also change the output equation by moving appropriate columns from D to C or vice versa. There is some freedom in this renaming of variables, which should be resolved by considering the application, and actually this step is not needed in some of the applications that we discuss below.

As a result of a possible reinterpretation of variables we obtain a remodeled system

$$\tilde{E}_1 \dot{\tilde{x}}(t) = \tilde{A}_1 \tilde{x}(t) + \tilde{B}_1 \tilde{u}(t) + \tilde{f}_1(t), \quad \tilde{x}(0) = \tilde{x}_0,$$

$$0 = \tilde{A}_2 \tilde{x}(t) + \tilde{B}_2 \tilde{u}(t) + \tilde{f}_2(t),$$

$$y(t) = \tilde{C} \tilde{x}(t) + \tilde{D} \tilde{u}(t),$$

where $\begin{bmatrix}\tilde{E}_1\\0\end{bmatrix}, \begin{bmatrix}\tilde{A}_1\\\tilde{A}_2\end{bmatrix} \in \mathbb{R}^{\tilde{n},\tilde{n}}, \begin{bmatrix}\tilde{B}_1\\\tilde{B}_2\end{bmatrix} \in \mathbb{R}^{\tilde{n},\tilde{m}}$, and $\tilde{n}$, $\tilde{m}$ are the numbers of state and input variables of the reinterpreted system, respectively.

It is also often useful to remove the feed-through term $\tilde{D}\tilde{u}(\cdot)$ in the output equation. This can be achieved by performing a row compression with an orthogonal matrix P such that $P^T\tilde{D} = \begin{bmatrix}\tilde{D}_1\\0\end{bmatrix}$ with $\tilde{D}_1 \in \mathbb{R}^{p_1,\tilde{m}}$ having full row rank. By setting (with an accordant partitioning)

$$P^T y(\cdot) = \begin{bmatrix}\tilde{y}_1(\cdot)\\\tilde{y}_2(\cdot)\end{bmatrix}, \quad P^T\tilde{C} = \begin{bmatrix}\tilde{C}_1\\\tilde{C}_2\end{bmatrix},$$

with $\tilde{C}_1 \in \mathbb{R}^{p_1,\tilde{n}}$, then we obtain a new system without feed-through term of the form

$$\overline{E}\dot{\overline{x}}(t) = \overline{A}\overline{x}(t) + \overline{B}\overline{u}(t) + \overline{f}(t), \quad \overline{x}(0) = \overline{x}_0, \tag{3.6a}$$

$$\overline{y}(t) = \overline{C}\overline{x}(t), \tag{3.6b}$$

with data

$$\overline{E} = \begin{bmatrix}\tilde{E}_1 & 0\\0 & 0\\0 & 0\end{bmatrix} \in \mathbb{R}^{\overline{n},\overline{n}}, \quad \overline{A} = \begin{bmatrix}\tilde{A}_1 & 0\\\tilde{A}_2 & 0\\\tilde{C}_1 & -I_{p_1}\end{bmatrix} \in \mathbb{R}^{\overline{n},\overline{n}}, \quad \overline{B} = \begin{bmatrix}\tilde{B}_1\\\tilde{B}_2\\\tilde{D}_1\end{bmatrix} \in \mathbb{R}^{\overline{n},\overline{m}},$$

$$\overline{C} = \tilde{C}_2 \in \mathbb{R}^{\overline{p},\overline{n}},$$

$$\overline{x}(\cdot) = \begin{bmatrix}\tilde{x}(\cdot)\\\tilde{y}_1(\cdot)\end{bmatrix}, \quad \overline{y}(\cdot) = \tilde{y}_2(\cdot), \quad \overline{u}(\cdot) = \tilde{u}(\cdot), \quad \overline{f}(\cdot) = \begin{bmatrix}\tilde{f}_1(\cdot)\\\tilde{f}_2(\cdot)\\0\end{bmatrix},$$

where $\overline{n} = \tilde{n} + p_1$, $\overline{m} = \tilde{m}$, and $\overline{p} = p - p_1$. The resulting system may not be of index at most one as a free system with $\overline{u} \equiv 0$. In this case, see [41, 78], one can construct a linear state feedback $\overline{u}(t) = K\overline{x}(t) + w(t)$, with $K \in \mathbb{R}^{\overline{m},\overline{n}}$ such that in the closed-loop system

$$\overline{E}\dot{\overline{x}}(t) = (\overline{A} + \overline{B}K)\overline{x}(t) + \overline{B}w(t) + \overline{f}(t), \quad \overline{x}(0) = \overline{x}_0, \tag{3.7a}$$

$$\overline{y}(t) = \overline{C}\overline{x}(t), \tag{3.7b}$$

the matrix $(\overline{A}_2 + \overline{B}_2K)\overline{S}_\infty$ is nonsingular, where $\overline{S}_\infty$ is a matrix that spans $\ker \overline{E}_1$. This implies that the DAE in (3.7a) is regular and of index at most one as a free system with $w \equiv 0$. A similar index reduction can also be constructed via output feedback [40, 41, 78], it would however require a change of basis in the state space, and thus a change of the physical meaning of the state variables. See Sect. 4 for more details. The flowchart given in Fig. 1 summarizes the regularization procedure.

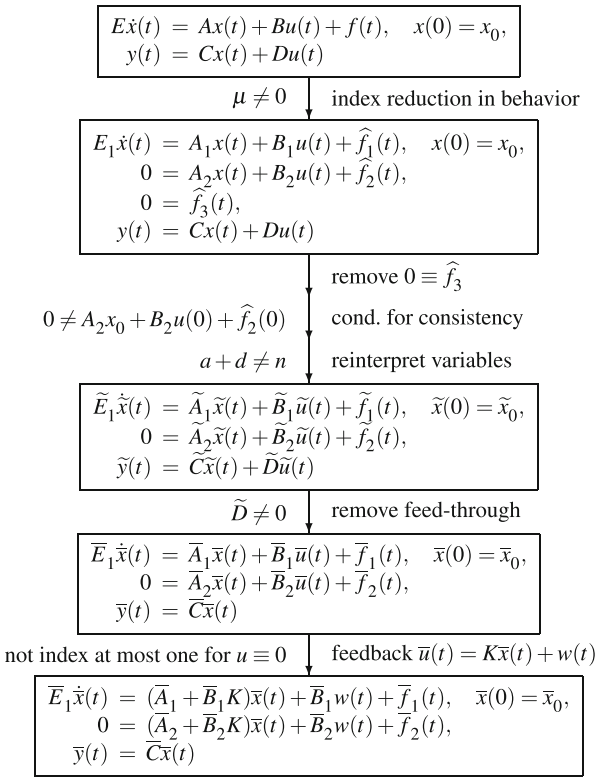

Fig. 1 Regularization procedure

Note that several of the steps in the regularization procedure may be void if the system has adequate properties and for some of the applications discussed below also a preliminary regularization may not be necessary. Note furthermore that in this procedure no changes have been performed in the state variables, except for the possible reinterpretation of variables or the extension of the state space in the case of feed-through removal. In any case, the original physical meaning of the state variables is still present in the system. This is of great importance and a clear advantage of the derivative array approach compared to the staircase forms that we discuss below.

Example 3.1 To illustrate the regularization procedure, consider the following example:

$$E = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad f(t) = \begin{bmatrix} f_1(t) \\ f_2(t) \\ f_3(t) \end{bmatrix},$$

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

The corresponding behavioral system (3.1) is given by

$$\mathscr{E} = \begin{bmatrix} 0\,1\,0\,0 \\ 0\,0\,0\,0 \\ 0\,0\,0\,0 \end{bmatrix}, \quad \mathscr{A} = \begin{bmatrix} 1\,0\,0\,1 \\ 0\,1\,0\,0 \\ 0\,0\,0\,0 \end{bmatrix}, \quad f(t) = \begin{bmatrix} f_1(t) \\ f_2(t) \\ f_3(t) \end{bmatrix}$$

The strangeness index of this system is 1. Thus we obtain the derivative array

$$(\mathscr{M}_1, \mathscr{N}_1) = \left(\begin{bmatrix} 0 & 1 & 0 & 0 & 0\,0\,0\,0 \\ 0 & 0 & 0 & 0 & 0\,0\,0\,0 \\ 0 & 0 & 0 & 0 & 0\,0\,0\,0 \\ -1 & 0 & 0 & -1 & 0\,1\,0\,0 \\ 0 & -1 & 0 & 0 & 0\,0\,0\,0 \\ 0 & 0 & 0 & 0 & 0\,0\,0\,0 \end{bmatrix}, \begin{bmatrix} 1\,0\,0\,1\,0\,0\,0\,0 \\ 0\,1\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0\,0 \end{bmatrix} \right).$$

With the notation of Theorem 3.1 we obtain

$$Z_2^T = \begin{bmatrix} 1\,0\,0\,0\,1\,0 \\ 0\,1\,0\,0\,0\,0 \end{bmatrix}, \quad Z_3^T = \begin{bmatrix} 0\,0\,1\,0\,0\,0 \\ 0\,0\,0\,0\,0\,1 \end{bmatrix}.$$

Then we obtain

$$\hat{A}_2 := Z_2^T \mathscr{N}_1 \begin{bmatrix} I_4 \\ 0 \end{bmatrix} = \begin{bmatrix} 1\,0\,0\,1 \\ 0\,1\,0\,0 \end{bmatrix},$$

and consequently we have

$$T_2 = \begin{bmatrix} -1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

We obtain $\mathscr{E} T_2 = 0$ and thus Z_1 is void. Overall, the strangeness-free system (3.2) reads

$$0 = \begin{bmatrix} 1\,0\,0\,1 \\ 0\,1\,0\,0 \\ 0\,0\,0\,0 \\ 0\,0\,0\,0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ u(t) \end{bmatrix} + \begin{bmatrix} f_1(t) + \dot{f}_2(t) \\ f_2(t) \\ f_3(t) \\ \dot{f}_3(t) \end{bmatrix},$$

$$y(t) = \begin{bmatrix} 1\,0\,0 \\ 0\,0\,1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u(t).$$

We see that in order for the system to be solvable at all, we need $f_3 \equiv 0$. Therefore, assume that $f_3 \equiv 0$. Then the reduced system (3.4) is given by

$$0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u(t) + \begin{bmatrix} f_1(t) + \dot{f}_2(t) \\ f_2(t) \end{bmatrix},$$

$$y(t) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u(t),$$

from which we can directly read off the condition for consistency by setting $t = 0$. Since $2 = a + d \neq n = 3$, a reinterpretation of variables is necessary. Thus, by setting $u_1(\cdot) := u(\cdot)$ and $u_2(\cdot) := x_3(\cdot)$, we obtain the square system

$$0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix} + \begin{bmatrix} f_1(t) + \dot{f}_2(t) \\ f_2(t) \end{bmatrix},$$

$$y(t) = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix}.$$

Now we remove the feed-through matrix D. Thus the feed-through-free system (3.6) reads

$$0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ y_1(t) \\ y_2(t) \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix} + \begin{bmatrix} f_1(t) + \dot{f}_2(t) \\ f_2(t) \\ 0 \\ 0 \end{bmatrix},$$

which is regular and of index one. Note that the output equation has vanished during the last step in the regularization procedure. In fact, it is included in the state equation.

4 Staircase Forms and Properties of Descriptor Systems

In this section we discuss the system theoretic properties of descriptor systems and present the staircase forms that allow us to check these properties. We focus on concepts like *controllability, stabilizability* and the related dual notions of *observability* and *detectability*. For brevity we only introduce these for the case of square systems and systems where the feed-through term D has been removed, so we assume that the system is already in the form (3.6) as generated by the regularization procedure of Sect. 3. Also, instead of defining these properties in system theoretic terms, we directly introduce equivalent algebraic characterizations.

These are very useful for numerically checking these properties. Note that there are several different concepts of controllability at infinity introduced in [102, 115] and compared in [22, 23, 41, 49, 50]. Furthermore, different observability notions are reviewed in [26].

Proposition 4.1 *Let $sE - A \in \mathbb{R}[s]^{n,n}$ be regular, $B \in \mathbb{R}^{n,m}$, $C \in \mathbb{R}^{p,n}$. Furthermore, let S_∞, T_∞ be matrices with* $\operatorname{range} S_\infty = \ker E$ *and* $\operatorname{range} T_\infty = \ker E^T$. *Then the triple (E, A, B) is called*

(i) behaviorally stabilizable *if* $\operatorname{rank}\begin{bmatrix}\lambda E - A & B\end{bmatrix} = n$ *for all* $\lambda \in \mathbb{C}^+ \cup \mathrm{i}\mathbb{R}$;
(ii) behaviorally controllable *if* $\operatorname{rank}\begin{bmatrix}\lambda E - A & B\end{bmatrix} = n$ *for all* $\lambda \in \mathbb{C}$;
(iii) impulse controllable *if* $\operatorname{rank}\begin{bmatrix}E & AS_\infty & B\end{bmatrix} = n$;
(iv) strongly stabilizable *if it is both behaviorally stabilizable and impulse controllable;*
(v) strongly controllable *if it is both behaviorally controllable and impulse controllable;*
(vi) completely controllable *if it is both behaviorally controllable and* $\operatorname{rank}\begin{bmatrix}E & B\end{bmatrix} = n$.

In a dual fashion, the triple (E, A, C) is called

(vii) behaviorally detectable *if* $\operatorname{rank}\begin{bmatrix}\lambda E^T - A^T & C^T\end{bmatrix} = n$ *for all* $\lambda \in \mathbb{C}^+ \cup \mathrm{i}\mathbb{R}$;
(viii) behaviorally observable *if* $\operatorname{rank}\begin{bmatrix}\lambda E^T - A^T & C^T\end{bmatrix} = n$ *for all* $\lambda \in \mathbb{C}$;
(ix) impulse observable *if* $\operatorname{rank}\begin{bmatrix}E^T & A^T T_\infty & C^T\end{bmatrix} = n$;
(x) strongly detectable *if it is both behaviorally detectable and impulse observable;*
(xi) strongly observable *if it is both behaviorally observable and impulse observable;*
(xii) completely observable *if it is both behaviorally observable and* $\operatorname{rank}\begin{bmatrix}E^T & C^T\end{bmatrix} = n$.

To check whether a given descriptor system satisfies these conditions one can use the staircase form of [40, 41, 113], which can be implemented as a sequence of orthogonal transformations to the system [28].

Theorem 4.2 ([40]) *If $E, A \in \mathbb{R}^{n,n}$, $B \in \mathbb{R}^{n,m}$, $C \in \mathbb{R}^{p,n}$, then there exist orthogonal matrices $U, V \in \mathbb{R}^{n,n}$, $W \in \mathbb{R}^{m,m}$, $Y \in \mathbb{R}^{p,p}$ such that*

$$U^T E V = \begin{array}{c} \\ t_1 \\ n - t_1 \end{array} \begin{array}{c} \begin{array}{cc} t_1 & n - t_1 \end{array} \\ \begin{bmatrix} \Sigma_E & 0 \\ 0 & 0 \end{bmatrix} \end{array},$$

$$
U^T A V = \begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \\ t_6 \end{array}
\begin{array}{c}
\begin{array}{cccccc} t_1 & s_2 & t_5 & t_4 & t_3 & s_6 \end{array} \\
\begin{bmatrix}
A_{11} & A_{12} & A_{13} & A_{14} & A_{15} & A_{16} \\
A_{21} & A_{22} & A_{23} & A_{24} & 0 & 0 \\
A_{31} & A_{32} & A_{33} & A_{34} & \Sigma_{35} & 0 \\
A_{41} & A_{42} & A_{43} & \Sigma_{44} & 0 & 0 \\
A_{51} & 0 & \Sigma_{53} & 0 & 0 & 0 \\
A_{61} & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\end{array},
$$

$$
U^T B W = \begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \\ t_6 \end{array}
\begin{array}{c}
\begin{array}{ccc} k_1 & k_2 & k_3 \end{array} \\
\begin{bmatrix}
B_{11} & B_{12} & 0 \\
B_{21} & 0 & 0 \\
B_{31} & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{bmatrix}
\end{array},
\tag{4.1}
$$

$$
Y^T C V = \begin{array}{c} \\ l_1 \\ l_2 \\ l_3 \end{array}
\begin{array}{c}
\begin{array}{cccccc} t_1 & s_2 & t_5 & t_4 & t_3 & s_6 \end{array} \\
\begin{bmatrix}
C_{11} & C_{12} & C_{13} & 0 & 0 & 0 \\
C_{21} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\end{array}.
$$

The matrices $\Sigma_E, \Sigma_{35}, \Sigma_{44}, \Sigma_{53}$ are nonsingular diagonal matrices, B_{12} has full column rank, C_{21} has full row rank, and the matrices

$$
\begin{bmatrix} B_{21} \\ B_{31} \end{bmatrix} \in \mathbb{R}^{k_1, k_1}, \quad \begin{bmatrix} C_{12} & C_{13} \end{bmatrix} \in \mathbb{R}^{l_1, l_1},
$$

with $k_1 = t_2 + t_3$ and $l_1 = s_2 + t_5$ are nonsingular.

For numerical examples for the above decomposition we refer to the appendix of [41]. Impulse controllability and observability and some further properties can be checked via the following corollary.

Corollary 4.3 ([40]) *Let E, A, B, C with regular $sE - A$ be given as in the condensed form (4.1). Then the following statements are satisfied.*

(i) The triple (E, A, B) is impulse controllable if and only if $t_6 = 0$, i.e., the last block row of A is void.

(ii) *The triple* (E, A, C) *is impulse observable if and only if* $s_6 = 0$*, i.e., the last block column of* A *is void.*

(iii) *The condition* $\operatorname{rank}\begin{bmatrix} E & B \end{bmatrix} = n$ *is satisfied if and only if* $t_4 = t_5 = t_6 = 0$.

(iv) *The condition* $\operatorname{rank}\begin{bmatrix} E^T & C^T \end{bmatrix} = n$ *is satisfied if and only if* $t_4 = t_3 = s_6 = 0$.

(v) *The triple* (E, A, B) *is completely controllable if and only if* $t_4 = t_5 = t_6 = 0$ *and the system is behaviorally controllable.*

(vi) *The triple* (E, A, C) *is completely observable if and only if* $t_4 = t_3 = s_6 = 0$ *and the system is behaviorally observable.*

If properly implemented, see [52–54, 113], these techniques determine the characteristic invariants of a *least generic* system within a tolerated perturbation, see [56, 57]. In this way the staircase form (4.1) presents an alternative way to check some of the properties compared to the derivative array as in Sect. 3. But the computation of the staircase form is much more subtle numerically, since the consecutive rank decisions of the transformed matrices have to be made in a proper way, see [113]. In contrast to the derivative array approach, two-sided transformations are used, i.e., also changes of basis in the state space. This allows us to check observability and controllability conditions simultaneously, but at the cost of changing the physical meaning of the state variables. This is clearly no problem when the data is produced from a realization or model reduction process [3, 4], where the state variables have no direct physical interpretation, but this may be a problem when the model directly arises from a physical model. In this case it is suggested to first perform the regularization procedure of Sect. 3 and then perform the staircase algorithm to check the properties.

If only the first step of the regularization via derivative arrays has been performed and the system (3.4) has been filtered out of the derivative array and $d + a = n$, i.e., no more reinterpretation of variables is necessary, then the system is already impulse controllable. If the system is not impulse observable, then this is critical because impulse observability cannot be achieved by removing equations and variables. In this case, impulses in the solution (that appear, e.g., for inconsistent initial values) cannot be observed and this is an indication of a problem in the modeling, see [42]. In some of the applications that we discuss below, the solvability depends on these properties and an alternative model should be created to ensure that they are satisfied.

If the system is impulse controllable and impulse observable, then the other properties, i.e., behavioral controllability or stabilizability and behavioral observability or detectability can be checked via the following *controllability/observability decompositions*, see [50, 113, 114]. Let Q_c, Z_c be real orthogonal matrices, such that

$$\begin{aligned} Q_c^T E Z_c &= \begin{bmatrix} E_c & * \\ 0 & E_{nc} \end{bmatrix}, \; Q_c^T A Z_c = \begin{bmatrix} A_c & * \\ 0 & A_{nc} \end{bmatrix}, \\ Q_c^T B &= \begin{bmatrix} B_c \\ 0 \end{bmatrix}, \qquad C Z_c = \begin{bmatrix} C_c & C_{nc} \end{bmatrix}, \end{aligned} \tag{4.2}$$

where the subsystem given by the matrices E_c, A_c, B_c, C_c contains the *controllable subsystem* of the original system, i.e., the triple (E_c, A_c, B_c) is behaviorally controllable. If the subpencil $sE_{nc} - A_{nc}$ corresponding to the uncontrollable part of the system has no finite eigenvalues with nonnegative real part, then the system is behaviorally stabilizable, otherwise it is not.

Similarly, one can determine an *observability decomposition*

$$\begin{aligned} Q_o^T E Z_o &= \begin{bmatrix} E_o & 0 \\ * & E_{no} \end{bmatrix}, \; Q_o^T A Z_o = \begin{bmatrix} A_o & 0 \\ * & A_{no} \end{bmatrix}, \\ Q_o^T B &= \begin{bmatrix} B_o \\ B_{no} \end{bmatrix}, \qquad C Z_o = \begin{bmatrix} C_o & 0 \end{bmatrix}, \end{aligned} \tag{4.3}$$

where Q_o, Z_o are orthogonal matrices and the subsystem given by the matrices E_o, A_o, B_o, C_o contains the *observable subsystem* of the original system, i.e., the system (E_o, A_o, C_o) is behaviorally observable. If the subpencil $sE_{no} - A_{no}$ corresponding to the unobservable part of the system has no finite eigenvalues with nonnegative real part, then the system is behaviorally detectable, otherwise it is not. Methods for the computation of these decompositions are described in [114] and implemented as `TG01HD`, `TG01ID` in the SLICOT library.[3]

For some applications, in particular those where the influence of the inputs to the outputs is crucial, it is not suitable to analyze the descriptor system in the time domain, i.e., in the form (1.2). Instead, one turns to the frequency domain. For this, assume that the system is square and that the pencil $sE - A$ is regular. Then we can apply the Laplace transformation to the functions $x(\cdot)$, $u(\cdot)$, and $y(\cdot)$ and under the assumption that $Ex(0) = 0$ we obtain the *transfer function*

$$G(s) := C(sE - A)^{-1}B + D \in \mathbb{R}(s)^{p,m}, \tag{4.4}$$

that directly maps the Laplace transformed inputs to the Laplace transformed outputs [50]. These transfer functions are typically associated with certain function spaces. Consider the normed spaces $\mathscr{RH}_\infty^{p,m}$ and $\mathscr{RL}_\infty^{p,m}$ of all *real-rational* $\mathbb{C}^{p,m}$-valued functions that are analytic and bounded in the open right half plane $\mathbb{C}^+$; and bounded on the imaginary axis $\mathrm{i}\mathbb{R}$, respectively. Obviously, the inclusion $\mathscr{RH}_\infty^{p,m} \subset \mathscr{RL}_\infty^{p,m}$ holds. For $G \in \mathscr{RL}_\infty^{p,m}$, the $\mathscr{L}_\infty$-norm is given by

$$\|G\|_{\mathscr{L}_\infty} = \sup_{\omega \in \mathbb{R}} \sigma_{\max}(G(\mathrm{i}\omega)),$$

where $\sigma_{\max}(\cdot)$ denotes the maximum singular value. For $G \in \mathscr{RH}_\infty^{p,m}$, the $\mathscr{L}_\infty$-norm is equal to the $\mathscr{H}_\infty$-norm. These norms play an important role in many applications, in particular as robustness measures in robust control. Details on this will be pointed out in Sects. 7 and 8.

[3] http://slicot.org/.

5 Even Matrix Pencils

After briefly introducing the basic concepts, some of the system theoretic properties and numerical methods to check these properties, we now turn to several important applications in control theory. As we will see later in the forthcoming sections, these are based on generalized eigenvalue problems for *even* matrix pencils. A matrix pencil $sN - M \in \mathbb{R}[s]^{n,n}$ is called even, if $N^T = -N$ and $M^T = M$. Besides the applications presented in this paper, even matrix pencils also play a role in linearized models that occur in the vibration analysis of buildings, machines, and vehicles [27, 61, 82, 87, 91, 111].

If the dimension of an even matrix pencil is even, i.e., $n = 2m$, then it is closely related to so-called *skew-Hamiltonian/Hamiltonian* matrix pencils [14, 90, 92, 95]. A matrix pencil $sS - H \in \mathbb{R}[s]^{2m,2m}$ is called *skew-Hamiltonian/Hamiltonian* if $\mathscr{J}_m(sS - H)$ is even, where

$$\mathscr{J}_m = \begin{bmatrix} 0 & I_m \\ -I_m & 0 \end{bmatrix},$$

that means S is skew-Hamiltonian (i.e., $(\mathscr{J}_m S)^T = -\mathscr{J}_m S$) and H is Hamiltonian (i.e., $(\mathscr{J}_m H)^T = \mathscr{J}_m H$).

Since even pencils are so closely related to skew-Hamiltonian/Hamiltonian pencils, it is easy to show that they exhibit the Hamiltonian spectral symmetry, i.e., if λ is a finite eigenvalue of $sN - M$, then $-\overline{\lambda}$ is an eigenvalue as well. This means that nonreal and nonimaginary finite eigenvalues of an even pencil appear in quadruples $(\lambda, -\lambda, \overline{\lambda}, -\overline{\lambda})$ while for purely real or purely imaginary eigenvalues they form pairs $(\lambda, -\lambda)$, $(\lambda, \overline{\lambda})$ on the real or imaginary axis, respectively. The only exceptions are the eigenvalues 0 and ∞. Furthermore, it is also well known that even pencils possess a structured Kronecker canonical form [110] as well as a corresponding staircase form under orthogonal congruence transformations [37, 43]. We briefly recall these forms within the next subsection. A structured Smith form is available as well [88]. The staircase form allows us to filter out a regular even pencil which has Kronecker blocks at ∞ of size at most one for which we can apply structure-preserving methods for skew-Hamiltonian/Hamiltonian eigenvalue problems. These are discussed in Sect. 5.2.

5.1 Structured Condensed Forms

Even pencils have a special Kronecker canonical form under congruence transformations which preserve the even structure, see [110]. This canonical form is described in the following theorem. Besides the usual invariants occurring in the Kronecker canonical form, the even Kronecker form has further invariants associ-

ated with each purely imaginary eigenvalue, called *sign-characteristics*. These arise due to the fact that congruence transformations preserve inertia.

Theorem 5.1 *If $sN - M \in \mathbb{R}[s]^{n,n}$ with $N = -N^T$ and $M = M^T$, then there exists a nonsingular matrix $X \in \mathbb{R}^{n,n}$ such that*

$$X^T(sN - M)X = \operatorname{diag}(\mathscr{K}_{\mathscr{S}}(s), \mathscr{K}_{\mathscr{I}}(s), \mathscr{K}_{\mathscr{Z}}(s), \mathscr{K}_{\mathscr{F}}(s)),$$

where

$$\begin{aligned}
\mathscr{K}_{\mathscr{S}}(s) &= \operatorname{diag}\left(\mathscr{S}_{\xi_1}(s), \ldots, \mathscr{S}_{\xi_k}(s)\right), \\
\mathscr{K}_{\mathscr{I}}(s) &= \operatorname{diag}\left(\mathscr{I}_{2\varepsilon_1+1}(s), \ldots, \mathscr{I}_{2\varepsilon_l+1}(s), \mathscr{I}_{2\delta_1}(s), \ldots, \mathscr{I}_{2\delta_m}(s)\right), \\
\mathscr{K}_{\mathscr{Z}}(s) &= \operatorname{diag}\left(\mathscr{Z}_{2\sigma_1+1}(s), \ldots, \mathscr{Z}_{2\sigma_r+1}(s), \mathscr{Z}_{2\rho_1}(s), \ldots, \mathscr{Z}_{2\rho_t}(s)\right), \\
\mathscr{K}_{\mathscr{F}}(s) &= \operatorname{diag}\left(\mathscr{R}_{\phi_1}(s), \ldots, \mathscr{R}_{\phi_u}(s), \mathscr{C}_{\psi_1}(s), \ldots, \mathscr{C}_{\psi_v}(s)\right)
\end{aligned}$$

and the blocks have the following properties:

(i) each $\mathscr{S}_{\xi_j}(s)$ is a $(2\xi_j + 1) \times (2\xi_j + 1)$ block ($\xi_j \in \mathbb{N}_0$) that combines a right singular block and a left singular block, both of minimal index ξ_j. It has the form

$$s\left[\begin{array}{ccc|ccc}
 & & & 1 & 0 & \\
 & & & & \ddots & \ddots \\
 & & & \ddots & \ddots & 1 \\
 & & & 1 & 0 & \\
\hline
 & & -1 & & & \\
 & \ddots & 0 & & & \\
-1 & \ddots & & & & \\
0 & & & & &
\end{array}\right] - \left[\begin{array}{ccc|ccc}
 & & & 0 & 1 & \\
 & & & & \ddots & \ddots \\
 & & & \ddots & \ddots & 0 \\
 & & & 0 & 1 & \\
\hline
 & & 0 & & & \\
 & \ddots & 1 & & & \\
0 & \ddots & & & & \\
1 & & & & &
\end{array}\right];$$

(ii) each $\mathscr{I}_{2\varepsilon_j+1}(s)$ is a $(2\varepsilon_j + 1) \times (2\varepsilon_j + 1)$ block ($\varepsilon_j \in \mathbb{N}_0$) that contains a single block corresponding to the eigenvalue $\lambda = \infty$ of size $2\varepsilon_j + 1$. It has the form

$$s\left[\begin{array}{ccc|ccc}
 & & & 1 & 0 & \\
 & & & & \ddots & \ddots \\
 & & & \ddots & \ddots & 1 \\
 & & & 1 & 0 & \\
\hline
 & & -1 & 0 & & \\
 & \ddots & 0 & & & \\
-1 & \ddots & & & & \\
0 & & & & &
\end{array}\right] - \left[\begin{array}{ccc|ccc}
 & & & 0 & 1 & \\
 & & & & \ddots & \ddots \\
 & & & \ddots & \ddots & 0 \\
 & & & 0 & 1 & \\
\hline
 & & 0 & \gamma & & \\
 & \ddots & 1 & & & \\
0 & \ddots & & & & \\
1 & & & & &
\end{array}\right],$$

where $\gamma \in \{1, -1\}$ is the sign-characteristic of the block;

(iii) each $\mathscr{I}_{2\delta_j}(s)$ is a $4\delta_j \times 4\delta_j$ block ($\delta_j \in \mathbb{N}$) that combines two $2\delta_j \times 2\delta_j$ blocks associated with $\lambda = \infty$. It has the form

$$s\begin{bmatrix} & & & 1 & 0 \\ & & & \ddots & \ddots \\ & & & 1 & \ddots \\ & & & 0 & \\ & & -1 & 0 & \\ & \ddots & \ddots & & \\ -1 & \ddots & & & \\ 0 & & & & \end{bmatrix} - \begin{bmatrix} & & & & 1 \\ & & & \ddots & \ddots \\ & & & 1 & \\ & & 1 & & \\ & \ddots & & & \\ 1 & & & & \end{bmatrix};$$

(iv) each $\mathscr{Z}_{2\sigma_j+1}(s)$ is a $(4\sigma_j + 2) \times (4\sigma_j + 2)$ block ($\sigma_j \in \mathbb{N}_0$) that combines two $(2\sigma_j + 1) \times (2\sigma_j + 1)$ Jordan blocks corresponding to the eigenvalue $\lambda = 0$. It has the form

$$s\begin{bmatrix} & & & 1 \\ & & \ddots & \ddots \\ & & 1 & \\ & -1 & & \\ \ddots & \ddots & & \\ -1 & & & \end{bmatrix} - \begin{bmatrix} & & & & 1 & 0 \\ & & & & \ddots & \ddots \\ & & & & 1 & \ddots \\ & & & & 0 & \\ & & 1 & 0 & & \\ & \ddots & \ddots & & & \\ 1 & \ddots & & & & \\ 0 & & & & & \end{bmatrix};$$

(v) each $\mathscr{Z}_{2\rho_j}(s)$ is a $2\rho_j \times 2\rho_j$ block ($\rho_j \in \mathbb{N}$) that contains a single Jordan block corresponding to the eigenvalue $\lambda = 0$. It has the form

$$s\begin{bmatrix} & & & 1 \\ & & \ddots & \ddots \\ & & 1 & \\ & -1 & & \\ \ddots & \ddots & & \\ -1 & & & \end{bmatrix} - \begin{bmatrix} & & & & 1 & 0 \\ & & & & \ddots & \ddots \\ & & & & 1 & \ddots \\ & & & \gamma & 0 & \\ & & 1 & 0 & & \\ & \ddots & \ddots & & & \\ 1 & \ddots & & & & \\ 0 & & & & & \end{bmatrix},$$

where $\gamma \in \{1, -1\}$ is the sign-characteristic of this block;

(vi) each $\mathscr{R}_{\phi_j}(s)$ is a $2\phi_j \times 2\phi_j$ block ($\phi_j \in \mathbb{N}$) that combines two $\phi_j \times \phi_j$ Jordan blocks corresponding to nonzero real eigenvalues a_j and $-a_j$. It has the form

$$s\begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{bmatrix} - \begin{bmatrix} & & 1 & a_j \\ & \iddots & \iddots & \\ 1 & \iddots & & \\ a_j & & & \end{bmatrix};$$

(vii) the entries $\mathscr{C}_{\psi_j}(s)$ take two slightly different forms:

(a) one possibility is that $\mathscr{C}_{\psi_j}(s)$ is a $2\psi_j \times 2\psi_j$ block ($\psi_j \in \mathbb{N}$) combining two $\psi_j \times \psi_j$ Jordan blocks corresponding to purely imaginary eigenvalues $\mathrm{i}b_j, -\mathrm{i}b_j$ ($b_j > 0$). In this case it has the form

$$s\begin{bmatrix} & & 1 \\ & \iddots & \\ -1 & & \end{bmatrix} - \gamma \begin{bmatrix} & & 1 & b_j \\ & \iddots & \iddots & \\ 1 & \iddots & & \\ b_j & & & \end{bmatrix},$$

where $\gamma \in \{1, -1\}$ is the sign-characteristic;

(b) the other possibility is that $\mathscr{C}_{\psi_j}(s)$ is a $4\psi_j \times 4\psi_j$ block ($\psi_j \in \mathbb{N}$) combining $\psi_j \times \psi_j$ Jordan blocks for each of the complex eigenvalues $a_j + \mathrm{i}b_j, a_j - \mathrm{i}b_j, -a_j + \mathrm{i}b_j, -a_j - \mathrm{i}b_j$ (with $a_j \neq 0$ and $b_j \neq 0$). In this case it has the form

$$s\begin{bmatrix} & & \Omega \\ & \iddots & \\ -\Omega & & \end{bmatrix} - \begin{bmatrix} & & \Omega & \Lambda_j \\ & \iddots & \iddots & \\ \Omega & \iddots & & \\ \Lambda_j & & & \end{bmatrix}$$

with $\Omega = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ *and* $\Lambda_j = \begin{bmatrix} -b_j & a_j \\ a_j & b_j \end{bmatrix}$.

This structured Kronecker canonical form is unique up to permutation of the blocks, i.e., the kind, size, and number of the blocks as well as the sign-characteristics are invariants of the pencil $sN - M$ under congruence transformations.

An even pencil is called *regular* if and only if no blocks of type (i) occur in the even Kronecker form. The *(Kronecker) index* of the pencil is the size of the largest block of type (ii) and (iii) in the even Kronecker form, thus a regular pencil is of *index*

at most one if and only if there are no blocks of type (iii) and the blocks of type (ii) are of size at most one. In some of the applications discussed below, it will be necessary to detect whether an even matrix pencil is regular and of index at most one and whether there exist *finite eigenvalues* with real part 0. In other applications the computation of the stable deflating subspace, i.e., the subspace spanned by the eigenvectors and generalized eigenvectors, associated with all eigenvalues in the open left-half plane is the goal. The structured Kronecker form reveals this information but usually it cannot be computed numerically, because arbitrary small perturbations may change the structural information and since the transformation matrices may be unbounded.

A computationally attractive alternative is the staircase form under orthogonal transformations. It allows us to check regularity and to determine the index within the usual limitations of rank computations in finite precision arithmetic, see [43] for a detailed discussion of the difficulties. This is an essential preparation for the computation of the eigenvalues and deflating subspaces.

Theorem 5.2 ([43]) *For every even pencil $sN - M \in \mathbb{R}[s]^{n,n}$, there exists a real orthogonal matrix $U \in \mathbb{R}^{n,n}$ such that*

$$
U^T N U = \begin{array}{c|cccc|c|cc}
s_1 & N_{1,1} & \cdots & \cdots & N_{1,w} & N_{1,w+1} & N_{1,w+2} & \cdots\ N_{1,2w}\ 0 \\
\vdots & \vdots & \ddots & & \vdots & \vdots & \vdots & \iddots\ \iddots \\
\vdots & \vdots & & \ddots & \vdots & \vdots & N_{w-1,w+2} & \iddots \\
s_w & -N_{1,w}^T & \cdots & \cdots & N_{w,w} & N_{w,w+1} & 0 & \\
\hline
l & -N_{1,w+1}^T & \cdots & \cdots & -N_{w,w+1}^T & N_{w+1,w+1} & & \\
\hline
q_w & -N_{1,w+2}^T & \cdots & -N_{w-1,w+2}^T & 0 & & & \\
\vdots & \vdots & \iddots & \iddots & & & & \\
q_2 & -N_{1,2w}^T & \iddots & & & & & \\
q_1 & 0 & & & & & &
\end{array}
$$

$$
U^T M U = \begin{array}{c|cccc|c|c}
s_1 & M_{1,1} & \cdots & \cdots & M_{1,w} & M_{1,w+1} & M_{1,w+2}\ \cdots\ \cdots\ M_{1,2w+1} \\
\vdots & \vdots & \ddots & & \vdots & \vdots & \vdots\ \iddots \\
\vdots & \vdots & & \ddots & \vdots & \vdots & \vdots\ \iddots \\
s_w & M_{1,w}^T & \cdots & \cdots & M_{w,w} & M_{w,w+1} & M_{w,w+2} \\
\hline
l & M_{1,w+1}^T & \cdots & \cdots & M_{w,w+1}^T & M_{w+1,w+1} & \\
\hline
q_w & M_{1,w+2}^T & \cdots & \cdots & M_{w,w+2}^T & & \\
\vdots & \vdots & \iddots & & & & \\
\vdots & \vdots & \iddots & & & & \\
q_1 & M_{1,2w+1}^T & & & & &
\end{array}, \tag{5.1}
$$

where $q_1 \geq s_1 \geq q_2 \geq s_2 \geq \ldots \geq q_w \geq s_w$, $l = r_{w+1} + a_{w+1}$, *and for* $i = 1, \ldots, w$, *we have* $N_{i,i} = -N_{i,i}^T$, $M_{i,i} = M_{i,i}^T$. *Furthermore,*

$$
\begin{aligned}
&N_{j,2w+1-j} \in \mathbb{R}^{s_j, q_j+1}, \quad 1 \leq j \leq w-1,\\
&N_{w+1,w+1} = \begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix}, \quad \Delta = -\Delta^T \in \mathbb{R}^{r_{w+1}, r_{w+1}},\\
&M_{j,2w+2-j} = \begin{bmatrix} \Gamma_j & 0 \end{bmatrix} \in \mathbb{R}^{s_j, q_j}, \quad \Gamma_j \in \mathbb{R}^{s_j, s_j}, \quad 1 \leq j \leq w,\\
&M_{w+1,w+1} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \quad \Sigma_{11} \in \mathbb{R}^{r_{w+1}, r_{w+1}}, \quad \Sigma_{22} \in \mathbb{R}^{a_{w+1}, a_{w+1}},\\
&M_{w+1,w+1} = M_{w+1,w+1}^T,
\end{aligned}
$$

and the blocks Σ_{22} *and* Δ *and* Γ_j, $j = 1, \ldots, w$ *(if they occur) are nonsingular.*

Production code implementations for the computation of these and other related structured staircase forms via a sequence of singular value decompositions have been presented in [37]. Since the staircase form uses congruence transformations, all the invariants of the even Kronecker canonical form are preserved, as discussed in the following corollary.

Corollary 5.3 ([43]) *Consider an even pencil and its staircase form 5.1.*

(i) The pencil is regular if and only if $s_i = q_i$ *for* $i = 1, \ldots, w$.
(ii) The pencil is regular and of index at most one if and only if $w = 0$.
(iii) The block $(N_{w+1,w+1}, M_{w+1,w+1})$ *contains the regular part associated with finite eigenvalues and blocks associated with the infinite eigenvalues of index at most one.*
(iv) The finite eigenvalues of the pencil are the eigenvalues of

$$s\Delta - \left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right).$$

(v) For every purely imaginary eigenvalue $\lambda_0 \in \mathrm{i}\mathbb{R}$, *satisfying*

$$\left(\lambda_0\Delta - \left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)\right) x_0 = 0$$

for $x_0 \in \mathbb{C}^{r_{w+1}} \setminus \{0\}$, *the sign-characteristic of* λ_0 *is given by the sign of the real number* $\mathrm{i}x_0^H \Delta x_0$.

Thus, once the staircase form has been computed, for the computation of eigenvalues and invariant subspaces one can restrict the methods to the middle regular index one block of the staircase form. We recall the appropriate methods in the next subsection.

5.2 Computing Eigenvalues and Deflating Subspaces of Regular Index One Even Pencils

For the computation of eigenvalues, eigenvectors, and deflating subspaces associated with finite eigenvalues of even pencils, we need eigenvalue methods for regular even pencils of index at most one that can be applied to the middle block in the staircase form 5.1

$$sN_{w+1,w+1} - M_{w+1,w+1} = s \begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}. \tag{5.2}$$

In the special case that this even pencil has no infinite eigenvalues, i.e., if the second block row and column are not occurring, and hence $a_{w+1} = 0$, then we have a pencil $s\Delta - \Sigma_{11}$, where Δ is nonsingular (and thus of even dimension). In this case one can perform a Cholesky-like decomposition, see [13, 39] of the form $\Delta = \mathscr{U}^T \mathscr{J}_{r_{w+1}/2} \mathscr{U}$ with an upper-triangular matrix $\mathscr{U}$. If the factorization is well conditioned and if $\mathscr{U}$ is well conditioned with respect to inversion, then one can turn this even eigenvalue problem into an eigenvalue problem for the Hamiltonian matrix $\mathscr{H} = \mathscr{J}^T_{r_{w+1}/2} \mathscr{U}^{-T} \Sigma_{11} \mathscr{U}^{-1}$ and apply the structure-preserving methods for Hamiltonian eigenvalue problems [48, 94]. If, however, the computation and inversion of $\mathscr{U}$ is ill-conditioned or if the pencil $sN_{w+1,w+1} - M_{w+1,w+1}$ has infinite eigenvalues, then it is better to proceed with the pencil formulation.

Recently, in [93], a new structure-preserving method to deflate the infinite eigenvalues via an orthogonal congruence transformation has been derived for the pencil case. Consider the even pencil $sN_{w+1,w+1} - M_{w+1,w+1}$ as in (5.2). This procedure works by using a rank-revealing QR-decomposition or a singular value decomposition to determine an orthogonal matrix V_{w+1} such that

$$\begin{bmatrix} \Sigma_{21} & \Sigma_{22} \end{bmatrix} V_{w+1} = \begin{bmatrix} 0 & \hat{\Sigma}_{22} \end{bmatrix},$$

with nonsingular $\hat{\Sigma}_{22}$. By forming

$$V_{w+1}^T \left(sM_{w+1,w+1} - N_{w+1,w+1} \right) V_{w+1} = s \begin{bmatrix} \tilde{\Delta}_{11} & \tilde{\Delta}_{12} \\ -\tilde{\Delta}_{12}^T & \tilde{\Delta}_{22} \end{bmatrix} - \begin{bmatrix} \tilde{\Sigma}_{11} & \tilde{\Sigma}_{12} \\ \tilde{\Sigma}_{12}^T & \tilde{\Sigma}_{22} \end{bmatrix},$$

partitioned accordingly, it has been shown in [93] that the eigenvalues of the even pencil $s\tilde{\Delta}_{11} - \tilde{\Sigma}_{11}$ are exactly the finite eigenvalues of $sN_{w+1,w+1} - M_{w+1,w+1}$ and also the eigenvectors and invariant subspaces can be easily recovered.

The detailed error analysis of this procedure in [93] analyzes when this deflation procedure is reliable and when it is more reasonable to proceed with the index one pencil formulation. In the following we assume that this decision has been made, and that we either proceed with an even pencil with only finite eigenvalues, which means that the dimension is even or with an index one even pencil. Since for skew-Hamiltonian/Hamiltonian pencils eigenvalue methods are well established and have been professionally implemented [9, 14, 20, 21, 59, 85, 92, 95], we just adapt these for the even pencil case. However, we suggest that in the long run these methods should be implemented to directly work for the even case, since it may happen that the middle block $sN_{w+1,w+1} - M_{w+1,w+1}$ in the even staircase form (i.e., the regular index one part) is of odd dimension. To apply the methods for skew-Hamiltonian/Hamiltonian pencils to this middle block in the odd-dimensional case, we consider an embedded $2k \times 2k$ pencil

$$sS - H = \mathcal{J}_k \left(s \begin{bmatrix} N_{w+1,w+1} & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} M_{w+1,w+1} & 0 \\ 0 & 1 \end{bmatrix} \right)$$

which has an additional eigenvalue ∞, right eigenvector e_{2k} (the $2k$th unit vector) and left eigenvector $\mathcal{J}_k^T e_{2k}$, which are orthogonal to all the other eigenvectors. So in the following, whenever an eigenvalue method for regular even pencils of index at most one is needed, then we can perform this embedding and employ a solver for the skew-Hamiltonian/Hamiltonian pencil $sS - H \in \mathbb{R}[s]^{2k,2k}$.

For the computation of the eigenvalues and deflating subspaces of skew-Hamiltonian/Hamiltonian pencils we make use of $\mathcal{J}_k$-congruence transformations of the form

$$s\tilde{S} - \tilde{H} := \mathcal{J}_k \mathcal{Q}^T \mathcal{J}_k^T (sS - H) \mathcal{Q}$$

with nonsingular matrices $\mathcal{Q}$, which preserve the skew-Hamiltonian/Hamiltonian structure. In general we would hope that we can compute an *orthogonal* matrix $\mathcal{Q}$ such that

$$\mathcal{J}_k \mathcal{Q}^T \mathcal{J}_k^T (sS - H) \mathcal{Q} = s \begin{bmatrix} S_{11} & S_{12} \\ 0 & S_{11}^T \end{bmatrix} - \begin{bmatrix} H_{11} & H_{12} \\ 0 & -H_{11}^T \end{bmatrix}$$

is in skew-Hamiltonian/Hamiltonian Schur form, i.e., the subpencil $sS_{11} - H_{11}$ is in generalized Schur form [62]. Unfortunately, not every skew-Hamiltonian/Hamiltonian pencil has this structured Schur form, since certain simple purely imaginary eigenvalues, or multiple purely imaginary eigenvalues with even algebraic multiplicity, but uniform sign-characteristic, cannot be represented in this structure. An embedding into a pencil of the double size solves this issue as follows.

We introduce the orthogonal matrices

$$\mathscr{Y} = \frac{\sqrt{2}}{2}\begin{bmatrix} I_{2k} & I_{2k} \\ -I_{2k} & I_{2k} \end{bmatrix}, \quad \mathscr{P} = \begin{bmatrix} I_k & 0 & 0 & 0 \\ 0 & 0 & I_k & 0 \\ 0 & I_k & 0 & 0 \\ 0 & 0 & 0 & I_k \end{bmatrix}, \quad \mathscr{X} = \mathscr{Y}\mathscr{P},$$

and define the matrix pencil

$$s\mathscr{B}_S - \mathscr{B}_H := \mathscr{X}^T \left(s\begin{bmatrix} S & 0 \\ 0 & S \end{bmatrix} - \begin{bmatrix} H & 0 \\ 0 & -H \end{bmatrix} \right) \mathscr{X} \in \mathbb{R}[s]^{4k,4k},$$

which is still regular and of index at most one.

It can be easily observed that $s\mathscr{B}_S - \mathscr{B}_H$ is again real skew-Hamiltonian/Hamiltonian with the same eigenvalues (now with double algebraic, geometric, and partial multiplicities, but with appropriate mixed sign-characteristic) as the pencil $sS-H$. To compute the eigenvalues of $s\mathscr{B}_S-\mathscr{B}_H$ one uses the generalized symplectic URV decomposition of $sS - H$, see [10, 11], i.e., there exist orthogonal matrices $Q_1,\ Q_2 \in \mathbb{R}^{4k,4k}$ such that

$$\begin{aligned} Q_1^T S \mathscr{J}_k Q_1 \mathscr{J}_k^T &= \begin{bmatrix} S_{11} & S_{12} \\ 0 & S_{11}^T \end{bmatrix}, \\ \mathscr{J}_k Q_2^T \mathscr{J}_k^T S Q_2 &= \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{11}^T \end{bmatrix}, \\ Q_1^T H Q_2 &= \begin{bmatrix} H_{11} & H_{12} \\ 0 & H_{22} \end{bmatrix}, \end{aligned} \tag{5.3}$$

where S_{12} and T_{12} are skew-symmetric and the formal matrix product $S_{11}^{-1}H_{11}T_{11}^{-1}H_{22}^T$ is in periodic Schur form [31, 67, 71].

Applying this result to the specially structured pencil $s\mathscr{B}_S-\mathscr{B}_H$, we can compute an orthogonal matrix $\mathscr{Q}$ such that

$$\mathscr{J}_{2k}\mathscr{Q}^T \mathscr{J}_{2k}^T (s\mathscr{B}_S - \mathscr{B}_H)\mathscr{Q} = s\left[\begin{array}{cc|cc} S_{11} & 0 & S_{12} & 0 \\ 0 & T_{11} & 0 & T_{12} \\ \hline 0 & 0 & S_{11}^T & 0 \\ 0 & 0 & 0 & T_{11}^T \end{array}\right] - \left[\begin{array}{cc|cc} 0 & H_{11} & 0 & H_{12} \\ -H_{22}^T & 0 & H_{12}^T & 0 \\ \hline 0 & 0 & 0 & H_{22} \\ 0 & 0 & -H_{11}^T & 0 \end{array}\right]$$

with $\mathscr{Q} = \mathscr{P}^T \begin{bmatrix} \mathscr{J}_k Q_1 \mathscr{J}_k^T & 0 \\ 0 & Q_2 \end{bmatrix} \mathscr{P}$.

Note, that for these computations we never explicitly construct the embedded pencils. It is sufficient to compute the necessary parts of the matrices in (5.3).

The eigenvalues of $sS - H$ can then be computed as $\pm i\sqrt{\lambda_j}$ where the λ_j, $j = 1, \ldots, k$, are the eigenvalues of the *formal matrix product* $S_{11}^{-1} H_{11} T_{11}^{-1} H_{22}^T$ which can be determined by evaluating the entries on the 1×1 and 2×2 diagonal blocks of the matrices only. In particular, the finite, purely imaginary eigenvalues correspond to the 1×1 diagonal blocks of this matrix product. Provided that the pairwise distance of the simple, finite, purely imaginary eigenvalues with mixed sign-characteristics is sufficiently large, they can be computed in a robust way without any error in the real part. This property of the algorithm plays an essential role for many of the applications that we will consider in subsequent sections.

If also the deflating subspaces associated with certain eigenvalues are desired, then one computes the real skew-Hamiltonian/Hamiltonian Schur form of the embedded pencil where the eigenvalues are reordered in such a way such that the desired ones appear in the leading principal subpencil. By determining also the sign-characteristics of the purely imaginary eigenvalues, one can (at least in exact arithmetics) check whether a Hamiltonian Schur form exists. It should be noted that if the problem has computed eigenvalues very close to the imaginary axis (within a strip of width $\sqrt{\mathbf{u}}$), then these may be the result of a perturbation of size $\mathbf{u}$ of a double purely imaginary eigenvalue with mixed sign-characteristic. This does not prevent the existence of a Hamiltonian Schur form, however, in the neighborhood of this problem there is then a problem with two simple purely imaginary eigenvalues of mixed sign-characteristic, but with no Hamiltonian Schur form, see [1].

The structure-preserving Algorithm 1 was introduced in [12] and has been updated and improved in [85]. It is available as the SLICOT subroutine `MB04BD`. While the classic unstructured QZ algorithm applied to the $2k \times 2k$ pencil would require $528k^3$ flops or $240k^3$ flops for the eigenvalues [62], this algorithm needs roughly 60 % of this, see [12]. Note that there are many more structure-exploiting algorithms for Hamiltonian and even eigenvalues problems in the dense but also in the sparse setting, see, e.g., [9, 17, 48, 72, 84, 94, 95, 107].

In later sections, when discussing applications for even pencils, we will always use the algorithm presented here, since the preservation of the spectral symmetry is essential for the robustness of the methods. For illustration, Fig. 2 from [21] plots the computed eigenvalues of a skew-Hamiltonian/Hamiltonian pencil that results from the stability analysis of a linearized gyroscopic system. A necessary condition for stability is that all eigenvalues are on the imaginary axis. The figure shows that the structure-preserving algorithm captures this behavior whereas the standard QZ algorithm fails to do so and therefore does not allow us to make any statement about stability.

Algorithm 1 Computation of stable eigenvalues and associated stable deflating subspaces of a real skew-Hamiltonian/Hamiltonian pencil

Input: A regular real skew-Hamiltonian/Hamiltonian pencil $sS - H \in \mathbb{R}[s]^{2k,2k}$ of index at most one.

Output: The eigenvalues of $sS - H$ and a matrix P_V^- whose columns form an orthogonal basis of the r-dimensional deflating subspace associated with the eigenvalues in the open left half plane.

1: Compute the generalized symplectic URV decomposition [85, Algorithm 2] of the pencil $sS - H$ and determine orthogonal matrices Q_1, Q_2 such that

$$Q_1^T S \mathcal{J}_k Q_1 \mathcal{J}_k^T = \begin{bmatrix} S_{11} & S_{12} \\ 0 & S_{11}^T \end{bmatrix},$$

$$\mathcal{J}_k Q_2^T \mathcal{J}_k^T S Q_2 = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{11}^T \end{bmatrix},$$

$$Q_1^T H Q_2 = \begin{bmatrix} H_{11} & H_{12} \\ 0 & H_{22} \end{bmatrix},$$

where the generalized matrix product $S_{11}^{-1} H_{11} T_{11}^{-1} H_{22}^T$ is in periodic Schur form.

2: Apply [85, Algorithm 3] to determine orthogonal matrices Q_3 and Q_4 such that

$$s\mathcal{S}_{11} - \mathcal{H}_{11} := Q_4^T \left(s \begin{bmatrix} S_{11} & 0 \\ 0 & T_{11} \end{bmatrix} - \begin{bmatrix} 0 & H_{11} \\ -H_{22}^T & 0 \end{bmatrix} \right) Q_3$$

is in generalized Schur form. Update

$$\mathcal{S}_{12} := Q_4^T \begin{bmatrix} S_{12} & 0 \\ 0 & T_{12} \end{bmatrix} Q_4, \quad \mathcal{H}_{12} := Q_4^T \begin{bmatrix} 0 & H_{12} \\ H_{12}^T & 0 \end{bmatrix} Q_4$$

and set

$$s\mathcal{B}_S - \mathcal{B}_H := s \begin{bmatrix} \mathcal{S}_{11} & \mathcal{S}_{12} \\ 0 & \mathcal{S}_{11}^T \end{bmatrix} - \begin{bmatrix} \mathcal{H}_{11} & \mathcal{H}_{12} \\ 0 & -\mathcal{H}_{11}^T \end{bmatrix}.$$

3: Apply the eigenvalue reordering method [85, Algorithm 4] to the pencil $s\mathcal{B}_S - \mathcal{B}_H$ to determine an orthogonal matrix $\hat{Q}$ such that

$$s\tilde{\mathcal{B}}_S - \tilde{\mathcal{B}}_H := \mathcal{J}_{2k} \hat{Q}^T \mathcal{J}_{2k}^T (s\mathcal{B}_S - \mathcal{B}_H) \hat{Q}$$

is still in structured Schur form, but the eigenvalues with negative real part of $s\tilde{\mathcal{B}}_S - \tilde{\mathcal{B}}_H$ are contained in the leading $2r \times 2r$ principal subpencil of $s\mathcal{S}_{11} - \mathcal{H}_{11}$.

4: Set

$$V = \begin{bmatrix} I_{2k} & 0 \end{bmatrix} \left(\mathcal{Y} \begin{bmatrix} \mathcal{J}_k Q_1 \mathcal{J}_k^T & 0 \\ 0 & Q_2 \end{bmatrix} \mathcal{P} \begin{bmatrix} Q_3 & 0 \\ 0 & Q_4 \end{bmatrix} \hat{Q} \right) \begin{bmatrix} I_{2r} \\ 0 \end{bmatrix}$$

and compute P_V^-, an orthonormal basis of range V, using any numerically stable orthogonalization scheme.

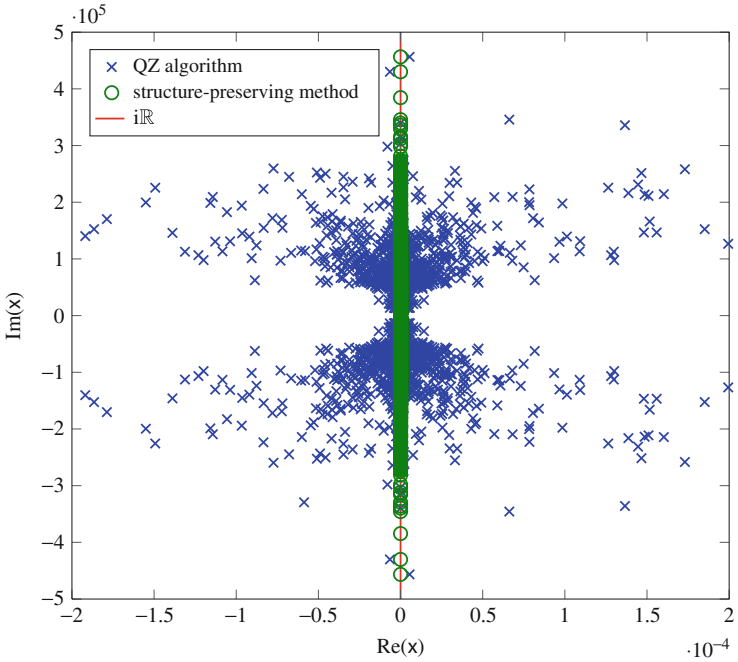

Fig. 2 Computed eigenvalues from a skew-Hamiltonian/Hamiltonian pencil with only purely imaginary eigenvalues resulting from a linearized gyroscopic system

6 Linear-Quadratic Optimal Control

In this section we consider the linear quadratic optimal control problem of minimizing

$$\mathscr{J}(x(\cdot),u(\cdot)) = \frac{1}{2}\int_0^\infty \left(x(t)^T Q x(t) + 2x(t)^T S u(t) + u(t)^T R u(t)\right) \mathrm{d}t \tag{6.1}$$

with $Q = Q^T \in \mathbb{R}^{n,n}$, $S \in \mathbb{R}^{n,m}$, and $R = R^T \in \mathbb{R}^{m,m}$ subject to the linear descriptor system of the form (1.2a) with initial value $x(0) = x_0$ and the stabilization condition $\lim_{t\to\infty} x(t) = 0$. If an output equation (1.2b) is also given, then the cost functional is usually given as $\tilde{\mathscr{J}}(y(\cdot),u(\cdot))$ which can then easily be transformed to the form given in (6.1) by inserting the output equation. This yields

$$\tilde{\mathscr{J}}(x(\cdot),u(\cdot)) = \frac{1}{2}\int_0^\infty \left(x(t)^T \tilde{Q} x(t) + 2x(t)^T \tilde{S} u(t) + u(t)^T \tilde{R} u(t)\right) \mathrm{d}t$$

with

$$\tilde{Q} := C^T Q C, \quad \tilde{S} := C^T Q D + C^T S, \quad \tilde{R} := D^T Q D + D^T S + S^T D + R. \tag{6.2}$$

Optimal control problems for equations of this form arise in mechanical multibody systems [64, 65, 108], electrical circuits [63], and many other applications like the linearization of general nonlinear systems along stationary trajectories [46].

In order for an optimal control to exist, for the initial value one needs the condition $E^+ E x_0 = x_0$ with the Moore-Penrose inverse E^+ of E, see [76]. Further additional assumptions are needed to ensure that (6.1) is bounded from below. A quite common assumption in the literature is

$$\begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \geq 0.$$

In [117] it has been further shown that for square systems, strong stabilizability and the feasibility of a certain linear matrix inequality are sufficient conditions for the boundedness of (6.1) from below, even in the case of an indefinite cost functional.

To solve this problem in the most general situation, we replace the DAE constraint by the *strangeness-free formulation*

$$\hat{E}\dot{x}(t) = \hat{A}x(t) + \hat{B}u(t), \tag{6.3}$$

where

$$\hat{E} = \begin{bmatrix} \hat{E}_1 \\ 0 \end{bmatrix}, \quad \hat{A} = \begin{bmatrix} \hat{A}_1 \\ \hat{A}_2 \end{bmatrix}, \quad \hat{B} = \begin{bmatrix} \hat{B}_1 \\ \hat{B}_2 \end{bmatrix},$$

with the additional property that the matrix

$$\begin{bmatrix} \hat{E}_1 & 0 \\ \hat{A}_2 & \hat{B}_2 \end{bmatrix}$$

has full row rank, see also Sect. 3. The necessary optimality system is then given by

$$\begin{bmatrix} 0 & \hat{E} & 0 \\ -\hat{E}^T & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \frac{\mathrm{d}}{\mathrm{d}t} \begin{bmatrix} \hat{\lambda}(t) \\ x(t) \\ u(t) \end{bmatrix} = \begin{bmatrix} 0 & \hat{A} & \hat{B} \\ \hat{A}^T & Q & S \\ \hat{B}^T & S^T & R \end{bmatrix} \begin{bmatrix} \hat{\lambda}(t) \\ x(t) \\ u(t) \end{bmatrix}, \tag{6.4}$$

with boundary conditions $x(0) = x_0$, and $\lim_{t\to\infty} \hat{E}^T \hat{\lambda}(t) = 0$. Solving this system will give the optimal input $u(\cdot)$, state $x(\cdot)$, and the Lagrange multiplier $\hat{\lambda}(\cdot)$.

Instead of first computing a strangeness-free formulation and forming the optimality system (6.4), we can instead directly form and solve the *formal optimality system* [7, 47, 76, 77, 81] given by

$$\begin{bmatrix} 0 & E & 0 \\ -E^T & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \frac{\mathrm{d}}{\mathrm{d}t} \begin{bmatrix} \lambda(t) \\ x(t) \\ u(t) \end{bmatrix} = \begin{bmatrix} 0 & A & B \\ A^T & Q & S \\ B^T & S^T & R \end{bmatrix} \begin{bmatrix} \lambda(t) \\ x(t) \\ u(t) \end{bmatrix}, \tag{6.5}$$

with boundary conditions $x(0) = x_0$, and $\lim_{t\to\infty} E^T\lambda(t) = 0$. One has the following relation between the true and the formal optimality system which we cite here for constant coefficient systems, for the general case of variable coefficient systems see [77].

Theorem 6.1 *Suppose that the formal necessary optimality system* (6.5) *has a solution* $\left[\lambda(\cdot)^T\ x(\cdot)^T\ u(\cdot)^T\right]^T$. *Then there exists a function* $\hat{\lambda}(\cdot)$ *replacing the function* $\lambda(\cdot)$ *such that* $\left[\hat{\lambda}(\cdot)^T\ x(\cdot)^T\ u(\cdot)^T\right]^T$ *solves the necessary optimality conditions* (6.4).

Theorem 6.1 shows that it is enough to solve the boundary value problem (6.5) in the original data, provided it is solvable. Since this is a homogeneous differential-algebraic system, the solvability of the boundary value problem depends on the consistency of the boundary conditions and the solvability of the linear system that relates initial and terminal conditions, see [5, 79, 80]. Since the boundary value problem is of the form

$$N\dot{z}(t) = Mz(t), \quad P_1z(0) = P_1z_0, \quad \lim_{t\to\infty} P_2z(t) = 0,$$

with $z(\cdot) = \left[\lambda(\cdot)^T\ x(\cdot)^T\ u(\cdot)^T\right]^T$, and some matrices P_1, and P_2, the simplest way to perform these computations is to apply the congruence transformation to even staircase form

$$U^TNU\dot{\tilde{z}}(t) = U^TMU\tilde{z}(t), \quad P_1U\tilde{z}(0) = P_1U\tilde{z}_0, \quad \lim_{t\to\infty} P_2U\tilde{z}(t) = 0,$$

with $\tilde{z}(\cdot) = U^Tz(\cdot)$, and $\tilde{z}_0 = U^Tz_0$.

This allows us to check the unique solvability by checking the regularity as in Corollary 5.3 and the consistency of the boundary conditions, see [43] for details. By partitioning $\tilde{z}(\cdot) = \left[\tilde{z}_1(\cdot)^T, \ldots, \tilde{z}_{2w+1}(\cdot)^T\right]^T$ analogous to (5.1), the last w blocks yield the consistency conditions $\tilde{z}_1 \equiv 0, \ \ldots, \ \tilde{z}_w \equiv 0$. The middle block system can be expressed as

$$N_{w+1,w+1}\dot{\tilde{z}}_{w+1}(t) = M_{w+1,w+1}\tilde{z}_{w+1}(t), \tag{6.6}$$

with appropriately transformed boundary conditions. This system is regular and has index at most one. If we make use of the semi-explicit form (5.2) and split

$$\tilde{z}_{w+1}(\cdot) = \begin{bmatrix} \eta(\cdot) \\ \zeta(\cdot) \end{bmatrix},$$

then we obtain

$$\begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{\eta}(t) \\ \dot{\zeta}(t) \end{bmatrix} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} \eta(t) \\ \zeta(t) \end{bmatrix}.$$

It follows that $\zeta(\cdot) = -\Sigma_{22}^{-1}\Sigma_{21}\eta(\cdot)$, which gives further consistency conditions on $\tilde{z}_{w+1}(\cdot)$ and

$$\Delta\dot{\eta}(t) = (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})\eta(t).$$

Then we can perform a factorization $\Delta = \mathscr{U}^T \mathscr{J}_{r_{w+1}/2}\mathscr{U}$ with nonsingular upper triangular matrix $\mathscr{U}$ [39]. If the factorization is well conditioned and the factor $\mathscr{U}$ is well conditioned with respect to inversion, then we set $\mathscr{H} := \mathscr{J}^T_{r_{w+1}/2}\mathscr{U}^{-T}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})\mathscr{U}^{-1}$ and $\xi(t) := \mathscr{U}\eta(t)$ to obtain the Hamiltonian boundary value problem

$$\dot{\xi}(t) = \mathscr{H}\xi(t). \tag{6.7}$$

With appropriate boundary conditions $\Pi_1\xi(0) = \Pi_1\xi_0$, and $\lim_{t\to\infty}\Pi_2\xi(t) = 0$. This system has the general solution $\xi(t) = \exp(\mathscr{H}t)\,\xi_0$ and therefore,

$$\tilde{z}_{w+1}(t) = \begin{bmatrix} \eta(t) \\ -\Sigma_{22}^{-1}\Sigma_{21}\eta(t) \end{bmatrix} = \begin{bmatrix} \mathscr{U}^{-1}\exp(\mathscr{H}t)\,\xi_0 \\ -\Sigma_{22}^{-1}\Sigma_{21}\mathscr{U}^{-1}\exp(\mathscr{H}t)\,\xi_0 \end{bmatrix}. \tag{6.8}$$

It is important to note that one does not have to compute the exponential function in (6.8) explicitly.

With a decomposition of the boundary value problem as

$$\begin{bmatrix} \dot{\xi}_1(t) \\ \dot{\xi}_2(t) \end{bmatrix} = \begin{bmatrix} \mathscr{H}_{11} & \mathscr{H}_{12} \\ \mathscr{H}_{21} & -\mathscr{H}_{11}^T \end{bmatrix} \begin{bmatrix} \xi_1(t) \\ \xi_2(t) \end{bmatrix}, \quad \mathscr{H}_{12} = \mathscr{H}_{12}^T, \quad \mathscr{H}_{21} = \mathscr{H}_{21}^T,$$

one rather computes the stable invariant subspace spanned by the matrix $\mathscr{V} = \begin{bmatrix} \mathscr{V}_1 \\ \mathscr{V}_2 \end{bmatrix} \in \mathbb{R}^{r_{w+1}, r_{w+1}/2}$ which satisfies

$$\begin{bmatrix} \mathscr{H}_{11} & \mathscr{H}_{12} \\ \mathscr{H}_{21} & -\mathscr{H}_{11}^T \end{bmatrix} \begin{bmatrix} \mathscr{V}_1 \\ \mathscr{V}_2 \end{bmatrix} = \begin{bmatrix} \mathscr{V}_1 \\ \mathscr{V}_2 \end{bmatrix} \tilde{\mathscr{H}}$$

with $\Lambda(\tilde{\mathscr{H}}) \subset \mathbb{C}^-$. The appropriate structure-preserving methods for this computation are outlined in [10] and implemented as the routine `MB03ZD` in SLICOT.

By defining $Y := -\mathscr{V}_2\mathscr{V}_1^{-1}$, $\tilde{\xi}_1(\cdot) := \xi_1(\cdot)$, and $\tilde{\xi}_2(\cdot) := Y\xi_1(\cdot) + \xi_2(\cdot)$ we get

$$\begin{aligned}\begin{bmatrix}\dot{\tilde{\xi}}_1(t)\\ \dot{\tilde{\xi}}_2(t)\end{bmatrix} &= \begin{bmatrix} I_{r_{w+1}/2} & 0\\ Y & I_{r_{w+1}/2}\end{bmatrix}\begin{bmatrix}\mathscr{H}_{11} & \mathscr{H}_{12}\\ \mathscr{H}_{21} & -\mathscr{H}_{11}^T\end{bmatrix}\begin{bmatrix} I_{r_{w+1}/2} & 0\\ -Y & I_{r_{w+1}/2}\end{bmatrix}\begin{bmatrix}\tilde{\xi}_1(t)\\ \tilde{\xi}_2(t)\end{bmatrix}\\ &= \begin{bmatrix}\tilde{\mathscr{H}}_{11} & \tilde{\mathscr{H}}_{12}\\ 0 & -\tilde{\mathscr{H}}_{11}^T\end{bmatrix}\begin{bmatrix}\tilde{\xi}_1(t)\\ \tilde{\xi}_2(t)\end{bmatrix},\end{aligned} \tag{6.9}$$

with $\Lambda(\tilde{\mathscr{H}}_{11}) \subset \mathbb{C}^-$ and the boundary conditions $\tilde{\xi}_1(0) = \tilde{\xi}_{1,0}$, $\lim_{t\to\infty}\tilde{\xi}_2(t) = 0$. Since $-\tilde{\mathscr{H}}_{11}^T$ is an unstable matrix, we obtain $\tilde{\xi}_2(\cdot) \equiv 0$ by backwards integration. This results in

$$\dot{\tilde{\xi}}_1(t) = \tilde{\mathscr{H}}_{11}\tilde{\xi}_1(t), \quad \tilde{\xi}_1(0) = \tilde{\xi}_{1,0},$$

which can be efficiently solved by a further transformation of $\tilde{\mathscr{H}}_{11}$ to Schur form. From that we can easily reconstruct $\tilde{z}_{w+1}(\cdot)$ as in (6.8).

This can be further used to determine $\tilde{z}_{w+2}(\cdot), \ldots, \tilde{z}_{2w+1}(\cdot)$ in terms of $\tilde{z}_{w+1}(\cdot)$, and the consistency conditions $\tilde{z}_1 \equiv 0, \ldots, \tilde{z}_w \equiv 0$ via a backward substitution process applied to the first w block rows of 5.1. This recursive process leads to

$$M_{w-j+1,w+j+1}\tilde{z}_{w+j+1}(t) = \left(\sum_{i=w+1}^{w+j} N_{w-j+1,i}\dot{\tilde{z}}_i(t) - \sum_{i=w+1}^{w+j} M_{w-j+1,i}\tilde{z}_i(t)\right), \tag{6.10}$$

which requires w differentiations to be carried out, see [43]. The complete procedure is graphically displayed in Fig. 3. Note that if $sN - M$ is regular, then $M_{w+j+1} = \Gamma_{w-j+1}$ is nonsingular and $\tilde{z}(\cdot)$ is uniquely determined. Otherwise, some of the variables remain undetermined and thus the optimal solution trajectory might not be unique.

We illustrate the whole procedure using the following example.

Example 6.1 Consider the linear-quadratic optimal control problem of minimizing

$$\mathscr{J}(x(\cdot), u(\cdot)) = \int_0^\infty \left(x_1(t)^2 + x_2(t)^2 + u(t)^2\right) \mathrm{d}t$$

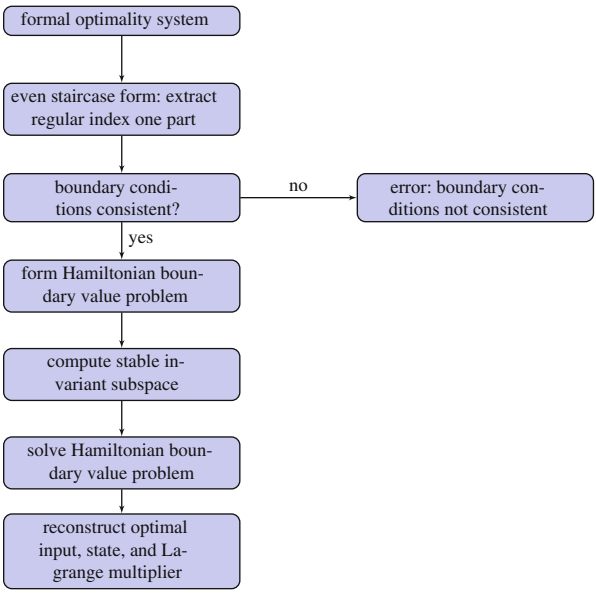

Fig. 3 Algorithm flowchart for solving linear-quadratic optimal control problems

subject to

$$\begin{bmatrix} 1&0&0&0&0\\ 0&1&0&0&0\\ 0&0&0&1&0\\ 0&0&0&0&0\\ 0&0&0&0&0 \end{bmatrix} \begin{bmatrix} \dot{x}_1(t)\\ \dot{x}_2(t)\\ \dot{x}_3(t)\\ \dot{x}_4(t)\\ \dot{x}_5(t) \end{bmatrix} = \begin{bmatrix} 1&1&0&0&0\\ -1&1&0&0&0\\ 0&0&1&0&0\\ 0&0&0&1&0\\ 0&0&0&0&0 \end{bmatrix} \begin{bmatrix} x_1(t)\\ x_2(t)\\ x_3(t)\\ x_4(t)\\ x_5(t) \end{bmatrix} + \begin{bmatrix} 1\\ 1\\ -1\\ 0\\ 0 \end{bmatrix} u(t),$$

$$\begin{bmatrix} x_{1,0}(t)\\ x_{2,0}(t)\\ x_{3,0}(t)\\ x_{4,0}(t)\\ x_{5,0}(t) \end{bmatrix} = \begin{bmatrix} 1\\ -1\\ 0\\ 0\\ 0 \end{bmatrix}.$$

with

$$Q = \begin{bmatrix} 1&0&0&0&0\\ 0&1&0&0&0\\ 0&0&0&0&0\\ 0&0&0&0&0\\ 0&0&0&0&0 \end{bmatrix}, \quad S = \begin{bmatrix} 0\\ 0\\ 0\\ 0\\ 0 \end{bmatrix}, \quad R = 1.$$

The formal optimality system is given by (6.5) with the boundary condition for the Lagrange multiplier given by $\lim_{t\to\infty} \lambda_1(t) = \lim_{t\to\infty} \lambda_2(t) = \lim_{t\to\infty} \lambda_3(t) = 0$. With the notation of Theorem 5.2, a reduction to even staircase form yields the structural characteristics

$$w = 1, \quad s_1 = 2, \quad q_1 = 4.$$

In particular, since $q_1 - s_1 = 2 \neq 0$, the formal optimality system is singular. Thus the transformed formal optimality system attains the form

$$\begin{bmatrix} N_{1,1} & N_{1,2} & 0 \\ -N_{1,2}^T & N_{2,2} & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{\tilde{z}}_1(t) \\ \dot{\tilde{z}}_2(t) \\ \dot{\tilde{z}}_3(t) \end{bmatrix} = \begin{bmatrix} M_{1,1} & M_{1,2} & M_{1,3} \\ M_{1,2}^T & M_{2,2} & 0 \\ M_{1,3}^T & 0 & 0 \end{bmatrix} \begin{bmatrix} \tilde{z}_1(t) \\ \tilde{z}_2(t) \\ \tilde{z}_3(t) \end{bmatrix}. \tag{6.11}$$

The regular index one part is given by

$$\begin{bmatrix} 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{z}_{2,1}(t) \\ \dot{z}_{2,2}(t) \\ \dot{z}_{2,3}(t) \\ \dot{z}_{2,4}(t) \\ \dot{z}_{2,5}(t) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & -1 & 0 \\ 1 & 0 & 1 & 0 & -1 \\ 0 & 1 & 1 & 1 & 0 \\ -1 & 0 & 1 & 0 & -1 \\ 0 & -1 & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} z_{2,1}(t) \\ z_{2,2}(t) \\ z_{2,3}(t) \\ z_{2,4}(t) \\ z_{2,5}(t) \end{bmatrix},$$

$$\begin{bmatrix} z_{2,1}(t) \\ z_{2,2}(t) \\ z_{2,3}(t) \\ z_{2,4}(t) \\ z_{2,5}(t) \end{bmatrix} = \begin{bmatrix} -x_1(t) \\ -\lambda_1(t) \\ -x_2(t) \\ -\lambda_2(t) \\ u(t) \end{bmatrix}.$$

A further reduction to a Hamiltonian boundary value problem yields

$$\begin{bmatrix} \dot{\xi}_{1,1}(t) \\ \dot{\xi}_{1,2}(t) \\ \dot{\xi}_{2,1}(t) \\ \dot{\xi}_{2,2}(t) \end{bmatrix} = \begin{bmatrix} 1 & 1 & -1 & -1 \\ -1 & 1 & -1 & -1 \\ -1 & 0 & -1 & 1 \\ 0 & -1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \xi_{1,1}(t) \\ \xi_{1,2}(t) \\ \xi_{2,1}(t) \\ \xi_{2,2}(t) \end{bmatrix},$$

$$\begin{bmatrix} \xi_{1,1}(0) \\ \xi_{1,2}(0) \end{bmatrix} = \begin{bmatrix} -x_1(0) \\ -x_2(0) \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \quad \lim_{t\to\infty} \begin{bmatrix} \xi_{2,1}(t) \\ \xi_{2,2}(t) \end{bmatrix} = \lim_{t\to\infty} \begin{bmatrix} -\lambda_1(t) \\ -\lambda_2(t) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

The stable invariant subspace of the Hamiltonian matrix is spanned by

$$\mathscr{V} = \begin{bmatrix} -0.2041 & -0.1727 \\ -0.2317 & -0.4438 \\ -0.9511 & 0.1548 \\ -0.0106 & -0.8656 \end{bmatrix},$$

and thus the Hamiltonian boundary value problem (6.9) reduces to

$$\begin{bmatrix} \dot{\tilde{\xi}}_{1,1}(t) \\ \dot{\tilde{\xi}}_{1,2}(t) \\ \dot{\tilde{\xi}}_{2,1}(t) \\ \dot{\tilde{\xi}}_{2,2}(t) \end{bmatrix} = \begin{bmatrix} -4.1813 & 1.4142 & -1.0000 & -1.0000 \\ -6.1813 & 1.4142 & -1.0000 & -1.0000 \\ 0 & 0 & 4.1813 & 6.1813 \\ 0 & 0 & -1.4142 & -1.4142 \end{bmatrix} \begin{bmatrix} \tilde{\xi}_{1,1}(t) \\ \tilde{\xi}_{1,2}(t) \\ \tilde{\xi}_{2,1}(t) \\ \tilde{\xi}_{2,2}(t) \end{bmatrix},$$

$$\begin{bmatrix} \tilde{\xi}_{1,1}(0) \\ \tilde{\xi}_{1,2}(0) \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \quad \lim_{t\to\infty} \begin{bmatrix} \tilde{\xi}_{2,1}(t) \\ \tilde{\xi}_{2,2}(t) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

which can now be solved by a numerical integrator. One first integrates the last two equations backward in time and stores the the trajectories either in discrete time points or in a collocation representation. Then the first two equations are integrated forward in time, making use of the already computed variables which then act as inhomogeneities. If the integrator needs this inhomogeneity in points different from the stored points, either these have to be recomputed or obtained by interpolation. It remains to determine $\tilde{z}_1(\cdot)$ and $\tilde{z}_3(\cdot)$ in (6.11). For this we have

$$\tilde{z}_1(t) = \begin{bmatrix} \lambda_3(t) \\ -x_4(t) \end{bmatrix} = 0,$$

which is consistent to the boundary conditions. For our example, (6.10) further yields

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} z_{3,1}(t) \\ z_{3,2}(t) \\ z_{3,3}(t) \\ z_{3,4}(t) \end{bmatrix} = - \begin{bmatrix} 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} z_{2,1}(t) \\ z_{2,2}(t) \\ z_{2,3}(t) \\ z_{2,4}(t) \\ z_{2,5}(t) \end{bmatrix}, \tag{6.12}$$

i.e., we have $x_3(t) = z_{3,1}(t) = z_{2,5}(t) = u(t)$, and $-\lambda_2(t) = z_{3,2}(t) = 0$ which are both in agreement to our state equation and the boundary conditions. Note that $z_{3,3}(t) = \lambda_5(t)$ and $z_{3,4}(t) = x_5(t)$ remain undetermined and thus the optimal solution is not unique. This observation is also in agreement to the state equation, since there $x_5(\cdot)$ is already free. To deal with this situation one either removes the last

equation and the variable $x_5(\cdot)$ already in the original state equation or one chooses another cost functional, see [73] for a discussion.

Remark 6.1 A similar decoupling procedure can also be constructed in the finite-time horizon problem by decoupling the forward and backward integration via the solution of a Riccati differential equation or by using other boundary value methods [5].

Remark 6.2 In [99, 117] the linear-quadratic optimal control problem has been solved by directly using the deflating subspaces of an even matrix pencil that is related to the boundary value problem (6.5), in particular in the context of singular control problems.

7 $\mathscr{H}_\infty$ Optimal Control

Our second application is the $\mathscr{H}_\infty$ optimal control problem which is one of the major tasks in robust control. We consider descriptor systems of the form

$$\mathbf{P}: \begin{cases} E\dot{x}(t) = Ax(t) + B_1 v(t) + B_2 u(t), \quad x(0) = x_0, \\ z(t) = C_1 x(t) + D_{11} v(t) + D_{12} u(t), \\ y(t) = C_2 x(t) + D_{21} v(t) + D_{22} u(t), \end{cases} \tag{7.1}$$

where $E, A \in \mathbb{R}^{n,n}$, $B_i \in \mathbb{R}^{n,m_i}$, $C_i \in \mathbb{R}^{p_i,n}$, and $D_{ij} \in \mathbb{R}^{p_i,m_j}$ for $i,j = 1,2$. In this system, $x : [0,\infty) \to \mathbb{R}^n$ is the state, $u : [0,\infty) \to \mathbb{R}^{m_2}$ is the control input, and $v : [0,\infty) \to \mathbb{R}^{m_1}$ is an exogenous input that may include noise, linearization errors, and unmodeled dynamics. The function $y : [0,\infty) \to \mathbb{R}^{p_2}$ contains measured outputs, while $z : [0,\infty) \to \mathbb{R}^{p_1}$ is a regulated output or an estimation error.

The $\mathscr{H}_\infty$ optimal control problem is typically formulated in the frequency domain. Its goal is to stabilize the system, while minimizing the $\mathscr{H}_\infty$-norm of the closed-loop transfer function $T_{zv}(\cdot)$ mapping noise or disturbance to error signals [122]. The value of $\|T_{zv}\|_{\mathscr{H}_\infty}$ is used as a measure for the worst-case influence of the disturbances v on the output z. A more rigorous formulation is given in the following definition [86].

Definition 7.1 (The Optimal $\mathscr{H}_\infty$ Control Problem) For the descriptor system (7.1), determine a controller (dynamic compensator)

$$\mathbf{K}: \begin{cases} \hat{E}\dot{\hat{x}}(t) = \hat{A}\hat{x}(t) + \hat{B}y(t), \\ u(t) = \hat{C}\hat{x}(t) + \hat{D}y(t), \end{cases} \tag{7.2}$$

with $\hat{E}, \hat{A} \in \mathbb{R}^{N,N}$, $\hat{B} \in \mathbb{R}^{N,p_2}$, $\hat{C} \in \mathbb{R}^{m_2,N}$, $\hat{D} \in \mathbb{R}^{m_2,p_2}$, and transfer function $K(s) = \hat{C}(s\hat{E} - \hat{A})^{-1}\hat{B} + \hat{D}$ such that the closed-loop system resulting from the combination of (7.1) and (7.2), given by

$$\begin{aligned} E\dot{x}(t) &= \left(A + B_2\hat{D}Z_1C_2\right)x(t) + B_2Z_2\hat{C}\hat{x}(t) + \left(B_1 + B_2\hat{D}Z_1D_{21}\right)v(t), \\ \hat{E}\dot{\hat{x}}(t) &= \hat{B}Z_1C_2x(t) + \left(\hat{A} + \hat{B}Z_1D_{22}\hat{C}\right)\hat{x}(t) + \hat{B}Z_1D_{21}v(t), \\ z(t) &= \left(C_1 + D_{12}Z_2\hat{D}C_2\right)x(t) + D_{12}Z_2\hat{C}\hat{x}(t) + \left(D_{11} + D_{12}\hat{D}Z_1D_{21}\right)v(t) \end{aligned} \tag{7.3}$$

with $Z_1 = \left(I_{p_2} - D_{22}\hat{D}\right)^{-1}$ and $Z_2 = \left(I_{m_2} - \hat{D}D_{22}\right)^{-1}$, has the following properties:

(i) System (7.3) is *internally stable*, i.e., the solution $\begin{bmatrix} x(\cdot) \\ \hat{x}(\cdot) \end{bmatrix}$ of the system with $v \equiv 0$ is *asymptotically stable*, in other words $\lim_{t\to\infty} \begin{bmatrix} x(t) \\ \hat{x}(t) \end{bmatrix} = 0$.

(ii) The closed-loop transfer function $T_{zv}(\cdot)$ from v to z satisfies $T_{zv} \in \mathscr{RH}_\infty^{p_1,m_1}$ and is minimized in the $\mathscr{H}_\infty$-norm.

Such an interconnection of a system with a controller is depicted in Fig. 4. Solving the optimal $\mathscr{H}_\infty$ control problem by trying to directly minimize the $\mathscr{H}_\infty$-norm of $T_{zv}(\cdot)$ over the complicated set of internally stabilizing controllers proves difficult or impossible by conventional optimization methods, since it is often unclear if a minimizing controller exists [122] and if one exists, it is typically not unique, there even exist infinitely many. So usually one studies two closely related optimization problems, the *modified optimal $\mathscr{H}_\infty$ control problem* and the *suboptimal $\mathscr{H}_\infty$ control problem* [15, 122].

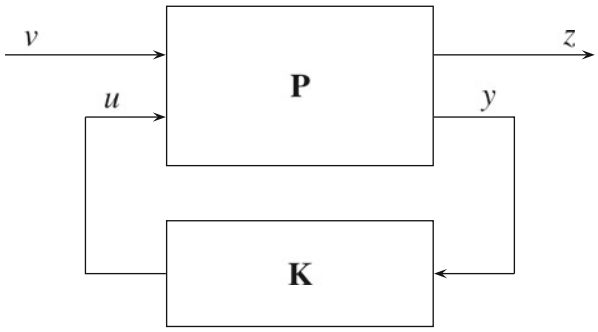

Fig. 4 Interconnection of a system **P** with a controller **K**

Definition 7.2 (The Modified Optimal $\mathscr{H}_\infty$ Control Problem) For the descriptor system (7.1), let Γ be the set of positive real numbers γ for which there exists an internally stabilizing dynamic controller of the form (7.2) so that the transfer function $T_{zv}(\cdot)$ of the closed-loop system (7.3) satisfies $T_{zv} \in \mathscr{RH}_\infty^{p_1,m_1}$ with $\|T_{zv}\|_{\mathscr{H}_\infty} < \gamma$. Determine $\gamma_{\mathrm{mo}} = \inf \Gamma$. If no internally stabilizing dynamic controller exists, we set $\Gamma = \emptyset$ and $\gamma_{\mathrm{mo}} = \infty$.

This problem is usually solved by an iterative process, which is often called the *γ-iteration*.

Definition 7.3 (The Suboptimal $\mathscr{H}_\infty$ Control Problem) For the descriptor system (7.1) and $\gamma \in \Gamma$ with $\gamma > \gamma_{\mathrm{mo}}$, determine an internally stabilizing dynamic controller of the form (7.2) such that the closed-loop transfer function satisfies $T_{zv} \in \mathscr{RH}_\infty^{p_1,m_1}$ with $\|T_{zv}\|_{\mathscr{H}_\infty} < \gamma$. We call such a controller *γ-suboptimal controller* or simply *suboptimal controller*.

To obtain an existence and uniqueness result we make the following assumptions:

(A1) The triple (E, A, B_2) is strongly stabilizable and the triple (E, A, C_2) is strongly detectable.

(A2) $\operatorname{rank}\begin{bmatrix} A - \mathrm{i}\omega E & B_2 \\ C_1 & D_{12} \end{bmatrix} = n + m_2$ for all $\omega \in \mathbb{R}$.

(A3) $\operatorname{rank}\begin{bmatrix} A - \mathrm{i}\omega E & B_1 \\ C_2 & D_{21} \end{bmatrix} = n + p_2$ for all $\omega \in \mathbb{R}$.

(A4) With matrices S_∞, $T_\infty \in \mathbb{R}^{n,n-r}$ satisfying $\operatorname{range} S_\infty = \ker E$, $\operatorname{range} T_\infty = \ker E^T$ and $r := \operatorname{rank} E$ we have

$$\operatorname{rank}\begin{bmatrix} T_\infty^T A S_\infty & T_\infty^T B_2 \\ C_1 S_\infty & D_{12} \end{bmatrix} = n + m_2 - r,$$

$$\operatorname{rank}\begin{bmatrix} T_\infty^T A S_\infty & T_\infty^T B_1 \\ C_2 S_\infty & D_{21} \end{bmatrix} = n + p_2 - r.$$

In Assumption **(A1)**, the conditions of impulse controllability and impulse observability are necessary to avoid impulsive solutions which cannot be controlled or observed. To check these conditions one can use the condensed forms of Theorem 4.2 with the characterization of Corollary 4.3. The property that the system is behavioral stabilizable and behavioral detectable is necessary for the existence of an internally stabilizing controller. To verify these conditions we use the decompositions (4.2) and (4.3) which can be computed via the codes `TG01HD`, `TG01ID` in the SLICOT library. These routines can also be used to check **(A2)** and **(A3)**.

To verify that assumption **(A4)** is satisfied, we check that the ranks of the extended matrices fulfill

$$\operatorname{rank}\begin{bmatrix} 0 & E & 0 \\ E^T & A & B_2 \\ 0 & C_1 & D_{12} \end{bmatrix} = n + m_2 + r,$$

and

$$\operatorname{rank}\begin{bmatrix} 0 & E & 0 \\ E^T & A & B_1 \\ 0 & C_2 & D_{21} \end{bmatrix} = n + p_2 + r.$$

This check is performed by applying a rank-revealing QR (RRQR) decomposition [62]. The corresponding routine `DGEQP3` is available in LAPACK.[4] For details on the implementation we refer to [29, 30, 55].

Once we have assured that the assumptions **(A1)**–**(A4)** hold, we can form the two even matrix pencils

$$sN_H - M_H(\gamma) = \left[\begin{array}{cc|ccc} 0 & -sE^T - A^T & 0 & 0 & -C_1^T \\ sE - A & 0 & -B_1 & -B_2 & 0 \\ \hline 0 & -B_1^T & -\gamma^2 I_{m_1} & 0 & -D_{11}^T \\ 0 & -B_2^T & 0 & 0 & -D_{12}^T \\ -C_1 & 0 & -D_{11} & -D_{12} & -I_{p_1} \end{array}\right], \tag{7.4}$$

and

$$sN_J - M_J(\gamma) = \left[\begin{array}{cc|ccc} 0 & -sE - A & 0 & 0 & -B_1 \\ sE^T - A^T & 0 & -C_1^T & -C_2^T & 0 \\ \hline 0 & -C_1 & -\gamma^2 I_{p_1} & 0 & -D_{11} \\ 0 & -C_2 & 0 & 0 & -D_{21} \\ -B_1^T & 0 & -D_{11}^T & -D_{21}^T & -I_{m_1} \end{array}\right]. \tag{7.5}$$

We determine the semi-stable deflating subspaces of both pencils, i.e., the deflating subspaces corresponding to the eigenvalues in the open left complex half-plane and a part of the deflating subspaces associated with the purely imaginary eigenvalues with even algebraic multiplicity and uniform sign-characteristic. Suppose that these

[4] http://www.netlib.org/lapack/.

subspaces are spanned by the columns of the matrices

$$X_H(\gamma) = \begin{bmatrix} X_{H,1}(\gamma) \\ X_{H,2}(\gamma) \\ X_{H,3}(\gamma) \\ X_{H,4}(\gamma) \\ X_{H,5}(\gamma) \end{bmatrix}, \quad X_J(\gamma) = \begin{bmatrix} X_{J,1}(\gamma) \\ X_{J,2}(\gamma) \\ X_{J,3}(\gamma) \\ X_{J,4}(\gamma) \\ X_{J,5}(\gamma) \end{bmatrix},$$

which are partitioned according to the block structure of the pencils $sN_H - M_H$ and $sN_J - M_J$.

We use the following result to solve the modified optimal $\mathscr{H}_\infty$ control problem.

Theorem 7.1 ([86]) *Consider system (7.1) and the even pencils $sN_H - M_H(\gamma)$ and $sN_J - M_J(\gamma)$ as in (7.4) and (7.5), respectively. Suppose that assumptions* **(A1)–(A4)** *hold.*

Then there exists an internally stabilizing controller such that the transfer function from v to z satisfies $T_{zv} \in \mathscr{RH}_\infty^{p_1,m_1}$ with $\|T_{zv}\|_{\mathscr{H}_\infty} < \gamma$ if and only if γ is such that the following conditions **C1)–C4)** *hold.*

C1) *The index of both pencils (7.4) and (7.5) is at most one.*

C2) *There exists a matrix $X_H(\gamma)$ such that*

- **C2.a)** *the space* $\operatorname{range} X_H(\gamma)$ *is a semi-stable deflating subspace of $sN_H - M_H(\gamma)$ and* $\operatorname{range} \begin{bmatrix} EX_{H,1}(\gamma) \\ X_{H,2}(\gamma) \end{bmatrix}$ *is an r-dimensional isotropic subspace of $\mathbb{R}^{2n}$;*
- **C2.b)** $\operatorname{rank} EX_{H,1}(\gamma) = r$.

C3) *There exists a matrix $X_J(\gamma)$ such that*

- **C3.a)** *the space* $\operatorname{range} X_J(\gamma)$ *is a semi-stable deflating subspace of $sN_J - M_J(\gamma)$ and* $\operatorname{range} \begin{bmatrix} E^T X_{J,1}(\gamma) \\ X_{J,2}(\gamma) \end{bmatrix}$ *is an r-dimensional isotropic subspace of $\mathbb{R}^{2n}$;*
- **C3.b)** $\operatorname{rank} E^T X_{J,1}(\gamma) = r$.

C4) *The matrix*

$$\mathscr{Y}(\gamma) = \begin{bmatrix} -\gamma X_{H,2}^T(\gamma) E X_{H,1}(\gamma) & X_{H,2}^T(\gamma) E X_{J,2}(\gamma) \\ X_{J,2}^T(\gamma) E^T X_{H,2}(\gamma) & -\gamma X_{J,2}^T(\gamma) E^T X_{J,1}(\gamma) \end{bmatrix}$$

is symmetric, positive semi-definite, and satisfies $\operatorname{rank} \mathscr{Y}(\gamma) = k_H + k_J$*, where* k_H *and* k_J *are such that for all sufficiently large* $\gamma_{H,1}, \gamma_{H,2}$*, and* $\gamma_{J,1}, \gamma_{J,2}$ *the conditions*

$$\begin{aligned} \operatorname{rank} E^T X_{H,2}(\gamma_{H,1}) &= \operatorname{rank} E^T X_{H,2}(\gamma_{H,2}) = k_H, \\ \operatorname{rank} E X_{J,2}(\gamma_{J,1}) &= \operatorname{rank} E X_{J,2}(\gamma_{J,2}) = k_J \end{aligned}$$

hold.

Furthermore, the set of values γ *satisfying the conditions* ***C1)–C4)*** *is nonempty.*

To check condition **C4)**, we make use of the LDL^T decomposition, described in [6] and implemented in LAPACK by `DSPTRF` which decomposes a real symmetric matrix A as $A = LDL^T$, where L is a product of permutation and lower triangular matrices, and D is symmetric and block diagonal with 1×1 and 2×2 diagonal blocks.

Using Theorem 7.1, we can use a bisection type algorithm to determine the suboptimal value γ_{mo}, see [85].

After completing the bisection process, one has the option to either use the result directly, or to perform a strong validation, by dividing the interval $(0, \gamma_{\mathrm{mo}})$ at a desired number of points and checking the four conditions **C1)–C4)** again at these points. If the conditions **C1)–C4)** are fulfilled for another $\gamma \in (0, \gamma_{\mathrm{mo}})$, we have obviously found a better value for γ_{mo}. We can either use this new value or continue with the γ-iteration to find an even better value. Once a satisfactory γ is found, it remains to compute the controller. The trick that we use to determine the controller is to compute an index-reducing static output feedback $u(t) = Fy(t) + \overline{u}(t)$, whose application leads to a new descriptor system of the form (7.1) with an index of at most one. It can be shown that the application of the feedback does not change the solution of the modified $\mathscr{H}_\infty$ optimal control problem [85, 86]. The feedback is computed using the condensed form (4.1) and the techniques presented in [40], which yield $s_2 = t_2$ and

$$F = \begin{bmatrix} F_{11} & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{m,p}, \quad F_{11} = \begin{bmatrix} B_{21} \\ B_{31} \end{bmatrix}^{-1} (I_{s_2} - A_{22}) \begin{bmatrix} C_{12} & C_{13} \end{bmatrix}^{-1}. \tag{7.6}$$

Note that due to the construction of the condensed form (4.1), the matrices

$$\begin{bmatrix} B_{21} \\ B_{31} \end{bmatrix}, \quad \begin{bmatrix} C_{12} & C_{13} \end{bmatrix}$$

can be kept in factored form as a product of an orthogonal and a diagonal matrix. So the computation of F can be carried out by the inversion of two diagonal matrices.

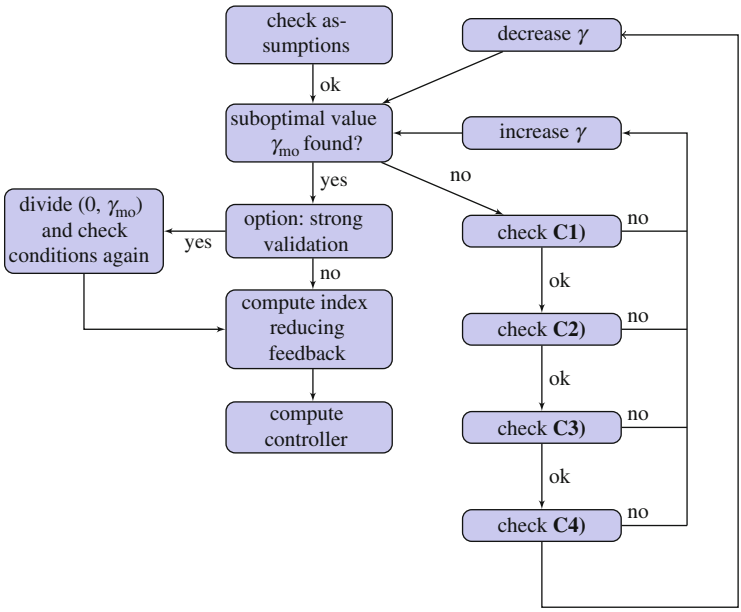

Fig. 5 Algorithm flowchart for solving $\mathscr{H}_\infty$ optimal control problems

We can use this new descriptor system to compute the controller. The controller formulas themselves and their derivation are rather involved. Therefore, we only refer to the robust controller formulas for the standard system case in [16], and based on that, the controller formulas for the descriptor system case in [85].

Figure 5 presents a flow chart for the solution of the optimal solution. First one checks the four assumptions **(A1)**–**(A4)**, using the condensed forms from Theorem 4.2, the decompositions (4.2) and performing some rank checks. Then one uses a bisection type algorithm to find the optimal value of γ, by checking the four conditions from Theorem 7.1 in each step by using the staircase form from Theorem 5.2, the computation of the semi-stable deflating subspaces using Algorithm 1, and the LDL^T decomposition from [6]. Here, the structure-preservation aspect of Algorithm 1 is very important, since using these methods, it *cannot* happen that eigenvalues from the left half plane move to the right half plane and vice versa due to round-off errors. Therefore, the computed subspaces are guaranteed to have the correct dimensions. Once the suboptimal value is found, one has the option to use a strong validation by checking the aforementioned four conditions again at a desired number of points. Then it remains to compute an index reducing feedback (7.6) and to compute the controller formulas given in [16, 85]. For an illustration of the method by numerical examples we refer to [16, 85].

8 $\mathscr{L}_\infty$-Norm Computation

In the previous section we have seen that the $\mathscr{H}_\infty$-norm of a transfer function is an important measure for the robustness of a linear system. This section is devoted to the actual computation of this norm. We will directly present this for the more general case of the $\mathscr{L}_\infty$-norm. Consider a square descriptor system (1.2) with regular pencil $sE - A$ and transfer function $G(\cdot)$ as in (4.4).

Before we can turn to the actual norm computation, we have to ensure that $G \in \mathscr{RL}_\infty^{p,m}$. First, we check whether the transfer function is *proper*, i.e., that $\lim_{\omega\to\infty} \|G(\mathrm{i}\omega)\| < \infty$. For this we make use of the following result of Benner et al. [18] and Voigt [116] in a modified formulation.

Theorem 8.1 *Consider a descriptor system* (1.2a) *given in the condensed form* (4.1)*. Then, $G(\cdot)$ is proper if and only if the subpencil*

$$s\begin{bmatrix}\Sigma_E & 0\\ 0 & 0\end{bmatrix} - \begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix}$$

is regular and of index at most one, i.e., if A_{22} is invertible.

Therefore, to check properness, we first reduce the system to the condensed form (4.1) and subsequently check A_{22} for invertibility, e.g., by employing condition estimators [62].

When we have checked the transfer function for properness, it remains to check whether $G(\cdot)$ has finite, purely imaginary poles. For this, we first determine the controllability and observability decompositions (4.2) and (4.3) to extract the controllable and observable subsystem. The finite eigenvalues of the pencil associated with this subsystem are poles of $G(\cdot)$ and we check whether there are eigenvalues that lie in a thin strip around the imaginary axis. The thickness of this strip depends on the multiplicity of the pole which is generally not known. In finite precision, eigenvalues in this region cannot be distinguished from eigenvalues on the imaginary axis. Generically, a pole will be simple and therefore, in the code we choose the thickness as a small multiple of machine precision. After we have ensured that $G \in \mathscr{RL}_\infty^{p,m}$, we can compute the norm value. For this we make use of the even matrix pencils

$$sN - M(\gamma) = \left[\begin{array}{cc|cc} 0 & sE - A & 0 & -B \\ -sE^T - A^T & 0 & -C^T & 0 \\ \hline 0 & -C & \gamma I_p & -D \\ -B^T & 0 & -D^T & \gamma I_m \end{array}\right]. \tag{8.1}$$

The following theorem connects the singular values of $G(\mathrm{i}\omega)$ with the finite, purely imaginary eigenvalues of $sN - M(\gamma)$, see [18, 19, 116] for details.

Theorem 8.2 *Assume that $sE - A$ has no purely imaginary eigenvalues, $G \in \mathscr{RL}_\infty^{p,m}$, $\gamma > 0$ and $\omega_0 \in \mathbb{R}$. Then γ is a singular value of $G(\mathrm{i}\omega_0)$ if and only if $\mathrm{i}\omega_0 N - M(\gamma)$ is singular.*

A direct consequence of Theorem 8.2 is the following result, see [18, 19].

Theorem 8.3 *Assume that $sE - A$ has no purely imaginary eigenvalues, $G \in \mathscr{RL}_\infty^{p,m}$ and suppose that $\gamma > \inf_{\omega\in\mathbb{R}} \sigma_{\max}(G(\mathrm{i}\omega))$. Then $\|G\|_{\mathscr{L}_\infty} \geq \gamma$ if and only if $sN - M(\gamma)$ in (8.1) has finite, purely imaginary eigenvalues.*

This directly yields an algorithm for the computation of the $\mathscr{L}_\infty$-norm, similarly as in [32–34]. Given an initial value of γ with $\inf_{\omega\in\mathbb{R}} \sigma_{\max}(G(\mathrm{i}\omega)) < \gamma < \|G\|_{\mathscr{L}_\infty}$, we check if $sN - M(\gamma)$ has purely imaginary eigenvalues. If yes, we denote these eigenvalues with positive imaginary part by $\mathrm{i}\omega_1, \ldots, \mathrm{i}\omega_q$. To obtain the next (larger) value of γ, we determine new test frequencies $m_j = \sqrt{\omega_j \omega_{j+1}}$, $j = 1, \ldots, q-1$. Then, the new value of γ is chosen as

$$\gamma = \max_{1\leq j\leq q-1} \sigma_{\max}(G(\mathrm{i}m_j)).$$

To check whether a prespecified relative error ε has already been achieved, we would have to check whether the pencil $sN - M(\hat{\gamma})$ with $\hat{\gamma} = \gamma(1 + 2\varepsilon)$ has no purely imaginary eigenvalues. To avoid the additional check in every step, we can directly incorporate this into the algorithm by always working with $\hat{\gamma}$ instead of γ when determining the eigenvalues of the even pencils.

It can be shown that this algorithm converges globally with a quadratic rate and a guaranteed relative error of ε when assuming exact arithmetics. We refer to [18, 19, 116] for details on the implementation and the algorithm properties. Note again that the decision about the existence of purely imaginary eigenvalues is crucial for a robust execution of this algorithm and does require a structured eigensolver as described in Sect. 5.2. A graphical interpretation is given in Fig. 6.

Note that when assuming that $G \in \mathscr{RL}_\infty^{p,m}$, the algorithm runs on the original data without performing any system reductions beforehand. However, $sE - A$ could still have uncontrollable or unobservable eigenvalues on the imaginary axis. If one does not perform the system reductions to extract the behavioral controllable and observable subsystem, then it remains to check whether $sE - A$ has no finite, purely imaginary eigenvalues. The complete procedure is summarized in Fig. 7. An illustrative numerical example can be found in [18], whereas in [19] one can find a more detailed analysis of the behavior of the algorithm.

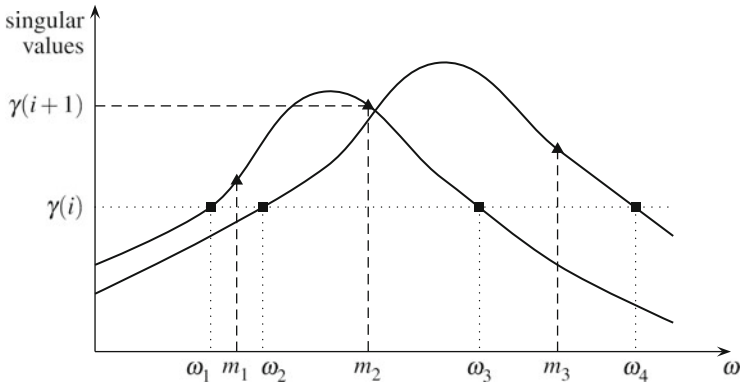

Fig. 6 Graphical interpretation of the algorithm for computing the $\mathscr{L}_\infty$-norm. Here, $\gamma(i)$ and $\gamma(i+1)$ denote the iterates at the ith and $(i+1)$st step, respectively

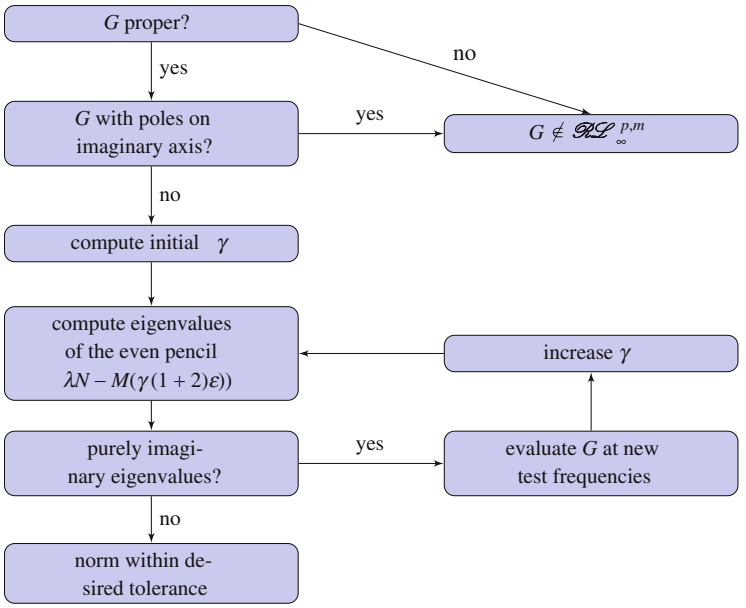

Fig. 7 Flowchart for computing the $\mathscr{L}_\infty$-norm

9 Dissipativity Check

The notion of dissipative systems is one of the most important concepts in systems and control theory, see for instance [118–120]. It naturally arises in many physical problems, especially when energy considerations are of importance. Roughly speaking, dissipative systems cannot internally generate energy. Equivalently, the

system cannot supply more energy to its environment than energy that has been supplied to the system. Typical areas where such systems appear are the modeling of electrical circuits [100] (where, e.g., resistors consume a part of the energy and transform it into heat), or thermodynamic processes (where a part of the energy is transformed into an increase of entropy due to the second law of thermodynamics).

When modeling real-world processes it is often desired or necessary to reflect the dissipative nature of the problem in the model structure. This is important in order to obtain physically meaningful results when performing simulations. This section presents a method to check a certain notion of dissipativity for linear time-invariant descriptor systems of the form (1.2) based on a spectral characterization for even pencils.

We first introduce a precise mathematical formulation of dissipativity. For this we need the notion of *supply rates* which measure the power supplied to the system at time t. In the following we restrict ourselves to quadratic supply functions of the form

$$s(u(t), y(t)) = \begin{bmatrix} y(t) \\ u(t) \end{bmatrix}^T \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \begin{bmatrix} y(t) \\ u(t) \end{bmatrix}, \tag{9.1}$$

where $Q = Q^T \in \mathbb{R}^{p,p}$, $S \in \mathbb{R}^{p,m}$, and $R = R^T \in \mathbb{R}^{m,m}$. Then the energy supplied to the system in a time interval $[t_0, t_1]$ is measured by

$$\int_{t_0}^{t_1} s(u(t), y(t)) \mathrm{d}t.$$

There are many different notions of dissipativity in the literature. In this survey, we stick to the notion of cyclo-dissipativity which has been introduced in [35, 36] in the context of behavior systems.

Definition 9.1 A descriptor system (1.2) is called *cyclo-dissipative* with respect to $s(\cdot, \cdot)$, if

$$\int_0^T s(u(t), y(t)) \mathrm{d}t \geq 0$$

for all $T \geq 0$ and all smooth trajectories $(u(\cdot), x(\cdot), y(\cdot))$ solving (1.2) with the boundary conditions $Ex(0) = Ex(T) = 0$.

Remark 9.1 Cyclo-dissipativity is only a property of the strongly controllable part of the system. A more general definition of dissipativity would require the existence of a *storage function* $\Theta : \operatorname{im} E \to \mathbb{R}$ with $\Theta(0) = 0$ such that the *dissipation inequality*

$$\Theta(Ex(t_1)) \leq \Theta(Ex(t_0)) + \int_{t_0}^{t_1} s(u(t), y(t)) \mathrm{d}t$$

is fulfilled for all $t_0 \leq t_1$ and all smooth solution trajectories $(u(\cdot), x(\cdot), y(\cdot))$ such that the supply rate is locally square-integrable, see [36]. If the system (1.2) is strongly controllable, then both definitions coincide. However, not every cyclo-dissipative system has to possess a storage function. A counter-example is given in [36].

Remark 9.2 In the definition of cyclo-dissipativity it is only required that trajectories that start at zero and return to zero in some finite time do not generate energy. A stronger definition, that would require all trajectories that start at zero not to generate energy, exists as well. Special cases of this stronger notion are passivity and contractivity (see below). Closely related to this is then nonnegativity of the storage function (if it exists). Unfortunately, its general treatment is much more involved. However, under the condition that the pencil $sE - A$ is regular, stable, and its Kronecker index is at most one, and Q is negative semidefinite, then this stronger definition coincides with Definition 9.1, see [38].

In practice, two particular cases for the choice of the supply rate are of great interest. If a descriptor system (1.2) is dissipative (in the sense of the stronger definition in Remark 9.2) with respect to the supply rate $s(u(t), y(t)) = u(t)^T y(t)$, i.e., if $k = n$, $p = m$ and

$$\begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 0 & I_m \\ I_m & 0 \end{bmatrix},$$

then the system is called *passive*. This situation typically arises in models for RLC circuits [2, 96–98].

The other special case is that the supply rate is given by $s(u(t), y(t)) = \|u(t)\|_2^2 - \|y(t)\|_2^2$, i.e., $k = n$ and

$$\begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} = \begin{bmatrix} -I_p & 0 \\ 0 & I_m \end{bmatrix}.$$

In this case, a dissipative system (in the sense of the stronger definition in Remark 9.2) is called *contractive*. Usually this structure occurs if (1.2) is a realization of scattering parameters [89], but similar structures also appear in $\mathscr{H}_\infty$ control, see Sects. 7 and 8.

For square systems (with $k = n$), a well-known relation of cyclo-dissipativity defined above between the time and frequency domain is given by the so-called *Popov function*

$$H(\xi, \zeta) := \begin{bmatrix} (\xi E - A)^{-1} B \\ I_m \end{bmatrix}^H \begin{bmatrix} \tilde{Q} & \tilde{S} \\ \tilde{S}^T & \tilde{R} \end{bmatrix} \begin{bmatrix} (\zeta E - A)^{-1} B \\ I_m \end{bmatrix},$$

with $\tilde{Q}$, $\tilde{S}$, and $\tilde{R}$ as in (6.2). One has the following theorem of [35, 36].

Theorem 9.1 *The square descriptor system* (1.2) *is cyclo-dissipative with respect to* $s(\cdot,\cdot)$ *if and only if* $H(\mathrm{i}\omega,\mathrm{i}\omega) \geq 0$ *for all* $\mathrm{i}\omega \notin \Lambda(E,A)$.

For the cases of passivity and contractivity we get more general relations. These are summarized in the following theorem [2].

Theorem 9.2 *Consider a square descriptor system of the form* (1.2) *with* $p = m$.

(i) The system is passive if and only if $G(\cdot)$ *is* positive real*, i.e.,*

(a) $G(\cdot)$ *is analytic in* $\mathbb{C}^+$*; and*
(b) $H(\lambda,\lambda) = G(\lambda) + G(\lambda)^H \geq 0$ *for all* $\lambda \in \mathbb{C}^+$.

(ii) The system is contractive if and only if $G(\cdot)$ *is* bounded real*, i.e.,*

(a) $G(\cdot)$ *is analytic in* $\mathbb{C}^+$*; and*
(b) $H(\lambda,\lambda) = I_m - G(\lambda)^H G(\lambda) \geq 0$ *for all* $\lambda \in \mathbb{C}^+$.

It is very important to note that similar equivalent conditions of Theorem 9.2 do in general *not* hold for systems that are dissipative in the sense of the stronger definition in Remark 9.2. A counterexample is given in [121]. There are many algebraic characterizations to check if a given system (1.2) is cyclo-dissipative. These are mainly based on solvability of certain linear matrix inequalities or matrix equations, see [66]. Instead we make use of the following spectral characterization of even matrix pencils. For this, we need the *sign-sum function* [35, 36, 38] of a Hermitian matrix T which is defined as

$$\eta(T) = \pi_+ + \pi_0 - \pi_-,$$

where π_+, π_0, and π_- are the numbers of positive, zero, and negative eigenvalues of T, respectively. Furthermore, we can define the rank of a polynomial matrix $P(s)$ over the field of real-rational functions (often called normal rank), given by

$$\operatorname{rank}_{\mathbb{R}(s)}(P(s)) := \max_{\lambda\in\mathbb{C}} \operatorname{rank}(P(\lambda)). \tag{9.2}$$

The maximum in (9.2) is attained for almost all values of $\lambda \in \mathbb{C}$, there is only a finite set of points, where the rank drops.

Theorem 9.3 ([36, Theorem 3.11]) *Consider the system* (1.2) *with supply rate* (9.1)*. Let*

$$r := \operatorname{rank}_{\mathbb{R}(s)} \left(\begin{bmatrix} sE - A & -B \end{bmatrix} \right) \tag{9.3}$$

and define $\ell := k + n + m + 2p$. *Consider the even pencil*

$$\mathscr{N}(s) = sN - M = \begin{bmatrix} 0 & 0 & 0 & sE - A & -B \\ 0 & 0 & I_m & -C & -D \\ 0 & I_m & Q & 0 & S \\ -sE^T - A^T & -C^T & 0 & 0 & 0 \\ -B^T & -D^T & S^T & 0 & R \end{bmatrix} \in \mathbb{R}[s]^{\ell,\ell}. \tag{9.4}$$

Then the system given by (1.2) *is cyclo-dissipative if and only if*

$$\eta(\mathscr{N}(\mathrm{i}\omega)) = k + n + m - 2r$$

for all $\omega \in \mathbb{R}$ *with* $\operatorname{rank}\left(\begin{bmatrix} \mathrm{i}\omega E - A & -B \end{bmatrix}\right) = r$.

To better understand this theorem, we present a visualization in terms of the so-called *spectral plot*. This plot is constructed by plotting the ℓ eigenvalues of $\mathscr{N}(\mathrm{i}\omega)$ depending on ω, see Fig. 8 for an example.

The general framework for checking cyclo-dissipativity then consists of two steps. First, we check if the assumptions of Theorem 9.3 are fulfilled. If the normal rank is unknown, then the GUPTRI form [53, 54, 70] is a suitable tool to compute it.

The next step consists in checking the sign-sum condition in Theorem 9.3. We exploit the fact that $\eta(\mathscr{N}(\mathrm{i}\omega))$ can only change at purely imaginary eigenvalues (of the regular index one part) and remains constant between two subsequent purely imaginary eigenvalues. We construct the pencil (9.4) and apply the even staircase algorithm from Theorem 5.2 to get the regular index one part $sN_{w+1,w+1} - M_{w+1,w+1}$. Then we compute its purely imaginary eigenvalues with positive imaginary part, denoted by $\mathrm{i}\omega_1, \ldots, \mathrm{i}\omega_q$, with $\omega_1 < \omega_2 < \ldots < \omega_q$. This is done using Algorithm 1.

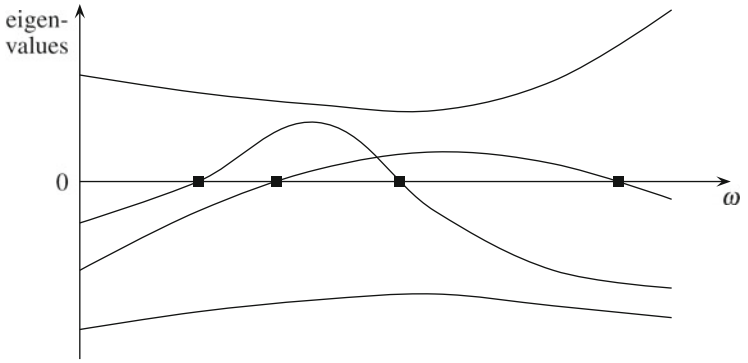

Fig. 8 Spectral plot. Here cyclo-dissipativity is violated, since the sign-sum function changes for varying ω

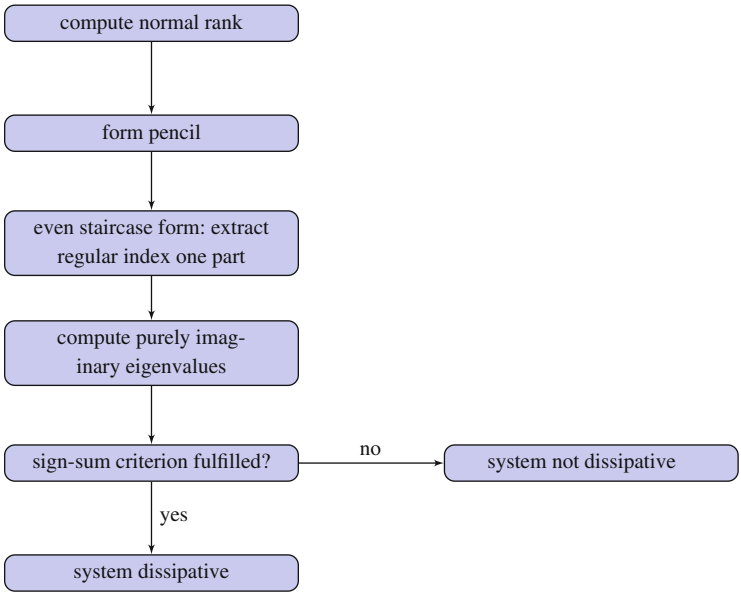

Fig. 9 Algorithm flowchart for dissipativity check

Next, we set $\omega_0 := 0$ and $\omega_{q+1} := \infty$. For $j = 0, \ldots, q$, we choose points $\alpha_j \in (\omega_j, \omega_{j+1})$ with rank $\left(\begin{bmatrix} \mathrm{i}\alpha_j E - A & -B \end{bmatrix}\right) = r$. Finally, for $j = 0, \ldots, q$ we compute the inertia $(\pi_+^j, \pi_0^j, \pi_-^j)$ of the Hermitian matrix $\mathscr{N}(\mathrm{i}\alpha_j)$ and thus obtain $\eta(\mathscr{N}(\mathrm{i}\alpha_j)) = \pi_+^j + \pi_0^j - \pi_-^j$. Then the system is dissipative if and only if $\eta(\mathscr{N}(\mathrm{i}\alpha_j)) = k + n + m - 2r$ for all j. Figure 9 summarizes the complete procedure in a diagram.

We further illustrate the algorithm by means of the following example.

Example 9.1 We consider a slightly modified circuit example from [116, Sect. 1.1.1] resulting from a modified nodal analysis given by the following data:

$$E = \begin{bmatrix} 0 & 0 & 0\,0 & 0 & 0\,0 \\ 0 & 10^4 & 0\,0 & 0 & 0\,0 \\ 0 & 0 & 0\,0 & 0 & 0\,0 \\ 0 & 0 & 0\,0 & 0 & 0\,0 \\ 0 & 0 & 0\,0 & 10^{-3} & 0\,0 \\ 0 & 0 & 0\,0 & 0 & 0\,0 \\ 0 & 0 & 0\,0 & 0 & 0\,0 \end{bmatrix}, \quad A = \begin{bmatrix} 0 & 0 & 0 & 0 & -1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & -10 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 10^{-2} & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 10 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$C = B^T = \begin{bmatrix} 0\,0\,0 & 0 & 0\,0 & -1 \\ 0\,0\,1 & -1 & 0\,0 & 0 \end{bmatrix}, \quad D = 0.$$

Due to the special structure of the matrices, the corresponding descriptor system is passive [96]. In particular, the system is cyclo-dissipative with respect to the supply rate defined by the weighting matrices

$$\begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} = \begin{bmatrix} 0 & I_2 \\ I_2 & 0 \end{bmatrix}.$$

We now verify this by checking the spectrum of the even pencil (9.4). First, we reduce the pencil to even staircase form. Using the notation of Theorem 5.2 we obtain the following structural information:

$$w = 3, \quad s_1 = 1, \quad s_2 = 1, \quad s_3 = 0, \quad q_1 = 1, \quad q_2 = 1, \quad q_3 = 1.$$

In particular, since $q_3 - s_3 = 1 \neq 0$, the pencil is singular. The regular index one part is given by

$$\begin{aligned} sN_{4,4} - M_{4,4} = -\operatorname{diag}\,(&1.9021, 1.6194, 1.414, 2.1756, 0.6217, \\ &-0.6144, -1.4142, -1.6167, -1.9021, 10.0504, \\ &1.9962, -0.0075, -1.1756, -1.9987, -10.0504) \in \mathbb{R}[s]^{15,15}, \end{aligned}$$

which has only (semisimple) infinite eigenvalues. Therefore, it is sufficient to evaluate the sign-sum function at a single point, for instance for $\omega = 0$ we obtain

$$\eta(\mathcal{N}(0)) = \eta(-M) = 2 = k + n + m - 2r.$$

Hence, it is confirmed that the system is cyclo-dissipative.

10 Conclusions

This paper provides a uniform treatment of differential-algebraic equations by methods from numerical linear algebra. First, we have presented the solution theory of such equations as well as regularization procedures. Based on that we have discussed several important applications from control and optimization of DAEs. These are based on the solution of even eigenvalue problems. We have presented several canonical forms of even pencils and discussed their properties. These canonical forms can be employed to numerically treat the presented applications in a uniform framework. The methods discussed here are usable for small-scale problems, i.e., to problems of size up to a few hundred. Here, the main computational bottleneck are the complexity and the storage requirements for solving even and skew-Hamiltonian/Hamiltonian eigenvalue problems.

Thus a big issue is the development of algorithms for large and sparse problems, which are widely unexplored. For instance, it is not clear how to determine *all* desired eigenvalues of a large-scale even pencil, e.g., the purely imaginary ones or how to approximate the complete subspace associated with all eigenvalues in the left half plane by a sparse representation.

Acknowledgements The author "Philip Losse" was supported by the DFG Research Center MATHEON in Berlin. The author "Volker Mehrmann" was supported by the European Research Council through ERC Advanced Grant ModSimConMP. Research of the author "Matthias Voigt" was supported in the framework of MATHEON project *C-SE1: Reduced order modeling for data assimilation* supported by the Einstein Foundation Berlin.

References

1. Alam, R., Bora, S., Karow, M., Mehrmann, V., Moro, J.: Perturbation theory for Hamiltonian matrices and the distance to bounded-realness. SIAM J. Matrix Anal. Appl. **32**(2), 484–514 (2011)
2. Anderson, B.D.O., Vongpanitlerd, S.: Network Analysis and Synthesis. Prentice Hall, Englewood Cliffs, NJ (1973)
3. Antoulas, A.C.: Approximation of Large-Scale Dynamical Systems. Adv. Des. Control. SIAM, Philadelphia, PA (2005)
4. Antoulas, A.C., Mayo, A.J.: A framework for the solution of the generalized realization problem. Linear Algebra Appl. **425**, 634–662 (2007)
5. Ascher, U.M., Mattheij, R., Russell, R.: Numerical Solution of Boundary Value Problems for Ordinary Differential Equations, 2nd edn. SIAM, Philadelphia, PA (1995)
6. Ashcraft, C., Grimes, R.G., Lewis, J.G.: Accurate symmetric indefinite linear equations solvers. SIAM J. Matrix Anal. Appl. **20**(2), 513–561 (1998)
7. Backes, A.: Extremalbedingungen für Optimierungs-Probleme mit Algebro-Differentialgleichungen. Dissertation, Institut für Mathematik, Humboldt-Universität zu Berlin (2006)
8. Bals, J., Hofer, G., Pfeiffer, A., Schallert, C.: Virtual iron bird – a multidisciplinary modelling and simulation platform for new aircraft system architectures. In: Deutscher Luft- und Raumfahrtkongress, Friedrichshafen (2005)
9. Benner, P., Effenberger, C.: A rational SHIRA method for the Hamiltonian eigenvalue problem. Taiwan. J. Math. **14**(6A), 805–823 (2010)
10. Benner, P., Mehrmann, V., Xu, H.: A new method for computing the stable invariant subspace of a real Hamiltonian matrix. J. Comput. Appl. Math. **86**(1), 17–43 (1997)
11. Benner, P., Mehrmann, V., Xu, H.: A numerically stable, structure preserving method for computing the eigenvalues of real Hamiltonian or symplectic pencils. Numer. Math. **78**(3), 329–358 (1998)
12. Benner, P., Byers, R., Mehrmann, V., Xu, H.: Numerical computation of deflating subspaces of embedded Hamiltonian pencils. Tech. Rep. SFB393/99-15, Fakultät für Mathematik, TU Chemnitz. Available from http://www.tu-chemnitz.de/sfb393/sfb99pr.html (1999)
13. Benner, P., Byers, R., Faßbender, H., Mehrmann, V., Watkins, D.: Cholesky-like factorizations of skew-symmetric matrices. Electron. Trans. Numer. Anal. **11**, 85–93 (2000)
14. Benner, P., Byers, R., Mehrmann, V., Xu, H.: Numerical computation of deflating subspaces of skew-Hamiltonian/Hamiltonian pencils. SIAM J. Matrix Anal. Appl. **24**(1), 165–190 (2002)
15. Benner, P., Byers, R., Mehrmann, V., Xu, H.: A robust numerical method for the γ-iteration in $\mathscr{H}_\infty$ control. Linear Algebra Appl. **425**(2–3), 548–570 (2007)

16. Benner, P., Byers, R., Losse, P., Mehrmann, V., Xu, H.: Robust formulas for optimal H_∞ controllers. Automatica **47**(12), 2639–2646 (2011)
17. Benner, P., Faßbender, H., Stoll, M.: A Hamiltonian Krylov-Schur-type method based on the symplectic Lanczos process. Linear Algebra Appl. **435**(3), 578–600 (2011)
18. Benner, P., Sima, V., Voigt, M.: $\mathscr{L}_\infty$-norm computation for continuous-time descriptor systems using structured matrix pencils. IEEE Trans. Autom. Control **57**(1), 233–238 (2012)
19. Benner, P., Sima, V., Voigt, M.: Robust and efficient algorithms for $\mathscr{L}_\infty$-norm computation for descriptor systems. In: Proceedings of 7th IFAC Symposium on Robust Control Design, Aalborg, pp. 195–200 (2012). doi:10.3182/20120620-3-DK-2025.00114
20. Benner, P., Sima, V., Voigt, M.: FORTRAN 77 subroutines for the solution of skew-Hamiltonian/Hamiltonian eigenproblems – Part I: algorithms and applications. Preprint MPIMD/13-11, Max Planck Institute Magdeburg. Available from http://www.mpi-magdeburg.mpg.de/preprints/2013/11/ (2013)
21. Benner, P., Sima, V., Voigt, M.: FORTRAN 77 subroutines for the solution of skew-Hamiltonian/Hamiltonian eigenproblems – Part II: implementation and numerical results. Preprint MPIMD/13-12, Max Planck Institute Magdeburg. Available from http://www.mpi-magdeburg.mpg.de/preprints/2013/12/ (2013)
22. Berger, T.: On differential-algebraic control systems. Dissertation, Fakultät für Mathematik und Naturwissenschaften, TU Ilmenau (2013)
23. Berger, T., Reis, T.: Controllability of linear differential-algebraic equations – a survey. In: Ilchmann, A., Reis, T. (eds.) Surveys in Differential-Algebraic Equations I, Differential-Algebraic Equations Forum, pp. 1–61. Springer, Berlin/Heidelberg (2013)
24. Berger, T., Trenn, S.: The quasi-Kronecker form for matrix pencils. SIAM J. Matrix Anal. Appl. **33**(2), 336–368 (2012)
25. Berger, T., Trenn, S.: Addition to "The quasi-Kronecker form for matrix pencils". SIAM J. Matrix Anal. Appl. **34**(1), 94–101 (2013)
26. Berger, T., Reis, T., Trenn, S.: Observability of differential-algebraic equations – a survey. Hamburger Beiträge zur Angewandten Mathematik 2015-13, Fachbereich Mathematik, Universität Hamburg. Available from http://preprint.math.uni-hamburg.de/public/papers/hbam/hbam2015-13.pdf (2015)
27. Betcke, T., Higham, N.J., Mehrmann, V., Schröder, C., Tisseur, F.: NLEVP: a collection of nonlinear eigenvalue problems. ACM Trans. Math. Softw. **39**(2), Article 7 (2013)
28. Binder, A., Mehrmann, V., Międ lar, A.: A MATLAB toolbox for the regularization of descriptor systems arising from generalized realization procedures (2014, in preparation)
29. Bischof, C.H., Quintana-Ortí, G.: Algorithm 782: codes for rank-revealing QR factorizations of dense matrices. ACM Trans. Math. Softw. **24**(2), 254–257 (1998). doi:http://doi.acm.org/10.1145/290200.287638
30. Bischof, C.H., Quintana-Ortí, G.: Computing rank-revealing QR factorizations of dense matrices. ACM Trans. Math. Softw. **24**(2), 226–253 (1998). doi:http://doi.acm.org/10.1145/290200.287637
31. Bojanczyk, A.I., Golub, G.H., Van Dooren, P.: The periodic Schur decomposition. Algorithms and applications. In: Luk, F.T. (ed.) Advanced Signal Processing Algorithms, Architectures, and Implementations III. Proceedings of SPIE, vol. 1770, pp. 31–42 (1992)
32. Boyd, S., Balakrishnan, V.: A regularity result for the singular values of a transfer matrix and a quadratically convergent algorithm for computing its L_∞-norm. Syst. Control Lett. **15**(1), 1–7 (1990)
33. Boyd, S., Balakrishnan, V., Kabamba, P.: A bisection method for computing the H_∞ norm of a transfer matrix and related problems. Math. Control Signals Syst. **2**(3), 207–219 (1989)
34. Bruinsma, N.A., Steinbuch, M.: A fast algorithm to compute the H_∞-norm of a transfer function matrix. Syst. Control Lett. **14**(4), 287–293 (1990)
35. Brüll, T.: Checking dissipativity of linear behaviour systems given in kernel representation. Math. Control Signals Syst. **23**(1–3), 159–175 (2011)

36. Brüll, T.: Dissipativity of linear quadratic systems. Dissertation, Institut für Mathematik, TU Berlin. Available from http://opus4.kobv.de/opus4-tuberlin/frontdoor/index/index/docId/2824 (2011)
37. Brüll, T., Mehrmann, V.: STCSSP: A FORTRAN 77 routine to compute a structured staircase form for a (skew-)symmetric/(skew-)symmetric matrix pencil. Preprint 31-2007, Institut für Mathematik, TU Berlin (2007)
38. Brüll, T., Schröder, C.: Dissipativity enforcement via perturbation of para-Hermitian pencils. IEEE Trans. Circuits Syst. Regul. Pap. **60**(1), 164–177 (2013)
39. Bunch, J.R.: A note on the stable decomposition of skew-symmetric matrices. Math. Comput. **38**(158), 475–479 (1982)
40. Bunse-Gerstner, A., Mehrmann, V., Nichols, N.K.: Regularization of descriptor systems by output feedback. IEEE Trans. Autom. Control **39**(4), 1742–1748 (1994)
41. Bunse-Gerstner, A., Byers, R., Mehrmann, V., Nichols, N.K.: Feedback design for regularizing descriptor systems. Linear Algebra Appl. **299**, 119–151 (1999)
42. Byers, R., Geerts, T., Mehrmann, V.: Descriptor systems without controllability at infinity. SIAM J. Control Optim. **35**(2), 462–479 (1997)
43. Byers, R., Mehrmann, V., Xu, H.: A structured staircase algorithm for skew-symmetric/symmetric pencils. Electron. Trans. Numer. Anal. **26**, 1–13 (2007)
44. Campbell, S.L.: Singular Systems of Differential Equations I. Pitman, San Francisco, CA (1980)
45. Campbell, S.L.: A general form for solvable linear time varying singular systems of differential equations. SIAM J. Math. Anal. **18**(4), 1101–1115 (1987)
46. Campbell, S.L.: Linearization of DAEs along trajectories. Z. Angew. Math. Phys. **46**(1), 70–84 (1995)
47. Campbell, S.L., Kunkel, P., Mehrmann, V.: Regularization of linear and nonlinear descriptor systems. In: Biegler, L.T., Campbell, S.L., Mehrmann, V. (eds.) Control and Optimization with Differential-Algebraic Constraints. Advances in Design and Control, Chap. 2, pp. 17–36. SIAM, Philadelphia, PA (2012)
48. Chu, D., Liu, X., Mehrmann, V.: A numerical method for computing the Hamiltonian Schur form. Numer. Math. **105**(3), 375–412 (2007)
49. Cobb, J.D.: Controllability, observability and duality in singular systems. IEEE Trans. Autom. Control **AC-29**(12), 1076–1082 (1984)
50. Dai, L.: Singular Control Systems. Lecture Notes in Control and Information Science, vol. 118. Springer, Berlin/Heidelberg (1989)
51. Datta, S., Mehrmann, V.: Computation of state reachable points of linear time invariant descriptor systems. Preprint 17/2014, Institut für Mathematik, Technische Universität Berlin. Submitted, available from http://www.math.tu-berlin.de/preprints/ (2014)
52. Demmel, J.W., Kågström, B.: Computing stable eigendecompositions of matrix pencils. Linear Algebra Appl. **88**, 139–186 (1987)
53. Demmel, J.W., Kågström, B.: The generalized Schur decomposition of an arbitrary pencil $\lambda A - B$, Part I. ACM Trans. Math. Softw. **19**, 160–174 (1993)
54. Demmel, J.W., Kågström, B.: The generalized Schur decomposition of an arbitrary pencil $\lambda A - B$, Part II. ACM Trans. Math. Softw. **19**, 185–201 (1993)
55. Drmač, Z., Bujanović, Z.: On the failure of rank-revealing QR factorization software. ACM Trans. Math. Softw. **35**(2), Article 12 (2008)
56. Edelmann, A., Elmroth, E., Kågström, B.: A geometric approach to perturbation theory of matrices and matrix pencils. Part I: versal deformations. SIAM J. Matrix Anal. Appl. **18**, 653–692 (1997)
57. Edelmann, A., Elmroth, E., Kågström, B.: A geometric approach to perturbation theory of matrices and matrix pencils. Part II: a stratification-enhanced staircase algorithm. SIAM J. Matrix Anal. Appl. **20**, 667–699 (1999)
58. Eich-Soellner, E., Führer, C.: Numerical Methods in Multibody Dynamics. Teubner, Stuttgart (1998)

59. Fassbender, H.: Symplectic Methods for the Symplectic Eigenproblem. Kluwer Academic/Plenum, New York (2000)
60. Gantmacher, F.R.: The Theory of Matrices, vol. 2. Chelsea, New York (1959)
61. Gohberg, I., Lancaster, P., Rodman, L.: Matrix Polynomials. Academic, New York (1982)
62. Golub, G.H., Van Loan, C.F.: Matrix Computations, 3rd edn. Johns Hopkins University Press, Baltimore (1996)
63. Günther, M., Rentrop, P.: Multirate ROW methods and latency of electric circuits. Appl. Numer. Math. **13**(1–3), 83–102 (1993)
64. Hagel, S.: Systematische Modellbildung mechanischer Systeme auf der Basis von Teilkomponenten. In: Müller, P.C., et al. (eds.) Notes of the Workshop "Identifizierung, Analyse und Entwurfsmethoden für mechanische Mehrkörpersysteme in Deskriptorform", pp. 67–72. Institut für Sicherheitstechnik, BUGH Wuppertal (1994)
65. Hahn, H., Wehage, R.: Dynamic simulation of terrain vehicles. In: Schiehlen, W. (ed.) Multibody Systems Handbook, pp. 491–503. Springer, Berlin (1990)
66. Hassibi, B., Sayed, A.H., Kailath, T.: Indefinite-Quadratic Estimation and Control: A Unified Approach to H^2 and H^∞ Theories. SIAM Studies in Applied and Numerical Mathematics. SIAM, Philadelphia, PA (1999)
67. Hench, J.J., Laub, A.J.: Numerical solution of the discrete-time periodic Riccati equation. IEEE Trans. Autom. Control **39**(6), 1197–1210 (1994)
68. Hiller, M., Hirsch, K.: Multibody system dynamics and mechatronics. Z. Angew. Math. Mech. **86**(2), 87–109 (2006)
69. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, Cambridge (1985)
70. Kågström, B.: RGSVD – An algorithm for computing the Kronecker structure and reducing subspaces of singular $A - zB$ pencils. SIAM J. Sci. Stat. Comput. **7**(1), 185–211 (1986)
71. Kressner, D.: An efficient and reliable implementation of the periodic QZ algorithm. In: Proceedings of IFAC Workshop on Periodic Control Systems (2001)
72. Kressner, D., Schröder, C., Watkins, D.S.: Implicit QR algorithms for palindromic and even eigenvalue problems. Numer. Algorithms **51**(2), 209–238 (2009)
73. Kunkel, P., Mehrmann, V.: Generalized inverses of differential-algebraic operators. SIAM J. Matrix Anal. Appl. **17**(2), 426–442 (1996)
74. Kunkel, P., Mehrmann, V.: A new class of discretization methods for the solution of linear differential-algebraic equations. SIAM J. Numer. Anal. **5**, 1941–1961 (1996)
75. Kunkel, P., Mehrmann, V.: Differential-Algebraic Equations. Analysis and Numerical Solution. EMS Publishing House, Zürich (2006)
76. Kunkel, P., Mehrmann, V.: Optimal control for unstructured nonlinear differential-algebraic equations of arbitrary index. Math. Control Signals Syst. **20**(3), 227–269 (2008)
77. Kunkel, P., Mehrmann, V.: Formal adjoints of linear DAE operators and their role in optimal control. Electron. J. Linear Algebra **22**, 672–693 (2011)
78. Kunkel, P., Mehrmann, V., Rath, W.: Analysis and numerical solution of control problems in descriptor form. Math. Control Signals Syst. **14**, 29–61 (2001)
79. Kunkel, P., Mehrmann, V., Stöver, R.: Multiple shooting for unstructured nonlinear differential-algebraic equations of arbitrary index. SIAM J. Numer. Anal. **42**(6), 2277–2297 (2004)
80. Kunkel, P., Mehrmann, V., Stöver, R.: Symmetric collocation for unstructured nonlinear differential-algebraic equations of arbitrary index. Numer. Math. **98**(2), 277–304 (2004)
81. Kurina, G.A., März, R.: On linear-quadratic optimal control problems for time-varying descriptor systems. SIAM J. Control Optim. **42**(6), 2062–2077 (2004)
82. Lancaster, P.: Lambda-Matrices and Vibrating Systems. International Series of Monographs in Pure and Applied Mathematics, vol. 94. Pergamon Press, Oxford (1966)
83. Landwehr, M., Lefarth, U., Wassmuth, E.: Parameter identification and optimization of nonlinear dynamic systems, exemplified by mechatronic systems. In: Computational Systems Analysis, pp. 257–262. Elsevier, Amsterdam (1992)
84. Lin, W.W., Xu, S.F.: Convergence analysis of structure-preserving doubling algorithms for Riccati-type matrix equations. SIAM J. Matrix Anal. Appl. **28**(1), 26–39 (2006)

85. Losse, P.: The $_\infty$ optimal control problem for descriptor systems. Dissertation, Fakultät für Mathematik, Technische Universität Chemnitz. Available from http://nbn-resolving.de/urn:nbn:de:bsz:ch1-qucosa-83628 (2012)
86. Losse, P., Mehrmann, V., Poppe, L., Reis, T.: The modified optimal $\mathscr{H}_\infty$ control problem for descriptor systems. SIAM J. Control Optim. **47**(6), 2795–2811 (2008)
87. Mackey, D.S., Mackey, N., Mehl, C., Mehrmann, V.: Structured polynomial eigenvalue problems: good vibrations from good linearizations. SIAM J. Matrix Anal. Appl. **28**(4), 1029–1051 (2006)
88. Mackey, D.S., Mackey, N., Mehl, C., Mehrmann, V.: Jordan structures of alternating matrix polynomials. Linear Algebra Appl. **432**, 867–891 (2010)
89. Mavaddat, R.: Network Scattering Parameters. Advanced Series in Circuits and Systems. World Scientific, River Edge, NJ (1996)
90. Mehl, C.: Condensed forms for skew-Hamiltonian/Hamiltonian pencils. SIAM J. Matrix Anal. Appl. **21**(2), 454–476 (2000)
91. Mehrmann, V., Voss, H.: Nonlinear eigenvalue problems: a challenge for modern eigenvalue methods. GAMM-Mitt. **27**, 121–151 (2005)
92. Mehrmann, V., Watkins, D.: Structure-preserving methods for computing eigenpairs of large sparse skew-Hamiltonian/Hamiltonian pencils. SIAM J. Sci. Comput. **22**(6), 1905–1925 (2001)
93. Mehrmann, V., Xu, H.: Structure preserving deflation of infinite eigenvalues in structured pencils. Preprint 1052, DFG Research Center MATHEON, Berlin. Available from http://opus4.kobv.de/opus4-matheon (2014)
94. Mehrmann, V., Schröder, C., Watkins, D.S.: A new block method for computing the Hamiltonian Schur form. Linear Algebra Appl. **431**, 350–368 (2009)
95. Mehrmann, V., Schröder, C., Simoncini, V.: An implicitly-restarted Krylov method for real symmetric/skew-symmetric eigenproblems. Linear Algebra Appl. **436**(10), 4070–4087 (2012)
96. Reis, T.: Circuit synthesis of passive descriptor systems – a modified nodal approach. Int. J. Circuit Theory Appl. **38**(1), 44–68 (2010)
97. Reis, T., Stykel, T.: PABTEC: passivity-preserving balanced truncation for electrical circuits. IEEE Trans. Circuits Syst. Regul. Pap. **29**(9), 1354–1367 (2010)
98. Reis, T., Stykel, T.: Passivity-preserving balanced truncation model reduction of circuit equations. In: Roos, J., Costa, L. (eds.) Scientific Computing in Electrical Engineering SCEE 2008. Math. Ind., vol. 14, pp. 483–490. Springer, Berlin/Heidelberg (2010)
99. Reis, T., Rendel, O., Voigt, M.: The Kalman-Yakubovich-Popov inequality for differential-algebraic systems. Hamburger Beiträge zur Angewandten Mathematik 2014-27, Fachbereich Mathematik, Universität Hamburg. Available from http://preprint.math.uni-hamburg.de/public/papers/hbam/hbam2014-27.pdf (2014)
100. Riaza, R.: Differential-Algebraic Systems. Analytical Aspects and Circuit Applications. World Scientific, Singapore (2008)
101. Roberson, R.E., Schwertassek, R.: Dynamics of Multibody Systems. Springer, Heidelberg (1988)
102. Rosenbrock, H.H.: Structural properties of linear dynamical systems. Int. J. Control **20**(2), 191–202 (1974)
103. Scherer, C.: $\mathscr{H}_\infty$-control by state-feedback and fast algorithms for the computation of optimal $\mathscr{H}_\infty$-norms. IEEE Trans. Autom. Control **35**(10), 1090–1099 (1990)
104. Schiehlen, W. (ed.): Advanced Multibody System Dynamics – Simulation and Sofware Tools. Solid Mechanics and its Applications, vol. 20. Kluwer, Dordrecht (1993)
105. Schlacher, K., Kugi, A.: Automatic control of mechatronic systems. Int. J. Appl. Math. Comput. Sci. **11**(1), 131–164 (2001)
106. Schlacher, K., Kugi, A., Scheidl, R.: Tensor analysis based symbolic computation for mechatronic systems. Math. Comput. Simul. **46**(5–6), 517–525 (1998)

107. Schröder, C.: Palindromic and even eigenvalue problems – analysis and numerical methods. Dissertation, Institut für Mathematik, TU Berlin. Available from http://opus4.kobv.de/opus4-tuberlin/frontdoor/index/index/docId/1770 (2008)
108. Simeon, B., Grupp, F., Führer, C., Rentrop, P.: A nonlinear truck model and its treatment as multibody system. J. Comput. Appl. Math. **50**(1–3), 523–532 (1994)
109. Stewart, G.W., Sun, J.G.: Matrix Perturbation Theory. Academic, New York (1990)
110. Thompson, R.C.: Pencils of complex and real symmetric and skew matrices. Linear Algebra Appl. **147**, 323–371 (1991)
111. Tisseur, F., Meerbergen, K.: The quadratic eigenvalue problem. SIAM Rev. **43**(2), 235–286 (2001). doi:10.1137/S0036144500381988. http://dx.doi.org/10.1137/S0036144500381988
112. Trautenberg, W.: SIMPACK 8.9. Available from www.simpack.com
113. Van Dooren, P.: The computation of Kronecker's canonical form of a singular pencil. Linear Algebra Appl. **27**, 103–121 (1979)
114. Varga, A.: Computation of irreducible generalized state-space realizations. Kybernetica **26**(2), 89–106 (1990)
115. Verghese, G.C., Lévy, B.C., Kailath, T.: A generalized state-space for singular systems. IEEE Trans. Autom. Control **AC-26**(4), 811–831 (1981)
116. Voigt, M.: $\mathscr{L}_\infty$-norm computation for descriptor systems. Diploma thesis, Chemnitz University of Technology, Faculty of Mathematics. Available from http://nbn-resolving.de/urn:nbn:de:bsz:ch1-201001050 (2010)
117. Voigt, M.: On linear-quadratic optimal control and robustness of differential-algebriac systems. Dissertation, Otto-von-Guericke-Universität Magdeburg, Fakultät für Mathematik (2015).
118. Willems, J.C.: Least squares stationary optimal control and the algebraic Riccati equation. IEEE Trans. Autom. Control **AC-16**(6), 621–634 (1971)
119. Willems, J.C.: Dissipative dynamical systems – Part I: general theory. Arch. Ration. Mech. Anal. **45**, 321–351 (1972)
120. Willems, J.C.: Dissipative dynamical systems – Part II: linear systems with quadratic supply rates. Arch. Ration. Mech. Anal. **45**, 352–393 (1972)
121. Willems, J.C.: On the existence of a nonpositive solution to the algebraic Riccati equation. IEEE Trans. Autom. Control **AC-19**(10), 592–593 (1974)
122. Zhou, K., Doyle, J.C., Glover, K.: Robust and Optimal Control. Prentice-Hall, Upper Saddle River, NJ (1995)

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

René Lamour, Roswitha März, and Ewa Weinmüller

Contents

R. Lamour • R. März (✉)
Department of Mathematics, Humboldt-University of Berlin, 10099 Berlin, Germany
e-mail: lamour@math.hu-berlin.de; maerz@math.hu-berlin.de

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

A. Ilchmann, T. Reis (eds.), *Surveys in Differential-Algebraic Equations III*,
Differential-Algebraic Equations Forum, DOI 10.1007/978-3-319-22428-2_4

Abstract We provide an overview on the state of the art concerning boundary-value problems for differential-algebraic equations. A wide survey material is analyzed, in particular polynomial collocation and shooting methods. Moreover, new developments are presented such as the theory of linear boundary-value problems for arbitrary-index differential-algebraic equations as counterpart of the well-known classical version.

Keywords Boundary-value problems • Differential-algebraic equation • Numerical methods • Well-posedness

AMS Subject Classification (2010): 34A09, 65L80, 34B05, 34B15

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 for explicit ordinary differential equations (ODEs), e.g., [13]. This applies even more for DAEs. We follow [13] in mainly concentrating on two-point BVPs.

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

Optimal control is one of the traditional sources of BVPs for DAEs. As 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 [86, 103, 104], 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 from the well-known classical ODE theory, but we also indicate 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 determined. 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 $(x_1(t), x_2(t))$ is 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

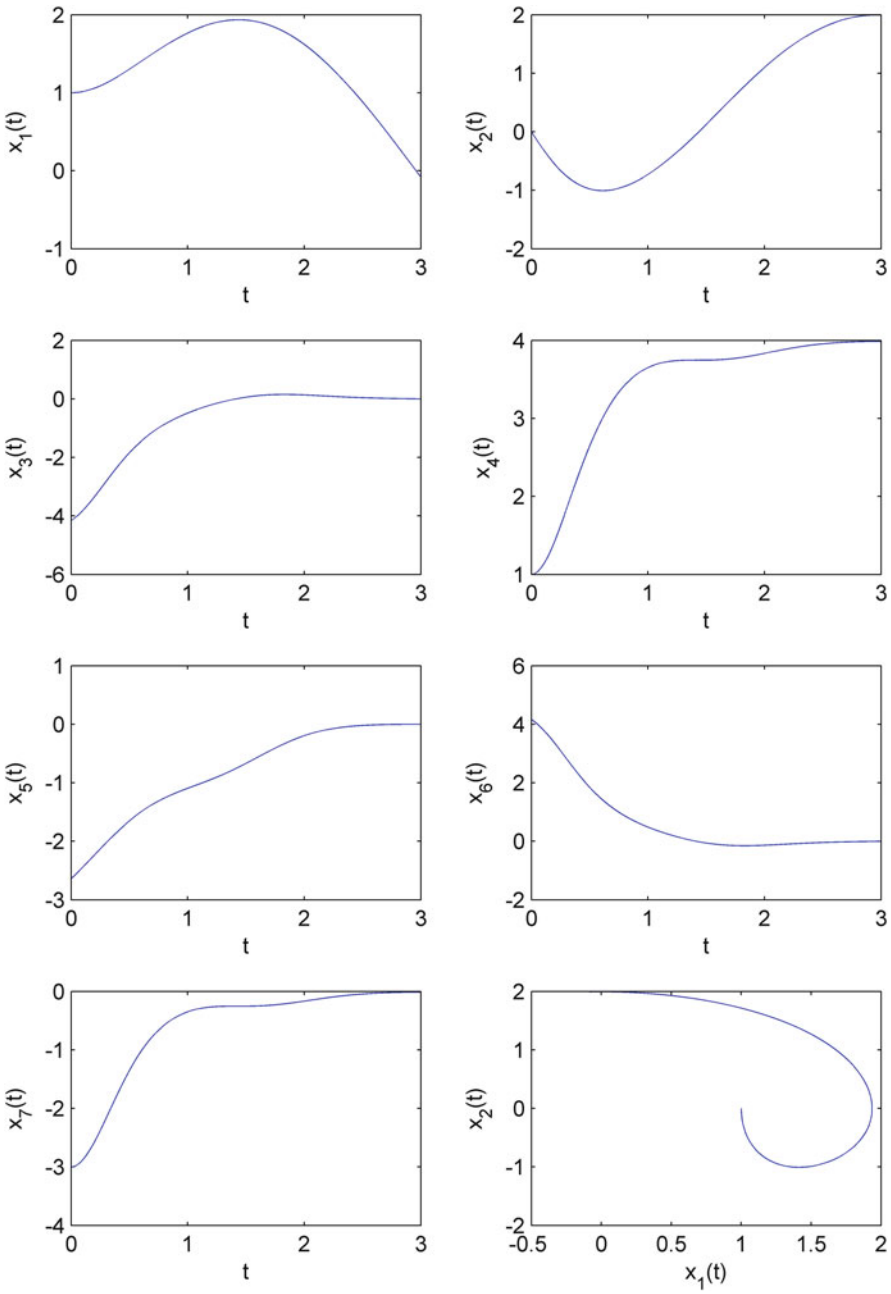

Fig. 1 Solution of the optimal BVP in Example 1.1

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 separate 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 additionally that $x_4(0) = 0$ is not a good idea.

We observe that any solutions 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 Fig. 2, where

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

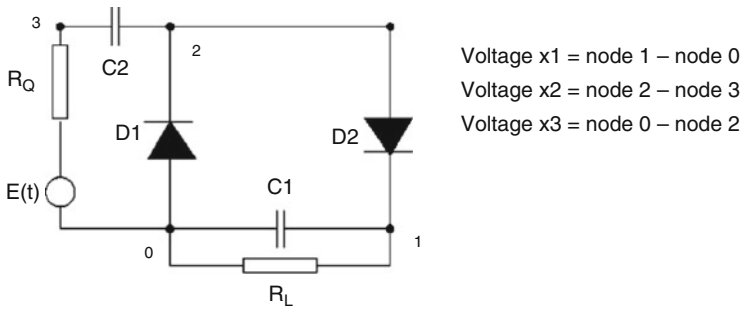

Fig. 2 Voltage doubling network in Example 1.2

and

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

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 is T-periodic. It is displayed in Fig. 3 and can only be provided numerically.

Replacing the above boundary condition by $x(0) = x(T)$ leads to a solvable BVP, but it is no longer well-posed because there are too many 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 of the DAE must reside in a restriction set, now in

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

Although 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 but vector fields on smooth manifolds is somewhat limited. Nevertheless, corresponding case studies are helpful to gain insights.

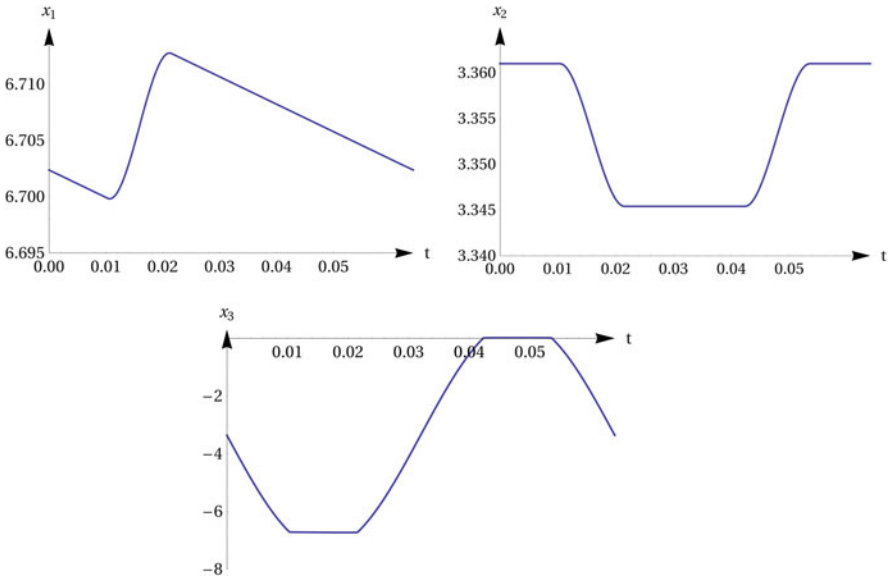

Fig. 3 T-periodic solution of the DAE in Example 1.2

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 conditions

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

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 turns out to be locally well-posed, no matter how α behaves, see Example 2.6.

Replacing the boundary conditions by the new ones

$$x_1(0) - x_1(2\pi) = 0,$$
$$x_2(0) - x_2(2\pi) = 0,$$

causes the situation to change. 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 observe a phenomenon known 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 also satisfies the BVPs and the unstable stationary solution

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

□

Example 1.4 The solutions of the DAE

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

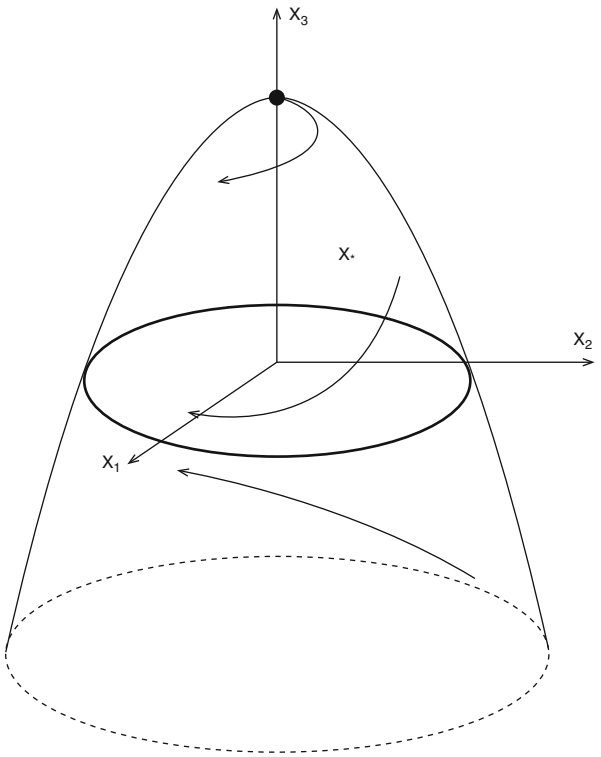

Fig. 4 Flow on the constraint set in Example 1.3 for identically vanishing α

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 $x_*(\cdot)$, 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

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

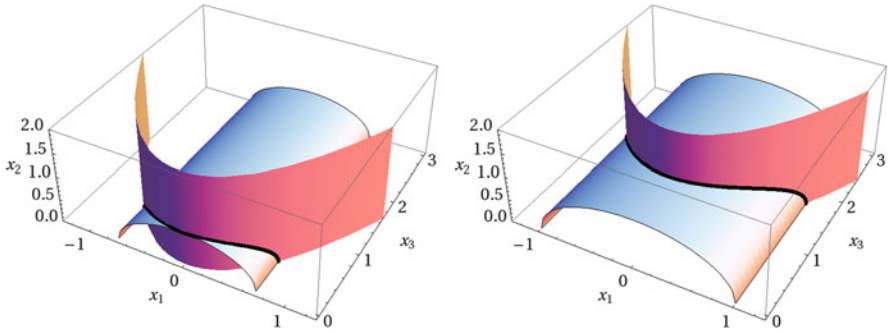

Fig. 5 Constraint set $\mathcal{M}_1$ at $t = 0$ and $t = \frac{1}{2}$ in Example 1.4

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 $t = \frac{1}{2}$.

Regarding the boundary condition

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

the BVP has 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 that it becomes well-posed in a special advanced setting, see Example 2.7. □

Usually, DAEs are given either in standard form

$$\mathfrak{f}(x'(t), x(t), t) = 0 \tag{1.1}$$

or in the advanced form

$$f((Dx)'(t), x(t), t) = 0, \tag{1.2}$$

with an extra matrix function D indicating which derivatives are actually involved. Most of the DAEs arising in applications originally show the latter form [35, 45, 103]. For large classes of DAEs of interest in the context of BVPs, for instance semi-explicit DAEs, Eq. (1.1) can be also written in the form (1.2) 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

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

In the present chapter we deal with DAEs of the form (1.2), which is more comfortable from the analytic point of view [86, 96]. Most results remain valid accordingly for the standard form (1.1).

The well-posed BVPs in Examples 1.1–1.3 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 for explicit ODEs, however, concerning perturbations of the right-hand side, the solution becomes discontinuous in the natural setting. We refer to the illustrative example [86, 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 [86]. 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 herein a twofold benefit of the projector-based analysis: it serves as an integrative framework for the wide survey material and, at the same time, as a source of new developments such as the linear BVP theory as a counterpart of the classical version in [13].

2 Analytical Theory

2.1 Basic Assumptions and Terminology

To tie in with the general discussion in [86, 96] we deal with DAEs of the form

$$f((Dx)'(t), x(t), t) = 0, \tag{2.1}$$

which exhibit 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

$$g(x(a), x(b)) = 0 \tag{2.2}$$

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 solutions of the DAE (2.1), that is, for functions from the function space

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

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 $x \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$, 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 $x \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ given on an interval $\mathcal{I} \ni t_0$ such that $x(t_0) = x_0$.

When dealing with BVPs (2.1), (2.2) we suppose the compact interval $\mathcal{I} = [a, b]$ and seek functions from $\mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ that satisfy the DAE (2.1) and, additionally, the boundary condition (2.2).

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

$$\ker f_y(y,x,t) \oplus \operatorname{im} D(t) = \mathbb{R}^n, \quad (y,x,t) \in \mathbb{R}^n \times D_f \times \mathcal{I}_f, \tag{2.3}$$

is valid.

Below, except for Sect. 3.4 on singular problems, we always assume the DAE (2.1) 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 (2.3) pointwise induces a projector matrix $R(t) \in \mathcal{L}(\mathbb{R}^n)$, the so-called *border projector*, such that

$$\operatorname{im} R(t) = \operatorname{im} D(t),\ \ker R(t) = \ker f_y(y,x,t),\ (y,x,t) \in \mathbb{R}^n \times \mathcal{D}_f \times \mathcal{I}_f. \tag{2.4}$$

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 [86, 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

$$D(t)D(t)^- D(t) = D(t),$$
$$D(t)^- D(t)D(t)^- = D(t)^-,$$
$$D(t)D(t)^- = R(t),$$
$$D(t)^- D(t) = P_0(t),$$

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 [86, 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,

$$\begin{aligned} x_1'(t) + k_1(x_1(t), x_2(t), t) &= 0, \\ k_2(x_1(t), x_2(t), t) &= 0, \end{aligned} \tag{2.5}$$

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 exists 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 (2.1), (2.2), $\mathcal{I} = [a, b]$. The BVP (2.1), (2.2) is said to be *well-posed locally* around x_* in its natural setting, if the slightly perturbed BVP

$$f((Dx)'(t), x(t), t) = q(t), \quad t \in \mathcal{I}, \tag{2.6}$$

$$g(x(a), x(b)) = \gamma, \tag{2.7}$$

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

$$\|x - x_*\|_{\mathcal{C}_D^1} \le \kappa(|\gamma| + \|q\|_\infty), \tag{2.8}$$

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

Instead of the inequality (2.8) one can use the somewhat simpler inequality

$$\|x - x_*\|_\infty \le \kappa(|\gamma| + \|q\|_\infty), \tag{2.9}$$

which is sometimes more convenient, see Remark 2.12.

The constant κ in the inequality (2.9) is called the *stability constant* of the BVP, e.g., in [10, 13, 51]. Here we do not use 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

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

it becomes evident that the linear BVP is well-posed if and only if $\mathcal{T}$ is bijective, and then κ in (2.8) is nothing but 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 $x_* \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ be a solution of the BVP (2.1), (2.2), $\mathcal{I} = [a, b]$. The BVP (2.1), (2.2) has *accurately stated boundary conditions* locally around x_* if the BVP with slightly perturbed boundary conditions

$$f((Dx)'(t), x(t), t) = 0, \tag{2.10}$$
$$g(x(a), x(b)) = \gamma, \tag{2.11}$$

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

$$\max_{t \in \mathcal{I}} |x(t) - x_*(t)| \le \kappa |\gamma|, \tag{2.12}$$

with a constant κ.

It is evident that x_* is locally the only solution of a BVP with accurately stated boundary conditions. On the 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. [13]), 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

$$\begin{aligned} x_1'(t) + x_3(t) &= 0, \\ x_2'(t) + x_3(t) &= 0, \\ x_2(t) - \sin(t - a) &= 0, \end{aligned} \tag{2.13}$$

and the different sets of boundary conditions

$$x_1(a) = 0, \qquad x_2(a) = 0, \qquad x_3(a) = 0, \tag{2.14}$$

$$x_1(a) = 0, \qquad x_2(a) = 0, \tag{2.15}$$

$$x_1(a) + \alpha x_2(a) + \beta x_3(a) = 0, \qquad \alpha, \beta \in \mathbb{R}, \tag{2.16}$$

$$x_2(a) = 0. \tag{2.17}$$

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 (2.13), (2.14) fails to be solvable, and the BVP (2.13), (2.17) is satisfied by all solutions with arbitrary c.

The BVP (2.13), (2.15) and the BVP (2.13), (2.16) 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 (2.13), (2.16) has accurately stated boundary conditions.

To check whether the BVP (2.13), (2.16) 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

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

cannot be estimated by an inequality (2.8). The BVP is ill-posed in its natural setting. □

Besides the original BVP (2.1), (2.2) we consider also the DAE linearized along the reference solution x_*,

$$A_*(t)(Dx)'(t) + B_*(t)x(t) = 0, \quad t \in \mathcal{I}, \tag{2.18}$$

with continuous coefficients

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

and the linearized boundary conditions

$$G_{*\,a}x(a) + G_{*\,b}x(b) = 0, \tag{2.19}$$

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 (2.18) inherits the properly stated leading term from the original DAE (2.1). The linearized BVP (2.18), (2.19) is said to be the *variational problem* for the original BVP (2.1), (2.2) at x_* (e.g., [13, 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. [13]).

Definition 2.4 A solution $x_* \in \mathcal{C}^1_D(\mathcal{I}, \mathbb{R}^m)$ of the BVP (2.1), (2.2) 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 is 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 (2.1), (2.2) is said to be *isolated* if the variational problem (2.18), (2.19) has the unique solution $x = 0$.

In the case of explicit ODEs, an isolated solution x_* of a BVP is locally unique and the BVP is well-posed if and only if the boundary conditions are accurately stated. The notion of isolatedness can be seen as a 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 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., [55, 90, 96, 111], cf. also Sect. 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

$$A(t)(Dx)'(t) + B(t)x(t) = q(t), \quad t \in \mathcal{I}, \tag{2.20}$$

which is regular with arbitrary tractability index $\mu \in \mathbb{N}$ in the sense of [86, 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 (2.1), then the products

$$\Pi_{can} := (I - \mathcal{H}_0)\Pi_{\mu-1}, \quad \Pi_{\mu-1} := P_0 \cdots P_{\mu-1}, \quad D\Pi_{can}D^- = D\Pi_{\mu-1}D^-, \tag{2.21}$$

with a coefficient $\mathcal{H}_0$ described in terms of the coefficients A, D, B in Appendix 6.1, are also projector-valued functions. In particular, Π_{can} has a special meaning independent of the choice of the corresponding factors (e.g., [86, Sect. 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., [86, Chap. 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

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

imply x to vanish identically.

The projector function Π_{can} is said to be the *canonical projector function* associated with the DAE (2.20). Π_{can} has constant rank; denote

$$l := \operatorname{rank} \Pi_{can}(t) = \operatorname{rank} \Pi_{\mu-1}(t), \quad t \in \mathcal{I}. \tag{2.22}$$

The rank l can be computed by means of the matrix function sequence supporting the regularity notion, see [86, Sect. 7.4] and 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 $\{AD, B\}$ [86, Sect. 1.4].

The canonical projector function depends strongly on the index. In particular, the canonical projector function of a regular index-1 DAE (2.20) 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 $\Pi_{can}(t)$ is often far from being symmetric, 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 then, if α 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 assume the DAE (2.20) 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 [86, Sect. 2.6]

$$A(t)(DX)'(t) + B(t)X(t) = 0, \quad t \in \mathcal{I}, \quad X(\bar{t}) = \Pi_{can}(\bar{t}). \tag{2.23}$$

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

$$A(t)(DX)'(t) + B(t)X(t) = 0, \quad t \in \mathcal{I}, \quad \Pi_{\mu-1}(\bar{t})(X(\bar{t}) - I) = 0, \tag{2.24}$$

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. [86]).

For DAEs, different kind of fundamental solution matrices make sense, in particular so-called *maximal size* and *minimal size* ones (cf. [28, 29, 86]). 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 ([86, Sect. 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., [86, Sect. 2.6])

$$\operatorname{im} X(t, \bar{t}) = \operatorname{im} \Pi_{can}(t), \quad \ker X(t, \bar{t}) = \ker \Pi_{can}(\bar{t}), \quad \operatorname{rank} X(t, \bar{t}) = l. \tag{2.25}$$

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

$$\begin{bmatrix} I_l & 0 \\ 0 & \mathcal{N} \end{bmatrix} x' + \begin{bmatrix} W & 0 \\ 0 & I_{m-l} \end{bmatrix} x = q, \tag{2.26}$$

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

$$X(t, \bar{t}) = \Pi_{can}(t)D(t)^{-}U(t, \bar{t})D(\bar{t})\Pi_{can}(\bar{t}), \tag{2.27}$$

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

$$u' - (D\Pi_{\mu-1}D^-)'u + D\Pi_{\mu-1}G_\mu^{-1}B_\mu D^- u = D\Pi_{\mu-1}G_\mu^{-1}q \tag{2.28}$$

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 and 6.2 below).

The generalized inverse $X(t,\bar{t})^-$ of $X(t,\bar{t})$ 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

$$X(t_1,t_2)^- = X(t_2,t_1), \quad X(t_1,t_2)X(t_2,t_3) = X(t_1,t_3). \tag{2.29}$$

The general solution of the DAE (2.20), with admissible right-hand side q, can now be expressed as

$$x(t) = X(t,\bar{t})c + x_q(t), \quad t \in \mathcal{I}, \tag{2.30}$$

whereby $x_q \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$ is the unique solution of the IVP [86, Theorem 2.52]

$$A(Dx)' + Bx = q, \quad \Pi_{can}(\bar{t})x(\bar{t}) = 0, \tag{2.31}$$

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

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

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 (2.31), which has a quite involved structure. Let q be admissible and the function u_q be the classical solution of the explicit ODE (2.28) that satisfies the initial condition $u(\bar{t}) = 0$. By means of fine decoupling projector functions we obtain the coefficients applied below ([86, Sect. 2.4], also Appendix 6.1.2) from the given coefficients A, D, B and then we

determine consecutively

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

Let us introduce further

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

We have $v_0 = \tilde{v}_0$ in the case of completely decoupling projector functions. We emphasize that to obtain $v_{\mu-2}$ 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 [86, Sect. 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

$$v_q := \tilde{v}_0 + v_1 + \cdots + v_{\mu-1}. \tag{2.32}$$

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 (2.20) reads

$$x(t) = X(t,\bar{t})c + \int_{\bar{t}}^{t} X(t,s)G_\mu^{-1}(s)q(s)\mathrm{d}s + v_q(t), \quad t \in \mathcal{I}, \tag{2.33}$$

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

$$x(\bar{t}) = \Pi_{can}(\bar{t})c + v_q(\bar{t}). \tag{2.34}$$

Comparing with the general solution of an explicit ODE, the first and second terms of the general DAE solution (2.33) 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 (2.20) is restricted to the time-varying affine subspace

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

which precisely consists of all consistent values at time t.

We recall that, in all higher-index cases, to obtain 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, see [86, 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 [86, 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 (2.20) (cf. [86, pp. 104–107] for the specification for semi-explicit systems).

Index-1 DAE: Let (2.20) 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

$$\begin{aligned}(Dx)' - R'Dx + DG_1^{-1}BD^-Dx &= DG_1^{-1}q,\\ Q_0x + Q_0G_1^{-1}BD^-Dx &= Q_0G_1^{-1}q,\end{aligned}$$

$$v_q = \tilde{v}_0 = Q_0 G_1^{-1} q,$$
$$x = (I - \mathcal{H}_0) D^- D x + Q_0 G_1^{-1} q.$$

We have here $u = Dx$, further $\mathcal{H}_0 = Q_0 G_1^{-1} B P_0$. 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 (2.20) 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 $\Pi_1 = P_0 P_1$, $B_1 := BP_0 - G_1 D^- (D\Pi_1 D^-)' D\Pi_0$, and $G_2 := G_1 + B_1 Q_1$. G_2 remains nonsingular. The DAE decoupling results in

$$(D\Pi_1 x)' - (D\Pi_1 D^-)' D\Pi_1 x + DG_2^{-1} B_1 D^- D\Pi_1 x = D\Pi_1 G_2^{-1} q,$$
$$v_1 = \Pi_0 Q_1 G_2^{-1} q,$$
$$\tilde{v}_0 = Q_0 P_1 G_2^{-1} q + Q_0 Q_1 D^- (D\Pi_0 Q_1 G_2^{-1} q)',$$
$$v_q = \tilde{v}_0 + v_1,$$
$$x = (I - \mathcal{H}_0) D^- D\Pi_1 x + v_q.$$

We have here $u = D\Pi_1 x$. The dynamical degree of freedom is $l = r_0 + r_1 - m$. The coupling coefficient $\mathcal{H}_0$ is now more elaborate,

$$\mathcal{H}_0 = Q_0 P_1 G_2^{-1} B \Pi_1 + Q_0 P_1 D^- (D\Pi_1 D^-)' D\Pi_1.$$

The DAE is solvable for precisely each arbitrary

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

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, $m_2 \le m_1$,

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

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

$$A(Dx)' + Bx = q, \quad G_a x(a) + G_b x(b) = \gamma. \tag{2.35}$$

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. [86, Sect. 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., [13]). We apply the general solution expression (2.33) with $\bar{t} = a$,

$$x(t) = X(t, a)c + \int_a^t X(t, s)G_\mu^{-1}(s)q(s)\mathrm{d}s + v_q(t), \quad t \in \mathcal{I}. \tag{2.36}$$

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

$$(G_a X(a,a) + G_b X(b,a))c = \hat{\gamma}, \tag{2.37}$$

$$\hat{\gamma} := \gamma - \gamma_q - G_b \int_a^b X(b,s) G_\mu(s)^{-1} q(s) \mathrm{d}s,$$

$$\gamma_q := G_a v_q(a) + G_b v_q(b).$$

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

$$S := G_a X(a,a) + G_b X(b,a) \tag{2.38}$$

actually plays the key role for solvability of the BVP. By construction, it holds that $\ker \Pi_{can}(a) \subseteq \ker S$. This fits the fact that the components $(I - \Pi_{can}(a))c$ do not matter at all 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

$$SS^-S = S, \quad S^-SS^- = S^-, \quad SS^- = I, \quad S^-S = \Pi_{can}(a), \tag{2.39}$$

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

$$\mathcal{G}(t,s) := \begin{cases} X(t,a)S^- G_a X(a,a) X(s,a)^-, & \text{if } t \geq s \\ -X(t,a)S^- G_b X(b,a) X(s,a)^-, & \text{if } t < s. \end{cases} \tag{2.40}$$

After the idea of conditioning constants for classical BVPs (e.g., [13]) 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 Sect. 2.2, the canonical projector function Π_{can} is an essential inherent feature of the DAE.

Theorem 2.1 *Let the DAE in (2.35) 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 statements hold:*

(1) *The BVP (2.35) 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*

$$\operatorname{im} [G_a\, G_b] = \mathbb{R}^l \quad \text{and} \quad \ker S = \ker \Pi_{can}(a) \tag{2.41}$$

are valid.

(2) *If (2.41) 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 $\bar{t} = a$ (2.23), the solvability matrix (2.38), Green's matrix function (2.40), the function v_q defined by (2.32), γ_q given in (2.37), and the matrix function G_μ constructed via Definition 6.1.

(3) *If (2.41) 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 (2.41) is satisfied, then the BVP (2.35) has accurately stated boundary conditions in the sense of Definition 2.3.*

(5) *Let (2.41) be satisfied. Then the BVP (2.35) is well-posed in its natural setting, if and only if $\mu = 1$, and ill-posed otherwise.*

We point out that the first condition in (2.41) 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 (2.41), 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)–(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 (2.41) serves as a criterion indicating whether the initial conditions are stated accurately.

Corollary 2.2 *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*

$$A(Dx)' + BX = q, \quad Cx(a) = \gamma \tag{2.42}$$

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

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 to let $\ker C = \ker \Pi_{can}(a)$ 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 (2.27) tempts us to consider the associated BVP induced for the IERODE (2.28).

We rewrite the solvability matrix S as

$$\begin{aligned} S &= G_a X(a,a) + G_b X(b,a) \\ &= G_a \Pi_{can}(a) D(a)^- U(a,a) D(a) \Pi_{can}(a) + G_b \Pi_{can}(b) D(b)^- U(b,a) D(a) \Pi_{can}(a) \\ &= \underbrace{\left(G_a \Pi_{can}(a) D(a)^- U(a,a) + G_b \Pi_{can}(b) D(b)^- U(b,a) \right)}_{=: S_{IERODE}} D(a) \Pi_{can}(a) \\ &=: S_{IERODE} D(a) \Pi_{can}(a). \end{aligned} \tag{2.43}$$

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 in

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

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

$$u' - (D\Pi_{\mu-1}D^-)'u + D\Pi_{\mu-1}G_\mu^{-1}B_\mu D^- u = D\Pi_{\mu-1}G_\mu^{-1}q, \tag{2.44}$$

$$C_a u(a) = 0, \tag{2.45}$$

$$G_a\Pi_{can}(a)D(a)^- u(a) + G_b\Pi_{can}(b)D(b)^- u(b) = \hat{\gamma} \tag{2.46}$$

is uniquely solvable and well-posed. This yields the further representation of the solution of the BVP (2.35), namely

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

with the solution u of the BVP (2.44)–(2.46). We summarize what we obtained in the next proposition.

Proposition 2.3 *Let the DAE in (2.35) 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 (2.35) 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 (2.44)–(2.46) 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 (2.44)–(2.46).

It is noteworthy that the IERODE (2.44) 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 $\Gamma_l \in \mathcal{C}^1(\mathcal{I}, \mathcal{L}(\mathbb{R}^n, \mathbb{R}^l))$ [86, Theorem 4.5]. Then, condition (2.45) becomes redundant and the boundary condition (2.46) transforms via $u = \Gamma_l^- \eta$.

In [102, Sect. 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 (2.44)–(2.46).

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. [13]) 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, although 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 the moderate size of κ_2. We are going to adapt these well-known classical results to the case of DAEs. The following lemma allows a useful scaling of the boundary conditions.

Lemma 2.4 *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

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

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 2.5 *Let the DAE in (2.35) 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 (2.41) 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 (2.23)), and S is the solvability matrix (2.38).

(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*

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

which leads to $\kappa_1 \leq 2\kappa_2$ 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

$$S = G_a X(a,a) + G_b X(b,a) = [G_a \Pi_{can}(a)\ G_b \Pi_{can}(b)] \begin{bmatrix} X(a,a) \\ X(b,a) \end{bmatrix} \tag{2.47}$$

makes clear that the factor $[B_a\, B_b] := [G_a \Pi_{can}(a)\ \ G_b \Pi_{can}(b)]$ also has full row-rank l. We apply Lemma 2.4 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 nonincreasing 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 [86, Definition 2.56].

Definition 2.6 The regular DAE (2.20) with index μ is said to be *dichotomic* if there are constants $K, \alpha, \beta \ge 0$ 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 $t, s \in \mathcal{I}$:

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

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 $P_{\phi,dich} = SP_{dich}S^- \in \mathcal{L}(\mathbb{R}^l)$:

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

Theorem 2.6 *Let the DAE in (2.35) 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 (2.41) be valid. Let the DAE be dichotomic and let the boundary condition be such that*

$$G_a\Pi_{can}(a)(I - P_{dich}) = 0, \quad G_b\Pi_{can}(b)P_{dich} = 0. \tag{2.48}$$

Then the Green's function satisfies the inequalities

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

Proof The conditions (2.48) 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

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

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

We emphasize that the concerns mentioned in [13] related to the fact that dichotomy of ODEs is thought for infinite intervals to feature the asymptotic flow behavior applied likewise for DAEs.

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

$$A(t)(Dx)'(t) + B(t)x(t) - q(t) = 0, \quad t \in \mathcal{I}, \tag{2.49}$$

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

$$Cx(a) = Cz, \quad z \in \mathbb{R}^m, \tag{2.50}$$

satisfies the condition $\ker C = \ker \Pi_{can}(a)$, then the solutions of all IVPs (2.49), (2.50) are given on the entire interval, e.g., by

$$x(t, z) = x_*(t) + X(t, a)(z - x_*(a)), \quad t \in \mathcal{I}. \tag{2.51}$$

The nonlinear regular DAE (2.1), that is,

$$f((Dx)'(t), x(t), t) = 0 \tag{2.52}$$

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

$$A_*(t)(Dx)'(t) + B_*(t)x(t) = 0, \quad t \in \mathcal{I}. \tag{2.53}$$

If the graph of the reference function x_* resides within an index-μ regularity region of the DAE, then the linear DAE (2.53) is also regular with index μ and shares with (2.52) further characteristics, see Appendix 6.1.3. We then denote by $\Pi_{*\,can}$ and $X_*(t, a)$ the canonical projector function associated with (2.53), and the fundamental solution matrix of (2.53) normalized by $X_*(a, a) = \Pi_{*\,can}(a)$.

After the idea of (2.51) we form the function

$$\tilde{x}(t, z) = x_*(t) + X_*(t, a)(z - x_*(a)), \quad t \in \mathcal{I}, \tag{2.54}$$

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

$$C_*\tilde{x}(a, z) = C_*z, \quad z \in \mathbb{R}^m, \tag{2.55}$$

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 (2.2), i.e.,

$$g(x(a), x(b)) = 0, \tag{2.56}$$

and introducing the solvability matrix of the linearized BVP,

$$S_* := G_{*\,a} X_*(a, a) + G_{*\,b} X_*(b, a), \tag{2.57}$$

one obtains 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 (2.55) 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, or 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 a sensitivity analysis. In the case of higher-index DAEs the hitherto known respective results are much too restrictive. We describe more details in the next two subsubsections.

As yet, there is a lack of precise general conditions ensuring the existence of solutions. 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 [74, 76]), 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 (2.52) 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. [86, 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 2.7 *Let $x_* \in \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m)$ satisfy the BVP (2.52), (2.56), $r := \operatorname{rank} D(a)$. 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 (2.52), 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 (2.52) and the BVP (2.52), (2.56) 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., [86, 89, 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 in

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

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 conditions 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 illustrated in Fig. 4, Example 1.3.

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 by

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

are 2π-periodic and satisfy the DAE. □

Theorem 2.7 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 [74, 76], 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 have been well understood for a long time. So the next perturbation results are nothing more than useful updates of [89, Theorem 4]. We refer to [86, Part II] for a recent elaborate exposition.

Theorem 2.8 *Let $x_* \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ satisfy the DAE (2.52), and the linearized along x_* DAE (2.53) be regular with index 1. Let $\Pi_{* can}$ denote the canonical projector function of the linear DAE (2.53). Let the matrix $C \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^l)$, $l = r$, be such that*

$$\ker C \cap \operatorname{im} \Pi_{* can}(a) = \{0\}. \tag{2.58}$$

Then the IVP

$$f((Dx)'(t), x(t), t) = 0, \quad t \in \mathcal{I}, \quad Cx(a) = Cz. \tag{2.59}$$

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 2.7. □

An *index-1 regularity region* $\mathcal{G}$ of the DAE (2.52) 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

$$\mathbb{R}^m = S(y, x, t) \oplus \ker D(t), \tag{2.60}$$
$$S(y, x, t) := \{z \in \mathbb{R}^m : f_x(y, x, t)z \in \operatorname{im} f_y(y, x, t)\},$$

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., [86, Part II]).

The decomposition (2.60) 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 (2.59) by a matrix C that has the same nullspace as $D(a)$ or choosing $C = D(a)$ makes condition (2.58) 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 2.8.

Corollary 2.9 *The assertions of Theorem 2.8 remain valid if the condition (2.58) is replaced by the easier condition*

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

For the further analysis the decoupled form (e.g., [55, 86])

$$u'(t) - R'(t)u(t) = D(t)\omega(u(t),t), \tag{2.61}$$

$$x(t) = D(t)^{-}u(t) + Q_0(t)\omega(u(t),t), \tag{2.62}$$

of the index-1 DAE (2.52) 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 $x_*(\cdot)$ 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 [86, 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 [86, Theorem 4.11]:

Theorem 2.10 *Given is the DAE (2.52) 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 $x_* \in C_D^1(\mathcal{I}_*,\mathbb{R}^m)$ 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 $x_* \in C_D^1(\mathcal{I}_*,\mathbb{R}^m)$ is locally unique.*

The solution x_* from Theorem 2.10 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 [86, Sect. 3.3].

2.5.2 BVPs Well-Posed in an Advanced Setting

By Theorem 2.7, 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 a quite straightforward generalization of index-2 results from [96, Sect. 4.3.3] and [86, Sect. 3.9] for the case of arbitrary higher index.

Definition 2.7 Let $x_* \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ be a solution of the BVP (2.1), (2.2), $\mathcal{I} = [a, b]$. 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 (2.1), (2.2) is said to be *well-posed in the advanced setting with image space* Y locally around x_*, if the slightly perturbed BVP

$$f((Dx)'(t), x(t), t) = q(t), \quad t \in \mathcal{I}, \tag{2.63}$$

$$g(x(a), x(b)) = \gamma, \tag{2.64}$$

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

$$\|x - x_*\|_{\mathcal{C}_D^1} \leq \kappa(|\gamma| + \|q\|_Y), \tag{2.65}$$

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

Instead of the inequality (2.65) one can use the somewhat simpler inequality

$$\|x - x_*\|_\infty \leq \kappa(|\gamma| + \|q\|_Y), \tag{2.66}$$

which is sometimes more convenient, see Remark 2.12.

Representing the linear BVP

$$A(t)(Dx)'(t) + B(t)x(t) = q(t), \ t \in \mathcal{I}, \quad G_a x(a) + G_b x(b) = \gamma, \tag{2.67}$$

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

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

$$T \in \mathcal{L}(\mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m), Y), \quad \mathcal{T} \in \mathcal{L}(\mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m), Y \times \mathbb{R}^l), \quad Y = \mathcal{C}^{ind\,\mu}(\mathcal{I}, \mathbb{R}^m),$$

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 (2.8) is nothing other than an upper bound of $\|\mathcal{T}^{-1}\|_{[Y \times \mathbb{R}^l \to \mathcal{C}_D^1]}$.

Theorem 2.11 *Let $x_* \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ satisfy the DAE (2.52). Let the linearized along x_* DAE (2.53) be regular with index μ, $\Pi_{*\,can}$ denote the canonical projector*

function of the linear DAE (2.53), 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*

$$x \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m),\ \|x - x_*\|_{\mathcal{C}_D^1} \le \rho \ \Rightarrow\ f(Dx)'(\cdot), x(\cdot), \cdot) \in Y_*. \tag{2.68}$$

Then the following assertions are valid:

(1) *Let* x_* *also satisfy the boundary condition (2.56). Then the BVP (2.53), (2.56) 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*

$$\ker C \cap \operatorname{im} \Pi_{*\,can}(a) = \{0\}. \tag{2.69}$$

Then the IVP

$$f((Dx)'(t), x(t), t) = 0, \quad t \in \mathcal{I}, \quad Cx(a) = Cz. \tag{2.70}$$

has a locally unique solution $x(\cdot; a, z) \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ *for each arbitrary* $z \in \mathbb{R}^m$ *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 $\mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$, *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 We again formulate the DAE (2.52) and the BVP (2.52), (2.56) as the operator equations $F(x) = 0$ and $\mathcal{F}(x) = 0$ in Banach spaces, this time, owing to condition (2.68), with definition domain $\operatorname{dom} F = \{x \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m) : \|x - x_*\|_{\mathcal{C}_D^1} < \rho\}$ and advanced image spaces,

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

Regarding condition (2.68) and Appendix 6.1.4, the operators F and $\mathcal{F}$ can be shown to be Fréchet-differentiable also in this setting, and

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

(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 a 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 2.7 excludes well-posedness in the natural setting.

The homogenous BVP linearized along x_*

$$\begin{bmatrix} 1 & 0 \\ 0 & x_{*2} \\ 0 & 0 \end{bmatrix} \Big(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} x \Big)' + \begin{bmatrix} 1 & 0 & 0 \\ 0 & x'_{*2} & 0 \\ 2x_{*1} & 2x_{*2} & 0 \end{bmatrix} x = 0,$$

$$x_1(0) - x_1(2) = 0,$$

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 $x \in \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^3)$ 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 2.11 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 2.11 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, Sect. 2].
(3) Condition (2.68) 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 Sect. 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 (2.49). The dynamical degree of freedom amounts to $l = r_0 + r_1 - m$. It becomes clear that the linear BVP (2.67) 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 (2.66) is obtained and the constant κ is called a *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 (2.53) 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 (2.52), we are aware of more transparent sufficient criteria for condition (2.68) to be valid. Namely, if the structural restriction (cf. [92], [86, Sect. 3.9.2])

$$W_{*0}(t)\{f(y,x,t) - f(0,P_0(t)x,t)\} \in \operatorname{im} W_{*0}(t)B_*(t)Q_0(t), \tag{2.71}$$

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

$$f(y,x,t) - f(0,P_0(t)x,t) \in \operatorname{im} G_{*1}(t), \tag{2.72}$$

is satisfied, then condition (2.68) is guaranteed. 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 instances of DAEs satisfying condition (2.71).

Formulating the initial condition in (2.70) by a matrix C that has the nullspace as $\ker C = \ker \Pi_{*\,can}(a) = \ker \Pi_{*\,\mu-1}(a)$ makes condition (2.69) trivially fulfilled. This means that the initial condition is directed promptly to the dynamical component and yields the following useful assertion.

Corollary 2.12 *The second assertion of Theorem 2.11 remains valid if the condition (2.69) is replaced by the simpler condition*

$$\ker C = \ker \Pi_{*\,\mu-1}(a) = \ker D(a) \oplus \ker G_{*\,1}(a). \tag{2.73}$$

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 2.9, in fact now also the matrix C in formula (2.73) depends on x_*. 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 [13], 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

$$l = m - \sum_{i=0}^{\mu-1}(m - r_i). \tag{2.74}$$

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. [86, Chap. 7], [38]).

In the present chapter we decide on mainly stating the boundary condition in $\mathbb{R}^l$ (following [8, 23, 38, 41], [13, 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., [4, 5, 55, 89, 90]) and so to emphasize that l has to be determined. Then, a consistency condition has to be respected. We address this topic in Sect. 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. [86]). In general, the canonical projector function is of great avail in theory, however, although 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 the integral condition by

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

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

Supposing a regular index-μ DAE

$$A(Dx)' + Bx = q \tag{2.75}$$

and applying the solution representation (2.30) 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 (2.75) has *accurately stated boundary condition* exactly if (cf. (2.41))

$$\operatorname{im} \Gamma = \mathbb{R}^l, \quad \ker S = \ker \Pi_{can}(a). \tag{2.76}$$

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. Sect. 2.5.1.

Nonlinear versions of those well-posed BVPs are treated, e.g., in [23, 41].

It might often be 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. [13]).

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

$$\begin{bmatrix} A & 0 \\ 0 & I \end{bmatrix} \left(\begin{bmatrix} D & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \right)' + \begin{bmatrix} B & 0 \\ -G & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} q \\ r \end{bmatrix}, \tag{2.77}$$

$$\begin{bmatrix} 0 & I \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x(a) \\ y(a) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} x(b) \\ y(b) \end{bmatrix} = \begin{bmatrix} \psi \\ \gamma \end{bmatrix}, \tag{2.78}$$

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 (2.76) 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 $[0, 1]$,

$$A_i \tfrac{1}{\eta_i - \eta_{i-1}} \tfrac{\mathrm{d}}{\mathrm{d}\tau}(D_i x_i) + B_i x_i = q_i, \quad i = 1, \ldots, s, \tag{2.79}$$

$$C_i(x_i(1) - x_{i+1}(0)) = 0,\ i = 1, \ldots, s-1, \quad \sum_{i=0}^{s-1} G_i x_{i+1}(0) + G_s x_s(1) = \gamma. \tag{2.80}$$

It is evident that the augmented DAE (2.79) is regular with index μ and the dynamical degree of freedom is sl, if the original DAE (2.75) 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 have the right number of boundary conditions. It is straightforward to prove that the boundary conditions (2.80) are accurately stated if the original BVP has accurately stated boundary condition, i.e., if

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

Replacing in (2.80) the matrices C_i by the identity $I \in \mathcal{L}(\mathbb{R}^m)$ and so requiring the $l + (s-1)m$ boundary conditions

$$x_i(1) = x_{i+1}(0), \quad i = 1, \ldots, s-1, \quad \sum_{i=0}^{s-1} G_i x_{i+1}(0) + G_s x_s(1) = \gamma. \tag{2.81}$$

leads to a consistent overdetermined problem.

BVPs for explicit ODEs with so-called *switching points* are discussed in [13]. In the case of DAEs, this corresponds in some sense to the BVP (2.79), (2.80) with unknown points $\eta_1, \ldots, \eta_{s-1}$. Up to 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

$$x(a) = x_a \in \mathbb{R}^m. \tag{2.82}$$

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 (2.75) with boundary conditions stated in $\mathbb{R}^m$,

$$\Gamma x = \gamma, \quad \Gamma : \mathcal{C}(\mathcal{I}, \mathbb{R}^m) \to \mathbb{R}^m, \tag{2.83}$$

whereby Γ is a linear bounded map describing initial, two-point boundary, multi-point boundary, and integral conditions as in Sect. 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)

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

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 (2.83) is said to be *accurately stated* if and only if

$$\operatorname{im} S = \operatorname{im} \Gamma, \quad \ker S = \ker \Pi_{can}(a). \tag{2.84}$$

Condition (2.84) requires $\operatorname{rank} S = l$ and $\dim \operatorname{im} \Gamma = l$. If the condition (2.84) is valid, then the operator $\mathcal{T}$ is a bijection between $\mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ 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 a 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

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

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 for $\mathbb{R}^l$.

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 [55, 88, 89, 91]. Linear and nonlinear multipoint BVPs for index-1 DAEs are studied in [4–6]. Recall that in several early papers after [55] one speaks of *transferable* DAEs instead of (regular) index-1 DAEs. In [4–6], the BVP for a transferable DAE is said to be *regular* if the condition (2.84) is satisfied, and *irregular* otherwise. We do not use 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 (2.84) no longer holds, 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 $\mathrm{ind}_{fredholm} = l - m = -(m - r_0) = -\dim\ker D(a)$. A representation of the general solution of such a BVP including the resulting consistency condition is developed in [4] by means of projectors onto $\ker \mathcal{T}$ and $\mathrm{im}\,\mathcal{T}$.

We emphasize that for higher-index DAEs this approach no longer works since then $\mathrm{im}\,\mathcal{T}$ 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 $Cz =: \gamma \in \mathrm{im}\,C$. The component of z belonging to the nullspace of C does not impact the solution of the IVP.

The condition (2.84) simplifies here to $\ker C \cap \mathrm{im}\,\Pi_{can}(a) = \{0\}$. 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

$$G_a x(a) + G_b x(b) = \gamma \tag{2.85}$$

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 2.6), 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. Sect. 5.2) can be applied.

If the boundary condition (2.85) 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 [13, Sect. 1.1]. For this one can utilize if either G_a or G_b is rank deficient.

Consider the BVP with boundary condition (2.85) 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

$$A(Dx)' + Bx = q. \tag{2.86}$$

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

$$\begin{bmatrix} A & 0 \\ 0 & I \end{bmatrix} \left(\begin{bmatrix} D & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} \right)' + \begin{bmatrix} B & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} = \hat{q} \tag{2.87}$$

is regular with index μ, too. The dynamical degree of freedom is $\hat{l} = l + l - s$. State for (2.87) separated boundary conditions

$$\begin{bmatrix} G_a & K \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x(a) \\ z(a) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ G_{b,2} & I \end{bmatrix} \begin{bmatrix} x(b) \\ z(b) \end{bmatrix} = \hat{\gamma}, \quad K = \begin{bmatrix} 0 \\ -I \end{bmatrix} \in \mathcal{L}(\mathbb{R}^{l-s}, \mathbb{R}^l). \tag{2.88}$$

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 (2.88) 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, the x-component of the solution of the BVP (2.87), (2.88) reproduces the solution of the original BVP (2.85), (2.86). If the boundary conditions of the original BVP are stated accurately, then so are the boundary conditions 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 (2.85), (2.86) to the augmented BVP on the half interval $[a, \frac{a+b}{2}]$,

$$\begin{bmatrix} A(t) & 0 \\ 0 & A(a+b-t) \end{bmatrix} \left(\begin{bmatrix} D(t) & 0 \\ 0 & D(a+b-t) \end{bmatrix} \hat{x}(t) \right)' + \begin{bmatrix} B & 0 \\ 0 & 0 \end{bmatrix} = \hat{q} \tag{2.89}$$

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

$$\begin{bmatrix} G_a & G_b \\ 0 & 0 \end{bmatrix} \hat{x}(a) + \begin{bmatrix} 0 & 0 \\ C_{\frac{a+b}{2}} & -C_{\frac{a+b}{2}} \end{bmatrix} \hat{x}(b) = \begin{bmatrix} \gamma \\ 0 \end{bmatrix}, \tag{2.90}$$

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

$$\hat{S} = \begin{bmatrix} G_a X(a,a) & G_b X(b,a) \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ C_{\frac{a+b}{2}} X(\frac{a+b}{2}, a) & -C_{\frac{a+b}{2}} X(\frac{a+b}{2}, a) \end{bmatrix} \in \mathcal{L}(\mathbb{R}^{2m}, \mathbb{R}^{2l}). \tag{2.91}$$

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 solutions. *Occasional* additional smoothness requirements concerning the data imply each existing $\mathcal{C}_D^1$-solution also belongs to class $\mathcal{C}^1$.

The *axiomatic* use of $\mathcal{C}^1$-solutions, e.g., in [74, 76], 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, one has to assume that B_{21}, B_{22} as well as q_2 are continuously differentiable. Therefore, in this approach, q_2 cannot serve as a control function, being only continuous.

Remark 2.2 (The Class of DAEs) Relations between DAEs in standard form (1.1) and DAEs showing a properly involved derivative (1.2) have been discussed at great length in [86, 96]. The setting with properly involved derivative indicates solutions from $\mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$. We emphasize that these are classical solutions; they satisfy the DAE at all points $t \in \mathcal{I}$. The present chapter is concerned with 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 [54, 56, 86, 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 [46, 57–59].

Remark 2.4 (Fundamental Solution Matrices) Given is a regular linear DAE (2.20) 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 [86] 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

$$\ker X(\bar{t}) \cap \operatorname{im} C = \{0\}, \tag{2.92}$$

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 (2.92) 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. To get one that is normalized at $\bar{t}$, we choose the special $C = X(\bar{t})^-$ which is the generalized inverse of $X(\bar{t})$ defined by

$$X(\bar{t})X(\bar{t})^-X(\bar{t}) = X(\bar{t})^-, \qquad X(\bar{t})^-X(\bar{t})X(\bar{t})^- = X(\bar{t})^-,$$
$$X(\bar{t})X(\bar{t})^- = \Pi_{can}(\bar{t}), \qquad X(\bar{t})^-X(\bar{t}) = I.$$

In fact, we have then $X(\bar{t})C = X(\bar{t})X(\bar{t})^- = \Pi_{can}(\bar{t})$.

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 [26, 28, 29]. 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 the *shooting matrix*. The shooting approach by maximal fundamental solution matrices to obtain solvability results is already applied for nonlinear index-1 DAEs in [55, 89] 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 [86], which is straightforward within this framework. We also address nonlinear DAEs.

Supposing in essence the solution structure (2.30) 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 solvability 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.

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 [76] well-posedness of BVPs for index-1 DAEs in reduced form (2.98), (2.99) 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 is 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 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.

Up to 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 Sect. 2.5, [86, 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 a locally unique solution of the original BVP. Unfortunately, this fine idea is not so easy to predicate on precise criteria. With the notion of a *regular solution of the BVP* the authors of [76] attempt to provide such a criterion. We take a closer look.

In [74, 76], nonlinear BVPs

$$\mathfrak{f}(x'(t), x(t), t) = 0, \quad t \in \mathcal{I} = [a, b], \tag{2.93}$$

$$g(x(a), x(b)) = 0, \tag{2.94}$$

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

$$\mathfrak{f}(x'_*(t), x_*(t), t) = 0, \quad t \in \mathcal{I}, \tag{2.95}$$

$$\mathfrak{f}_\mu(\mathcal{P}_*(t), x_*(t), t) = 0, \quad t \in \mathcal{I}, \tag{2.96}$$

$$g(x_*(a), x_*(b)) = 0, \tag{2.97}$$

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

$$\hat{\mathfrak{f}}_1(x'(t), x(t), t) = 0, \tag{2.98}$$

$$\hat{\mathfrak{f}}_2(x(t), t) = 0, \quad t \in \mathcal{I}, \tag{2.99}$$

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* [76], 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 [76] that *the original BVP* (2.93), (2.94) *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 Sects. 2.2 and 2.3.

Remark 2.8 (Segregation of Solution Subspaces by Means of the Adjoint Equation) Similarly as 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, 30, 100, 101]. 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 (2.49) 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 $\mathfrak{L} \subset C_D^1(\mathcal{I}, \mathbb{R}^m)$ is a k-dimensional affine-linear subspace of solutions of the DAE if and only if it is described by

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

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

$$-D^*(A^*Y)' + B^*Y = 0,$$
$$v' + Y^*q = 0.$$

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

Remark 2.9 (Conditioning Constants) In [13], dealing with BVPs for explicit ODEs, the constant κ in the inequality (2.9) and the quantities κ_1, κ_2 introduced in Sect. 2.3 are called *conditioning constants*. If they have moderate size, then one speaks of *well-conditioned* BVPs.

Our presentations in Sects. 2.3 and 2.4 are generalizations of the results in [87, 88] by means of the projector-based analysis from [86].

It should be emphasized once again that the conditioning constants for index-1 DAEs and higher-index DAEs have essentially different meanings. For index-1 DAEs, the quantities $\kappa_1, \kappa_2, \kappa_3$ can be seen as specifications 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 exists, but $\kappa_1, \kappa_2, \kappa_3$ 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, and their sensitivity with respect to several small parameters are described in [114]. The conditioning of BVPs for index-2 Hessenberg systems is addressed 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 2.11) Consider the nonlinear DAE (2.52). If the reference solution x_* resides within an index-μ regularity region, then the linearized along x_* DAE (2.53) 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 (2.53) 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 (2.53) is regular with index $\mu \geq 2$, but there is no regularity region housing the graph of x_*, e.g., [86, 95].

If $\mu = 2$, then the additional condition (2.68) 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 could objectionable condition (2.68) be replaced by the requirement that x_* resides within a regularity region? Up to now, no answer 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 Sect. 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 solution in sight.

Remark 2.12 (Inequalities (2.9) and (2.8)) In the context of BVPs for explicit ODEs (e.g., [13]), 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 (2.8) by the simpler inequality (2.9) by the following arguments: The inequality (2.8) immediately implies (2.9), that is,

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

Conversely, (2.8) follows from (2.9). Namely, for $x \in \mathcal{B}_{\mathcal{C}_D^1}(x_*, \rho)$, the identities

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

imply

$$A_{[x,x_*]}(t)(Dx - Dx_*)'(t) + B_{[x,x_*]}(t)(x(t) - x_*(t)) = q(t), \quad t \in \mathcal{I}, \tag{2.100}$$

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 (2.100) 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 (2.9) we finally obtain

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

The same arguments also apply 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 [12] and its later modification COLNEW [13, 21] can be used to solve multipoint BVPs 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 done in [62] for semi-explicit DAEs

$$x_1'(t) + k_1(x_1(t), x_2(t), t) = 0, \tag{3.1}$$

$$k_2(x_1(t), x_2(t), t) = 0, \tag{3.2}$$

with index 1. The package COLDAE [11] also uses this approach, but for a wider class of DAEs. The MATLAB code BVPSUITE [16] is designed to solve systems of implicit ODEs of arbitrary order including order zero, which includes an implicit version of (3.1).

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

As pointed out in Sect. 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,

$$\mathfrak{f}_1(x'(t), x(t), t) = 0, \tag{3.3}$$

$$\mathfrak{f}_2(x(t), t) = 0, \tag{3.4}$$

which inter alia arises by reduction from derivative array systems (e.g., [76]). In addition to the regular DAEs we also consider 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 (3.1), (3.2) indicates the different smoothness of the first and second components, which can be reasonably resumed for their approximations (e.g., [11, 23, 42, 62, 73]). A useful generalization of this class of DAEs is given by DAEs with properly involved derivatives

$$f((Dx)'(t), x(t), t) = 0, \tag{3.5}$$

which satisfy the basic assumption from Sect. 2.1, and, additionally,

$$\operatorname{im} D(t) = \mathbb{R}^n, \quad t \in [a, b] = \mathcal{I}, \tag{3.6}$$

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

$$f(u'(t), x(t), t) = 0, \tag{3.7}$$

$$u(t) - D(t)x(t) = 0 \tag{3.8}$$

features partitioned variables. For each solution x_* of (3.5), 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 (3.5).

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

$$u'(t) = D(t)\omega(u(t), t) \tag{3.9}$$

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

With the boundary condition

$$g(x(a), x(b) = 0, \tag{3.10}$$

a well-posed BVP (3.5), (3.10), yields a well-posed BVP (3.7), (3.8), (3.10), and vice versa.

As pointed out in [61], [86, Chap. 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 step sizes $h_i := t_{i+1} - t_i$, $i = 0, \ldots, N-1$. 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

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

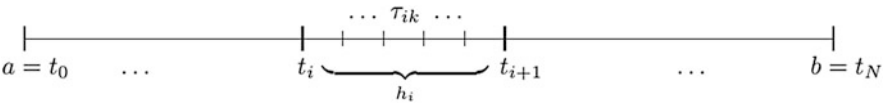

Fig. 6 The computational grid

A grid with equidistant interior collocation points is illustrated in Fig. 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 $\leq s$. To be precise, we agree upon right continuity at the mesh points $t_0, \ldots, t_{N-1}$.

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

$$\int_0^1 t^j \prod_{i=1}^{s} (t - \rho_i) \mathrm{d}t = 0, \quad j = 0, \ldots, s_+ - s, \tag{3.11}$$

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 are a variety of possible collocation approaches for DAEs. As emphasized in [11], collocating the differential components by continuous piecewise polynomial functions and allowing generally discontinuous piecewise polynomial functions for the algebraic components is most natural, see Sects. 3.1.1 and 3.3. Alternative approaches suppose continuous (Approach A in Sect. 3.1.2, Approach C in Sect. 3.1.4, and Sect. 3.2) or discontinuous (Approach B in Sect. 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, which means $x_\pi(t_i) \in \mathcal{M}_0(t_i)$.

3.1 BVPs Well-Posed in the Natural Setting

Let the BVP (3.5), (3.10), satisfy the basic assumptions described in Sect. 2.1, let the DAE have a properly involved derivative, and let (3.6) be valid. Let x_* denote the required solution, and $u_* = Dx_*$.

Theorem 2.7 provides precise 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 (3.5), (3.10) by means of the enlarged version (3.7), (3.8), (3.10). 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

$$u_\pi(t_i^-) = u_\pi(t_i), \quad i = 1, \ldots, N-1, \tag{3.12}$$

and therefore, we have to determine $n(s+1)N + msN - n(N-1) = (n+m)sN + n$ remaining unknowns. The boundary condition (3.10) yields

$$g(x_\pi(a), x_\pi(b)) = \gamma, \tag{3.13}$$

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

$$f(u'_\pi(\tau_{ik}), x_\pi(\tau_{ik}), \tau_{ik}) = 0, \tag{3.14}$$

$$u_\pi(\tau_{ik}) - D(\tau_{ik}) x_\pi(\tau_{ik}) = 0, \quad k = 1, \ldots, s,\ i = 0, \ldots, N-1. \tag{3.15}$$

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 the left derivative at $\tau_{is} = t_{i+1}$.

By means of the decoupling function the scheme (3.14), (3.15) transforms to

$$x_\pi(\tau_{ik}) = D(\tau_{ik})^- u_\pi(\tau_{ik}) + Q_0(\tau_{ik}) \omega(u_\pi(\tau_{ik}), \tau_{ik}), \tag{3.16}$$

$$u'_\pi(\tau_{ik}) = D(\tau_{ik}) \omega(u_\pi(\tau_{ik}), \tau_{ik}), \quad k = 1, \ldots, s,\ i = 0, \ldots, N-1. \tag{3.17}$$

On the other hand, we are given the solution representation (cf. (2.61), (2.62))

$$x_*(t) = D(t)^- u_*(t) + Q_0(t) \omega(u_*(t), t), \tag{3.18}$$

$$u'_*(t) = D(t) \omega(u_*(t), t), \quad t \in [a, b]. \tag{3.19}$$

In particular, u_* satisfies the IERODE (3.9). Obviously, the collocation scheme (3.14), (3.15), (3.12) 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 (3.16).

The next theorem represents a straightforward extension of [23, Theorem 3.2] which concerns semi-explicit index-1 DAEs. It can be proved analogously.

Theorem 3.1 *Let the BVP (3.5), (3.10) be well-posed locally around its solution x_* in the natural setting. Let condition (3.6) hold and the data of the DAE be sufficiently smooth for respective order conditions.*

Then, for the collocation scheme (3.14), (3.15), (3.13), (3.12), 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 (3.11), 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 $\mathcal{M}_0(t_i)$, means which 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 (3.5), (3.10) by means of the enlarged version (3.7), (3.8), (3.10). 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

$$u_\pi(t_i^-) = u_\pi(t_i), \quad x_\pi(t_i^-) = x_\pi(t_i), \quad i = 1, \dots, N-1, \tag{3.20}$$

and we have to determine $(n+m)(s+1)N-(n+m)(N-1)=(n+m)(sN+1)$ further coefficients. The boundary condition (3.10) contains n equations. We now apply the $(n+m)sN$ collocation equations and the boundary conditions

$$f(u'_\pi(\tau_{ik}),x_\pi(\tau_{ik}),\tau_{ik})=0, \tag{3.21}$$

$$u_\pi(\tau_{ik})-D(\tau_{ik})x_\pi(\tau_{ik})=0,\quad k=1,\dots,s,\ i=0,\dots,N-1, \tag{3.22}$$

$$g(x_\pi(a),x_\pi(b))=\gamma, \tag{3.23}$$

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:

$$D(a)x_\pi(a)-u_\pi(a)=0,\quad \tilde{W}(u'_\pi(a),x_\pi(a),a)f(u'_\pi(a),x_\pi(a),a)=0. \tag{3.24}$$

Observe that $\rho_1=0$ would lead to $\tau_{01}=t_0=a$ and cause the second part of the consistency condition (3.24) and the collocation (3.21) for $i=0,k=1$ to become redundant.

If the DAE is given with separated derivative-free equations

$$f_1((D(t)x(t))',x(t),t)=0,$$
$$f_2(x(t),t)=0,$$

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, Eqs. (3.21), (3.22) can be decoupled,

$$x_\pi(\tau_{ik})=D(\tau_{ik})^-u_\pi(\tau_{ik})+Q_0(\tau_{ik})\omega(u_\pi(\tau_{ik}),\tau_{ik}), \tag{3.25}$$

$$u'_\pi(\tau_{ik})=D(\tau_{ik})\omega(u_\pi(\tau_{ik}),\tau_{ik}),\quad k=1,\dots,s,\ i=0,\dots,N-1. \tag{3.26}$$

Therefore, the related equations from (3.21), (3.22), (3.20), (3.23) 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 (3.25), (3.24) together with the continuity conditions (3.20).

The following theorem is a byproduct of the investigations in [43, 72] 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 (3.24) 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 (3.21), (3.22), (3.23), (3.20), (3.24) 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 that*

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

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 (3.11) 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 reported that the error behaves consistently as $O(h^2)$. Tables 1, 2, 3, 4, 5, and 6 show the results generated by the collocation scheme (3.20)–(3.23), for $s = 1, 2, 3$, each with uniform and Gauss collocation. gex_π and geu_π denote the maximal global errors in the mesh points, and gex_{unif} and geu_{unif} are discrete maxima taken over 1000 equidistributed points. □

Table 1 Example 3.1, $s = 1$, uniform collocation

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 2 Example 3.1, $s = 1$, Gauß collocation

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 3 Example 3.1, $s = 2$, uniform collocation

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 4 Example 3.1, $s = 2$, Gauß collocation

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 5 Example 3.1, $s = 3$, uniform collocation

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 6 Example 3.1, $s = 3$, Gauß collocation

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

3.1.3 Uniform Approach B

Any regular index-1 DAE in standard form

$$E(t)x'(t) + F(t)x(t) = q(t) \tag{3.27}$$

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

$$A(t)(Dx)'(t) + B(t)x(t) = q(t) \tag{3.28}$$

by means of a proper factorization $E = AD$, and $B = E - AD'$. The BVP for (3.27) and the boundary condition

$$G_a x(a) + G_b x(b) = \gamma \tag{3.29}$$

is well-posed in the natural setting exactly if this is the case for the BVP (3.28), (3.29).

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

$$A(\tau_{ik})(Dx_\pi)'(\tau_{ik}) + B(\tau_{ik})x_\pi(\tau_{ik}) = q(\tau_{ik}), \quad k = 1, \ldots, s,\ i = 0, \ldots, N-1, \tag{3.30}$$

$$D(t_i)(x_\pi(t_i^-) - x_\pi(t_i)) = 0, \quad i = 1, \ldots, N-1, \tag{3.31}$$

$$x_\pi(t_i) \in \mathcal{M}_0(t_i), \quad i = 0, \ldots, N-1, \tag{3.32}$$

$$G_a x_\pi(a) + G_b x_\pi(b) = \gamma, \tag{3.33}$$

which consists of the usual *smN* collocation conditions (3.30), $(N-1)n$ continuity conditions applying only to the component Dx_π which approximates the smooth solution component Dx_*, further, $N(m-n)$ consistency conditions (3.32), 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 (3.30) already contains the condition $x_\pi(\tau_{01}) = x_\pi(t_0) \in \mathcal{M}_0(t_0)$ and the equation (3.32) 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 $x_\pi(t_1), \ldots, x_\pi(t_N)$ are consistent by construction. In the case of $\rho_s = 1$, in particular for Radau IIa, x_π is continuous in $t_1, \ldots, t_N$.

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

$$E(\tau_{ik})x'_\pi(\tau_{ik}) + F(\tau_{ik})x_\pi(\tau_{ik}) = q(\tau_{ik}), \qquad k = 1, \ldots, s,\ i = 0, \ldots, N-1, \tag{3.34}$$

$$E_1(t_i)(x_\pi(t_i^-) - x_\pi(t_i)) = 0, \qquad i = 1, \ldots, N-1, \tag{3.35}$$

$$F_2(t_i)x_\pi(t_i) - q_2(t_i) = 0, \qquad i = 0, \ldots, N-1, \tag{3.36}$$

$$G_a x_\pi(a) + G_b x_\pi(b) = \gamma, \tag{3.37}$$

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, Eqs. (3.31)–(3.33) coincide with (3.35)–(3.37), 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 (3.30) and (3.34) coincide. The next theorem is a consequence of [23, Theorem 5.11].

Theorem 3.3 *Let the linear BVP (3.27), (3.29) 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 (3.14)–(3.37), 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 that*

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

(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 [23] 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' [23, 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:

$$\mathfrak{f}(x'_\pi(\tau_{ik}), x_\pi(\tau_{ik}), \tau_{ik}) = 0, \quad k = 1, \dots, s, \quad i = 0, \dots, N-1, \tag{3.38}$$

$$E_1(x'_\pi(\tau_{ik}), x_\pi(\tau_{ik}), \tau_{ik})(x_\pi(t_i^-) - x_\pi(t_i)) = 0, \quad i = 1, \dots, N-1, \tag{3.39}$$

$$F_2(x_\pi(\tau_{ik}), \tau_{ik}) = 0, \quad i = 0, \dots, N-1, \tag{3.40}$$

$$g(x_\pi(a), x_\pi(b)) = 0. \tag{3.41}$$

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 retained, see [23, 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 (3.27), (3.29) by a continuous piecewise polynomial function $x_\pi \in \mathcal{B}^m_{\pi,s} \cap \mathcal{C}(\mathcal{I}, \mathbb{R}^m)$ using the system

$$E(\tau_{ik})x'_\pi(\tau_{ik}) + F(\tau_{ik})x_\pi(\tau_{ik}) = q(\tau_{ik}), \qquad k = 1, \ldots, s,\ i = 0, \ldots, N-1, \tag{3.42}$$

$$x_\pi(t_i^-) - x_\pi(t_i) = 0, \qquad i = 1, \ldots, N-1, \tag{3.43}$$

$$F_2(a)x_\pi(a) - q_2(a) = 0, \tag{3.44}$$

$$G_a x_\pi(a) + G_b x_\pi(b) = \gamma, \tag{3.45}$$

or, equivalently (cf. Sect. 3.1.3), by

$$A(\tau_{ik})(Dx_\pi)'(\tau_{ik}) + B(\tau_{ik})x_\pi(\tau_{ik}) = q(\tau_{ik}), \qquad k = 1, \ldots, s,\ i = 0, \ldots, N-1, \tag{3.46}$$

$$x_\pi(t_i^-) - x_\pi(t_i) = 0, \qquad i = 1, \ldots, N-1, \tag{3.47}$$

$$x_\pi(a) \in \mathcal{M}_0(a), \tag{3.48}$$

$$G_a x_\pi(a) + G_b x_\pi(b) = \gamma. \tag{3.49}$$

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 (3.27), (3.29) be well-posed in the natural setting. Let E and F be twice continuously differentiable.*

Then, for the collocation scheme (3.42)–(3.45), 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 (3.3), (3.4) 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 [74–76] 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 (3.3), (3.4), with n and $m-n$ equations, and the boundary condition (3.10) 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

$$E_1(t)x'(t) + F_1(t)x(t) = q_1(t), \tag{3.50}$$

$$F_2(t)x(t) = q_2(t), \tag{3.51}$$

is treated in [75] by means of the symmetric scheme

$$E_1(\tau_{ik})x'_\pi(\tau_{ik}) + F_1(\tau_{ik})x_\pi(\tau_{ik}) = q_1(\tau_{ik}), \quad k = 1,\dots,s,\ i = 0,\dots,N-1, \tag{3.52}$$

$$F_2(\tau^L_{ik})x_\pi(\tau^L_{ik}) = q_2(\tau^L_{ik}), \quad k = 0,\dots,s,\ i = 0,\dots,N-1, \tag{3.53}$$

$$T_2(t_i)^*(x_\pi(t_i^-) - x_\pi(t_i)) = 0, \quad i = 1,\dots,N-1, \tag{3.54}$$

$$G_a x_\pi(a) + G_b x_\pi(b) = \gamma \tag{3.55}$$

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 [75, Theorems 3.1–3.3].

Theorem 3.5 *Let the linear BVP (3.50), (3.51), (3.10) 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 (3.52) and (3.53), the boundary condition (3.10) as well as the consistency conditions (3.54).*

(2) *It holds that*

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

(3) *If the data of the DAE is sufficiently smooth, then superconvergence order holds,*

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

Condition (3.54) is no longer mentioned in Theorem 3.5. It only ensures the continuity of the differential component, similar to condition (3.35). In fact, (3.54) could be replaced by the easier conditions (3.35). Namely, for each fixed $1 \le i \le N-1$, one has from (3.53) 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 [74, 76] to nonlinear BVPs with partitioned DAEs (3.3), (3.4) by means of the scheme

$$\mathfrak{f}_1(x'_\pi(\tau_{ik}), x_\pi(\tau_{ik}), \tau_{ik}) = 0, \quad k = 1, \ldots, s, \quad i = 0, \ldots, N-1, \tag{3.56}$$

$$\mathfrak{f}_2(x_\pi(\tau^L_{ik}), \tau^L_{ik})) = 0, \quad k = 0, \ldots, s, \quad i = 0, \ldots, N-1, \tag{3.57}$$

$$g(x_\pi(a), x_\pi(b)) = 0. \tag{3.58}$$

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 [74–76] 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 [74, 76] 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 2.7. Fortunately, for a large class of index-2 DAEs, the BVPs with accurately stated boundary conditions become well-posed in the advanced setting, see Sect. 2.5.2. In this case, the associated inequality (2.66) reads:

$$\|x - x_*\|_\infty \le \kappa \, (|\gamma| + \|q\|_\infty + \|(DQ_{*1}G_{*2}^{-1}q)'\|_\infty). \tag{3.59}$$

The linear Hessenberg system of m_1 and $m_2 \le m_1$ equations,

$$\begin{aligned} x_1'(t) + B_{11}(t)x_1(t) + B_{12}(t)x_2(t) &= q_1(t), \\ B_{21}(t)x_1(t) \qquad\qquad\quad &= q_1(t), \end{aligned}$$

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 (3.59) simplifies to

$$\begin{aligned} \|x - x_*\|_\infty &\le \kappa \, (|\gamma| + \|q\|_\infty + \|(B_{12}(B_{21}B_{12})^{-1}q_2)'\|_\infty) \\ &\le \tilde{\kappa} \, (|\gamma| + \|q\|_\infty + \|q_2'\|_\infty). \end{aligned} \tag{3.60}$$

A direct investigation of the linear index-2 DAE by 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 [9], 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 (3.60) applies to the nonlinear index-2 DAE

$$x_1'(t) + b_1(x_1(t), x_2(t), t) = 0, \tag{3.61}$$

$$b_2(x_1(t), t) \qquad = 0, \tag{3.62}$$

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 (3.62) can be significantly amplified, at least by a factor h^{-1}. Therefore, it is a good idea to keep the defects in this equation reasonable small. For this purpose, so-called *projected Runge–Kutta methods* and *projected collocation* are introduced in [8, 9].

It is proposed to complete the standard collocation methods locally at fixed time points by an additional backward projection onto the constraint given by Eq. (3.62). 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

$$b_2(x_{l,1}^{new}, t_l) = 0. \tag{3.63}$$

This is accomplished by the ansatz

$$x_{l,1}^{new} := x_{l,1} + B_{12}(x_{l,1}^{new}, x_{l,2}^{new}, t_l)\lambda_l, \tag{3.64}$$

$$x_{l,2}^{new} := x_{l,2}, \tag{3.65}$$

where $B_{ij} := \frac{\partial b_i}{\partial x_j}$. If the given approximations are sufficiently accurate, then the values x_l^{new} and λ_l are locally uniquely determined by (3.63)–(3.65). A Newton step starting from the initial guess $x_l^{new,(0)} = x_l$, $\lambda_l^{(0)} = 0$ yields

$$x_{l,1}^{new,(1)} = x_{l,1} - F_l b_2(x_{l,1}, x_{l,2}, t_l), \tag{3.66}$$

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 (3.66) 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 only the particular Ω-component is affected. 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 additional 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 (3.63)–(3.65) 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 [9, 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

$$\|x_{*,1} - x_{\pi,1}\|_\infty = O(h^{\min(s+1,\, s+)}), \quad \|x_{*,2} - x_{\pi,2}\|_\infty = O(h^s),$$
$$|x_{*,1}(t_i) - x_{\pi,1}(t_i)| = O(h^{s+}), \quad i = 0, \ldots, N,$$

hold. In contrast to the results for index-1 problems, now $x_{\pi,1}$ is generally discontinuous due to the backward projection.

The projected collocation is extended to some more general semi-explicit index-2 DAEs [7, 11], whereby the components to be changed by projections are locally identified by means of a singular value decomposition. This procedure is called *selective projected collocation*.

The package COLDAE [11] 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 effort has been put into the analysis and numerical treatment of BVPs in ODEs which can exhibit singularities (e.g., [13, 15, 17, 43, 68, 72] and references therein). Such problems are typically given as

$$t^\alpha u'(t) = M(t)u(t) + h(u(t), t), \quad t \in (0, 1], \quad g(u(0), u(1)) = 0, \tag{3.67}$$

with $\alpha \geq 1$. For $\alpha = 1$, one refers to a *singularity of the first kind*. For instance, a singularity of the first kind may arise 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 is the case in the following two examples.

Example 3.2 The DAE taken from [72],

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix} \left(\begin{bmatrix} 1 & -1 \end{bmatrix} x \right)'(t) + \begin{bmatrix} 2 & 0 \\ 0 & t+2 \end{bmatrix} x(t) = 0, \tag{3.68}$$

has index 1 on the interval $(0, 1]$ and yields there the inherent ODE

$$t\,u'(t) = -(2t+4)u(t), \quad u(t) = x_1(t) - x_2(t), \tag{3.69}$$

showing a singularity of the first kind. The inherent ODE (3.69) 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 (3.68)

$$\Pi_{can}(t) = I - Q_{can}(t) = \begin{bmatrix} 1+\frac{2}{t} & -\frac{2+t}{t} \\ \frac{2}{t} & 1-\frac{2+t}{t} \end{bmatrix}$$

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]),

$$\begin{bmatrix} t & 0 \\ 0 & t \\ 0 & 0 \end{bmatrix}\left(\begin{bmatrix}1&0&0\\0&1&0\end{bmatrix}x\right)'(t) + \begin{bmatrix}1&0&1\\1&0&0\\0&\frac{1}{2}&\frac{1}{2}\end{bmatrix}x(t) = q(t), \tag{3.70}$$

has index 1 on the interval $(0, 1]$ and yields the inherent ODE

$$t\,u'(t) = \begin{bmatrix}-1&1\\-1&0\end{bmatrix}u(t) + \begin{bmatrix}q_1(t) - 2q_3(t)\\ q_2(t)\end{bmatrix}, \quad u(t) = \begin{bmatrix}x_1(t)\\x_2(t)\end{bmatrix}. \tag{3.71}$$

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 (3.70) 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 (3.71) correspond to the bounded solutions of the DAE (3.70). □

In the context of classical singular BVPs (3.67), seeking a solution that is 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 the case of DAEs, this procedure becomes a much tougher 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

$$f((D(t)x(t))', x(t), t) = 0, \quad t \in (0, 1], \tag{3.72}$$

$$G_a x(0) + G_b x(1) = \gamma, \tag{3.73}$$

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

$$\ker f_y(y, x, t) = \{0\}, \quad (y, x, t) \in \mathbb{R}^n \times \mathcal{D} \times (0, 1], \tag{3.74}$$

$$\operatorname{im}(D(t)) = \mathbb{R}^n, \quad t \in [0, 1]. \tag{3.75}$$

Conditions (3.74) and (3.75) 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 is the case for the DAE (3.70). The structural conditions (3.74) and (3.75) guarantee that the system (3.72) 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 (3.73) 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.

Let $\mathcal{I} = [0, 1]$. We are looking for a solution of the BVP (3.72), (3.73) which belongs at least to the function space $\mathcal{C}(\mathcal{I}, \mathbb{R}^m) \cap \mathcal{C}^1_D((0, 1], \mathbb{R}^m)$.

The further structure of the boundary conditions (3.73) which is necessary and sufficient for the BVP (3.72)–(3.73) to become well-posed in a special sense will be specified in the course of the discussion. Here, we argue without function space

settings, but adopt the understanding of well-posed BVPs common in the framework of singular ODEs (e.g., [68]). Although first existence and uniqueness results are given for a special class of singular DAEs in [98], more general solvability statements justifying well-posedness of BVPs in appropriate function spaces are missing up to now. As we will see, well-posedness in this special sense incorporates aspects of well-conditioning.

We define

$$N_0(t) := \ker D(t), \quad t \in [0,1],$$

and note that

$$\begin{aligned} \ker f_y(y,x,t)D(t) &= N_0(t), \quad (y,x,t) \in \mathbb{R}^n \times \mathcal{D} \times (0,1], \\ \ker f_y(y,x,t)D(t) &\supset N_0(t), \quad (y,x,t) \in \mathbb{R}^n \times \mathcal{D} \times \{0\}. \end{aligned}$$

Below, the pointwise generalized inverse D^- of D is defined as in the regular case in Sect. 2.1.

In [43, 72], 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 [43, 72].

3.4.1 Linear Case

Following the lines of [72] we first decouple the DAE in order to formulate sufficient conditions ensuring a singularity of the first kind for the inherent ODE and the well-posed boundary conditions. Consider the linear DAE

$$A(t)(Dx)'(t) + B(t)x(t) = q(t), \quad t \in (0,1], \tag{3.76}$$

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 (3.74), (3.75) that

$$\ker A(t) = \{0\}, \qquad t \in (0,1], \tag{3.77}$$

$$\operatorname{im} D(t) = \mathbb{R}^n, \qquad t \in [0,1]. \tag{3.78}$$

We decouple the DAE on the interval $(0, 1]$ as described in Sect. 2.2. With Q_0 being a continuous projector function onto $\ker D$, and $P_0 := I - Q_0$ we form

$$G_0(t) := A(t)D(t), \quad t \in [0, 1], \tag{3.79}$$

$$G_1(t) := G_0(t) + B(t)Q_0(t), \quad t \in [0, 1]. \tag{3.80}$$

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 [86, 97, 103, 105], 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 (3.76) exhibits a singularity of the first kind. To this end, we decouple the solution of DAE (3.76) on $(0, 1]$ into the differential component Dx and the algebraic component Q_0x. While $u = Dx$ satisfies the inherent explicit ODE,

$$u'(t) + D(t)G_1(t)^{-1}B(t)D(t)^{-}u(t) = D(t)G_1(t)^{-1}q(t), \quad t \in (0, 1], \tag{3.81}$$

the algebraic component is given by

$$Q_0(t)x(t) = -Q_0(t)G_1(t)^{-1}B(t)D(t)^{-}u(t) + Q_0(t)G_1(t)^{-1}q(t), \quad t \in (0, 1]. \tag{3.82}$$

If $u(t)$ represents the general solution of the inherent ODE (3.81), then the general solution of the DAE (3.76) can be expressed as

$$x(t) = D(t)^{-}u(t) + Q_0(t)x(t) = \Pi_{can}(t)D(t)^{-}u(t) + Q_0(t)G_1(t)^{-1}q(t), \quad t \in (0, 1],$$

whereby

$$\Pi_{can}(t) = I - Q_0(t)G_1(t)^{-1}B(t), \quad t \in (0, 1],$$

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 (3.81), typical for the singularity of the first kind, is observed when $G_1(0)$ is singular but $tG_1(t)^{-1}$ has a continuous extension on $[0, 1]$. Then, we can rewrite the matrix $D(t)G_1(t)^{-1}B(t)D(t)^{-}$ and obtain

$$D(t)G_1(t)^{-1}B(t)D(t)^{-} =: -\frac{1}{t}M(t), \tag{3.83}$$

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 (3.81) by $p(t)$ we arrive at the inherent explicit ODE of the form

$$u'(t) = \frac{1}{t} M(t)u(t) + p(t), \quad t \in (0,1]. \tag{3.84}$$

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 eigenstructure 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 $G_1(t)^{-1}q(t)$ and thus $p(t)$ are continuous on the whole interval $[0,1]$. Then, by [40], the bounded solutions of the ODE (3.84) can be represented in the form

$$u(t) = Ec + tf(t), \quad t \in [0,1], \tag{3.85}$$

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 (3.82) is also bounded.

Proposition 3.6 *Let the DAE (3.76) be regular with index 1 on* $(0,1]$ *and satisfy conditions (3.77), (3.78), and let the problem data be sufficiently smooth.*

Let $G_1(0)$ *be singular, but the matrix functions*

$$tG_1(t)^{-1}, \quad G_1(t)^{-1}q(t), \quad Q_0(t)G_1(t)^{-1}B(t)D(t)^{-}E, \quad t \in (0,1], \tag{3.86}$$

have continuous extensions on the closed interval $[0,1]$*,*

$$[tG_1(t)^{-1}]^{ext}, \quad [G_1(t)^{-1}q(t)]^{ext}, \quad [Q_0(t)G_1(t)^{-1}B(t)D(t)^{-}E]^{ext}.$$

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

$$x(t) = [\Pi_{can}(t)D(t)^{-}E]^{ext}c + [t\Pi_{can}(t)]^{ext}D(t)^{-}f(t) + Q_0(t)[G_1(t)^{-1}q(t)]^{ext},$$

$t \in [0,1]$*, 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 (3.86) 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 a critical point of type 0, see Example 3.3, then also the term $Q_0(t)G_1(t)^{-1}B(t)D(t)^{-}E = Q_{can}(t)D(t)^{-}E$ has the continuous extension.

The inherent ODE (3.84) is augmented by the boundary conditions

$$B_a u(0) + B_b u(1) = \gamma. \tag{3.87}$$

These boundary conditions have to be chosen such that a well-posed singular BVP results for u. In [72], the attention is focused on BVPs for singular ODE systems (3.84) which can equivalently be expressed as a well-posed IVP 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 (3.83), see [69, 70] 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 IVP posed at $t^* = 0$ for the differential equation (3.84) 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, 70].

Under the assumptions of Proposition 3.6, polynomial collocation methods are analyzed in [71, 72]. 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 Sect. 3.1.1. The numerical scheme defining x_π and u_π has the form

$$A(\tau_{ik})u'_\pi(\tau_{ik}) + B(\tau_{ik})x_\pi(\tau_{ik}) = q(\tau_{ik}), \tag{3.88}$$

$$D(\tau_{ik})x_\pi(\tau_{ik}) - u_\pi(\tau_{ik}) = 0, \qquad k = 1, \ldots, s,\ i = 0, \ldots, N-1, \tag{3.89}$$

$$B_0 u_\pi(0) + B_1 u_\pi(1) = \gamma. \tag{3.90}$$

As in Sect. 3.1.1, further conditions are necessary to close the system for the numerical computations. We choose these additional conditions as

$$B(0)x_\pi(0) - q(0) \in \lim_{t \to 0+} \operatorname{im}(A(t)), \quad u_\pi(0) = D(0)x_\pi(0), \tag{3.91}$$

or

$$B(1)x_\pi(1) - q(1) \in \operatorname{im}(A(1)), \quad u_\pi(1) = D(1)x_\pi(1). \tag{3.92}$$

The convergence results in the case of a singular inherent ODE are quite similar to the regular index-1 DAE case. Owing to the assumptions of Proposition 3.6, for arbitrary collocation points, stage order s uniformly in t is ensured in the 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 BVP, the analysis in [72] additionally requires

$$Q_{can} \in \mathcal{C}([0,1], \mathcal{L}(\mathbb{R}^m)) \tag{3.93}$$

to ensure this optimal convergence behavior. If the assumptions (3.86) and (3.93) are violated, order reductions in the algebraic components might occur. In particular, 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 [71, 72]. 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:

$$A(Dx)' + \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22}(t) \end{bmatrix} x(t) = q(t), \tag{3.94}$$

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

$$Q_{can}(t) = Q_0 G_1(t)^{-1} B(t) = \begin{bmatrix} 0 & 0 \\ B_{22}(t)^{-1} B_{21} & I \end{bmatrix} \tag{3.95}$$

is unbounded on (0, 1]. Moreover, it holds that

$$DG_1(t)^{-1}B(t)D^- = B_{11} - B_{12}B_{22}(t)^{-1}B_{21} = -\frac{1}{t}\begin{bmatrix} -7 & -12 \\ 8 & 13 \end{bmatrix} =: -\frac{1}{t}M.$$

Since M is nonsingular, we have $E = 0$, and the matrix functions $Q_0G_1(t)^{-1}B(t)D^-$ $E = 0$ 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 (3.86) 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 (3.94) subject to the boundary conditions

$$x_1(1) = \sin(1), \quad x_2(1) = \mathrm{e}.$$

The additional conditions

$$x_1(1) + x_2(1) + x_3(1) = q_3(1),$$
$$2x_1(1) + 3x_2(1) + \frac{1}{5}x_4(1) = q_4(1)$$

are consistent boundary conditions for the algebraic components to complete the collocation scheme used in BVPSUITE. These conditions 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 For $\ell = 3$, $\gamma = 1$, $\delta = 1$ all solution components are smooth.

Problem 2 For $\ell = \frac{5}{2}$, $\gamma = \frac{6}{5}$, $\delta = \frac{5}{2}$ the differential components x_1 and x_2 become unsmooth.

The numerical results obtained from BVPSUITE for this example are given in Table 7. For more details see [71, Tables 192–200, 228–236]. For the case when the differential solution components, $u(t)$, are smooth no order reduction is observed, although the projection matrix (3.95) is unbounded for $t \to 0$.

Table 7 Problems 1 and 2: experimentally observed convergence rates for different collocation schemes with $s = 3, 4$, cf. [72] for details

		$s = 3$				$s = 4$			
		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

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

In Problem 2 we observe order reductions due to the fact that the canonical projector (3.95) 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 loses 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 (3.72), (3.73). We assume the DAE to be regular with index 1 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 BVPs 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 (3.86) for linear BVPs, it is assumed that

$$tG_1(y, x, t)^{-1} \tag{3.96}$$

has a continuous extension for $t \to 0$, where

$$G_0(y, x, t) := f_y(y, x, t)D(t),$$
$$G_1(y, x, t) := G_0(y, x, t) + f_x(y, x, t)Q_0(t), \quad (y, x, t) \in \mathbb{R}^n \times \mathcal{D} \times [0, 1].$$

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 $\mathrm{im} f_y(y,x,t)^{\perp}$, 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

$$\mathrm{rank}\begin{bmatrix} W^{ext}(y,x,0)f_x(y,x,0) \\ D(0) \end{bmatrix} = m. \tag{3.97}$$

An inspection of Examples 3.2 and 3.3 confirms this criterion.

To apply standard linearization arguments, the BVP (3.72)–(3.73) is supposed to possess a solution $x_\star \in \mathcal{C}_D^1([0,1],\mathbb{R}^m)$ and the linearization of the DAE (3.72) along $x_\star$,

$$A_\star(t)(D(t)z(t))' + B_\star(t)z(t) = 0, \quad t \in (0,1], \tag{3.98}$$

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 (3.98) is regular with tractability index 1 on the interval $(0,1]$. Thus the linearized BVP can be treated as in Sect. 3.4.1.

In analogy to Definition 2.5, one says that the solution $x_\star$ of the BVP (3.72)–(3.73) is *isolated* if and only if its linearization

$$\begin{aligned} &A_\star(t)(D(t)z(t))' + B_\star(t)z(t) = 0, \quad t \in (0,1], \\ &B_0D(0)z(0) + B_1D(1)z(1) = 0, \end{aligned}$$

has only the trivial solution. In this case, as common in the theory of singular explicit ODEs (e.g., [68, 70]), also the nonlinear BVP (3.72), (3.73) 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., (2.61), (2.62)) of the nonlinear DAE (3.72),

$$u'(t) = D(t)\omega(u(t),t), \quad t \in (0,1], \tag{3.99}$$

$$x(t) = D(t)^- u(t) + Q_0(t)\omega(u(t),t), \quad t \in (0,1], \tag{3.100}$$

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

$$D(t)\omega(u,t) = \frac{1}{t}M(t)u + q(u,t), \quad u \in \mathcal{D}_\omega, \quad t \in (0,1], \tag{3.101}$$

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 (3.96), (3.101), as well as to feature a bounded canonical projector function.

This yields the BVP

$$u'(t) = \frac{1}{t}M(t)u(t) + q(u(t),t), \quad t \in (0,1], \tag{3.102}$$

$$B_0 u(0) + B_b u(1) = \gamma. \tag{3.103}$$

In turn, the linearization of the last BVP reads,

$$\zeta'(t) = D(t)\omega_u(u_\star(t),t)\zeta(t) = \frac{1}{t}M_\star(t)\zeta(t), \quad t \in (0,1], \tag{3.104}$$

$$B_0\zeta(0) + B_1\zeta(1) = 0, \tag{3.105}$$

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 (3.104), (3.105) to have only the trivial solution. It was shown in [40] that the form of the boundary conditions (3.105) which guarantee that (3.104), (3.105) has only the trivial solution depends on the spectral properties of the coefficient matrix $M_\star(0)$. Note that (3.101) 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 (3.104) 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

$$U := R_+ + Q_M, \quad V := I - U, \tag{3.106}$$

The BVP (3.104), (3.105) is well-posed if and only if the boundary conditions (3.105) can equivalently be written as [40]

$$V\zeta(0) = 0, \quad R_+\zeta(1) = 0, \quad Q_M\zeta(0) = 0, \text{ or } Q_M\zeta(1) = 0. \tag{3.107}$$

The first set of homogeneous initial conditions specified in (3.107) are necessary and sufficient for ζ to be continuous on the closed interval $[0, 1]$.

The polynomial collocation methods (uniform approach A) described in Sect. 3.1.2 are used in [43] to approximate the solution of well-posed singular nonlinear BVPs (3.72), (3.73). The basic collocation scheme

$$\begin{aligned}
u_\pi(t_i^-) - u_\pi(t_i) &= 0, \quad i = 1, \ldots, N-1, \\
x_\pi(t_i^-) - x_\pi(t_i) &= 0, \quad i = 1, \ldots, N-1, \\
f(u'_\pi(\tau_{ik}), x_\pi(\tau_{ik}), \tau_{ik}) &= 0, \quad k = 1, \ldots, s, \ i = 0, \ldots, N-1 \\
u_\pi(\tau_{ik}) - D(\tau_{ik}) x_\pi(\tau_{ik}) &= 0, \quad k = 1, \ldots, s, \ i = 0, \ldots, N-1, \\
B_0 u_\pi(a) + B_1 u_\pi(b)) &= \gamma,
\end{aligned}$$

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 (3.25), and the consistency conditions. This results in

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

3.5 *Defect-Based a posteriori Error Estimation for Index-1 DAEs*

When designing error estimation procedures, one usually 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 double the number of subintervals. This procedure is used often in software for BVPs in ODEs and DAEs, for instance, in COLNEW, COLDAE, see [11]. Since in the context of collocation methods this procedure 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

$$A(t)(Dx)'(t) + B(t)x(t) = q(t), \quad t \in [a, b], \tag{3.108}$$

satisfying the general assumptions in Sect. 2.1, and, additionally, condition (3.6) 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 (3.7), (3.8) of the DAE under consideration.

We consider a well-posed BVP (cf. Sect. 2.3) for the DAE (3.108) and the collocation equations

$$A(\tau_{ik})u'_\pi(\tau_{ik}) + B(\tau_{ik})x_\pi(\tau_{ik}) = q(\tau_{ik}), \tag{3.109}$$

$$Dx_\pi(\tau_{ik}) - u_\pi(\tau_{ik}) = 0, \quad k = 1, \ldots, s, \ i = 0, \ldots, N-1, \tag{3.110}$$

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, 86] 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 Sect. 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 (3.108), 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 [15, 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 [15] 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

$$F(x(t)) = 0, \quad t \in [a, b], \tag{3.111}$$

on a grid π. Define the *defect* $d = d(t)$ by interpolating ξ_π by a continuous piecewise polynomial function $p(t)$ of degree $\le s$ and substituting $p(t)$ into (3.111),

$$d(t) := F(p(t)), \quad t \in [a, b]. \tag{3.112}$$

Obviously, $p(t)$ is the exact solution to a *neighboring problem*

$$F(x(t)) = d(t) \tag{3.113}$$

related to the original problem (3.111). 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 $\tilde{x}^{def}_\pi$ for both the original and neighboring problems on the grid π, i.e., $\tilde{F}(\tilde{x}_\pi) = 0$ and $\tilde{F}(\tilde{x}^{def}_\pi) = d_\pi$, where d_π is an appropriate restriction of $d(t)$ to the grid π.

Since (3.111) and (3.113) differ only by the (presumably) small defect d, we expect that

$$\varepsilon_\pi := \tilde{x}^{def}_\pi - \tilde{x}_\pi \tag{3.114}$$

is a good estimate for the global error

$$e_\pi := \xi_\pi - x^*_\pi. \tag{3.115}$$

In other terms,

$$\begin{aligned} e_\pi := \xi_\pi - x^*_\pi &\approx F^{-1}(d) - F^{-1}(0) \\ &\approx \tilde{F}^{-1}(d_\pi) - \tilde{F}^{-1}(0) = \tilde{x}^{def}_\pi - \tilde{x}_\pi = \varepsilon_\pi. \end{aligned} \tag{3.116}$$

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 [15] 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 (3.112) 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 Sect. 3.5.1 to the linear DAE (3.108). 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

$$A(\tau_{ik}) \frac{D\varepsilon_{ik} - D\varepsilon_{i,k-1}}{h_{ik}} + B(\tau_{ik})\varepsilon_{ik} = \bar{d}_{ik}, \tag{3.117}$$

with homogeneous initial condition $\varepsilon_{0,0} = 0$ and the backward Euler scheme playing the role of $\tilde{F}$. According to (3.112), the straightforward, classical way to define the defect $\bar{d}_{ik}$ would be to substitute $x_\pi(t)$ into (3.108) in the pointwise sense,

$$d(t) := A(t)(Dx_\pi)'(t) + B(t)x_\pi(t) - q(t), \quad t \in [a, b], \tag{3.118}$$

and using the pointwise defect $\bar{d}_{ik} := d(\tau_{ik})$ in (3.117). However, as has been pointed out in [15] in the ODE context, this procedure does not lead to successful results. For collocation this is obvious: Since, by definition of the collocation solution (3.109), 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 a slight variation of the procedure introduced in [15], we now define a modified defect via the integral means

$$\bar{d}_{ik} := \sum_{l=0}^{s} \alpha_{kl} d(\tau_{il}) = \frac{1}{h_{ik}} \int_{\tau_{i,k-1}}^{\tau_{ik}} d(t)\mathrm{d}t + \mathcal{O}(h^{s+1}), \tag{3.119}$$

for $i = 0, \ldots, N-1$, $k = 1, \ldots, s$, where the α_{kl} are quadrature coefficients for the integral means in (3.119), i.e.,

$$\alpha_{kl} = \frac{1}{\rho_k - \rho_{k-1}} \int_{\rho_{k-1}}^{\rho_k} L_l(t)dt, \quad k = 1 \ldots s,\ l = 0 \ldots s, \tag{3.120}$$

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 (3.119).

The following result is shown in [18].

Theorem 3.7 *While the global error of the collocation method (3.109) is of order h^s, i.e.,*

$$e(t) = x_\pi(t) - x_*(t) = \mathcal{O}(h^s), \tag{3.121}$$

the error estimate of the global error (3.121) based on the modified defect (3.119) and the auxiliary scheme (3.117) is asymptotically correct, i.e.,

$$\varepsilon_{ij} - e(\tau_{ij}) = \mathcal{O}(h^{s+1}). \tag{3.122}$$

Example 3.5 We consider the IVP

$$\begin{bmatrix} \mathrm{e}^t \\ \mathrm{e}^t \end{bmatrix} \left(\begin{bmatrix} 1 & 0 \end{bmatrix} x \right)'(t) + \begin{bmatrix} \mathrm{e}^t(1+\cos^2 t) & \cos^2 t \\ \mathrm{e}^t(-1+\cos^2 t) & -\cos^2 t \end{bmatrix} x(t) = \begin{bmatrix} \sin^2 t(1-\cos t) - \sin t \\ \sin^2 t(-1-\cos t) - \sin t \end{bmatrix}, \tag{3.123}$$

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 Fig. 7.

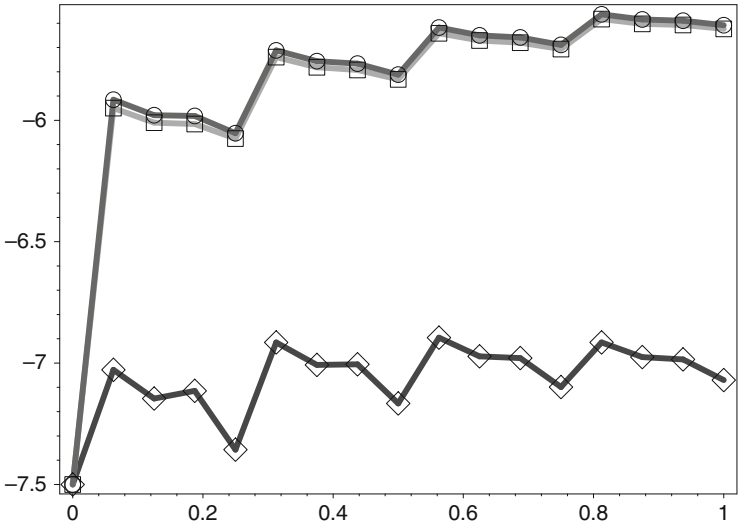

Fig. 7 $\log_{10}$-plot for first solution component, $N = 4$ of Example 3.5. *Open circle*: error $|e_1(t)| = |x_{\pi,1}(t) - x_{*,1}(t)|$; *open square*: error estimate $|\varepsilon_1(t)|$; *open diamond*: error of error estimate $|\varepsilon_1(t) - e_1(t)|$

- 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 $_e$	$\varepsilon - e$	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 [41, 42, 73] 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 (3.5), which allows to restrict the consideration directly to $u_* = Px_*$, $v_* = (I - P)x_*$ and their approximations. Degenhardt [41, Theorem 4.13] provides superconvergence of 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 as do the error estimates and mesh control as well.

As noted, e.g., in [55, 89], 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 extent those effects can be resolved.

Concerning the different collocation approaches to DAEs, up to 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 a DAE might be caused by a singular inherent ODE as in Sect. 3.4, but also by the other components of a DAE, see [86, 102, 103]. 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 there might be so-called harmless ones [44, 86], 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 shown 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 also work for general regular index-2 DAEs (3.5) satisfying (3.6). Possibly, this way, projected collocation for the corresponding BVPs could work well.

4 Shooting Methods

The *shooting method* or *initial-value adjusting method*—a description used in the very first publications—is a classical method to solve two-point boundary value problems (TPBVP) but also multi-point BVPs for ODEs and DAEs.

The first papers dealing with DAEs and shooting methods are [38, 55, 79, 89]. The idea is to imbed the BVP into a family of IVPs, with unknown initial values, and then to seek among them for the true one.

We consider the TPBVP

$$f((Dx)'(t), x(t), t) = 0, \quad t \in [a, b] \tag{4.1}$$

$$g(x(a), x(b)) = 0, \tag{4.2}$$

and we assume that the DAE (4.1) is regular with index μ and that the TPBVP has a locally unique solution x_*. Set $z_* := x_*(a)$.

As 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 Sect. 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

$$C(x(a) - z) = 0, \quad z \in \mathcal{N}_* \tag{4.3}$$

with an appropriate singular matrix $C \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^l)$. As shown in Sect. 2, linear IVPs have unique solutions existing on $[a, b]$, if C is such that

$$\ker C = \ker \Pi_{can} = \ker \Pi_{\mu-1}, \tag{4.4}$$

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 (4.1), (4.3) are uniquely solvable on $[a, b]$, then one looks for a z such that the boundary condition (4.2) is satisfied. This is the basic idea of the shooting method.

4.1 Solution of Linear DAEs

We consider the linear TPBVP (2.35) and a related IVP

$$A(Dx)' + Bx = q, \tag{4.5}$$

$$C(x(a) - z) = 0, \tag{4.6}$$

and C is chosen fulfilling (4.4) with given value z. We assume that the DAE (4.5) is regular with index μ. The solution of the IVP is represented in (2.33) 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 using the structure of X (cf. (2.27)) from

$$X(t,a)z = X(t,a)D(a)^- D(a)\Pi_{\mu-1}(a)z,$$

we discover 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 (4.5), (4.6) by $x(t;a,\xi)$. This means that we implicitly assume, for the moment, that we also know the solution 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 (2.36) 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 (4.5) and to solve the TPBVP (2.35) the boundary condition also has to be fulfilled. The relation to determine ξ is given by

$$\begin{aligned} G_a x(a) + G_b x(b) = G_a(X(a,a)D(a)^-\xi + v_q(a)) + G_b(X(b,a)D(a)^-\xi \\ + \int_a^b X(b,s)G_\mu^{-1}(s)q(s)\mathrm{d}s + v_q(b)) = \gamma. \end{aligned} \tag{4.7}$$

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. (2.37)). Theorem 2.1 provides a unique initial value $D(a)^-\xi$. The solution of the IVP (4.5) 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 $C(x(a)-D(a)^-\xi) = D(a)\Pi_{\mu-1}(a)(x(a)-D(a)^-\xi) = 0$ it follows that $D(a)\Pi_{\mu-1}(a)x(a) = D(a)\Pi_{\mu-1}(a)D(a)^-\xi = \xi$.

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 usually differs 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 (4.2). Additionally, consistent initial values are very helpful to start the 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. Lamour [79] and England and Lamour [47] propose for index-1 DAEs the use of the tractability index concept. Amodio and Mazzia [2], Brown et al. [36], and Kiehl [67] assume a semi-explicit structure of the DAE, which makes the computation much easier. Gerdts [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

$$ADy + B(D^-\xi + v) = q(\bar{t}). \tag{4.8}$$

Because $\operatorname{rank} D = r_0$ 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. (2.74)) 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 on a known ξ as

$$\begin{pmatrix} y \\ v \end{pmatrix} = \mathcal{I}_\mu(\xi, \bar{t}). \tag{4.9}$$

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 $v := (I - P_0(\bar{t}))x(\bar{t}) = Q_0(\bar{t})x(\bar{t})$. Equation (4.8) looks like

$$ADy + B(D^-\xi + v) = q(\bar{t}), \tag{4.10}$$

$$Q_0 y + P_0 v = 0. \tag{4.11}$$

Equation (4.11) 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 (4.10), (4.11) 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

$$v = Q_0 G_1^{-1}(q - BD^-\xi) \tag{4.12}$$

we obtain $\frac{\partial v}{\partial \xi} = -Q_0 G_1^{-1} B D^- = -\mathcal{H}_0$ for index 1 (cf. (6.9)).

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. [86, Chaps. 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 (4.5) by the projector W_1 projecting along $\operatorname{im} G_1$ and we obtain with $W_1 B Q_0 = 0$

$$W_1 B \underbrace{D^-(Dx)'}_{=y} + (W_1 B D^-)' Dx = (W_1 q)'.$$

This leads to the system

$$ADy + B(D^-\xi + v) = q(\bar{t}), \tag{4.13}$$

$$W_1 B y + (W_1 B D^-)'(\xi + Dv) = (W_1 q)'(\bar{t}), \tag{4.14}$$

$$Q_0 y + \Pi_1 v = 0. \tag{4.15}$$

We solve Eqs. (4.13)–(4.15) explicitly for a given value $\xi = D\Pi_1 D^- \xi$. Multiplying (4.13) by $Q_1 G_2^{-1}$ provides using the relations $v = (I - \Pi_1)v$ from (4.15) and the admissible projector $Q_1 = Q_1 G_2^{-1} B_1$, which realizes a fine decoupling,

$$Q_1 G_2^{-1} B v = Q_1 G_2^{-1} q(\bar{t}) - \underbrace{Q_1 G_2^{-1} B D^- D}_{Q_1} P_1 D^- \xi \quad \text{and we obtain}$$

$$Q_1 v = Q_1 G_2^{-1} q(\bar{t}),$$

i.e., $P_0 v = \underbrace{\Pi_1 v}_{=0} + P_0 Q_1 v = P_0 Q_1 G_2^{-1} q(\bar{t})$. The multiplication of (4.14) by $Q_1 G_2^{-1}$ results because $Q_1 G_2^{-1} W_1 = Q_1 G_2^{-1}$ and $Dv = DP_0 v$

$$\underbrace{Q_1 G_2^{-1} B P_0}_{=Q_1} y = Q_1 G_2^{-1}((W_1 q)'(\bar{t}) - (W_1 B D^-)'(\xi + Dv)).$$

From Eq. (4.13) we obtain by scaling with G_2^{-1}

$$\begin{aligned} G_2^{-1} G_0 y + G_2^{-1} B Q_0 v &= G_2^{-1}(q(\bar{t}) - BD^-(\xi + Dv)), \\ (\Pi_1 - Q_0 Q_1) y + Q_0 v &= G_2^{-1}(q(\bar{t}) - BD^-(\xi + Dv)), \\ \Pi_1 y + Q_0 v &= G_2^{-1}(q(\bar{t}) - BD^-(\xi + Dv)) + Q_0 Q_1 y. \end{aligned} \tag{4.16}$$

Multiplying Eq. (4.16) by Π_1 respectively Q_0, we obtain

$$\begin{aligned} \Pi_1 y &= \Pi_1 G_2^{-1}(q(\bar{t}) - BD^-(\xi + Dv)), \text{ respectively} \\ Q_0 v &= Q_0 G_2^{-1}(q(\bar{t}) - BD^-(\xi + Dv)) + Q_0 Q_1 y. \end{aligned}$$

Summarizing the components of y and v we obtain

$$\begin{aligned} y &= \Pi_1 G_2^{-1}(q(\bar{t}) - BD^-(\xi + Dv)) + P_0 Q_1 G_2^{-1}((W_1 q)'(\bar{t}) - (W_1 B D^-)'(\xi + Dv)), \\ v &= (I - \Pi_1) G_2^{-1} q(\bar{t}) - Q_0 G_2^{-1} B D^-(\xi + Dv) + Q_0 Q_1 y. \end{aligned} \tag{4.17}$$

Later on we will need the relation between v and ξ. With (4.17) we obtain

$$\frac{\partial v}{\partial \xi} = -Q_0 G_2^{-1} B D^- - Q_0 Q_1 G_2^{-1}(W_1 B D^-)' = -\mathcal{H}_0 \tag{4.18}$$

for the index-2 case (cf. Appendix (6.9)).

Lemma 4.1 *The linear DAE (4.8) has index 2 and let v be the solution of Eqs. (4.13)–(4.15). We choose the fine decoupling projector $Q_1 = Q_1 G_2^{-1} B_1$ and assume that $Q_0 Q_1 G_2^{-1}, Q_0 Q_1 D^- \in \mathcal{C}^1$ then $(D^- - \frac{\partial v}{\partial \xi}) D \Pi_1 = \Pi_{can,2}$.*

Proof Using Eq. (4.18) 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 $Q_0G_2^{-1}B\Pi_1 = Q_0(P_1 + \underbrace{Q_1)G_2^{-1}B\Pi_1}_{=0} = Q_0P_1G_2^{-1}B\Pi_1$ and

$$\begin{aligned} Q_0Q_1G_2^{-1}(W_1BD^-)'D\Pi_1 &= Q_0((Q_0Q_1D^-)' - (Q_0Q_1G_2^{-1})'W_1BD^-)D\Pi_1D^-D\Pi_1, \\ &= -\underbrace{Q_0Q_1D^-}_{=-Q_0P_1D^-}(D\Pi_1D^-)'D\Pi_1 \\ &\quad - Q_0(Q_0Q_1G_2^{-1})'W_1G_2\underbrace{Q_1G_2^{-1}BD^-D\Pi_1}_{=0} \end{aligned}$$

because of $W_1 = W_1G_2Q_1G_2^{-1}$. Now we have with (4.18) 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 (4.5). We choose fine decoupling projectors $Q_0, Q_1, \ldots, Q_{\mu-1}$ (cf. Sect. 6.1.2) then*

$$\frac{\partial \mathcal{I}_{\mu,v}}{\partial \xi}(\xi, \bar{t})) = \frac{\partial \mathcal{I}_{\mu,v}}{\partial \xi}(\xi, \bar{t}))D(\bar{t})D(\bar{t})^- \quad \textit{and} \tag{4.19}$$

$$(D(\bar{t})^- - \frac{\partial \mathcal{I}_{\mu,v}}{\partial \xi}(\xi, \bar{t}))D(\bar{t})\Pi_{\mu-1}(\bar{t}) = \Pi_{can}(\bar{t}). \tag{4.20}$$

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 (6.9) which explicitly represent the components v_i, $i = 0, \ldots, \mu - 1$. For a fine decoupling (6.9) 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 (4.19). 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. Sect. 6.1.2) the proof is done. □

The realization of algorithms to compute consistent initial values using (6.9) is very expensive. For higher-index systems it would be helpful to take advantage of 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 (2.35) and a related IVP

$$A(Dx)' + Bx = q, \tag{4.21}$$

$$C(x(a) - z) = 0. \tag{4.22}$$

The solution of the IVP (4.21), (4.22) at $t = b$ is applying (2.33) 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 (2.33) $x(a) = X(a, a)D(a)^{-}\xi + v_q(a)$. The boundary condition fixes the unknown ξ we are looking for

$$G_a(X(a, a)D(a)^{-}\xi + v_q(a)) + G_b x(b; a, \xi) = 0, \tag{4.23}$$

$$(I - D(a)\Pi_{\mu-1}(a)D(a)^{-})\xi = 0 \tag{4.24}$$

and (4.24) fixes that $\xi \in \operatorname{im} D(a)\Pi_{\mu-1}(a)$. But for a realization of (4.23) we have to know $v_q(a)$ too. Therefore we combine (4.23) with the equations describing consistent initial values at $t = a$ (cf. (4.9))

$$\begin{pmatrix} y \\ v \end{pmatrix} - \mathcal{I}_\mu(\xi, a) = 0. \tag{4.25}$$

Lemma 4.3 *Let the BVP (2.35) be uniquely solvable and the admissible projectors Q_i, $0 \le i \le \mu - 1$ realize a fine decoupling. The Jacobian matrix of (4.23)–(4.25) 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 (2.35) 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 (4.19). 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 in $z_y = 0,\ z_v = 0$. □

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 a 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 (4.21)–(4.22) and obtain the solution $x(b;a,u_0)$. The corrections $\Delta\xi, \Delta v$ are the solutions of the linear system

$$\begin{pmatrix} (G_a + G_b X(b,a))D(a)^- & G_a \\ I - D(a)\Pi_{\mu-1}(a)D(a)^- & 0 \\ \mathcal{H}_0 & I \end{pmatrix} \begin{pmatrix} \Delta\xi \\ \Delta v \end{pmatrix} = \begin{pmatrix} G_a(D(a)^-\xi_0 + v_0) + G_b x(b;a,\xi_0) - \gamma \\ 0 \\ 0 \end{pmatrix}. \tag{4.26}$$

The solution of the TPBVP (2.35) 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

$$(G_a X(a,a) + G_b X(b,a)) D(a)^- \Delta\xi = SD(a)^- \Delta\xi$$
$$= G_a(D(a)^- \xi_0 + v_0) + G_b x(b; a, \xi_0) - \gamma.$$

The rectangular coefficient matrix can be arranged in such a way that it may handle quadratic matrices. We have to combine the first two rows of equation system (4.26) (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 IVP 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. (4.26)).

We discuss here the case of multiple forward (parallel) shooting only. Methods shooting in different directions 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 matching points t_j we require continuity of the dynamic component u of the solution (cf. Sect. 2.2). We obtain

$$D\Pi_{\mu-1}(t_j)(D(t_j)^- \xi_j - x(t_j; t_{j-1}, \xi_{j-1})) = 0, \quad 1 \le j \le N-1 \quad \text{or shorter} \tag{4.27}$$

$$u_j - D\Pi_{\mu-1}(t_j) x(t_j; t_{j-1}, \xi_{j-1}) = 0, \tag{4.28}$$

with $\xi_j := u(t_j)$ and from the boundary condition

$$G_a(D^- \xi_0 + v_0) + G_b x(t_N; t_{N-1}, \xi_{N-1}) = 0. \tag{4.29}$$

The unknowns are $(\xi_0, \xi_1, \ldots, \xi_{N-1}, v_0, y_0)$, i.e., we have to extend the system by the computation of consistent initial values at t_0 to determine v_0,

$$\begin{pmatrix} y_0 \\ v_0 \end{pmatrix} - \mathcal{I}_\mu(\xi_0, t_0) = 0 \tag{4.30}$$

and the restriction of ξ_i to the subspace $\operatorname{im} D\Pi_{\mu-1}(t_i)$, $i = 0, \ldots, N-1$ by

$$(I - D(t_i)\Pi_{\mu-1}(t_i)D(t_i))^- \xi_i = 0 \tag{4.31}$$

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 for the first time in [38].)

Below we apply the notations

$$\begin{aligned} \pi_{\mu-1}(t) &:= D(t)\Pi_{\mu-1}(t)D(t)^- = D(t)\Pi_{can}(t)D(t)^-, \\ Y(t,\bar{t}) &:= D(t)\Pi_{\mu-1}(t)X(t,\bar{t})D(\bar{t})^- = D(t)X(t,\bar{t})D(\bar{t})^- \\ &= \pi_{\mu-1}(t)U(t,\bar{t})\pi_{\mu-1}(\bar{t}) = U(t,\bar{t})\pi_{\mu-1}(\bar{t}). \end{aligned}$$

We refer to (2.21) and [86, (2.82)] for the properties used

We obtain the following Jacobian matrix of the system (4.28)–(4.31) with respect to $\xi_0, \ldots, \xi_{N-1}, y_0, v_0$

$$J_\mu = \begin{bmatrix} G_a D(t_0)^- & & & & & & G_b X(t_N, t_{N-1}) D(t_{N-1})^- & 0 & G_a \\ -Y(t_1, t_0) & \pi_{\mu-1}(t_1) & & & & & & & \\ & -Y(t_2, t_1) & \pi_{\mu-1}(t_2) & & & & & & \\ & & \ddots & \ddots & & & & & \\ & & & -Y(t_{N-1}, t_{N-2}) & \pi_{\mu-1}(t_{N-1}) & & & & \\ I - \pi_{\mu-1}(t_0) & & & & & & & & \\ & I - \pi_{\mu-1}(t_1) & & & & & & & \\ & & I - \pi_{\mu-1}(t_2) & & & & & & \\ & & & \ddots & & & & & \\ & & & & I - \pi_{\mu-1}(t_{N-1}) & & & & \\ \frac{\partial \mathcal{I}_{\mu,y_0}}{\partial \xi_0} & & & & & & & I & 0 \\ \frac{\partial \mathcal{I}_{\mu,v_0}}{\partial \xi_0} & & & & & & & 0 & I \end{bmatrix}. \tag{4.32}$$

There is, for practical reasons, the possibility to compress J_μ by mixing (4.27) with (4.31). We obtain

$$\bar{J}_\mu = \begin{bmatrix} G_a D(t_0)^- & & & & & G_b X(t_N, t_{N-1}) D(t_{N-1})^- & 0 & G_a \\ -Y(t_1, t_0) & I & & & & & & \\ & -Y(t_2, t_1) & I & & & & & \\ & & \ddots & \ddots & & & & \\ & & & -Y(t_{N-1}, t_{N-2}) & & I & & \\ I - \pi_{\mu-1}(t_0) & & & & & & & \\ \frac{\partial \mathcal{I}_{\mu, v_0}}{\partial u_0} & & & & & & I & 0 \\ \frac{\partial \mathcal{I}_{\mu, v_0}}{\partial u_0} & & & & & & 0 & I \end{bmatrix} \tag{4.33}$$

The use of (4.28) for computing the Jacobian matrix leads immediately to (4.33).

As for the single shooting method, we show a regularity condition for the matrix (4.32).

Lemma 4.4 *Let the BVP (2.35) 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 (4.32) 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 $\pi_{\mu-1}(t_i) z_i = D(t_i) X(t_i, t_{i-1}) D(t_{i-1})^- z_{i-1}$ for $i = 1, \ldots, N-1$, the second to the Nth 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 like

$$(G_a + G_b X(t_N, t_0)) D(t_0)^- z_0 + G_a z_v = 0$$

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

$$(G_a X(t_0, t_0) + G_b X(t_N, t_0)) D(t_0)^- z_0 = S D(t_0)^- z_0 = 0.$$

From (2.39) we have that then $\Pi_{\mu-1}(t_0) D(t_0)^- z_0 = 0$, therefore $z_0 = 0$ and successively using (4.31) $z_i = 0$, $i = 1, \ldots, N-1$ it at last follows that $z_v = 0$, $z_y = 0$. □

We are interested in the relation of the multiple shooting method of a DAE with the inherent ODE. For that we use the v-component of (4.30) in (4.29) and we consider the system (4.27) and (4.29). Its Jacobian matrix looks like

$$S_{mult} = \begin{bmatrix} G_a X(t_0,t_0) D(t_0)^- & & & & G_b X(t_N,t_{N-1}) D(t_{N-1})^- \\ 0_{n-l,n} & & & & \\ -Y(t_1,t_0) & \pi_{\mu-1}(t_1) & & & \\ & -Y(t_2,t_1) & \pi_{\mu-1}(t_2) & & \\ & & \ddots & \ddots & \\ & & -Y(t_{N-1},t_{N-2}) & \pi_{\mu-1}(t_{N-1}) & \end{bmatrix}. \tag{4.34}$$

For (4.34) we have the representation $S_{mult} = S_{mult,ODE} \Pi_r$ with

$$S_{mult,ODE} = \begin{bmatrix} G_a \Pi_{can}(t_0) D(t_0)^- & & & & G_b \Pi_{can}(t_N) D(t_N)^- U(t_N,t_{N-1}) \\ C_a & & & & \\ -U(t_1,t_0) & I & & & \\ & \ddots & \ddots & & \\ & & -U(t_{N-1},t_{N-2}) & & I \end{bmatrix},$$

$$\Pi_r = \begin{bmatrix} \pi_{\mu-1}(t_0) & & & \\ & \pi_{\mu-1}(t_1) & & \\ & & \ddots & \\ & & & \pi_{\mu-1}(t_{N-1}) \end{bmatrix}.$$

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 (2.44)–(2.46). 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 (2.40) and for S_{IERODE} (2.43)).

In [13, Sect. 4.3], for classical BVPs in ODEs, it is pointed out that the single shooting matrix may have a very large condition number (in the block row sum norm) when unstable IVPs have to be integrated. The responsible factor is $e^{L(b-a)}$, with a positive constant L given by the original data. In contrast, roughly speaking, the condition number of the multiple shooting matrix depends on terms of e^{Lh}, with $h = \max\limits_{1 \leq i \leq N} t_i - t_{i-1}$.

The BVPs in DAEs inherit the difficulties of the single shooting, but also the advantages of the multiple shooting. The canonical projector $\Pi_{can}(t)$ is uniformly bounded in $[a,b]$ and so is $\pi_{\mu-1}(t)$. If $\pi_{\mu-1}(t)$ is bounded, then there is a bound K of Π_r independent of N. K is a factor in the bounds of both the single and the multiple shooting matrices. Comparing the condition numbers, the size of K does not matter.

This makes clear that we can reduce the condition number by the multiple shooting approach as in the ODE case (cf. [13]).

Theorem 4.5 *A reflexive inverse S_{mult}^- of the multiple shooting matrix S_{mult} is given by*

$$S_{mult}^- = \Pi_r S_{mult,ODE}^{-1} = diagD \begin{bmatrix} X(t_0,t_0)S^- & \mathcal{G}(t_0,t_1) & \cdots & \mathcal{G}(t_0,t_{N-1}) \\ \vdots & \vdots & & \vdots \\ X(t_{N-1},t_0)S^- & \mathcal{G}(t_{N-1},t_1) & \cdots & \mathcal{G}(t_{N-1},t_{N-1}) \end{bmatrix} diagD^-$$

with $diagD := \operatorname{diag}(D(t_0),\ldots,D(t_{N-1}))$ *and* $diagD^- := \operatorname{diag}(D(t_0)^-,\ldots,D(t_{N-1})^-)$.

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 $S_{mult}^-S_{mult} = \Pi_r S_{mult,ODE}^{-1} S_{mult,ODE}\Pi_r = \Pi_r$ and obtain the required relations. □

The relation of Theorem 4.5 was shown for index-1 DAEs in [88].

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 by applying the differentiation index concept [38] or the strangeness index concept [74, 111]. In the latter, the realization of the shooting procedure is strongly interlocked with the reduction from the derivative array system. Gerdts [53] investigated a special structured index-2 DAE, which is reduced to index 1 by differentiation. See also Remark 2.5. In [25, 37, 47, 79, 80] the shooting method is investigated for index-1 DAEs. We find many papers considering very special applications. Lamour [80] and Baiz [25] focus on periodic BVPs. The necessary conditions of optimal control problems are investigated and shooting methods applied in [32, 37, 53, 63]. A lot of papers deal with single problems in science and techniques which are then solved by shooting methods.

We consider the TPBVP (2.1), (2.2). 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

$$f(D(\bar{t})y, P_0(\bar{t})\bar{\alpha} + Q_0 v, \bar{t}) = 0, \tag{4.35}$$

$$Q_0 y + P_0 v = 0, \tag{4.36}$$

which have in the index-1 case the nonsingular Jacobian matrix (cf. for a related proof [86, Lemma 4.12])

$$\begin{bmatrix} f_y D & f_x Q_0 \\ Q_0 & P_0 \end{bmatrix}.$$

The matching conditions are given by

$$D(t_i)(D(t_i)^-\xi_i - x(t_i; t_{i-1}, \xi_{i-1})) = 0 \quad \text{for } i = 1, \ldots, N-1. \tag{4.37}$$

The system to solve consists of (2.2) as

$$g(D(t_0)^-\xi_0 + v_0, x(b; t_{N-1}, \xi_{N-1})) = 0, \tag{4.38}$$

the matching conditions (4.37) and the determination of v_0 using (4.35), (4.36) 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 (4.32) with the linearization (cf. Sect. 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 for solving nonlinear overdetermined systems are applicable. As mentioned above a formulation as a square system is also 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 $R(t) \equiv I$. This holds because of the nonsingularity of DD^T ($D = DD^-D \Rightarrow DD^- = I$), i.e., all equations related to (4.31) vanish.

Very often a semi-explicit structure of f is assumed (see (2.5)). 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 also has full rank.

A semi-explicit structure of the DAE is assumed, e.g., in [63, 67, 106].

4.3 Further References, Comments, and Open Questions

Remark 4.1 (Take Advantage of (Partially) Separated Boundary Conditions) If the boundary conditions $G_a x(a) + G_b x(b) = \gamma$ are structured such that a part is separated at $t = a$ we should take advantage of such explicitly required initial values. This can be done by using for shooting an adapted initial 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 semi-explicit index-1 DAEs, a special way to avoid the computation of consistent initial values at every shooting point is proposed. The following DAE is considered:

$$\begin{aligned} y' &= f(t, y, u, p) \\ 0 &= g(t, y, u, p). \end{aligned}$$

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 [83].

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., [25, 51, 107, 113]. 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 and 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., [51, 80]).

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 [30, 101]. 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

$$x'(t) + B(t)x(t) = 0, \quad t \in [a, b], \tag{5.1}$$

$$C_a x(a) = 0, \tag{5.2}$$

with $C_a \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^k)$, $\operatorname{rank} C_a = k \leq m$, 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

$$y'(t) - B(t)^* y(t) = 0, \quad t \in [a, b], \tag{5.3}$$

$$y(a) = C_a^*. \tag{5.4}$$

The subspace $\mathcal{M}(t) = \ker y(t)^* = (\operatorname{im} y(t))^\perp$ has dimension $m - k$. BVPs for (5.1) and separated boundary conditions

$$C_a x(a) = 0, \quad C_b x(b) = 0 \tag{5.5}$$

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 unique solvability.

Generally the adjoint ODE is not easier to integrate than the original ODE. The idea behind the Abramov transfer method [1] consists of a continuous orthogonalization by demanding $y^* y' = 0$ and turning to the nonlinear equation

$$y'(t) - (I - y(t)(y(t)^* y(t))^{-1} y(t)^*) B(t)^* y(t) = 0, \quad t \in [a, b], \tag{5.6}$$

instead of (5.3). Equation (5.6) 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 [13, 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 mentioned in [13], 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

$$f((Dx)'(t), x(t), t) = 0, \quad t \in [a, b],$$
$$g(x(a), x(b) = 0,$$

on a grid $\pi : a = t_0 < \cdots < t_N = b$ have been studied 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 an 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

$$f((Dx)'(t), x(t), t) = 0, \quad t \in [a, b] = \mathcal{I}, \tag{5.7}$$
$$g(x(a), x(b) = 0, \tag{5.8}$$

can be formulated as an operator equation (cf. the proofs of Theorems 2.7 and 2.11). Let $\mathcal{D}_F \subseteq \mathcal{D}_f$ be open. We associate with the DAE (5.7) the nonlinear operator

$$F : \operatorname{dom} F \subseteq \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m) \to \mathcal{C}(\mathcal{I}, \mathbb{R}^m),$$
$$\operatorname{dom} F := \{x \in C_D^1(\mathcal{I}, \mathbb{R}^m) : x(t) \in \mathcal{D}_F \text{ for all } t \in \mathcal{I}\},$$
$$(Fx)(t) := f((Dx)'(t), x(t), t), \quad t \in \mathcal{I}, \quad x \in \operatorname{dom} F, \tag{5.9}$$

such that the DAE (5.7) is represented as the operator equation

$$Fx = 0. \tag{5.10}$$

F is said to be a *nonlinear differential-algebraic* operator. The operator equation (5.10) reflects the classical view of a DAE: the solutions belong to $\mathcal{C}^1_D(\mathcal{I}, \mathbb{R}^m)$ 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

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

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

$$A_*(Dx)' + B_*x = q. \tag{5.11}$$

The composed operator

$$\begin{aligned} \mathcal{F} &: \operatorname{dom} F \subseteq \mathcal{C}^1_D(\mathcal{I}, \mathbb{R}^m) \to \mathcal{C}(\mathcal{I}, \mathbb{R}^m) \times \mathbb{R}^{m-l}, \\ \mathcal{F}x &:= (Fx,\ g(x(a), x(b))), \quad x \in \operatorname{dom} F, \end{aligned} \tag{5.12}$$

is Fréchet differentiable since F is so. The equation $\mathcal{F}x = 0$ represents the BVP (5.7), (5.8), whereas the equation $\mathcal{F}x = (q, \gamma)$ is the operator form of the perturbed BVP

$$f((D(t)x(t))', x(t), t) = q(t),\ t \in \mathcal{I}, \quad g(x(a), x(b)) = \gamma. \tag{5.13}$$

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

$$x_{k+1} = x_k - \mathcal{F}'(x_k)^{-1}\mathcal{F}(x_k), \quad k \geq 0, \tag{5.14}$$

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

$$\mathcal{F}'(x_k)z = -\mathcal{F}(x_k), \quad k \geq 0, \tag{5.15}$$

and, having the solution z_{k+1} of the linear problem (5.15), one puts

$$x_{k+1} = x_k + z_{k+1}. \tag{5.16}$$

The linear problem (5.15) represents the linear BVP

$$\begin{aligned} f_y((Dx_k)'(t), x_k(t), t))(Dz)'(t) + f_x((Dx_k)'(t), x_k(t), t))z(t) &= -f((Dx_k)'(t), x_k(t), t)), \\ & \qquad t \in \mathcal{I}, \\ Ga(x_k(a), x_k(b))z(a) + G_b(x_k(a), x_k(b))z(b) &= -g(x_k(a), x_k(b)), \end{aligned}$$

with partial derivatives G_a, G_b of the function g with respect to its first and second arguments.

A damping parameter is usually incorporated, and instead of (5.16) one applies

$$x_{k+1} = x_k + \alpha_{k+1} z_{k+1}, \quad \text{with} \quad \alpha_{k+1} \in (0, 1]. \tag{5.17}$$

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,

$$\begin{aligned} & \|\mathcal{F}(x)\|_{res} := \|\mathcal{F}(x)\| = \|F(x)\|_\infty + |g(x(a), x(b))| \\ \text{and} \quad & \|\mathcal{F}(x)\|_{res}^2 := \|F(x)\|_{L^2}^2 + |g(x(a), x(b))|^2. \end{aligned}$$

Sufficient conditions for the composed operator $\mathcal{F}$ to be a local diffeomorphism in the natural setting are described in [96, Sect. 4.3.2]. Then the BVP is well-posed in the natural setting and the DAE has index 1, see Sect. 2.5.1.

In [96, Sect. 4.3.3] and Sect. 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

$$J(x) := \frac{1}{2}\|F(x)\|_{L^2}^2 + \frac{1}{2}|g(x(a), x(b))|^2, \quad x \in \operatorname{dom} \mathcal{F}. \tag{5.18}$$

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

$$\begin{aligned} J'(x)z = & (F'(x)z, F(x))_{L^2} \\ & + \langle\, b_a(x(a), x(b))z(a) + b_e(x(a), x(b))z(b),\, b(x(a), x(b)) \,\rangle. \end{aligned}$$

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,

$$\mathcal{F}'(x^0)z = -\mathcal{F}(x^0), \quad k \geq 0, \tag{5.19}$$

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 the 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 linear BVPs, with differing success. Although the linear BVPs could be solved successfully, the intermediate processing to prepare the next iteration could not be managed in an efficient way. Although 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 has been reported up to 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 (5.15) resp. (5.19) is essentially ill-posed and might not be solvable. Changing to least-squares solutions does not make a great deal of 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 [86, Sect. 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 (5.14) one uses a bounded outer inverse. Such an outer inverse is provided by [96, Theorem 4.2]. It seems that no practical experience is available in this context up to now.

6 Appendix

6.1 Basics Concerning Regular DAEs

We collect basic facts on the DAE

$$f((Dx)'(t), x(t), t) = 0, \tag{6.1}$$

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 [86] for proofs, motivation, and more details.

6.1.1 Regular DAEs, Regularity Regions

The DAE (6.1) 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 (2.3) pointwise induces the continuously differentiable (see [86, Lemma A.20]) *border projector* $R : \mathcal{D}_f \times \mathcal{I}_f \to \mathcal{L}(\mathbb{R}^n)$ given by

$$\operatorname{im} R(x, t) = \operatorname{im} D(t), \ \ker R(x, t) = \ker f_y(y, x, t), \ (y, x, t) \in \mathbb{R}^n \times \mathcal{D}_f \times \mathcal{I}_f. \tag{6.2}$$

Next we depict the notion of regularity regions of a DAE (6.1). For this aim we introduce *admissible matrix function sequences* and associated projector functions (cf. [86]). 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 (2.3) 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 $t \in \mathcal{I}_f$

$$\operatorname{im} Q_0(t) = N_0(t) := \ker D(t), \quad \Pi_0(t) := P_0(t) := I - Q_0(t). \tag{6.3}$$

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

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

The resulting function D^- is continuous, if P_0 is continuously differentiable then so also is D^-.

Definition 6.1 Let the DAE (6.1) have a properly involved derivative and let $\mathcal{G} \subseteq \mathcal{D}_f \times \mathcal{I}_f$ 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 (6.1) 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 $i \geq 1$:

$$G_i := G_{i-1} + B_{i-1} Q_{i-1}, \tag{6.4}$$

$$N_i := \ker G_i, \quad \widehat{N}_i := (N_0 + \cdots + N_{i-1}) \cap N_i,$$

find a complement X_i such that $N_0 + \cdots + N_{i-1} = \widehat{N}_i \oplus X_i$,

choose a projector Q_i such that $\operatorname{im} Q_i = N_i$ and $X_i \subseteq \ker Q_i$,

set $P_i := I - Q_i$, $\Pi_i := \Pi_{i-1} P_i$,

$$B_i := B_{i-1} P_{i-1} - G_i D^- (D \Pi_i D^-)' D \Pi_{i-1}, \tag{6.5}$$

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 $G_0, \ldots, G_\kappa$ 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 [86] 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 the 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

$$r_0 \le \cdots \le r_{\mu-1} < r_\mu = m. \tag{6.6}$$

Thereby, neither the factorization nor the special choice of admissible projectors 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 [86, Corollary 1.32].

For linear DAEs with time-varying coefficients, the term $(\cdot)'$ in (6.5) means the derivative in time, and all matrix functions are functions in time. In general, the term $(\cdot)'$ in (6.5) 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 [86, Sect. 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 this 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 ourselves to the wording f *is sufficiently smooth*.

The next definition ties regularity to the inequalities (6.6) 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 (6.1) have a properly involved derivative. Let $\mathcal{G} \subseteq \mathcal{D}_f \times \mathcal{I}_f$ be an open, connected subset. The DAE (6.1) 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 (6.6) 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) applies).

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 (6.1) 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 (6.1), 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 etc., see examples in [82, 86, 95]. No doubt, much further research is needed to elucidate these phenomena.

6.1.2 The Structure of Linear DAEs

The general DAE (6.1) captures linear DAEs

$$A(t)(Dx)'(t) + B(t)x(t) - q(t) = 0 \tag{6.7}$$

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 of *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 a triviality; however, there is a continuing profound debate about some related questions, cf. [96, Sect. 4.4].

If the linear DAE (6.7) is regular on the interval $\mathcal{I}$, then (see [86, Sect. 2.4]) it can be decoupled by admissible projector functions into an *IERODE*

$$u' - (D\Pi_{\mu-1}D^-)'u + D\Pi_{\mu-1}G_\mu^{-1}B_\mu D^- u = D\Pi_{\mu-1}G_\mu^{-1}q \tag{6.8}$$

and a triangular subsystem of several equations including differentiations

$$\begin{bmatrix} 0 & \mathcal{N}_{01} & \cdots & \mathcal{N}_{0,\mu-1} \\ & 0 & \ddots & \vdots \\ & & \ddots & \mathcal{N}_{\mu-2,\mu-1} \\ & & & 0 \end{bmatrix} \begin{bmatrix} 0 \\ (Dv_1)' \\ \vdots \\ (Dv_{\mu-1})' \end{bmatrix} \tag{6.9}$$

$$+ \begin{bmatrix} I & \mathcal{M}_{01} & \cdots & \mathcal{M}_{0,\mu-1} \\ & I & \ddots & \vdots \\ & & \ddots & \mathcal{M}_{\mu-2,\mu-1} \\ & & & I \end{bmatrix} \begin{bmatrix} v_0 \\ v_1 \\ \vdots \\ v_{\mu-1} \end{bmatrix} + \begin{bmatrix} \mathcal{H}_0 \\ \mathcal{H}_1 \\ \vdots \\ \mathcal{H}_{\mu-1} \end{bmatrix} D^- u = \begin{bmatrix} \mathcal{L}_0 \\ \mathcal{L}_1 \\ \vdots \\ \mathcal{L}_{\mu-1} \end{bmatrix} q.$$

The subspace $\operatorname{im} D\Pi_{\mu-1}$ is an invariant subspace for the IERODE (6.8).

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

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

$$
\begin{aligned}
\mathcal{L}_0 &:= Q_0P_1\cdots P_{\mu-1}G_\mu^{-1}, \\
\mathcal{L}_i &:= \Pi_{i-1}Q_iP_{i+1}\cdots P_{\mu-1}G_\mu^{-1}, \qquad i = 1,\ldots,\mu-2, \\
\mathcal{L}_{\mu-1} &:= \Pi_{\mu-2}Q_{\mu-1}G_\mu^{-1}, \\
\mathcal{H}_0 &:= Q_0P_1\cdots P_{\mu-1}\mathcal{K}\Pi_{\mu-1}, \\
\mathcal{H}_i &:= \Pi_{i-1}Q_iP_{i+1}\cdots P_{\mu-1}\mathcal{K}\Pi_{\mu-1}, \qquad i = 1,\ldots,\mu-2, \\
\mathcal{H}_{\mu-1} &:= \Pi_{\mu-2}Q_{\mu-1}\mathcal{K}\Pi_{\mu-1},
\end{aligned}
$$

with

$$
\mathcal{K} := (I-\Pi_{\mu-1})G_\mu^{-1}B_{\mu-1}\Pi_{\mu-1} + \sum_{l=1}^{\mu-1}(I-\Pi_{l-1})(P_l-Q_l)(D\Pi_lD^-)'D\Pi_{\mu-1},
$$

$$
\mathcal{M}_j := \sum_{k=0}^{j-1}(I-\Pi_k)\{P_kD^-(D\Pi_kD^-)' - Q_{k+1}D^-(D\Pi_{k+1}D^-)'\}D\Pi_{j-1}Q_lD^-,
$$

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

The IERODE is always uncoupled from the second subsystem, but the latter is tied to the IERODE (6.8) if among the coefficients $\mathcal{H}_0,\ldots,\mathcal{H}_{\mu-1}$ there is at least one which 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 (6.7) is regular and the original data are sufficiently smooth, then the DAE (6.7) is called *fine*. Fine DAEs always possess fine and complete decouplings, see [86, Sect. 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 $Q_0,\ldots,Q_{\mu-1}$ 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 $\{E,F\}$. In 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 (6.1) along x_*, that

is, the linearized DAE

$$A_*(t)(Dx)'(t) + B_*(t)x(t) = q(t), \quad t \in \mathcal{I}_*, \tag{6.10}$$

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 (6.10) inherits from the nonlinear DAE (6.1) the properly stated leading term.

We denote by $\mathcal{C}^m_{ref}(\mathcal{G})$ 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 also have $x_* \in \mathcal{C}^1_D(\mathcal{I}_{x_*}, \mathbb{R}^m)$. 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 [86, Theorem 3.33] to our somewhat simpler situation.

Theorem 6.1 *Let the DAE (6.1) 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 (6.1) is regular on $\mathcal{G}$ if the linearized DAE (6.10) along each arbitrary reference function $x_* \in \mathcal{C}^m_{ref}(\mathcal{G})$ is regular, and vice versa.*
(2) *If the DAE (6.1) 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 (6.10) 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 (6.10) along reference functions $x_* \in \mathcal{C}^m_{ref}(\mathcal{G})$ are regular, then they have uniform index and characteristics, and the nonlinear DAE (6.1) is also regular on $\mathcal{G}$, with the same index and characteristics.*

Corollary 6.2 *Let the DAE (6.1) 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 (6.1) is regular on $\mathcal{D} \times \mathcal{I}$ if the linearized DAE (6.10) 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 (6.1) 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 (6.10) 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 (6.10) 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 (6.1) is also regular on $\mathcal{D} \times \mathcal{I}$, with the same index and characteristics.*

Proof Statement (1) is a consequence of 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) is proved along the lines of [86, Theorem 3.33] by means of so-called widely orthogonal projector functions. The proof given in [86] also works if one supposes solely compact individual intervals $\mathcal{I}_{x_*}$.

By Lemma 6.3 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 6.3 *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 (6.7) 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 [86, Sect. 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 $\mu > 1$. In an advanced setting we choose

$$Y = \mathcal{C}^{\operatorname{ind}\mu}(\mathcal{I}, \mathbb{R}^m)$$

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. Sect. 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

$$\|Tx\|_{\operatorname{ind}\mu} := \|Tx\|_\infty + \|(D\Pi_{\mu-2}Q_{\mu-1}x)'\|_\infty + \cdots + \|(D\Pi_0 Q_1 x)'\|_\infty.$$

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 required inequality $\|Tx\|_{\operatorname{ind}\mu} \le k_{\operatorname{ind}\mu}\|x\|_{\mathcal{C}_D^1}$.

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 $\mathcal{C}^{\operatorname{ind}\mu}(\mathcal{I}, \mathbb{R}^m)$ as well as its norm $\|.\|_{\operatorname{ind}\mu}$ 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 [86, Sect. 3.9.1], [96, Sect. 2]. It seems to be impossible to capture errors in practical computational procedures using these 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 inverse
K^+	Orthogonal generalized (Moore–Penrose) inverse
dom K	Definition domain of the map K
ker K	Nullspace (kernel) of the operator K
im K	Image (range) of the operator K
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
$\lVert\cdot\rVert$	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. Abramov, A.A.: On transfer of boundary conditions for systems of linear ordinary differential equations (a variant of transfer method). USSR Comput. Math. Math. Phys. **1**(3), 542–544 (1961)
2. Amodio, P., Mazzia, F.: Numerical solution of differential algebraic equations and computation of consistent initial/boundary conditions. J. Comput. Appl. Math. **87**, 135–146 (1997)
3. Amodio, P., Mazzia, F.: An algorithm for the computation of consistent initial values for differential-algebraic equations. Numer. Algorithms **19**, 13–23 (1998)
4. Anh, P.K.: Multipoint boundary-value problems for transferable differential-algebraic equations. I–linear case. Vietnam J. Math. **25**(4), 347–358 (1997)
5. Anh, P.K.: Multipoint boundary-value problems for transferable differential-algebraic equations. II–quasilinear case. Vietnam J. Math. **26**(4), 337–349 (1998)

6. Anh, P.K., Nghi, N.V.: On linear regular multipoint boundary-value problems for differential algebraic equations. Vietnam J. Math. **28**(2), 183–188 (2000)
7. Ascher, U., Lin, P.: Sequential regularization methods for nonlinear higher index DAEs. SIAM J. Sci. Comput. **18**, 160–181 (1997)
8. Ascher, U.M., Petzold, L.R.: Numerical methods for boundary value problems in differential-algebraic equations. In: Byrne, G.D., Schiesser, W.E. (eds.) Recent Developments in Numerical Methods and Software for ODEs/DAEs/PDEs, pp. 125–135. World Scientific, London/Singapore (1992)
9. Ascher, U.M., Petzold, L.R.: Projected collocation for higher-order higher-index differential-algebraic equations. J. Comput. Appl. Math. **43**, 243–259 (1992)
10. Ascher, U.M., Petzold, L.R.: Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations. SIAM, Philadelphia (1998)
11. Ascher, U., Spiteri, R.: Collocation software for boundary value differential-algebraic equations. SIAM J. Sci. Comput. **15**, 938–952 (1994)
12. Ascher, U., Christiansen, J., Russell, R.: Collocation software for boundary value ODEs. ACM Trans. Math. Softw. **7**(209–222) (1981)
13. Ascher, U., Mattheij, R., Russell, R.: Numerical Solution of Boundary Value Problems for Ordinary Differential Equations. Prentice Hall, Englewood Cliffs, NJ (1988)
14. Auzinger, W., Kneisl, G., Koch, O., Weinmüller, E.: SBVP 1.0 – A MATLAB solver for singular boundary value problems. ANUM Preprint 2/02, Vienna University of Technology (2002)
15. Auzinger, W., Koch, O., Weinmüller, E.: Efficient collocation schemes for singular boundary value problems. Numer. Algorithms **31**, 5–25 (2002)
16. Auzinger, W., Kneisl, G., Koch, O., Weinmüller, E.: A collocation code for boundary value problems in ordinary differential equations. Numer. Algorithms **33**, 27–39 (2003)
17. Auzinger, W., Koch, O., Weinmüller, E.: Analysis of a new error estimate for collocation methods applied to singular boundary value problems. SIAM J. Numer. Anal. **42**, 2366–2386 (2005)
18. Auzinger, W., Lehner, H., Weinmüller, E.: Defect-based a-posteriori error estimation for Index-1 DAEs. ASC Technical Report 20, Vienna University of Technology (2007)
19. Auzinger, W., Lehner, H., Weinmüller, E.: An efficient asymptotically correct error estimator for collocation solution to singular index-1 DAEs. BIT Numer. Math. **51**, 43–65 (2011)
20. Backes, A.: Extremalbedingungen für Optimierungs-Probleme mit Algebro-Differentialgleichungen. Logos, Berlin (2006). Dissertation, Humboldt-University Berlin (October 2005/January 2006)
21. Bader, G., Ascher, U.: A new basis implementation for a mixed order boundary value ODE solver. SIAM J. Sci. Stat. Comput. **8**, 483–500 (1987)
22. Bai, Y.: A perturbated collocation method for boundary value problems in differential-algebraic equations. Appl. Math. Comput. **45**, 269–291 (1991)
23. Bai, Y.: Modified collocation methods for boundary value problems in differential-algebraic equations. Ph.D. thesis, Fachbereich Mathematik, Philipps-Universität, Marburg/Lahn (1991)
24. Bai, Y.: A modified Lobatto collocation for linear boundary value problems of differential-algebraic equations. Computing **49**, 139–150 (1992)
25. Baiz, A.: Effiziente Lösung periodischer differential-algebraischer Gleichungssysteme in der Schaltungssimulation. Ph.D. thesis, Fachbereich Informatik, Technische Universität, Darmstadt. Shaker, Aachen (2003)
26. Balla, K.: Differential-algebraic equations and their adjoints. Dissertation, Doctor of the Hungarian Academy of Sciences, Hungarian Academy of Sciences, Budapest (2004)
27. Balla, K., März, R.: Transfer of boundary conditions for DAEs of index 1. SIAM J. Numer. Anal. **33**(6), 2318–2332 (1996)
28. Balla, K., März, R.: Linear differential-algebraic equations of index 1 and their adjoints. Results Math. **37**, 13–35 (2000)
29. Balla, K., März, R.: A unified approach to linear differential-algebraic equations and their adjoints. J. Anal. Appl. **21**(3), 783–802 (2002)

30. Balla, K., März, R.: Linear boundary value problems for differential-algebraic equations. Miskolc Math. Notes **5**(1), 3–18 (2004)
31. Barz, B., Suschke, E.: Numerische behandlung eines Algebro-Differentialgleichungssystems. RZ-Mitteilungen, Humboldt-Universität, Behandlung, Berlin (1994)
32. Bell, M., Sargent, R.: Optimal control of inequality constrained DAE systems. Comput. Chem. Eng. **24**, 2385–2404 (2000)
33. Biegler, L., Campbell, S., Mehrmann, V.: Control and Optimization with Differential-Algebraic Constraints. SIAM, Philadelphia (2011)
34. Bock, H., Eich, E., Schlöder, J.: Numerical solution of constrained least squares boundary value problems in differential-algebraic equations. In: Strehmel, K. (ed.) Numerical Treatment of Differential Equations, NUMDIFF-4. Teubner Texte zur Mathematik, vol. 104. Teubner, Leipzig (1987)
35. Brenan, K., Campbell, S., Petzold, L.: Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations. North Holland, New York (1989)
36. Brown, P., Hindmarsh, A., Petzold, L.: Consistent initial condition calculation for differential-algebraic systems. SIAM J. Sci. Comput. **19**(5), 1495–1512 (1998)
37. Callies, R.: Entwurfsoptimierung und optimale Steuerung. Differential-algebraische Systeme, Mehrgitter-Mehrzielansätze und numerische Realisierung. Habilitation, Technische Universität, München (2000)
38. Clark, K.D., Petzold, L.R.: Numerical solution of boundary value problems in differential-algebraic systems. SIAM J. Sci. Stat. Comput. **10**, 915–936 (1989)
39. de Boor, C., Swartz, B.: Collocation at Gaussian points. SIAM J. Numer. Anal. **10**, 582–606 (1973)
40. de Hoog, F., Weiss, R.: Difference methods for boundary value problems with a singularity for the first kind. SIAM J. Numer. Anal. **13**, 775–813 (1976)
41. Degenhardt, A.: A collocation method for boundary value problems of transferable differential-algebraic equations. Preprint (Neue Folge) 182, Humboldt-Universität zu Berlin, Sektion Mathematik (1988)
42. Degenhardt, A.: Collocation for transferable differential-algebraic equations. In: Griepentrog, E., Hanke, M., März, R. (eds.) Berlin Seminar on Differential-Algebraic Equations, Seminarberichte, vol. 92-1, pp. 83–104. Fachbereich Mathematik, Humboldt-Universität zu, Berlin (1992)
43. Dick, A., Koch, O., März, R., Weinmüller, E.: Convergence of collocation schemes for boundary value problems in nonlinear index-1 DAEs with a singular point. Math. Comput. **82**(282), 893–918 (2013)
44. Dokchan, R.: Numerical integration of differential-algebraic equations with harmless critical points. Ph.D. thesis, Institute of Mathematics, Humboldt-University, Berlin (2011)
45. Eich-Soellner, E., Führer, C.: Numerical Methods in Multibody Dynamics. B.G. Teubner, Stuttgart (1998)
46. Engl, H.W., Hanke, M., Neubauer, A.: Tikhonov regularization of nonlinear differential-algebraic equations. In: Sabatier, P.C. (ed.) Inverse Methods in Action, pp. 92–105. Springer, Berlin/Heidelberg (1990)
47. England, R., Lamour, R., Lopez-Estrada, J.: Multiple shooting using a dichotomically stable integrator for solving DAEs. Appl. Numer. Math. **42**, 117–131 (2002)
48. Estévez Schwarz, D., Lamour, R.: The computation of consistent initial values for nonlinear index-2 differential-algebraic equations. Numer. Algorithms **26**(1), 49–75 (2001)
49. Estévez Schwarz, D., Lamour, R.: Monitoring singularities while integrating DAEs. In: Progress in Differential-Algebraic Equations. Descriptor 2013, pp. 73–96. Differential-Algebraic Equations Forum. Springer, Heidelberg (2014)
50. Estévez Schwarz, D., Lamour, R.: Diagnosis of singular points of properly stated DAEs using automatic differentiation. Numer. Algorithms (2015, to appear)
51. Franke, C.: Numerical methods for the investigation of periodic motions in multibody dynamics. A collocation approach. Ph.D. thesis, Universität Ulm. Shaker, Aachen (1998)

52. Gear, C.W.: Maintaining solution invariants in the numerical solution of ODEs. SIAM J. Sci. Stat. Comput. **7**, 734–743
53. Gerdts, M.: Direct shooting method for the numerical solution of higher-index DAE optimal control problems. J. Optim. Theory Appl. **117**(2), 267–294 (2003)
54. Gerdts, M.: A survey on optimal control problems with differential-algebraic equations. In: Ilchmann, A., Reis, T. (eds.) Surveys in Differential-Algebraic Equations II. Springer, Heidelberg (2015)
55. Griepentrog, E., März, R.: Differential-Algebraic Equations and Their Numerical Treatment. Teubner-Texte zur Mathematik, vol. 88. BSB B.G. Teubner Verlagsgesellschaft, Leipzig (1986)
56. Hanke, M.: On a least-squares collocation method for linear differential-algebraic equations. Numer. Math. **54**, 79–90 (1988)
57. Hanke, M.: Beiträge zur Regularisierung von Randwertaufgaben für Algebro-Differentialgleichungen mit höherem Index. Dissertation(B), Habilitation, Institut für Mathematik, Humboldt-Universität zu Berlin (1989)
58. Hanke, M.: On the regularization of index 2 differential-algebraic equations. J. Math. Anal. Appl. **151**, 236–253 (1990)
59. Hanke, M.: Asymptotic expansions for regularization methods of linear fully implicit differential-algebraic equations. Zeitschrift für Analysis und ihre Anwendungen **13**, 513–535 (1994)
60. Higueras, I., März, R.: Differential algebraic equations with properly stated leading term. Comput. Math. Appl. **48**, 215–235 (2004)
61. Higueras, I., März, R., Tischendorf, C.: Stability preserving integration of index-1 DAEs. Appl. Numer. Math. **45**(2–3), 175–200 (2003)
62. Ho, M.D.: A collocation solver for systems of boundary-value differential/algebraic equations. Comput. Chem. Eng. **7**, 735–737 (1983)
63. Houska, B., Diehl, M.: A quadratically convergent inexact SQP method for optimal control of differential algebraic equations. Optim. Control Appl. Methods **34**, 396–414 (2013)
64. Kalachev, L.V., O'Malley, R.E.: Boundary value problems for differential-algebraic equations. Numer. Funct. Anal. Optim. **16**, 363–378 (1995)
65. Keller, H.: Approximation methods for nonlinear problems with application to two-point boundary value problems. Math. Comput. **29**, 464–474 (1975)
66. Keller, H.B., White Jr., A.B.: Difference methods for boundary value problems in ordinary differential equations. SIAM J. Numer. Anal. **12**(5), 791–802 (1975)
67. Kiehl, M.: Sensitivity analysis of ODEs and DAEs – theory and implementation guide. Optim. Methods Softw. **10**, 803–821 (1999)
68. Koch, O.: Asymptotically correct error estimation for collocation methods applied to singular boundary value problems. Numer. Math. **101**, 143–164 (2005)
69. Koch, O., Weinmüller, E.: The convergence of shooting methods for singular boundary value problems. Math. Comput. **72**(241), 289–305 (2003)
70. Koch, O., Kofler, P., Weinmüller, E.: 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)
71. Koch, O., März, R., Praetorius, D., Weinmüller, E.: Collocation for solving DAEs with singularities. ASC Report 32/2007, Vienna University of Technology, Institute for Analysis and Scientific Computing (2007)
72. Koch, O., März, R., Praetorius, D., Weinmüller, E.: Collocation methods for index-1 DAEs with a singularity of the first kind. Math. Comput. **79**(269), 281–304 (2010)
73. Kopelmann, A.: Ein Kollokationsverfahren für überführbare Algebro-Differentialbleichungen. Preprint (Neue Folge) 151, Humboldt-Universität zu Berlin, Sektion Mathematik (1987)
74. Kunkel, P., Mehrmann, V.: Differential-Algebraic Equations - Analysis and Numerical Solution. EMS Publishing House, Zürich (2006)

75. Kunkel, P., Stöver, R.: Symmetric collocation methods for linear differential-algebraic boundary value problems. Numer. Math. **91**, 475–501 (2002)
76. Kunkel, P., Mehrmann, V., Stöver, R.: Symmetric collocation methods for unstructured nonlinear differential-algebraic equations of arbitrary index. Numer. Math. **98**, 277–304 (2004)
77. Lamour, R.: A shooting method for fully implicit index-2 differential-algebraic equations. SIAM J. Sci. Comput. **18**(1), 94–114 (1997)
78. Lamour, R.: Bestimmung optimaler Integrationsrichtungen beim Mehrfachschießverfahren zur Lösung von Zwei-Punkt-Randwertproblemen (1984). Wiss. Beitr., Martin-Luther-University Halle Wittenberg 1984/24(M 33), 66–70 (1984)
79. Lamour, R.: A well–posed shooting method for transferable DAEs. Numer. Math. **59** (1991)
80. Lamour, R.: Oscillations in differential–algebraic equations. In: Seminarbericht Nr. 92–1. Fachbereich Mathematik der Humboldt, Universität zu Berlin (1992)
81. Lamour, R.: Index determination and calculation of consistent initial values for DAEs. Comput. Math. Appl. **50**(2), 1125–1140 (2005)
82. Lamour, R., März, R.: Detecting structures in differential-algebraic equations: computational aspects. J. Comput. Appl. Math. **236**(16), 4055–4066 (2012). Special Issue: 40 years of Numerical Math
83. Lamour, R., Mazzia, F.: Computation of consistent initial values for properly stated index-3 DAEs. BIT Numer. Math. **49**, 161–175 (2009)
84. Lamour, R., März, R., Winkler, R.: How floquet theory applies to index-1 differential-algebraic equations. J. Appl. Math. **217**(2), 372–394 (1998)
85. Lamour, R., März, R., Winkler, R.: Stability of periodic solutions of index-2 differential algebraic systems. J. Math. Anal. Appl. **279**, 475–494 (2003)
86. Lamour, R., März, R., Tischendorf, C.: Differential-algebraic equations: a projector based analysis. In: Ilchman, A., Reis, T. (eds.) Differential-Algebraic Equations Forum. Springer, Berlin/Heidelberg/New York/Dordrecht/London (2013)
87. Lentini, M., März, R.: Conditioning and dichotomy in differential-algebraic equations. SIAM J. Numer. Anal. **27**(6), 1519–1526 (1990)
88. Lentini, M., März, R.: The condition of boundary value problems in transferable differential-algebraic equations. SIAM J. Numer. Anal. **27**(4), 1001–1015 (1990)
89. März, R.: On difference and shooting methods for boundary value problems in differential-algebraic equations. Zeitschrift für Angewandte Mathematik und Mechanik **64**(11), 463–473 (1984)
90. März, R.: On correctness and numerical treatment of boundary value problems in DAEs. Zhurnal Vychisl. Matem. i Matem. Fiziki **26**(1), 50–64 (1986)
91. März, R.: Numerical methods for differential-algebraic equations. Acta Numer. 141–198 (1992)
92. März, R.: On linear differential-algebraic equations and linearizations. Appl. Numer. Math. **18**, 267–292 (1995)
93. März, R.: 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. März, R.: 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). http://www2.mathematik.hu-berlin.de/publ/pre/2011/M-11-16.html
95. März, R.: Notes on linearization of DAEs and on optimization with differential-algebraic constraints. In: Biegler, L.T., Campbell, S.L., Mehrmann, V. (eds.) Control and Optimization with Differential-Algebraic Constraints. Advances in Design and Control, pp. 37–58. SIAM, Philadelphia (2012)
96. März, R.: Differential-algebraic equations from a functional-analytic viewpoint: a survey. In: Ilchmann, A., Reis, T. (eds.) Surveys in Differential-Algebraic Equations II. Springer, Heidelberg (2015)

97. März, R., Riaza, R.: Linear differential-algebraic equations with properly leading term: a-critical points. Math. Comput. Model. Dyn. Syst. **13**, 291–314 (2007)
98. März, R., Weinmüller, E.B.: Solvability of boundary value problems for systems of singular differential-algebraic equations. SIAM J. Math. Anal. **24**(1), 200–215 (1993)
99. Moszyński, K.: A method of solving the boundary value problem for a system of linear ordinary differential equations. Algorithmy **11**(3), 25–43 (1964)
100. Petry, T.: On the stability of the Abramov transfer for differential-algebraic equations of index 1. SIAM J. Numer. Anal. **35**(1), 201–216 (1998)
101. Petry, T.: Realisierung des Newton-Kantorovich-Verfahrens für nichtlineare Algebro-Differentialgleichungen mittels Abramov-Transfer. Ph.D. thesis, Humboldt-Universität zu, Berlin. Logos, Berlin (1998)
102. Rabier, P., Rheinboldt, W.: Theoretical and numerical analysis of differential-algebraic equations. In: Ciarlet, P.G., et al. (eds.) Handbook of Numerical Analysis, vol. VIII. Techniques of Scientific Computing (Part 4), pp. 183–540. North Holland/Elsevier, Amsterdam (2002)
103. Riaza, R.: Differential-Algebraic Systems. Analytical Aspects and Circuit Applications. World Scientific, Singapore (2008)
104. Riaza, R.: DAEs in Circuit Modelling: a survey. In: Ilchmann, A., Reis, T. (eds.) Surveys in Differential-Algebraic Equations I. Differential-Algebraic Equations Forum. Springer, Heidelberg (2013)
105. Riaza, R., März, R.: Linear index-1 DAEs: regular and singular problems. Acta Appl. Math. **84**, 29–53 (2004)
106. Schulz, V.H., Bock, H.G., Steinbach, M.C.: Exploiting invariants in the numerical solution of multipoint boundary value problems for DAE. SIAM J. Sci. Comput. **19**, 440–467 (1998)
107. Selting, P., Zheng, Q.: Numerical stability analysis of oscillating integrated circuits. J. Comput. Appl. Math. **82**, 367–378 (1997)
108. Shampine, L.: Conservative laws and the numerical solution of ODEs. Comput. Math. Appl. **12**, 1287–1296 (1986)
109. Simeon, B.: Computational flexible multibody dynamics. In: A Differential-Algebraic Approach. Differential-Algebraic Equations Forum. Springer, Heidelberg (2013)
110. Stetter, H.: The defect correction principle and discretization methods. Numer. Math. **29**, 425–443 (1978)
111. Stöver, R.: Numerische Lösung von linearen differential-algebraischen Randwertproblemen. Ph.D. thesis, Universität Bremen. Doctoral thesis, Logos, Berlin (1999)
112. Trenn, S.: Solution concepts for linear DAEs: a survey. In: Ilchmann, A., Reis, T. (eds.) Surveys in Differential-Algebraic Equations I, pp. 137–172. Springer, Heidelberg (2013)
113. Wernsdorf, B.: Ein Kollokationsverfahren zur numerischen Bestimmung periodischer Lösungen von nichtlinearen algebro-differentialgleichungen. Ph.D. thesis, Sektion Mathematik, Humboldt-Universität zu Berlin (1984)
114. Wijckmans, P.M.E.J.: Conditioning of differential-algebraic equations and numerical solutions of multibody dynamics. Ph.D. thesis, Technische Universiteit, Eindhoven (1996)
115. Zadunaisky, P.: On the estimation of errors propagated in the numerical integration of ODEs. Numer. Math. **27**, 21–39 (1976)

Index

A. Ilchmann, T. Reis (eds.), *Surveys in Differential-Algebraic Equations III*,
Differential-Algebraic Equations Forum, DOI 10.1007/978-3-319-22428-2